ZERO-PHASE INVERSE FILTERING FOR EXTRACTION OF SOURCE CHARACTERISTICS

T.V. Ananthapadmanabha and B. Yegnanarayana

Department of Electrical Communication Engineering
Indian Institute of Science
Bangalore-560012

ABSTRACT

The instants at which significant excitation of vocal tract take place during voicing are referred to as epochs. Epochs and strengths of excitation pulses at epochs are useful in characterizing voice source. Epoch filtering technique proposed by the authors determine epochs from speech waveform. In this paper we propose zero-phase inverse filtering to obtain strengths of excitation pulses at epochs. Zero-phase inverse filter compensates the gross spectral envelope of short-time spectrum of speech without affecting phase characteristics. Linear prediction analysis is used to realize the zero-phase inverse filter. Source characteristics that can be derived from speech using this technique are illustrated with examples.

I. INTRODUCTION

Extraction of voice source information from speech is an important problem in speech analysis. It is generally difficult to obtain glottal pulse shape from speech waveform. Alternatively one can think of parameters characterising voice source which can more easily be extracted from speech waveform. Epoch, the instant of significant excitation of vocal-tract cavity, is one such parameter [1]. Several techniques have been proposed for identifying epochs [2-4]. Source characteristics can more accurately be described by considering epochs as well as strengths of impulses at epochs. Zero-phase inverse filtering technique is proposed to obtain these characteristics. The technique is illustrated with examples of vowel sounds.

II. EPOCH CHARACTERISATION

Consider a finite duration pulse $g(t)$. Let $g^{(n)}(t)$ be the nth derivative of $g(t)$. $g^{(n+1)}(t)$ will contain impulses at the instants of discontinuities of $g^{(n)}(t)$. If these impulses are removed from $g^{(n+1)}(t)$ and the residual signal differentiated, $g^{(n+2)}(t)$ will contain impulses at discontinuities of the residual $g^{(n+1)}(t)$. Let the instant at which residual $g^{(i)}(t)$ is discontinuous be denoted by t_{ij}, where j is the serial number of the discontinuity. Then Laplace transform of the pulse $g(t)$ can be expanded into a polynomial in s^{-1} as,

$$G(s) = \sum_{i=1}^{\infty} E_i(s)/s^i \qquad (1)$$

where

$$E_i(s) = \sum_j g^{(i)}(t_{ij}) e^{-st_{ij}} . \qquad (2)$$

A single pitch period of glottal pulse is a finite duration signal. It is not necessary to assume glottal flow to be zero beyond closure. Then Laplace transform of glottal pulse can be expressed in the form (1). A glottal pulse cannot have a waveform discontinuity as otherwise it would imply infinite velocity of vocal-folds. Hence $E_1(s)$ is identically zero. Eq. (1) reduces to

$$G(s) = \sum_{i=2}^{\infty} E_i(s)/s^i . \qquad (3)$$

According to source-system model [5] voiced speech is considered as the response of vocal-tract system $V(s)$ to quasi-periodic sequence of glottal pulses (Fig. 1a). Laplace transform $S(s)$ of voiced speech signal $s(t)$ is given by

$$S(s) = V(s)G(s) = \sum_{i=2}^{\infty} V(s)E_i(s)/s^i$$

$$= \sum_{i=2}^{\infty} V_i(s)E_i(s) \qquad (4)$$

where

$$V_i(s) = V(s)/s^i \quad . \qquad (5)$$

From Eq.(4) we get

$$s(t) = \sum_i v_i(t) * e_i(t). \qquad (6)$$

According to Eq. (6) s(t) is a superposition of responses of systems $V_i(s)$ to excitation signals $e_i(t)$ which contain only impulses. A model based on (6) is shown in Fig. 1b. Defining $e_i(t)$ as i th order epoch signal, it follows from (5) that higher order epochs become less significant at higher frequencies. Ideally voice source is completely characterised by the set $\{e_i(t), i=2, \infty\}$. Generally $e_2(t)$ adequately represents the glottal pulse except for a fixed second order roll-off. When there are no slope discontinuities $e_2(t)$ will be zero and $e_3(t)$ can be used to specify voice source, and so on. The instants of slope discontinuities together with the value of impulses at these instants are referred to as epoch characteristics.

The adequacy of representation of a pulse g(t) by its epoch characteristics is illustrated by considering a few typical glottal pulse shapes [6]. The glottal pulse, its second derivative and its epoch characteristics are given in Fig. 2 for three cases. The weighted log spectrum of g(t) and log spectrum of $e_2(t)$ are also shown in Fig. 2. The log-spectrum of g(t) is given a 12 db/octave high frequency emphasis to compare it with the log-spectrum of $e_2(t)$. It can be seen that the spectral characteristics of g(t) are adequately represented by $e_2(t)$. The mean-square error between $g^{(2)}(t)$ and $e_2(t)$ normalised with respect to energy of the pulse g(t) for the three cases are .00003, .000022, 0.0. Thus the error caused by omitting higher derivatives appears to be negligible. Also, this error is mostly concentrated within 100 Hz in the spectrum.

III. ZERO-PHASE INVERSE FILTERING

Epoch filtering technique has been proposed by the authors for identification of epochs [4]. Epoch filter computes Hilbert envelope of band-pass filtered signal. The band-pass filtering is realised using a frequency domain window of about 1.25 KHz width located in the high frequency range. The output of epoch filter gives peaks at epoch locations. However, the amplitudes of peaks at epochs depend on the vocal-tract

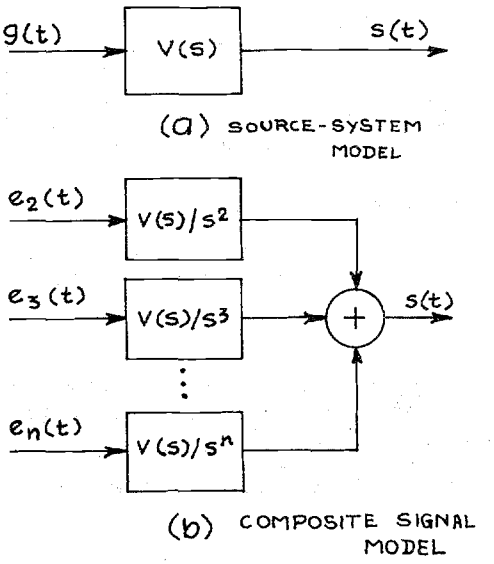

Fig.1

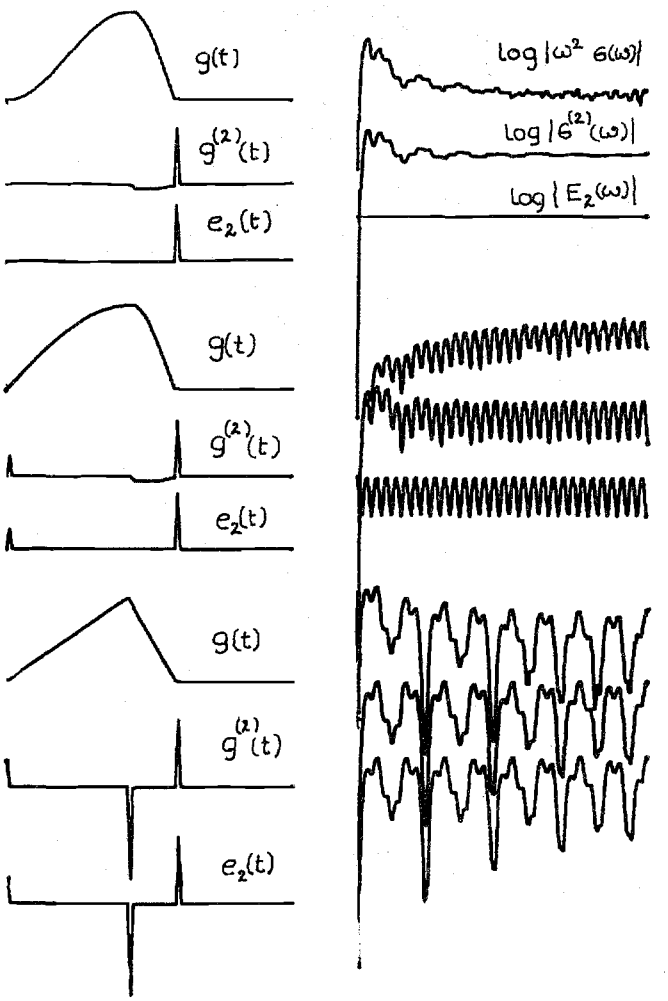

Fig.2

transfer function and spectral roll-off due to glottal pulse. Hence the amplitudes of peaks at epochs cannot be related to source. According to the model proposed the Fourier transform of speech signal can be written as

$$S(w) \simeq E_2(w)\ V(w)/(jw)^2$$

$$\simeq \left|\frac{V(w)}{w^2}\right| \left(\sum_k a_k e^{-jwt_k}\right) e^{j\phi_v(w)} \qquad (7)$$

where $\phi_v(w)$ is the phase transform of $V(s)$ and a_k and t_k are used in place of $g^{(2)}(t_{ij})$ and t_{ij} respectively. Assume that speech signal is passed through a filter

$$H(w) = \left|\frac{V(w)}{w^2}\right|^{-1} \qquad (8)$$

to obtain the output $r(t)$. If the epoch filtering is performed on $r(t)$, the amplitude of peaks at epochs is independent of $|V(w)/w^2|$.

$H(w)$ is an inverse filter to the vocal-tract system with zero-phase characteristics. It may be recalled that the gross-spectral envelope of short-time spectrum of speech signal corresponds to vocal-tract system. It is well known that linear prediction (LP) technique can be used as a whitening filter [7]. The LP technique uses only the spectral information of speech signal. The phase characteristics of the digital inverse filter corresponds to a minimum-phase signal. In general, the inverse filter cannot be expected to compensate for phase characteristics of vocal-tract system, thus introducing additional phase changes at the output of digital inverse filter. However zero-phase digital inverse filter (ZPDIF) does not affect the phase-characteristics of input speech. Analysing a four resonator model, it has been found that retaining the original phase characteristics do not affect the amplitudes of peaks at the output of epoch filter for $r(t)$. The pass-band of epoch filter operating on $r(t)$ is chosen to be between 2 and 4 KHz. Since the first formant frequency usually has a sharp peak, ZPDIF does not exactly cancel the first formant. As a consequence, first formant ripples affect the output if this region is enclosed in the pass-band. The frequency range beyond 4 KHz is not used to avoid noise and aliasing errors. It may be noted that no low pass filtering was used prior to sampling since the fast transition of the filter affects stability. The block diagram of the method is shown in Fig. 3.

IV. EXAMPLES

Waveform of vowel sound /a/, the estimated glottal pulse shape, squared LP error, and epoch filter output of whitened signal for two speakers are given in Figs. 4 and 5. It is observed that the main excitation occurs at peaks of glottal flow in both cases. For case A, there is also an excitation at closure. The relative strengths of excitation at peak and closure are in the ratio 1.3:1.

The differentiated waveform, the reconstructed waveform obtained using LPCs and epoch characteristics and the reconstructed waveform by conventional LP synthesis are shown in Fig. 6. It may be noted that by using epoch characteristics a waveform nearly resembling the original signal is obtained.

V. CONCLUSIONS

It is known that LP residual excited vocoder gives high quality speech. This is because information pertaining to multiple excitations are retained in the LP residual. However source parameters as described in this paper can also be used in vocoders. Usually the major excitation is assumed to take place at glottis closure. However we have observed several cases in these studies that major excitation occurs at the peak of glottis opening rather than at the closure. The technique can be used for studying variations in glottal pulse characteristics for continuous speech and for studies relating to voice disorders.

REFERENCES

1. J.L. Flanagan, Speech Analysis Synthesis and Perception, New York: Springer-Verlag, 1965; Second ed.1972,Ch.5,p.186.

2. J.N. Holmes, 'An investigation of the volume-velocity waveform at larynx during speech by means of an inverse filter,' in Proc. Stockholm Speech Commn.Seminar, Royal Inst.Technol., Stockholm, Sweden, Sept. 1962.

3. H.W. Strube, 'Determination of the instant of glottal closure from the speech wave,' J.Acoust.Soc.Amer.,Vol. 56, pp. 1625-1629, Nov.1974.

4. T.V. Ananthapadmanabha and B. Yegnanarayana, 'Epoch Extraction of Voiced Speech,' IEEE Trans. Acoust. Speech and Signal Processing, Vol.ASSP-23, pp. 562-570, Dec. 1975.

5. G. Fant, Acoustic Theory of Speech Production, S-Gravenhage: Mounton and Co., 1960.

6. A.E. Rosenberg, 'Effect of glottal pulse shape on the quality of natural vowels,' J.Acoust.Soc.Amer., Vol.49, pp. 583-590,Feb. 1971.

7. J. Makhoul, 'Linear Prediction: A Tutorial Review,' Proc. IEEE, Vol. 63, pp. 561-580, Apr. 1975.

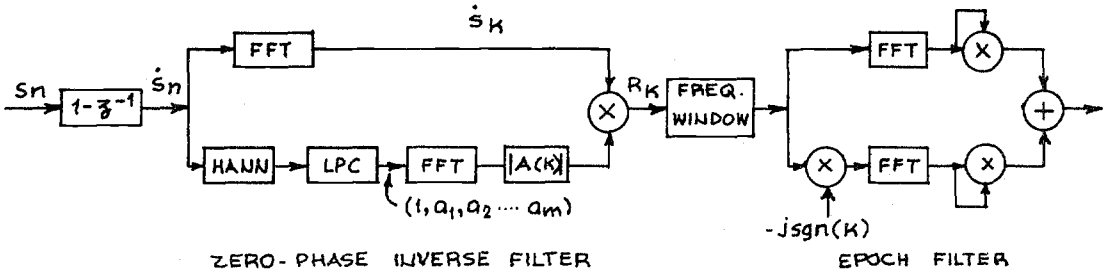

Fig.3

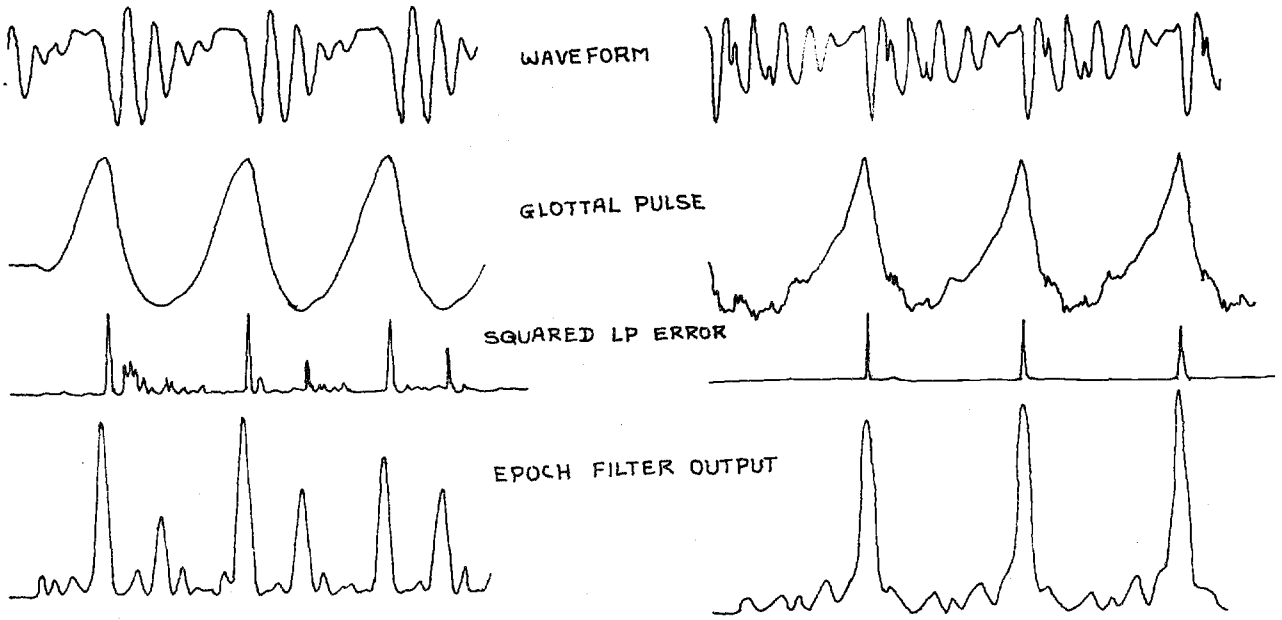

Fig.4 CASE A

Fig.5 CASE B

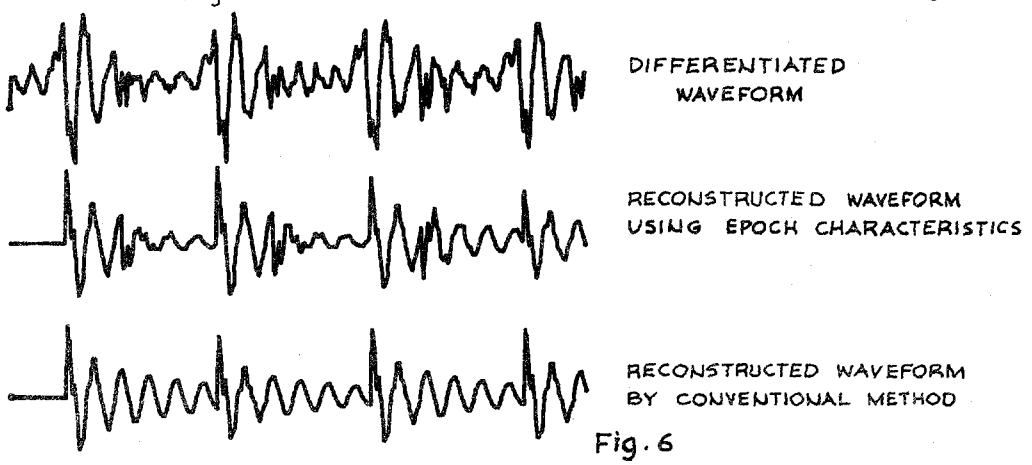

Fig.6