Correspondence

Source-System Windowing for Speech Analysis and Synthesis

B. Yegnanarayana and P. Satyanarayana Murthy

Abstract—**In this correspondence, a new method of analysis of speech is proposed that will bring out variations in vocal tract system characteristics in short (2–4 ms) segments. In this method, the source and system components of the speech signal are suitably windowed to reduce the effects of truncation of conventional waveform windowing.**

I. Introduction

Speech is produced as a result of excitation of the time-varying vocal tract system. In speech production, both excitation and the vocal tract change continuously with time. One objective in speech analysis is to derive the time-varying characteristics of the speech production mechanism from the speech signal. For analysis purposes, a linear source-system model is assumed for speech production, and the source and system characteristics are assumed quasistationary in the analysis interval [1]. Obviously, this simple model does not give an accurate representation of speech in each frame. Even in the steady voiced sounds, excitation characteristics change within each pitch period due to glottal vibrations, and the vocal tract system changes due to coupling and decoupling of the trachea during open and closed phases of the glottal excitation, respectively [2]. Linear prediction coefficients (LPC's) [3] capture only the averaged behavior over the analysis frame. The detail lost in LPC modeling cannot easily be compensated for by using a glottal pulse model [4] for excitation. The main difficulty is in determining the characteristics of the vocal tract system from short (2–4 msec) segments of the speech signal, at least in the two distinct phases in each pitch period, namely, the closed and open glottis regions. The problem in short-segment analysis is that the samples outside the chosen interval are either assumed to be zero (short-time spectrum, LP analysis by the autocorrelation method [5]) or assumed to belong to the same stationary region (covariance method [6], [7]).

The aim of this work is to explore methods to reduce the effects of the short window in the analysis. It is to be noted that windowing the signal causes discontinuity at the edges of the window. At the same time, windowing is essential for capturing the dynamic characteristics of the source and system in the speech production mechanism. Several methods have been proposed earlier for windowing the signal directly; however, we propose a new approach in this work, which we call *source-system windowing* [8], [9]. The central idea is to explore methods where the source and system components of a speech signal are independently modified to confine their effects to the selected analysis window region. The windowed components are then used to regenerate a speech signal corresponding to the source and system in that region, although the generated signal itself may extend beyond the selected window length. The generated signal is analyzed to extract the system characteristics more accurately. In this correspondence, we show that even an approximate decomposition of the original speech signal into source and system components will be adequate to implement the proposed source-system windowing.

The proposed method is presented in Section II. Results of analysis of different vowels and other types of speech segments are presented in Section III to demonstrate the effectiveness of the method for obtaining an estimation of system information from short segments of speech. Finally, the significance of this analysis for speech synthesis is discussed briefly in Section IV.

II. Source-System Windowing

In waveform windowing, the shape of the signal waveform is altered. Consequently, the source and system characteristics derived from the signal may not properly represent the speech production system. The effects of waveform windowing will be severe when the window size is small (2–4 ms). In these cases, the system model tries to fit the zero-value samples outside the window, assuming they are the natural extension of the samples within the window region. This leads to bias in the estimated autocorrelation or spectral values, and consequently results in errors in the parameters of the model derived from these values. The bias is due to correlation between samples in natural speech. The high correlation between signal samples can be seen from the autocorrelation function of a segment of speech. The errors in correlation estimates get worse as window size is reduced.

Correlation between samples is reduced significantly in the residual signal derived from LP analysis. The values of the normalized autocorrelation function of the residual are small, and they remain small even when window size is reduced. That is, the short window effects are much less severe for the residual than for the original signal. The system characteristics in the signal not captured in the LPC's appear in the LP residual. Moreover, these characteristics are typically reflected in small durations of the residual due to the finite impulse response nature of the inverse filter. So selecting a short window in the residual and reexciting the all-pole system would generate a signal whose characteristics in the selected window will be similar to those in the corresponding window in the original speech signal. Due to all-pole filtering, the signal so generated extends beyond the chosen window with its natural decay, even though there is no excitation. This signal can be called a *source-windowed signal*. Using a nonrectangular tapered window on the residual will reduce the edge effects without significantly affecting the source characteristics. Note that the generated signal beyond the residual window region is due to all-pole filtering. Since the signal generated beyond the residual window region is not influenced by the excitation, we are not likely to get significantly new information other than what is present in the all-pole filter.

In order to reduce the dominance of the system (all-pole filter) on the residual excited waveform (source windowed signal), we propose a modification of the system in the regions outside the chosen window. We call this *bandwidth (BW) windowing*, by which we mean a modification of bandwidths of the resonances to produce a tapering window effect. A bandwidth function is used to increase the bandwidth of the poles of the original all-pole system significantly beyond the selected window region. The resulting waveform will have nearly the same vocal tract system characteristics as those of

Manuscript received June 4, 1994; revised October 19, 1995. The associate editor coordinating the review of this paper and approving it for publication was Dr. Douglas D. O'Shaughnessy.

The authors are with the Department of Computer Science and Engineering, Indian Institute of Technology, Madras, India (e-mail: yegna@iitm.ernet.in).

Publisher Item Identifier S 1063-6676(96)02449-2.

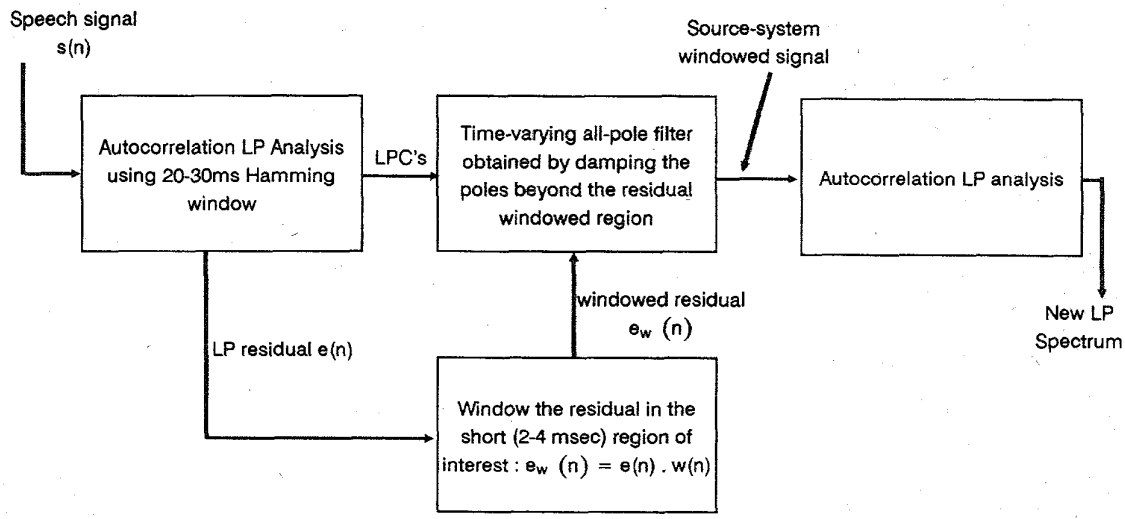

Fig. 1. Steps in the algorithm for source-system windowing. For source windowing alone, no change in damping is made in the all-pole filter.

the original speech signal within the window. The samples beyond the selected region taper faster than the signal in the short region of interest, thus enhancing the characteristics of the signal within the window. The steps in the algorithm for source-system windowing are given in Fig. 1. BW windowing may reduce the frequency resolution slightly, but would still bring out the characteristics of the system in the analysis window. The results of analysis of a synthetic signal, generated by exciting a tenth-order all-pole model by a periodic sequence of Liljencrants–Fant (LF) model glottal pulses [4], are shown in Fig. 2. The signal is preemphasized before analysis. For the closed glottis region the all-pole model spectrum in Fig. 2(a) is used, and for the open glottis region the same model with its poles damped, as shown in Fig. 2(b), is used. Since all the poles are damped, some higher formants may be lost in the LP analysis of the damped signal for the open phase region. The need for bandwidth windowing is evident from Fig. 2(e) and (f), which show the signal generated using source windowing in the open glottis interval and the corresponding tenth-order LP spectrum, respectively. Clearly, the oscillations due to the first formant outside the selected window dominate the analysis. This influence is reduced significantly when BW windowing is used, as shown in Fig. 2(k). Comparison with covariance analysis is illustrated in Fig. 2(l) and (m) for the closed and the open glottis regions, respectively. The covariance analysis gives the correct LP spectrum, as shown in Fig. 2(l), in the closed phase, as it does not have any effects of glottal source. Fig. 2(m) shows that the covariance analysis fails in the open glottis region.

We have also observed that the significant features, such as formant peaks and their bandwidths, are preserved to a large extent even for very sharp changes in the bandwidth function. There will be slight increases in the bandwidths of the formants as the bandwidth window is sharpened. The effectiveness of BW windowing is illustrated through the results given in Fig. 3. The symmetric Itakura distances [10] are computed between the spectra in the closed and open glottis regions in a pitch period of a natural voiced speech signal, for cases with and without bandwidth windowing. The distances are also computed between spectra for three different bandwidth windows. The three bandwidth windows (denoted as A, B, C in Fig. 3) differ in their sharpness of the taper outside the duration of the residual window, with sharpness increasing from A to C. The spectrum obtained in the closed glottis region without BW windowing is consistent with the spectra obtained using BW windowing for the three BW windows, as seen by the closeness of the distances d_{13}, d_{35}, and d_{57} to unity. The spectrum obtained in the open glottis region without BW windowing is dominated by the first formant (F_1) of the original LP spectrum and is different from the spectrum obtained using BW windowing for the same region. Hence, the distance d_{24} is large while the distances d_{46} and d_{68} are close to unity. The fact that both d_{46} and d_{68}, and similarly d_{35} and d_{57} are all close to unity even though the BW window is progressively sharpened from A to C, demonstrates that the sharpness with which the BW window is tapered is not very critical. The large distances d_{34}, d_{56}, and d_{78} confirm that there is a significant change in the spectra for the closed and the open glottis regions.

III. Short Window Analysis of Speech Segments

In this section we consider analysis of several types of speech segments using source-system windowing. In all these cases we consider a source window of size 3 msec which includes a 0.5 ms taper on either side and a BW window of the shape shown in Fig. 2(g). Our observation of the results of analysis are as follows: Analysis of the signal for nasal /m/ shows that the low sharp first formant due to the nasal tract is not significantly influenced by the glottal opening and closure. For unvoiced speech, source-system windowing does not show significant differences when compared to the conventional LP spectrum. This is because the vocal tract system is generally steady during the production of unvoiced speech.

Results of analysis of different natural vowels using the source-system windowing in the closed and open glottis regions of a pitch cycle are shown in Fig. 4. In these cases, the closed phase region is identified as the region after the largest peak in the LP residual. The open phase region is identified as the region just preceding the largest peak in the LP residual. In the open glottis region, we observe a significant increase in the bandwidth of the first formant and an increase in the value of the first formant (F_1) in some cases [11]–[13]. These observations are confirmed by the Itakura distances between the spectra in the open and closed regions. As in Fig. 2, some higher formants are lost in a few cases due to LP analysis of short data records.

For some speech segments (e.g., the vowel /u/ in high-pitched female speech), the source-system windowing brings out clearly the differences in spectra in the closed and open glottis regions. However,

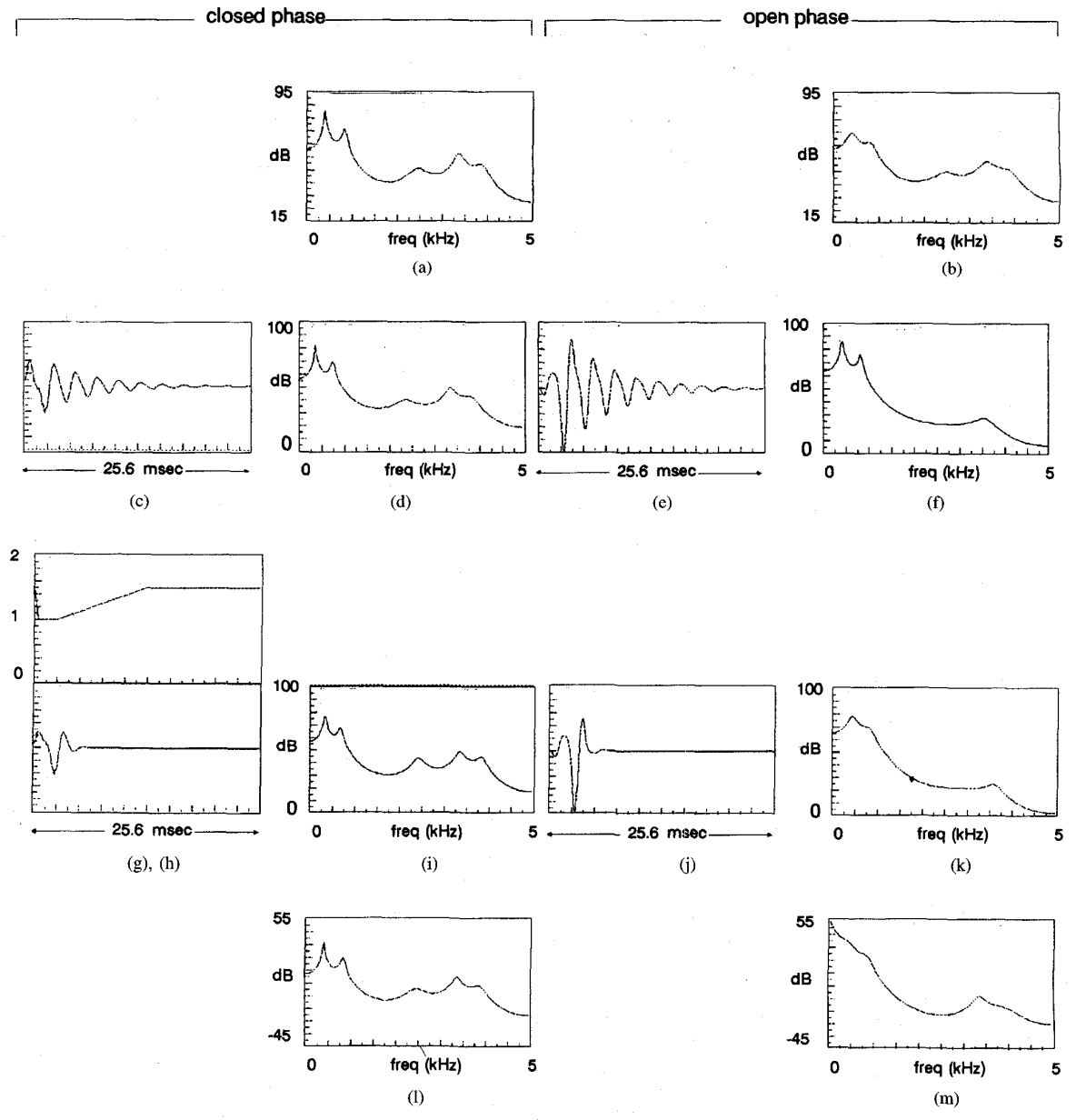

Fig. 2. Analysis of a synthetic signal using source-system windowing. (a), (b) Tenth-order all-pole model spectra used for synthesis in the closed and open glottis intervals, respectively. (c) Signal generated without BW windowing in the closed glottis interval. (d) Tenth-order LP spectrum of the signal in (c). (e) Signal generated in the open glottis interval without BW windowing. (f) The corresponding tenth-order LP spectrum. (g) BW window function—(h), (i), (j), and (k) are the figures corresponding to (c), (d), (e), and (f) for the case with BW windowing. (l) Tenth-order covariance analysis LP spectrum in closed glottis region of the signal. (m) Tenth-order covariance analysis LP spectrum in the open glottis region of the signal.

it is generally difficult to identify the closed and open glottis regions, even approximately, from either the waveform or the LP residual.

In CV transition regions, the characteristics of the vocal tract system exhibit rapid temporal and spectral changes. The spectra obtained using source-system windowing in one such transition for the sound /ca/ are shown in Fig. 5. The LP spectra obtained in the consonant region (shown by $2, 3, \cdots, 10$ in the figure) are different from those obtained in the vowel region (shown by 11 in the figure), while the conventional LP spectrum (shown by 1) exhibits the behavior for the entire frame. These results show that the new method of windowing indeed helps in extracting the system characteristics in the open and closed glottis regions of voiced speech. The effects of these differences in the vocal tract system on the quality of synthetic speech is examined briefly in the next section.

IV. Discussion

Voiced parts of speech primarily dictate the naturalness of synthetic speech [13], although the overall quality depends on both voiced and unvoiced speech. The quality of speech synthesized from LPC's depends on modeling the excitation source for voiced speech. Using a glottal pulse model for excitation and the LPC's for system will not reflect the dynamics of the vocal tract system within a pitch period. Through informal listening, we have noticed

1: Without BW windowing in the closed phase (CP)
2: Without BW windowing in the open phase (OP)
3: With BW window A in the CP 4: With BW window A in the OP
5: With BW window B in the CP 6: With BW window B in the OP
7: With BW window C in the CP 8: With BW window C in the OP

d_{ij} is the symmetric Itakura distance between LP spectra for the cases i and j.

$d_{13} = 1.027$ $d_{24} = 1.541$ $d_{34} = 3.449$
$d_{35} = 1.012$ $d_{46} = 1.010$ $d_{56} = 3.935$
$d_{57} = 1.117$ $d_{68} = 1.105$ $d_{78} = 3.007$

Fig. 3. Comparison of different bandwidth windows for a natural voiced speech signal. The distances shown below are the symmetric Itakura distances for the cases mentioned in the figure.

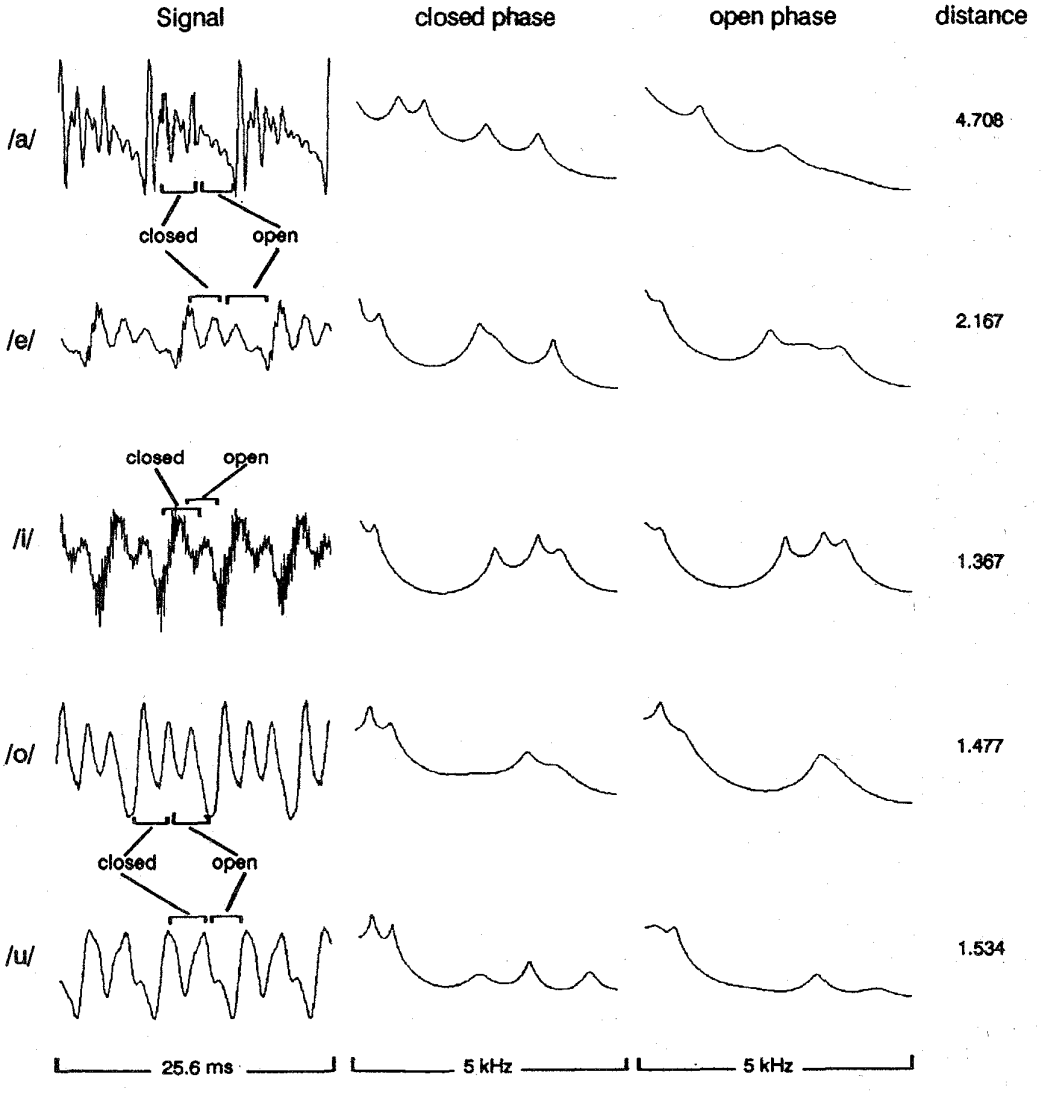

Fig. 4. Analysis of natural vowels using source-system windowing in the closed and open glottis regions of a pitch period. For each vowel the signal waveform, LP spectrum for closed phase, LP spectrum for open phase and the symmetric Itakura distance between these LP spectra are given along a row.

that synthesizing speech using separate LPC's for open and closed glottis regions produces a more natural sounding speech compared to the conventional LPC synthesis.

In this correspondence, we have shown that, using suitable residual and bandwidth windows for source and system components of a speech signal, it is possible to derive the characteristics of the vocal

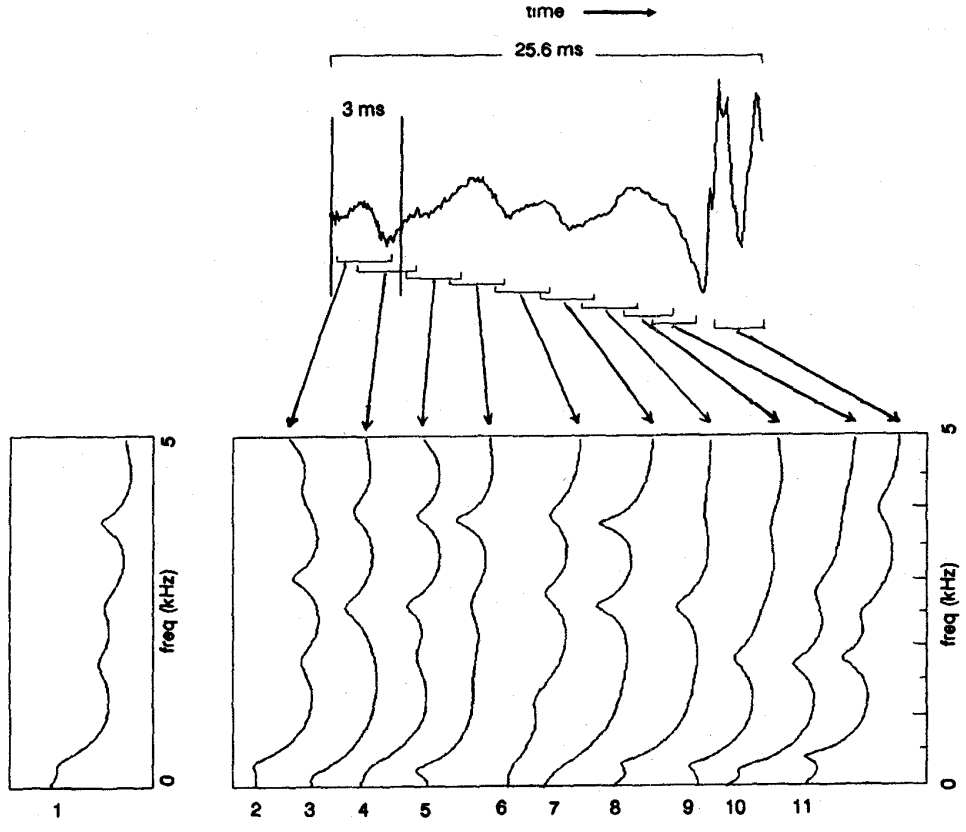

Fig. 5. Analysis of a CV transition region /ca/ using source-system windowing. The spectrum indicated as 1 is the conventional tenth order LP spectrum using a 25.6-ms Hamming window. The LP spectra labeled $2, 3, \cdots, 11$ are obtained using source-system windowing.

tract system in the closed and open glottis regions within each pitch period. Thus, to some extent, source-system windowing overcomes the limitations of short window analysis. This type of representation of the vocal tract system may help in generating natural sounding synthetic speech. However, the performance of this analysis depends on the positioning of the window on the residual signal. If the system characteristics change significantly within the analysis window, then it is difficult to interpret the results.

REFERENCES

[1] D. O' Shoughnessy, *Speech Communication—Human and Machine.* New York: Addison-Wesley, 1987.
[2] A. K. Krishnamurthy, "Glottal source estimation using a sum-of-exponentials model," *IEEE Trans. Signal Processing*, vol. 40, no. 3, pp. 682–686, Mar. 1992.
[3] J. Makhoul, "Linear prediction: A tutorial review," in *Proc. IEEE*, Apr. 1975, vol. 63, pp. 561–580.
[4] G. Fant, "Glottal flow: models and interaction," *J. Phonetics*, vol. 14, pp. 393–399, 1986.
[5] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals.* Englewood Cliffs, NJ: Prentice-Hall, 1978.
[6] A. A. Giordano and F. M. Hsu, *Least Square Estimation With Applications to Digital Signal Processing.* New York: Wiley, 1985.
[7] S. L. Marple Jr., *Digital Spectral Analysis With Applications.* Englewood Cliffs, NJ: Prentice-Hall, 1987.
[8] B. Yegnanarayana and P. Satyanarayana Murthy, "On windowing speech data for analysis," in *Proc. III ICAPRDT, ISI*, Calcutta, India, Dec. 1993, pp. 334–345.
[9] B. Yegnanarayana, P. Satyanarayana Murthy, and J. H. Eggen, "Source-system windowing for speech analysis," Inst. for Perception Research, Eindhoven, The Netherlands, no. 28, pp. 53–58, 1993.
[10] P. E. Papamichalis, *Practical Approaches to Speech Coding.* Englewood Cliffs, NJ: Prentice-Hall, 1987.
[11] D. H. Klatt and L. C. Klatt, "Analysis, synthesis and perception of voice quality variations among male and female talkers," *J. Acoust. Soc. Amer.*, vol. 87, no. 2, Feb. 1990.
[12] G. Fant, "Some problems in voice source analysis," *Speech Commun.*, vol. 13, no. 1–2, pp. 7–22, Oct. 1993.
[13] D. G. Childers and C. F. Wong, "Measuring and modeling vocal source-tract interaction," *IEEE Trans. Biomed. Eng.*, vol. 41, no. 7, pp. 663–671, July 1994.