ROBUST FEATURES FOR SPEECH RECOGNITION SYSTEMS

Aruna Bayya[*] *and B. Yegnanarayana*[**]
* Rockwell Semiconductor Systems, Newport Beach, CA, USA.
** Indian Institute of Technology, Madras, INDIA.

ABSTRACT

In this paper we propose a set of features based on group delay spectrum for speech recognition systems. These features appear to be more robust to channel variations and environmental changes compared to features based on Melspectral coefficients. The main idea is to derive cepstrum-like features from group delay spectrum instead of deriving them from power spectrum. The group delay spectrum is computed from modified auto-correlation-like function. The effectiveness of the new feature set is demonstrated by the results of both speaker-independent (SI) and speaker-dependent (SD) recognition tasks. Preliminary results indicate that using the new features, we can obtain results comparable to Mel cepstra and PLP cepstra in most of the cases and a slight improvement in noisy cases. More optimization of the parameters is needed to fully exploit the nature of the new features.

1. INTRODUCTION

Feature extraction is the most crucial step in the speech recognition process. Several feature extraction methods resulting in different types of representations have been explored in the context of both SD and SI recognition development. However, complete robustness still remains to be an issue for all speech recognition problems.

Many of the features proposed in the past have not been computed with consideration to robustness. Robustness is achieved by applying various weighting schemes to the features and by introducing either pre-processing on the speech signal in the time-domain or by post-processing of the features. The pre-processing includes high-pass filtering, band-pass filtering the time trajectories of the spectrum in various frequency bands [5]. The post-processing techniques include cepstral normalization, cepstral mean subtraction [1] and using weighting functions for cepstral vectors.

The pre-processing techniques, if not carefully applied may distort the speech spectrum in a way that can affect the feature extraction process. The post-processing techniques on the other hand can not recover the information lost during the feature extraction. Therefore, it is very important to select a feature set that eliminates this additional processing either in the time domain or spectral/cepstral domain while maintaining its robustness.

In this paper, we propose a set of features that are derived from group delay spectrum in which the important features related to formant information are preserved even in the presence of noise [3][6]. These features are not influenced by the dynamic range of the spectrum and hence the dynamic range fluctuations due to environmental changes will have less effect on the performance of the recognition system.

In the following sections, we discuss the process of extracting the features and the results of using the new features in recognition studies, as compared to the features extracted by conventional methods [2][4]. A description of the new feature extraction method is outlined in Section 2. In Section 3, an analysis of intermediate steps is given to provide insight into the new approach for feature extraction. A brief overview of the experimental set up for recognition studies and the results are presented in Section 4 followed by a discussion of the results in the last section.

2. FEATURE EXTRACTION

The steps involved in computing the new features are shown in the form of a flow diagram in Figure 1.

As in the case of other signal processing techniques, the speech signal $s(n)$ is pre-emphasized before feature-extraction for removing the dc component and to spectrally flatten the signal. In this case, the conventional

pre-emphasis is replaced by a difference operation.

$$\tilde{s}(n) = s(n) - s(n-1) \qquad 1 \le n < N$$

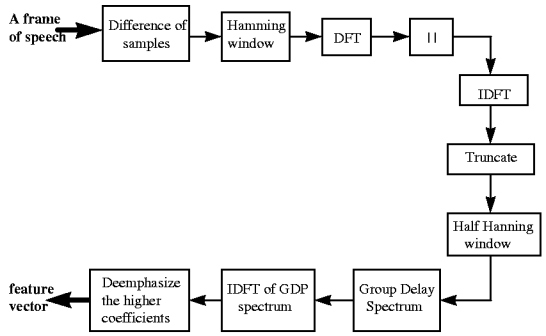

Figure 1. Feature Extraction

This operation can be performed on each frame of the signal without loss of accuracy. Then, as shown in the figure, each frame of speech $\tilde{s}(n)$ is multiplied by a Hamming window. If m is the frame number and N is of the number of samples in each frame,

$$\hat{s}_m(n) = \tilde{s}_m(n)h(n) \quad \text{for} \quad n = 0, \cdots N-1$$

where *h(n)* are the samples of Hamming window.

The windowing is followed by the computation of an autocorrelation like function derived from the magnitude of the Fourier Transform (FT).

$$\tilde{r}_m(n) = DFT^{-1}\left[\left|DFT\left[\hat{s}_m(n)\right]\right|\right]$$
$$0 \le n \le K-1$$

where K is the DFT size typically chosen as $2*N$. $\tilde{r}_m(n)$ is truncated to smooth the finer details and to obtain spectral envelope.

$$\hat{r}(n) = \begin{cases} \tilde{r}_m(n) & n \le L \\ 0 & otherwise \end{cases}$$

where L is *16-24*. Then, the truncated sequence is multiplied by a tapering window such as a Half Hanning window to eliminate discontinuities at the ends of the sequence.

$$r(n) = \hat{r}(n)h(n) \quad \text{for} \quad 0 \le n \le L$$

The group delay spectrum of *r(n)* is then computed as:

$$GD[k] = \frac{R_R(k)D_R(k) + R_I(k)D_I(k)}{|R(k)|^2}$$

where $R(k)$ is the $FT[r(n)]$, $R_R(k) = real\{FT[r(n)]\}$, $R_I(k) = imaginary\{FT[r(n)]$, $D_R(k) = real\{FT[nr(n)]$, and $D_I(k) = imaginary\{FT(nr(n)]\}$.

Finally, the features are computed in a manner similar to the cepstral coefficients as the inverse DFT of sampled group-delay spectrum.

$$c(n) = IDFT\left[GD(k)\right] \qquad 0 \le n \le M$$

the typical values for M being *12-16*.

3. NATURE OF THE GROUP DELAY SPECTRUM

The figures provided in this section illustrate the effect of the proposed technique on a speech signal recorded under varying noise conditions and recording media.

Figure 2 shows the log spectrum of the truncated sequence *r(n)*, the log power spectrum of *s(n)*, and the group delay spectrum of *r(n)* for one frame of speech recorded over the telephone. Plotted in Figure 3 are the same three spectra for open microphone. For comparison, we also show the log spectrum of *r(n)*, the log spectrum of *s(n)* and the group delay spectrum for speech recorded with high quality microphone, in Figure 4. The speech samples are extracted from databases collected under these different conditions.

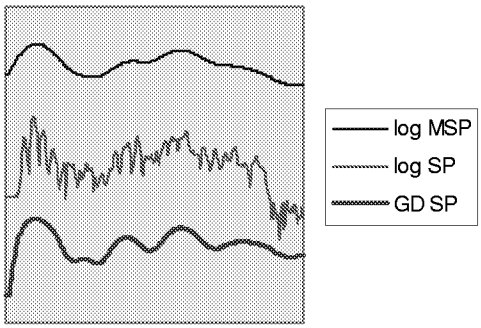

Figure 2. Spectra for clean telephone speech

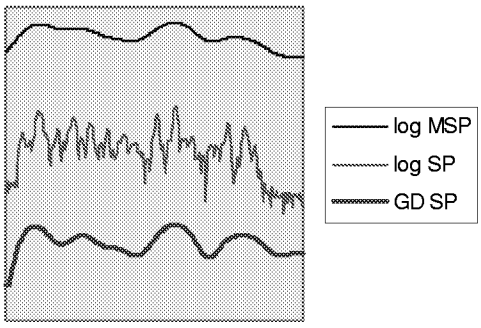

Figure 3. Spectra for open microphone speech

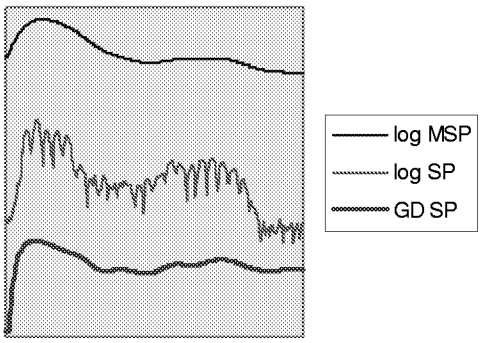

Figure 4. Spectra for close microphone speech

As the figures indicate, while the log power spectrum and the group delay spectrum for the high quality speech look somewhat alike, the advantage of the processing in group-delay domain can be clearly noted from the spectra of telephone speech and more so in open microphone speech where the background noise is noticeable.

These observations are confirmed by the results of the recognition experiments carried out on speech databases for each of these conditions.

4. RECOGNITION EXPERIMENTS

The above described feature extraction is incorporated into the SD as well as the SI isolated-word recognition systems. An overview of each of these systems' implementation is given prior to presenting the results on various recognition tasks.

Both the SI and the SD recognition algorithms are based on Hidden Markov Modeling using Maximum Likelihood criterion. Continuous Gaussian density, single-mixture HMMs with diagonal covariance matrix are used to represent word models. The number of states in each model is determined from the average duration of each vocabulary word which in turn is obtained from the training data.

The SI task is the recognition of digits spoken in isolation. These tests were performed on two data sets, each one representing a different quality of speech. The first one (TI digit database) is recorded using close-speaking microphone resulting in high quality speech. The vocabulary is made of 11 digits (0-9 and 'oh'). For test purposes, the database which consists of speech from 222 speakers was divided into training and test sets. The training set includes 56 male speakers and 56 female speakers and the test set includes 110 speakers (55 males and 55 females). Each of the 11 digits were spoken twice by each speaker. The second database is recorded using an inexpensive open microphone typically used in speakerphones. This data set consists of 10 repetitions of each digit spoken in isolation by 20 speakers (10 males and 10 females). As before, the data set is equally divided between training and recognition tasks.

All of the SI recognition experiments are repeated for LPC feature set, PLP feature set , Mel cepstral feature set and the new feature set where each set is made of feature vectors with 12 cesptral coefficients, 4 delta cepstral coefficients, delta energy and delta delta energy.

The results of the SI recognition tests are summarized in Table 1.

	LPC	PLP	MEL	GDP
Close Mic.	97.4%	98.2%	98.25%	97.0%
Open Mic.	90.0%	89.1%	90.2%	90.7%

Table 1. Results of SI recognition tests

It can be concluded from the above examples that the recognition performance with new features is comparable to LPC and slightly lower than the PLP and MEL features in case of high quality speech. However, with no additional processing

the new feature set yields slightly better performance than the rest in the case of noisy speech.

In order to verify the performance of new features in recognition systems with limited amount of training, we apply the same feature set in SD speech recognition. The task is to recognize words from a 20-word vocabulary set. Ten speakers were asked to train the system with 20 names (with no restriction imposed on the type of names or lengths of names). Each person repeated the 20 names 16 times each (8 repetitions were recorded with low-quality open microphone and the rest were recorded over different telephone channels).

The HMM models for each of these words were built from one token of the word. The rest of the tokens were used in recognition. The experiments consist of matched training and recognition conditions as well as mismatched training and test conditions.

The following table provides the performance figures resulting from the SD tests.

	LPC	PLP	MEL	GDP
train & test on mic.	96.2%	97.2%	97.0%	97.0%
train & test on tel.	90.2%	93.4%	93.7%	94.5%
train on mic. & test on tel.	89.6%	91.0%	91.0%	90.0%

Table 2. Results of SD recognition tests

From the table, it is clear that while the performance of the system using group delay features is slightly worse under the mismatched conditions, in the other two cases, it is about the same or better than the performance of the other features.

5. DISCUSSION

The superior performance of the group delay spectral features can be attributed to the fact that they represent formant information which is robust to different channel conditions. Moreover, the group delay spectrum does not depend on the dynamic range and the slope of the short-time spectrum due to high resolution and additive properties of group delay spectrum [7]. It is also interesting to note that all the important features of spectral envelope are retained in the group delay spectrum as the frequency scale is linear, unlike the Mel frequency scale where the spectral details in the high frequency region are lost in averaging.

It should be noted that in all these experiments, for the given model architecture, the PLP, LPC parameters were optimized while no systematic optimization was done for the new feature set. The encouraging preliminary results indicate the potential of the new technique. With better optimization of the parameters, its performance in recognition can be further improved.

6. REFRENCES

[1] A. Acero, *Acoustical and Environmental Robustness in Automatic Speech recognition*, PhD thesis, CMU, 1990.

[2] S.B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences", *IEEE Trans. ASSP*, pp. 357-366, August, 1980.

[3] G. Duncan, B. Yegnanarayana and Hema A. Murthy, "A nonparametric method of formant estimation using group delay spectra", *Proc. ICASSP89*, Glasgow, Scotland, pp.572-575, May 23-26, 1989.

[4] Hynek Hermansky, "Perceptual linear predictive (PLP) analysis of speech", *JASA*, vol. 87, no. 4, pp. 1738-1752, 1990.

[5] H. Hirsh, P. Meyer and H.W. Ruehl, "Improved speech recognition using high-pass filtering of sub-band envelopes", *Proc. EUROSPEECH 91*, pp.413-416, Genova, Italy.

[6] Hema A. Murthy and B. Yegnanarayana, "Formant extraction from group delay function", *Speech communication*, vol.10, no.3, pp.209-267, Aug. 1991.

[7] B. Yegnanarayana, "Formant extraction from linear prediction phase spectra", *JASA*, vol.63, no.5, pp.1638-1640, May 1978.