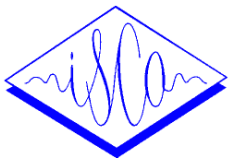

First International Conference on
Spoken Language Processing (ICSLP 90)
Kobe, Japan
November 18-22, 1990

SPEECH ENHANCEMENT USING GROUP DELAY FUNCTIONS

B. Yegnanarayana, Hema A. Murthy and V. R. Ramachandran
Department of Computer Science and Engineering
Indian Institute of Technology, Madras-600036, India.

ABSTRACT

A method of processing noisy speech to enhance spectral features corresponding to the vocal tract system is presented. A new method of extracting pitch from noisy speech data is also presented. These methods depend on processing the Fourier transform phase through group delay functions. The basic idea used in these methods is that the features of noise and an all-pole system are distinct in the group delay function. To enhance the spectral features of the vocal tract system a modified group delay function is derived by suppressing the features corresponding to noise from the standard group delay function. Similar ideas are used to extract the periodic component corresponding to pitch from the magnitude spectrum. Performance of these methods is demonstrated for noisy speech data.

I. INTRODUCTION

One of the major problems in speech analysis is extraction of features corresponding to the vocal tract system and excitation. This problem is difficult because of the fact that the characteristics of both the system and excitation source are varying with time during production of speech. The problem is compounded by the fact that speech signals are also corrupted by noise. Many successful methods have been developed to process clean speech data. The most effective method is based on an all pole model for the vocal tract system. Performance of speech analysis techniques based on this approach degrades significantly when the speech data is noisy. Techniques based on all pole modelling of degraded speech are only partially successful[1]. More general pole-zero models are difficult to use for analysis. For processing noisy speech, methods based on direct spectral subtraction were proposed in the literature[2]. While performance of these methods have been demonstrated on short individual segments of data, significant empirical manipulation is required to process continuous speech. Moreover, most of these methods use long term statistics of noise, either assumed or computed, for reducing the effects of noise. For a quasistationary signal like speech, it is difficult to take long enough data without losing the short time spectral details.

The objective of this paper is to present new methods of processing noisy speech signals to extract information corresponding to the vocal tract system and excitation source. The methods are based on the fact that noise and signal components have distinct characteristics even in short(20-30 msec) segments of data. In particular, we exploit the properties of sample to sample uncorrelation for noise and correlation for signal. These properties stand out distinctly in group delay functions. This forms the basis for the ideas presented in this paper[3]. We describe the characteristics of group delay functions of noise and signals in Section II. We derive a modified group delay function to enhance the signal characteristics by reducing the effects of noise. In Section III we show how the modified group delay function can be used to extract the formant information from noisy speech and in Section IV we show how the modified group delay function can be used to derive the pitch information.

II. GROUP DELAY FUNCTIONS

In this section we develop the theory and discuss a technique for enhancing the characteristics of the vocal tract system from noisy speech. The characteristic we are looking for are the formants(resonances) of the vocal tract system and the pitch period of the glottal excitation. We ignore for the time being the effects of data windows.

We define our problem as follows:
Given a noisy signal

$$x(n) = e(n)*h(n) + u(n) \qquad (1)$$

where h(n) is the impulse response of the all-pole system G/A(z) and e(n) is either a periodic train of pulses or random noise sequence, determine the resonances of the all-pole system and periodicities of the excitation signal.

Equation (1) can be expressed in terms of z-transform as

$$X(z) = E(z)H(z) + U(z) \qquad (2)$$

$$H(z) = G/A(z) \qquad (3)$$

$$A(z) = 1 + \sum_{k=1}^{p} a_k z^{-k} \qquad (4)$$

The frequency responce is given by

$$X(\omega) = V(\omega)/A(\omega) \qquad (5)$$

$$V(\omega) = GE(\omega) + A(\omega)U(\omega) \qquad (6)$$

The group delay function is defined as the negative derivative of the Fourier transform(FT) phase of a signal. Let $\tau_x(\omega)$, $\tau_v(\omega)$ and $\tau_A(\omega)$ represent the group delay functions corresponding

to $X(\omega)$, $V(\omega)$ and $A(\omega)$, respectively. Then

$$\tau_X(\omega) = \tau_V(\omega) - \tau_A(\omega) \quad (7)$$

For a given sequence x(n), the group delay function is computed as follows:

Let $X(\omega)$ and $Y(\omega)$ be the Fourier transforms of the sequences x(n) and nx(n), respectively. Then

$$\tau_X(\omega) = \frac{X_R(\omega)Y_R(\omega) + X_I(\omega)Y_I(\omega)}{|X(\omega)|^2} \quad (8)$$

where $X_R(\omega)$ and $Y_R(\omega)$ correspond to the real parts and $X_I(\omega)$ and $Y_I(\omega)$ correspond to the imaginary parts of $X(\omega)$ and $Y(\omega)$, respectively.

The group delay functions of the impulse response of an all-pole system and a noise sequence are shown in Figs. 1 and 2, respectively. The peaks in Fig. 1 correspond to the resonances of the system[4]. This function is generally smoother than the group delay function of a noise sequence (Fig. 2). The group delay function of a noise sequence has its z-transform roots distributed randomly close to the unit circle in the z-plane. The roots may lie both inside and outside the unit circle. In the group delay function(Fig. 3) of the the combined response of noise and the all-pole system, the characteristics of the system are completely masked by the large spikes due to noise. The combined response is obtained by convolving the excitation signal with the impulse response of the system. Even when there is additive noise as in equation (2), we can show that $\tau_V(\omega)$ in (7) is mainly due to large spikes contributed either by excitation or the additive noise or by both[5]. The group delay function $(-\tau_A(\omega))$ due to the all-pole system is masked by the noise term.

In order to bring out the details of the system, the contribution due to noise must be suppressed. This can be done if we know the locations and amplitudes of the spikes. We can take advantage of the behaviour of the FT magnitude spectrum with nearly flat spectral envelope. We call this spectrum as zero spectrum. By multiplying the group delay function $\tau_X(\omega)$ of the signal with the estimated zero spectrum, we obtain an estimate $-\hat{\tau}_A(\omega)$ of the group delay function of the all-pole system. This modified group delay function and its smoothed version for a segment of voiced speech is shown in Fig. 5. The peaks of the smoothed contour correspond to formants[4].

III. FORMANT EXTRACTION FROM NOISY SPEECH

The modified group delay function brings out spectral features even at high noise levels. Therefore by smoothing the modified group delay function we can obtain formant information from noisy speech. We computed the locations of the peaks for successive overlapping frames from an all voiced utterance. The resulting raw formant contour data for the clean signal(Fig. 7) is shown in Fig. 8. By adding random noise we have generated a noisy speech signal(Fig. 10) with an overall signal to noise ratio(SNR) of 3 dB. (The SNR as a function of time is shown by dotted curve in Fig. 12.). The raw formant contour data obtained from the noisy speech signal is shown in Fig. 11. It is interesting to note that most of the formant information is captured in the modified group delay function even when the SNR is low(< 0 dB) for some segments. The random fluctuations in the raw formant data occur in the regions where the segments are either unvoiced or the segments have a low (<-5 dB) SNR values.

IV. PITCH EXTRACTION FROM NOISY SPEECH

In this section we show that the characteristics of the modified group delay function can be used to derive the periodicity(pitch) in the excitation signal. For this it is important to note that for a sinusoidal signal in noise the modified group delay function gives a sharp peak at the frequency of the sinusoid. The width of the peak depends on the width of the time domain signal. Therefore the modified group delay functions can be used to estimate sinusoids in noise. For a voiced speech segment the excitation is periodic with some period T_o. Let us consider the z-transform of two impulses separated by T_o

$$E(z) = 1 + z^{-T_o} \quad (9)$$

Then

$$|E(\omega)|^2 = 2 + 2\cos\omega T_o \quad (10)$$

In the frequency domain $|E(\omega)|^2$ has a periodic component with period $1/T_o$ (pitch frequency). If a zero spectrum, corresponding to the FT magnitude spectrum with a flat spectral envelope, is derived for a voiced speech segment, then the resulting spectrum contains a sinusoidal component with period $1/T_o$. If there is an additive noise component, then pitch estimation from noisy speech can be viewed as estimating the frequency of a sinusoid in noise. We can consider the low frequency(upto 1kHz) portion of the spectrum as a noisy signal and derive a modified group delay function for this signal. The resulting signal will have a peak at time T_o, which corresponds to the pitch period. Fig. 6 illustrates the modified group delay function for the zero spectrum of the voiced speech segment of Fig. 4. The strong peak at the pitch period in the modified group delay shows that pitch can be estimated accurately. Figs. 9 and 12 show the raw pitch contours obtained using the modified group delay functions of the zero spectra for clean and noisy(overall SNR=3dB) speech data, respectively. The figures demonstrate that pitch contours can be obtained from noisy speech data even at high noise levels. Fig. 12 also shows the SNR as a function of time. The pitch period contour has random fluctuations in the regions where the segments are either unvoiced or have low(<-5 dB) SNR values.

V. CONCLUSIONS

We have proposed methods for obtaining formant contours and pitch contours from noisy speech data. The methods are based on the

properties of a modified group delay function which enhances the signal characterstics by reducing the effects of noise. These studies demonstrate that FT phase can be processed through group delay functions. They also demonstrate that separation of signal from noise can be done better through the FT phase function rather than through the FT magnitude function. Currently we are developing an analysis-synthesis system for noisy speech based on the ideas presented in this paper.

REFERENCES

[1] J.S.Lim and A.V.Oppenheim, "Enhancement and bandwidth compression of noisy speech", *Proc. IEEE*, vol.67, no.12, pp.1586-1604, December 1979.

[2] S.Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol.ASSP-27, no.2, pp.113-120, April 1979.

[3] B.Yegnanarayana, D.K.Siakia and T.R.Krishnan, "Significance of group delay functions in signal reconstrunstruction from spectral magnitude or phase," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol.ASSP-32, no.3, pp.610-623, June 1984.

[4] B.Yegnanarayana, "Formant extraction from linear prediction phase spectra", *Journal of Acoustical Society of America*, vol.63, pp.1638-1640, May 1978.

[5] B.Yegnanarayana and Hema A.Murthy, "Spectrum estimation using Fourier transform phase," *Proceedings of Workshop on Signal Processing, Communications and Networking*, IISc, Bangalore, India, pp.44-53, July 23-26, 1990.

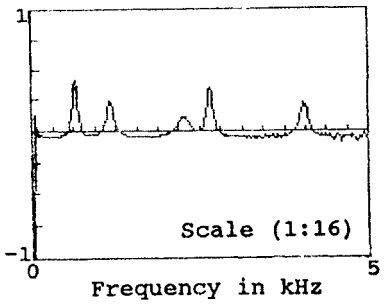

Fig.1. Group delay function for an all-pole system

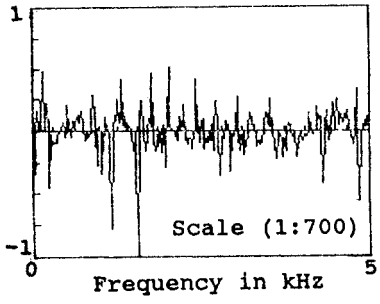

Fig.2. Group delay function for a noise sequence

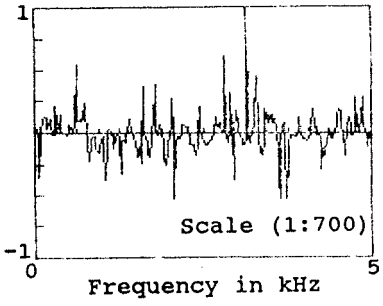

Fig.3. Group delay function of the output of an all-pole system exited by a noise sequence

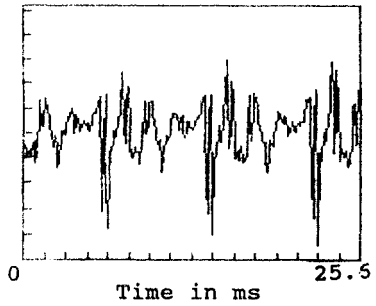

Fig.4. A segment (25.6 ms) of voiced speech

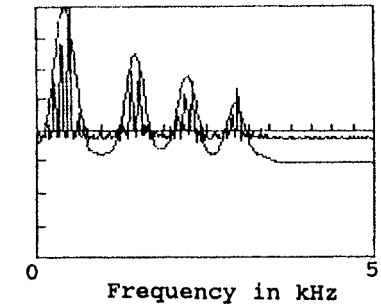

Fig.5. Modified group delay function of the voiced segment given in Fig.4. A smoothed modified group delay function is also plotted in the figure to illustrate the locations of formants.

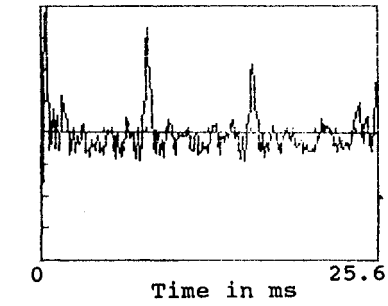

Fig.6. Modified group delay function of the zero spectrum corresponding to the voiced segment given in Fig.4. The location of the peak indicates the pitch period.

Fig.7. Speech signal corresponding to the utterance "We were away a year ago".

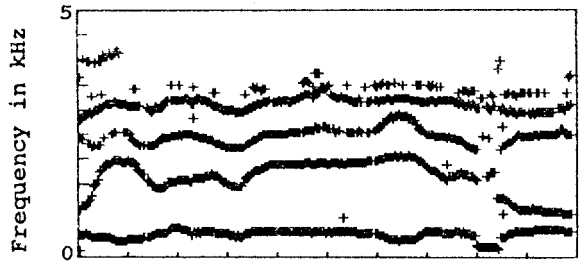

Fig.8. Raw formant data for the signal in Fig.7 derived for each frame (25.6 ms) from the peaks of the smoothed modified group delay function shown in Fig.5.

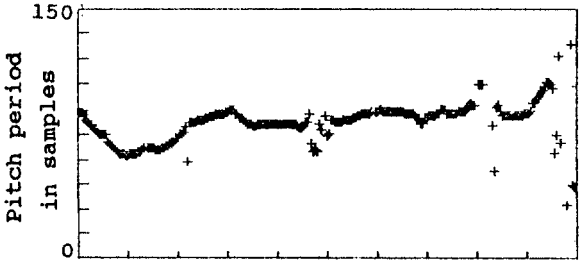

Fig.9. Raw pitch data for the signal in Fig.7. derived for each frame (51.2 ms) from the peaks in the modified group delay function (Fig.6.) of the zero spectrum of the signal.

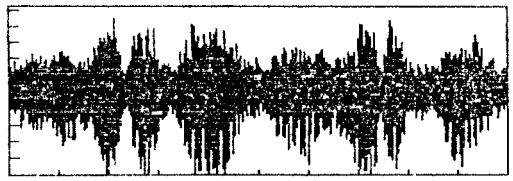

Fig.10. Noisy (overall SNR = 3dB) speech signal corresponding to the utterance "We were away a year ago".

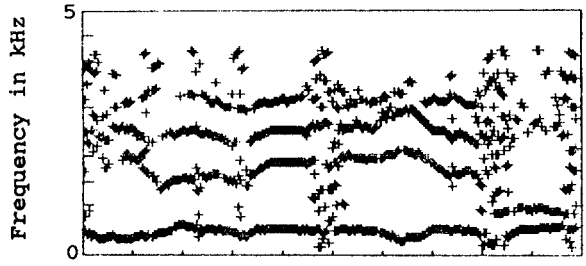

Fig.11. Raw formant data for the signal in Fig.10. derived for each frame (25.6 ms) from the peaks of the modified group delay function shown in Fig. 5. Note that SNR will be different for different frames as shown in Fig.12. below. The regions of random fluctuation of the contour correspond to unvoiced or low SNR (<-5 dB) frames.

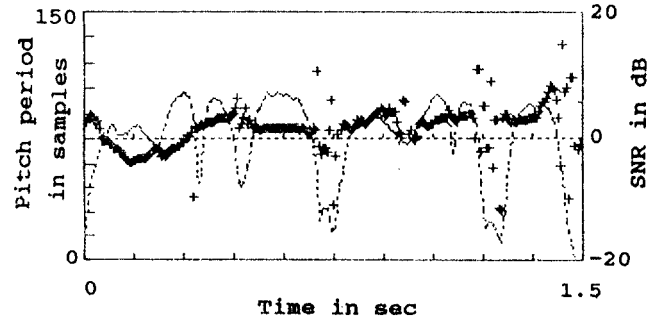

Fig.12. Raw pitch data for the signal in Fig.10 derived for each frame (51.2 ms) from the peaks in the modified group delay function (Fig.6.) of the zero spectrum of the signal. The SNR as a function of time is superimposed over the pitch contour. The regions of random fluctuations of the pitch contour correspond to unvoiced or low SNR (<-5 dB) frames.