PROCESSING OF NOISY SPEECH USING MODIFIED GROUP DELAY FUNCTIONS

B. Yegnanarayana, Hema A. Murthy and V. R. Ramachandran
Department of Computer Science and Engineering
Indian Institute of Technology, Madras - 600 036, India.

ABSTRACT

The objective of this paper is to present a new method of processing noisy speech. The method exploits the properties of the negative derivative of the Fourier transform (FT) phase spectrum (group delay function) to derive the features of the vocal tract system and the excitation from the speech signal. The key idea used is that the properties of group delay functions for noise and a stable all-pole filter are distinct. Estimation of the spectrum of the vocal tract system and fundamental frequency (F_0) are treated as problems of spectrum estimation from noisy data. The component due to noise is suppressed in the group delay function to obtain reliable estimates for vocal tract response spectrum and F_0. Results of our studies show that intelligible speech can be synthesised from parameters derived from noisy data with an overall signal-to-noise (SNR) as low as 3dB.

INTRODUCTION

Processing of speech involves the estimation of (a) the time-varying parameters of the vocal tract system and (b) the parameters corresponding to the source excitation. The source parameters are Voiced/Unvoiced decisions and the frequency F_0 of the fundamental. The system parameters are the location of the resonances of the vocal tract (formants) and their bandwidths or the vocal tract response spectrum. The problem of estimating parameters from the speech signal is compounded by the fact that speech signal is always corrupted by noise. In this paper estimation of both source and system parameters from the noisy speech signal is treated as a problem of spectrum estimation from noisy data[1,2].

Our research objective is to show that the FT phase of a signal contains significant information which can be extracted by suitable processing [3]. It is possible to extract information about the embedded sinusoids or autoregressive process in noise by using the FT phase also. In fact the FT phase may resolve some uncertainty in the spectral information derived from the FT magnitude, or may give some new information which cannot be derived from the FT magnitude. Note that in general, the FT phase is independent from the FT magnitude.

There are many important features of phase which can be exploited for processing in many applications [4,5,6,7,8,9]. The most important feature is the additive property of the phase. The characteristics of the FT phase and its relation to the FT magnitude using group delay functions (negative derivative of the phase function) are discussed in [3]. Both the FT phase and group delay function are not directly amenable for spectrum estimation. Noise and window effects contribute large changes in the FT phase values. The true phase values are masked in the computed phase due to the inevitable wrapping. Although the computed group delay function does not suffer from the wrapping problem, large phase changes manifest as spikes in the computed group delay function. The required spectral information is buried in the phase or group delay function. As the objective in spectrum estimation is the estimation of the parameters that characterise (a) sinusoidal components of the signal of interest or (b) autoregressive process, we show that the required information can be estimated by reducing the effects of noise.

In this paper, the problem of estimation of parameters corresponding to the vocal tract is treated as a problem of the estimation of the parameters that characterise an autoregressive process in noise. Likewise the problem of estimation of the fundamental frequency (F_0) is treated as a problem of the estimation of sinusoids in noise. Since the characteristics of the source and system vary continuously, it is necessary to process the signal over short durations (10-30msec). The problem is further complicated by the combined effects of (a) additive noise (b) side lobe effects in the frequency domain, (c) finite duration and (d) time varying characteristics of source and system. Methods based on group delay functions are proposed to determine the vocal tract response spectrum and fundamental frequency F_0 from a segment of voiced speech. The key idea is that properties of the group delay functions for noise and a stable all-pole filter are distinct. The most important property of noise used in our study is that the noise and signal samples are uncorrelated. In addition, additive noise introduces spectral zeros close to the unit circle in the z-plane. These properties are used to derive a modified group delay function [10,11] where the effects of noise and finite duration are suppressed.

CH2977-7/91/0000-0945 $1.00 © 1991 IEEE

In Section II we discuss the basis for the proposed method of analysis of noisy speech. We present a method for estimation of the vocal tract response spectrum from noisy speech in Section III and a method for pitch extraction in Section IV. Finally in Section V we discuss a procedure for synthesis of speech from parameters extracted from noisy speech.

II. BASIS FOR THE PROPOSED METHOD

In this Section we discuss the basis for the proposed method for extracting the characteristics of the vocal tract system from noisy speech. The characteristics that we are looking for are the spectral features corresponding to the resonances of the vocal tract system. For the time being we ignore the effects of data windows. Assuming a source-filter model (see Fig.1) for speech production, the z-transform representation of speech data x(n) is given by

$$X(z) = \frac{GE(z)}{A(z)} + U(z) = \frac{V(z)}{A(z)} \qquad (1)$$

The corresponding frequency domain representation is given by

$$X(\omega) = \frac{GE(\omega)}{A(\omega)} + U(\omega) = \frac{V(\omega)}{A(\omega)} \qquad (2)$$

$$\log X(\omega) = \log V(\omega) - \log A(\omega)$$

$$= \log|V(\omega)| - j\theta_V(\omega) - \log|A(\omega)| + j\theta_A(\omega) \qquad (3)$$

where

$$V(\omega) = V(z)\big|_{z=e^{j\omega}} = |V(\omega)|e^{-j\theta_V(\omega)} \qquad (4)$$

and

$$A(\omega) = A(z)\big|_{z=e^{j\omega}} = |A(\omega)|e^{-j\theta_A(\omega)} \qquad (5)$$

Since all the coefficients in V(z) and A(z) are real, we can write

$$V(z) = V_1(z).V_2(z). \ \ldots \ .V_q(z) \qquad (6)$$

and

$$A(z) = A_1(z).A_2(z). \ \ldots \ .A_p(z) \qquad (7)$$

where each of these $V_i(z)$ and $A_i(z)$ are either first order or second order polynomials with real coefficients. The roots of V(z)=0 may be either inside or outside the unit circle. V(z) is mainly contributed by noise or periodic sequence, since in frequency regions where $|A(\omega)|$ is small $GE(\omega)$ dominates, and in frequency regions where $|A(\omega)|$ is large, $U(\omega)$ dominates [9]. Therefore the roots of V(z) are distributed randomly around the unit circle with most of the roots close to the unit circle. Roots close to the unit circle in the z-domain produce sharp spikes in group delay function $\tau_V(\omega)$ of $V(\omega)$.

On the other hand, all the roots of A(z) are well inside the unit circle in the z-plane, and they contribute to a relatively smooth curve in the group delay function $-\tau_A(\omega)$ of $1/A(\omega)$. The location of the peaks in the group delay function $-\tau_A(\omega)$ correspond to resonances [4]. Note also that, due to the additive nature of the group delay functions, the influence of one resonance peaks on the other is negligible [4].

The group delay functions of the impulse response of an all-pole system and noise sequence are shown in Figs.2 and 3. In the group delay function (Fig.4) of the combined response of noise and an all-pole system, the characteristics of the system are completely masked by the large spikes due to noise. The combined response was obtained by convolving the excitation signal with the impulse response of the system. In the combined response the spikes in the group delay function are due to the excitation function. In the context of additive noise, the spikes may be due to (i) excitation alone or (ii) noise alone or (iii) both excitation and noise.

To extract vocal tract (VT) information from the group delay function $\tau_X(\omega)$ of $X(\omega)$, we have to suppress the effects of spikes in the group delay function due to $V(\omega)$. To do this, it is necessary to know the locations of the roots of V(z). Since at these roots $|X(\omega)|^2$ has sharp nulls, we can derive a spectrum $|V(\omega)|^2$ (called zero spectrum) having an approximately flat spectral envelope, and multiply the group delay function $\tau_X(\omega)$ with $|\hat{V}(\omega)|^2$ to obtain an estimate of the group delay function corresponding to $1/A(\omega)$. The resulting modified group delay function shows peaks corresponding to the resonances and can be used to obtain an estimate of VT response spectrum[11].

III. ESTIMATION OF THE VOCAL TRACT SYSTEM RESPONSE FROM MODIFIED GROUP DELAY FUNCTION

In this Section we describe a procedure to compute the modified group delay function from which the vocal tract system response can be estimated. Given a segment of speech signal, x(n), n = 0,1,...,N-1, the group delay function is computed as follows:

Let X(k) and Y(k) be the discrete Fourier transforms of the sequences x(n) and nx(n), respectively. The samples of the group delay function are given by [7]

$$\tau_X(k) = \frac{X_R(k)Y_R(k) + X_I(k)Y_I(k)}{X_R^2(k) + X_I^2(k)} \qquad (9)$$

where the subscripts R and I refer to the real and imaginary parts, respectively.

Let $|\hat{V}(k)|^2$ be an estimate of the the zero spectrum. This is derived by flattening the magnitude spectrum either by linear prediction or by cepstrum analysis. The modified group delay function is given by [11]

$$\hat{\tau}(k) = \tau_x(k)|\hat{V}(k)|^2 \qquad (10)$$

The envelope of $\hat{\tau}(k)$ contains peaks corresponding to the resonances of the vocal tract system.

Fig.5 shows a segment (25.6 msec) of voiced speech and Fig.6 the corresponding modified group delay function. Fig.6 also shows the smoothed modified group delay function. Cepstral coefficients are estimated from the modified group delay function. The cepstral coefficients are then used to obtain an estimate of the vocal tract response spectrum (Fig.7). Figs.9 and 10 show the modified group delay function and estimated vocal tract response spectrum for a noisy segment (Fig.8) of speech signal (overall SNR = 3dB).

IV. PITCH EXTRACTION FROM MODIFIED GROUP DELAY FUNCTION

In this Section we describe a procedure to obtain pitch information from noisy speech. The periodic glottal pulse excitation in voiced segments manifests as a sinusoid in the frequency domain. In other words, the magnitude spectrum $|X(\omega)|^2$ of a voiced segment contains a sinusoidal component corresponding to pitch, besides peaks due to the VT response and random fluctuations due to additive noise. The problem of pitch extraction is simply estimation of the frequency of the sinusoids in $|X(\omega)|^2$ in the presence of the distortions and noise. The high signal-to-noise ratio (SNR) portions of $|X(\omega)|^2$ can be taken as the signal and the modified group delay function of the same can be computed. The distance between peaks in the modified group delay function corresponds to the pitch period (or $1/F_0$). Fig.11 shows the smoothed modified group delay function for the zero spectrum of the voiced segment of Fig.5. Fig.12 shows the smoothed modified group delay function of the zero spectrum of the noisy segment of Fig.8. Generally we have noticed that in the unvoiced and noisy regions the locations of the peaks are distributed randomly in successive frames. The gain can be derived from noisy speech using the high SNR regions of $|X(\omega)|^2$.

V. SPEECH SYNTHESIS

Using the F_0 information obtained in Section IV and the vocal tract information in Section III the speech signal is synthesised. Informal listening tests show that the synthetic speech signal generated is noise free. Although intelligible, there is a significant degradation of naturalness.

CONCLUSION

In this paper a new procedure of analysis and synthesis of noisy speech is presented. Properties of group delay functions are used to estimate the vocal tract system response and fundamental frequency corresponding to the excitation. We have been able to reduce the noise significantly with some degradation in the quality of synthesised speech.

References

[1] S.M.Kay, *Modern Spectrum Estimation,* Englewood Cliffs, NJ:Prentice Hall, 1988.

[2] S.L. Marple, *Digital Spectral Analysis,* Englewood Cliffs, NJ:Prentice Hall,1987.

[3] B.Yegnanarayana, D.K.Saikia and T.R.Krishnan, "Significance of Group Delay functions in Signal Reconstruction from Spectral Magnitude or Phase", Vol.ASSP-32, No.3,pp.610-623, June,1984.

[4] B.Yegnanarayana, "Formant extraction from linear prediction phase spectra", JASA, Vol.63, pp.1638-1640, 1978.

[5] B.Yegnanarayana and D.Raj Reddy, "A Distance Measure Based on the Derivative of Linear Prediction Phase Spectrum", *Proc.* ICASSP-79, pp.744-747, 1979.

[6] B.Yegnanarayana, "Design of ARMA Digital Filters by pole-zero decomposition", Vol.ASSP-29, No.3, pp.433-439, June, 1981.

[7] B.Yegnanarayana, J.Sreekanth and Anand Rangarajan, " Waveform Estimation Using Group Delay Processing", Vol.ASSP-33, No.4, pp.832-836, August, 1985.

[8] B.Yegnanarayana, George Duncan and Hema A. Murthy, " Improving Formant Extraction From Speech Using Minimum Phase Group Delay Spectra", *Proc.* EUSIPCO-88, pp.447-450, 1988.

[9] B.Yegnanarayana, Hema A.Murthy and V.R.Ramachandran, "Speech Enhancement Using Group Delay functions", *Proc.* ICSLP-90, 1990.

[10] Hema A.Murthy K.V.Madhu Murthy and B.Yegnanarayana, "Formant Extraction from Phase Using Weighted Group Delay Function", *Electronic Letters,* Vol.25, No.23, pp.1609-1611, 1989.

[11] Hema A.Murthy and B.Yegnanarayana, "Speech processing using group delay functions", (to appear in *Signal Processing,* Vol.22, No.3, 1991).

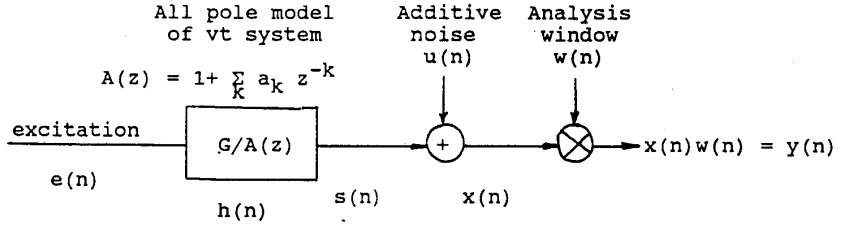

Fig.1 Model of speech data generation

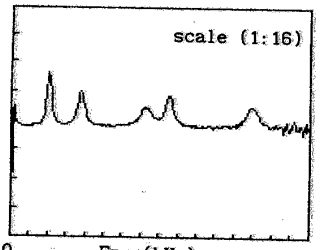

Fig.2 Group delay function of an all-pole system.

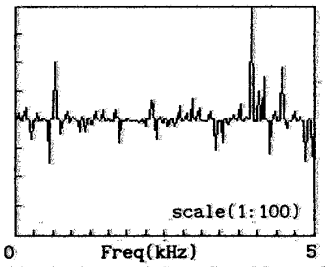

Fig.3 Group delay function of random noise.

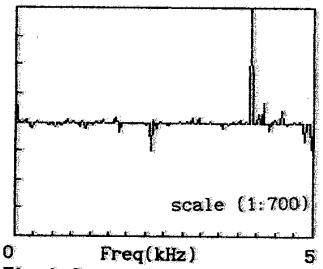

Fig.4 Group delay function of combined response of all-pole system and random noise.

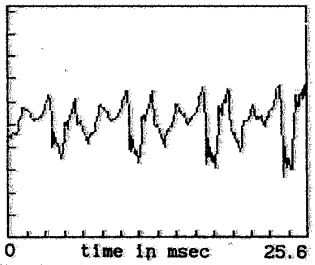

Fig.5 A segment of voiced speech.

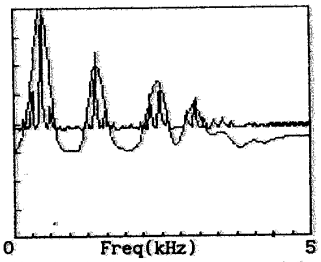

Fig.6 Modified group delay function for Fig.5.

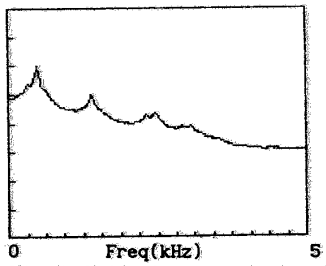

Fig.7 Estimated vocal tract spectrum.

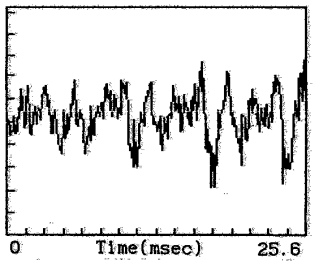

Fig.8 A segment of noisy speech.

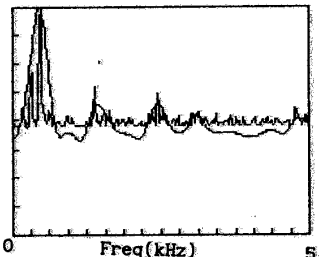

Fig.9 Modified group delay function for Fig.8.

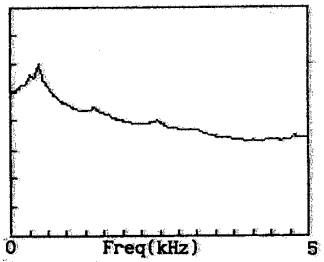

Fig.10 Estimated vocal tract spectrum.

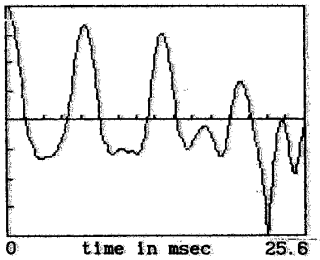

Fig.11. Smoothed modified group delay function of the zero-spectrum of voiced segment of Fig.5.

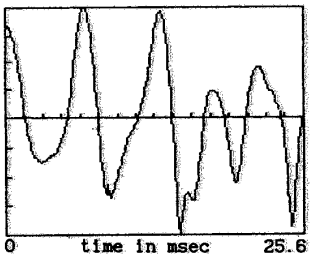

Fig.12. Smoothed modified group delay function of the zero-spectrum of voiced segment of Fig.8.