NOISE-ININVARIANT REPRESENTATION FOR SPEECH SIGNALS
Aruna Bayya* and B. Yegnanarayana**
* Rockwell Semiconductor Systems, Newport Beach, CA, USA.
** Indian Institute of Technology, Madras, INDIA.

ABSTRACT
A new group-delay based spectral domain is explored for representation of speech signals and for extraction of robust features. The spectrum is computed using the group-delay functions defined on the autocorrelation of a short segment of speech. The features derived from this spectrum are easy to compute and are robust to the background noise. The invariance of the spectral shape to noise in this domain is demonstrated by comparing the group-delay spectrum to the Discrete Fourier transform (DFT) based spectrum and the LPC-derived spectrum. The new domain representation can be applied for parameter estimation as well as speech recognition. In this paper we present preliminary results of using such features in Speaker-Dependent (SD) as well Speaker-Independent (SI) recognition systems.

1. INTRODUCTION
Mismatches between expected operating environment and the actual operating environment are one of the challenges for developing a robust speech processing system. The mismatch may be introduced by various sources, the background noise being a prominent one.

Several approaches have been proposed to deal with the background noise in speech processing systems. Success of these approaches depends on the objective. For the purpose of quality enhancement, several techniques including variations of spectral subtraction [1] and adaptive filtering of signal in the time domain and filtering of time trajectories of the spectral band energies have been proposed [2]. More recent approaches process the information in several domains [3]. All these techniques manipulate the power spectrum and hence overestimation or underestimation may either introduce distortion or may not reduce the ill-effects of noise effectively.

Pre-processing approaches such as spectral subtraction, high-pass filtering and band-pass filtering, feature normalization techniques such as cepstral weighting, cepstral mean subtraction and transformation of feature space, and model adaptation techniques such as parameter transformation have been suggested for reducing the mismatch in speech recognition applications [4].

While all the above-mentioned approaches are being applied to speech recognition as well as quality enhancement, the best approach to handle mismatches is to select proper domain for representation of signals. Using features that are invariant to noise will not only save additional processing, but also will eliminate sources of side effects and distortions.

Current methods of representation involve either approximation in power spectral domain (LPC), projection of power spectrum onto a more physically meaningful space (MEL) or transforming the power spectrum to match physical models in these projected spaces (PLP). Each of these approaches has been optimized for a given task such as speech recognition. However, none of them have proven to be completely noise resistant.

In this paper, we explore a different domain for representing a speech signal. We show that in this domain, the important features of the envelope of the spectrum are more pronounced than the other representations even in the presence of noise. Hence we contend that it is more robust to background noise than the previously proposed representations for speech recognition. Reconstruction of signals from robust representation also will result in better quality enhancement.

In our previous work [5], we have shown that analysis using group-delay functions results in features that are useful in robust recognition of speech. The current analysis, also a group-delay based approach, is performed on shorter segments of speech. It also enhances the peaks and valleys of the spectrum and makes it immune to additive noise [6]. It requires only few simple steps to obtain the new representation of the signal and to derive the features that are useful in speech recognition.

In the remaining of this paper, we show how the new representation is derived from the signal. We also discuss some of the implementation issues before providing some sample applications.

2. GROUP-DELAY FUNCTION AND ITS COMPUTATION
Due to its additive and high resolution properties [6,7], the negative derivative of a sequence preserves the spectral features very well. Hence, we propose to represent the signal in the group-delay (GD) domain in place of power spectral domain. The GD function in this paper is defined for
the normalized autocorrelation of a short segment of a signal [5]. If \( r(n) \) is the normalized autocorrelation of a short segment of a signal, then,

\[
\log(1 + \sum_{n=1}^{\infty} r(n)e^{-jn\omega}) = \log R(\omega)
\]

\[
= \log |R(\omega)|e^{j\theta R(\omega)}
\]

\[
= \log |R(\omega)| + j\theta R(\omega)
\]  

(1)

where \( R(\omega) \) is Fourier transform of \( \{r(n)\} \) with \( r(0) = 1 \). For large energy (i.e., autocorrelation with zero lag), we have \( r(n) \ll 1 \). Therefore, we can write

\[
\log(1 + \sum_{n=1}^{\infty} r(n)e^{-jn\omega}) \approx \sum_{n=1}^{\infty} r(n)e^{-jn\omega}
\]

\[
= \sum_{n=1}^{\infty} r(n) \cos n\omega - j\sum_{n=1}^{\infty} r(n) \sin n\omega
\]  

(2)

From (1) and (2) we get

\[
\Theta_R(\omega) = -\sum_{n=1}^{\infty} r(n) \sin n\omega
\]  

(3)

The group-delay function is given by

\[
\tau(\omega) = -\frac{\partial \Theta_R(\omega)}{\partial \omega} = \sum_{n=1}^{\infty} nr(n) \cos n\omega
\]  

(4)

Due to the asymptotic decay of \( r(n) \), the first \( p \) (10-30) coefficients can be used in the above equation to obtain a reasonably good approximation of the GD spectrum.

Since the energy term is removed from the summation in equation (1), the shape of group-delay spectrum is not influenced by variations in the autocorrelation coefficient with zero lag, thus making \( \tau(\omega) \) less sensitive to additive noise. In addition to removing the effect of additive noise, the GD spectrum enhances the low amplitude and high bandwidth resonances. The high bandwidth resonances are enhanced relative to low bandwidth resonances by evaluating the group-delay function on a circle with radius greater than 1 in the \( z \)-plane. This will also reduce the dominance of the low bandwidth resonances in the GD spectrum. Also, to reduce the truncation effects of the rectangular window on the autocorrelation sequence, it is multiplied by a half-Hann window. Thus, the resulting modified group-delay function is given by:

\[
\tau(\omega) = \sum_{n=1}^{P} w(n) nr(n) \alpha^n \cos n\omega
\]

where \( \alpha < 1 \) and \( w(n) \) is the half-Hann window.

In all our examples, normalized autocorrelation is computed for every frame of 5 ms with each frame shifted by 1 ms. The modified autocorrelation sequence is then averaged over 20 frames and then the modified group-delay function is computed as defined above.

### 3. INVARIANCE OF GROUP DELAY SPECTRUM TO NOISE

The invariance of group delay spectrum to noise is illustrated by the following examples. In figure 1, we show the group delay spectrum of one frame of speech recorded with a high quality microphone in a quiet environment. For comparison, we also show in the same figure, the log magnitude of DFT as well as LPC spectrum. As can be seen from the plots, both the group-delay spectrum and the LPC spectrum follow the formants well. To study the effect of noise, noise recorded in a car environment was added to the clean speech signal to yield an average SNR of 0dB.

![Figure 1. Spectra for clean speech](image)

In figure 2 the same spectra as in figure 1 are shown for the noisy signal. Figure 2 clearly illustrates the noise robustness of group-delay spectrum even at 0dB SNR level at which the LPC spectrum fails to preserve the peaks and valleys of the original clean spectrum.

![Figure 2. Spectra for noisy speech](image)
The above example is given to demonstrate the noise-invariance of the new representation at a frame level. For this example, the clean and noisy speech samples are time aligned. To achieve the time alignment, the noise is digitally added to clean speech to obtain the noisy speech. In the following, we provide our observations on more realistic data where the recording was done both in the absence and in the presence of background noise. Since time alignment is not possible in this case, we show long-term average of the spectrum of an utterance in both recordings.

Figure 3. Average Spectra for clean speech

Figure 3 shows the average FFT spectrum and GD spectrum for an utterance spoken in quiet environment. The spectra are displayed in figure 4 for the same utterance spoken by the same speaker with high way noise being played in the background. Finally, Figure 5 shows the spectra of the same utterance spoken in a simulated environment of an automobile driven on a city street.

Figure 4. Av. Spectra for speech with highway noise

Once again, in the group-delay domain, the formants can be tracked very well even in the presence of background noise.

4. APPLICATIONS

As illustrated above the representation of signal in group-delay domain is more resistant to noise. The above displayed behavior of the group delay spectra under various noise conditions is a clear indication of its potential for various applications. We explored two such applications: 1. The parameter estimation 2. Speech Recognition. Some preliminary results of using the new representation in parameter estimation and speech recognition are presented in the following.

4.1 Parameter Estimation

In this example, we show how the formant frequencies of a voice speech segment can be estimated in the presence of noise. We compute the group-delay spectrogram of vowel /a/ in various noise conditions. Shown in figure 6 are the GD spectrogram of the clean signal and the GD spectrogram of the same signal in the presence of noise at SNR levels 10dB and 5dB. As illustrated in figure 6, the group-delay spectrum can discriminate the formants at different bandwidths even in the presence of noise.

4.2 Speech Recognition

The robustness of group-delay spectrum to noise offers an incentive to apply the processing in this domain to speech recognition. Since the group delay spectrum is derived from
normalized autocorrelation coefficients, these coefficients are directly used as features in speech recognition.

Figure 6. Group delay spectrograms of vowel /a/. (a) Clean signal. (b) Spectrogram of clean signal. (c) Spectrogram of noisy signal with SNR=10dB. (d) Spectrogram of noisy signal with SNR=5dB.

In this paper, we present the results of SI and SD recognition tests on databases consisting of speech samples of varying quality. Both SD and SI recognition systems are based on Hidden Markov Modeling. In the SI isolated-digit recognition task, the recognition accuracy using the group-delay spectral features is 95% compared to the 97% accuracy obtained using LPC cepstral features. However, the same tests on digit database corrupted by noise yield better results with new features than those given by LPC cepstral features.

Similar performance figures were obtained when the features were used in SD recognition tasks. In matched conditions where the training data and test data are recorded in the same environment, the recognition rate was 96.5% when LPC cepstral features were used as compared to 94.5% for the group-delay spectral features. In mismatched conditions where the training data was obtained in a quiet environment and the test data was obtained in a simulated automobile environment, similar observations were made. The LPC cepstral features yielded slightly superior performance. We believe that the reason for slightly lower performance of new features is the large dynamic range of the GD spectrum. Reducing the dynamic range will improve the performance significantly.

5. CONCLUSIONS

A new representation of signals using group-delay functions is proposed. Various examples provided in the paper indicate the robustness of the new representation to high levels of background noise.

Several applications were suggested for which the new representation is more suitable than the other representations. Even though at a first glance, the results of the recognition experiments may suggest less than satisfactory performance, it should be pointed out that reducing the dynamic range, some smoothing of the features and also optimization of parameters may provide better performance. Even though the application for speech enhancement has not been discussed in this paper, we currently exploring ways to improve the quality by processing in the new domain.

6. REFERENCES