The Modified Group Delay Feature: A New Spectral Representation of Speech

Rajesh M. Hegde, Hema A. Murthy
Department of Computer Science and Engineering
Indian Institute of Technology, Madras, Chennai.
{rajesh,hema}@lantana.tenet.res.in

Gadde V. Ramana Rao
STAR Laboratory, SRI International,
333, Ravenswood Avenue, Menlo Park,
CA 94025
{rao}@speech.sri.com

Abstract

Automatic recognition of speech by machines begins with extraction of meaningful features from the speech signal. Conventional features like the MFCC are derived from the Fourier transform magnitude spectrum, while totally ignoring the phase spectrum. The importance of the Modified group delay feature (MODGDF) derived from the Fourier transform phase spectrum for speaker and phoneme recognition has been presented in our previous efforts. In this paper we try to analyse the feature theoretically and provide justifications in terms of de-correlation, robustness to convolutional and white noise, cluster structures, separability in lower dimensional space, task independence and class separability. The results of speaker identification and continuous speech recognition using the MODGDF as the front end are also presented. Joint features derived from the MODGDF and MFCC gave significant improvements in recognition performance for both speaker and continuous speech recognition tasks. Using the analytical results in the first half of the paper and the results of performance evaluation in the second half, the MODGDF is proposed as an alternative spectral representation of speech.

1. Introduction

Spectral representation of speech is complete when both the magnitude and phase spectrum of Fourier transform are used to derive features. But the phase spectrum is ignored in all conventional feature extraction techniques. This is primarily because unwrapping the phase is a complex task. Early psychoacoustic experiments conducted by Helmholtz [1] and more recent perception studies by Liu [2] have also claimed that short time phase spectrum conveys no information about the intelligibility of the speech signal. These studies are based on acoustic experiments on limited representative speech data and do not take into account the complex functional principles on which speech is produced. Also, the misconception that the future of speech recognition lies in designing efficient classifiers and the hesitation to look beyond the magnitude spectrum has held back efforts to derive new features, although efforts to derive features based on simulation of properties of the human auditory system have been made. Group delay functions [3] defined as the negative derivative of the phase spectrum have been used for various speech processing tasks, including spectrum estimation and formant extraction. In our previous efforts we have proved the importance of the features derived from the modified group delay function, called the MODGDF for phoneme recognition [3] and also for automatic speaker identification [4]. In this paper we theoretically and experimentally evaluate the MODGDF for de-correlation, robustness to convolutional and white noise, cluster structures [5], separability in lower dimensional space, task independence and class separability [6]. The first half of the paper includes an introduction to the MODGDF and its theoretical evaluation for the aforementioned criteria. In the second half we present results of performance evaluation of the MODGDF for speaker identification and continuous speech recognition [7, 8]. Based on the theoretical justifications and performance evaluation, the MODGDF is proposed as a new phase spectral representation of speech.

2. The Modified Group Delay Feature

The group delay function [3], defined as the negative derivative of phase, can be effectively used to extract various system parameters when the signal under consideration is a minimum phase signal. The group delay function is defined as

$$\tau(\omega) = -\frac{d\theta(\omega)}{d\omega}$$  \hspace{1cm} (1)

where $\theta(\omega)$ is the unwrapped phase function. The group delay function can also be computed from the speech signal as in [3] using

$$\tau_x(\omega) = \frac{X_R(\omega)Y_R(\omega) + Y_I(\omega)X_I(\omega)}{|X(\omega)|^2}$$  \hspace{1cm} (2)

where the subscripts $R$ and $I$ denote the real and imaginary parts of the Fourier transform. $X(\omega)$ and $Y(\omega)$ are the Fourier transforms of $x(n)$ and $nx(n)$, respectively. The group delay function requires that the signal be minimum phase or that the poles of the transfer function be well within the unit circle for it to be well behaved. This has been clearly illustrated in [3]. It is also important to note that the denominator term $|X(\omega)|^2$ in equation 2 becomes zero, at zeros that...
are located close to the unit circle. The spiky nature of the group delay spectrum can be overcome by replacing the term $|X(\omega)|$ in the denominator of the group delay function with its cepstrally smoothed version, $S(\omega)$. Further it has been established in [3] that peaks at the formant locations are very spiky in nature. To reduce these spikes two new parameters $\gamma$ and $\alpha$ are introduced. The new modified group delay function as in [3] is defined as

$$\tau_m(\omega) = \left(\frac{\tau(\omega)}{|\tau(\omega)|}\right) (|\tau(\omega)|)^\alpha$$  \hspace{1cm} (3)$$

where

$$\tau(\omega) = \frac{X_R(\omega)Y_R(\omega) + Y_I(\omega)X_I(\omega)}{S(\omega)^2}$$  \hspace{1cm} (4)$$

where $S(\omega)$ is the smoothed version of $|X(\omega)|$. The new parameters $\alpha$ and $\gamma$ introduced vary from 0 to 1 where (0 < $\alpha$ < 1.0) and (0 < $\gamma$ < 1.0). The algorithm for computation of the modified group delay function is explicitly dealt with in [3]. To convert the modified group delay function to some meaningful parameters, the group delay function is converted to cepstra using the Discrete Cosine Transform (DCT).

$$c(n) = \sum_{k=0}^{k=N_f} \tau_x(k) \cos(n(2k + 1)\pi/N_f)$$  \hspace{1cm} (5)$$

where $N_f$ is the DFT order and $\tau_x(k)$ is the group delay function. The second form of the DCT, DCT-II is used, which has asymptotic properties to that of the Karhunen Loeve Transformation (KLT) as in [3]. The DCT acts as a linear decorrelator, which allows the use of diagonal co-variances in modellng the speaker vector distribution.

3. Feature Evaluation of the MODGDF

In this section we evaluate the MODGDF against five important feature evaluation criteria.

3.1. Decorrelation

From the previous section it is evident that, assuming the correlation matrix of the input data exhibits toeplitz structure and neglecting boundary effects, always gives decorrelated features when a cosine tranform is used. As a proof of concept we compute the measured correlation matrix for the MODGDF and visualize it, to show that the matrix is diagonal like and therefore allows the use of diagonal covariances in modellng the feature distribution. The complete decorrelation obtained by the MODGDF is shown as a 3 dimensional plot in Figure 1 where the first two dimensions correspond to the feature component and the third dimension is the correlation coefficient.

3.2. Robustness to convolutional and white noise

Features that are invariant to noise save additional processing like cepstral mean subtraction, and eliminate sources of side effects and distortion. Representation of speech in the group delay domain enhances the important features of the envelope of the spectrum making it immune to noise. Assuming a source system model of speech production, the clean speech

\[ x_c(n), \] its Fourier transform and the corresponding group delay function [3] is given by:

\[ x_c(n) = \sum_{k=1}^{p} a_k x_c(n-k) + G e(n) \]  \hspace{1cm} (6)$$

\[ X_c(\omega) = \frac{G E(\omega)H(\omega)}{A(\omega)} \]  \hspace{1cm} (7)$$

\[ \tau_c(\omega) = G \tau_e(\omega) + \tau_h(\omega) - \tau_a(\omega) \]  \hspace{1cm} (8)$$

where $e(n)$ is the excitation signal and $G$ the gain. The term $\tau_e(\omega)$ is due to the periodicity of the pulse train and delay between successive formants of the speech signal. Similarly the noisy speech signal and its Fourier transform is given by:

\[ x_n(n) = x(n) * h(n) + w(n) \]  \hspace{1cm} (9)$$

\[ X_n(\omega) = X(\omega)H(\omega) + W(\omega) \]  \hspace{1cm} (10)$$

where $h(n)$ is the time invariant channel response and $w(n)$ additive white noise. Taking Fourier transform of equation 6 and substituting in equation 8, $X_n(\omega)$ and the corresponding group delay function is given by:

\[ X_n(\omega) = \frac{G E(\omega)H(\omega) + A(\omega)W(\omega)}{A(\omega)} \]  \hspace{1cm} (11)$$

\[ \tau_n(\omega) = \tau_{\text{numerator}}(\omega) - \tau_a(\omega) \]  \hspace{1cm} (12)$$

where $\tau_{\text{numerator}}(\omega)$ is the is the group delay function corresponding to that of $G E(\omega)H(\omega) + A(\omega)W(\omega)$. Further $\tau_{\text{numerator}}(\omega)$ dominates in high SNR regions and $\tau_a(\omega)$ dominates in low SNR regions. From the discussion so far it is justified to state that the excitation makes the group delay spectrum spiky and distorted primarily due to zeros that are close to the unit circle in the $Z$ domain. White noise can be easily handled in the group delay domain while the excitation can be dealt with by pushing all zeros close to the unit circle in the $Z$ domain, well inside the unit circle by appropriately selecting values for the two parameters $\alpha$, $\gamma$ as defined in equation 3 and 4. The relevance of $\alpha$ and $\gamma$ is also highlighted in [3].

3.3. Task Independence

Features should capture contradictory information for speaker and speech recognition tasks. The former should
capture speaker related information only, while the latter should carry information related to the phonetic unit only with no speaker related information. In order to illustrate the importance of the MODGDF for speech recognition we cite our work in [3]. In our previous work on speaker identification [4] we have shown that the MODGDF is effective for the task. To reinforce its relevance, we use the illustrations in Figure 2 and 3, surface plots of the MODGDF and MFCC respectively, for one phrase of a speaker picked from the NTIMIT database. It is emphasised here that both the spectra are reconstructed from the cepstral coefficients. It is clearly evident from the Figure 2 that higher formants are more pronounced in the MODGD spectrum, when compared to the spectrum of MFCC as in Figure 3. This opens up an exciting possibility of combining the MODGDF and MFCC to derive joint features that can capture information pertinent to both speaker and speech recognition tasks. Indeed the results of performance evaluation in Section 4 reinforce this fact.

![Figure 2: Surface plot of the formants derived using MODGDF for one phrase of a speaker from the NTIMIT database](image1)

![Figure 3: Surface plot of the formants derived using MFCC for one phrase of a speaker from the NTIMIT database](image2)

### 3.4. Class separability

The objective of determining the optimal feature set for classification is to minimize the classification error rate, which is the Bayes error. But statistics literature indicates that Bayes error is never easy to estimate. To circumvent the complex problem of estimating Bayes error directly several techniques have evolved to measure class separability. The most widely used measures among them are the geometrically intuitive measures like the F ratio and mathematical measures like the Chernoff and the Bhattacharya bound [6]. The Bhattacharya bound which is a special case of the Chernoff bound is a probabilistic error measure and relates more closely to the likelihood maximization classifiers that we use for performance evaluation. The Bhattacharya bound is primarily defined for a two class problem and extension to a multi-class case is a combination of pairwise bounds. The MODGDF shows good separability on a pairwise basis as in [4] and therefore the Bhattacharya distance measure is very apt to investigate class separability criteria. We consider fifty speakers from the NTIMIT database and compute a 16 dimensional codebook of size 32 for each speaker. The separability criterion based on the Bhattacharya distance measure is then calculated. The separability criterion and the cumulative separability criterion versus feature dimension for both the MODGDF and the MFCC is illustrated in Figure 4 and Figure 5 respectively. From Figures 4 and 5 it is evident that the MODGDF outperforms MFCC with respect to class separability, as the separability curve corresponding to MODGDF is well above that of MFCC.

![Figure 4: Class separability of MODGDF and MFCC feature sets using Bhattacharya distance](image3)

![Figure 5: Cumulative class separability of MODGDF and MFCC feature sets using Bhattacharya distance](image4)

#### 3.5. Cluster structure and Separability in low dimensional feature space

It is a fact that speaker identification is more of a discrimination problem while continuous speech recognition is a representational problem. Compactness of the cluster structure assumes importance in the latter problem while separability is the crucial for solving the former problem. We therefore reduce the multidimensional MODGDF vectors to two dimensional vectors using the non linear dimensionality reduction technique, Sammon mapping [5], as the MODGDF exhibits a non linear structure in the higher dimensional space. The cluster structures of two syllables (fricatives) /sa/ and
In the two dimensional space are shown in Figure 6 while the cluster structures for two speakers are illustrated in Figure 7. It is clear from Figure 6 that the clusters are tightly bound and from Figure 7 that two speakers are well separated.

![Figure 6: Cluster structures of two fricatives /sa/ and /shA/ in the 2 dimensional feature space](image1)

![Figure 7: Cluster structures of two female speakers in the 2 dimensional feature space](image2)

### 4. Performance Evaluation

#### 4.1. Speaker Identification

The results of speaker identification experiments using MODGDF and joint features (MODGDF+MFCC) for clean (TIMIT), noisy telephone (NTIMIT) speech for 120 tests are presented in Table 1. The baseline system uses a likelihood maximization scheme (GMM).

<table>
<thead>
<tr>
<th>Feature</th>
<th>Database</th>
<th>Classifier</th>
<th>Recognition</th>
</tr>
</thead>
<tbody>
<tr>
<td>MODGDF</td>
<td>TIMIT</td>
<td>GMM</td>
<td>98.5%</td>
</tr>
<tr>
<td>MODGDF</td>
<td>NTIMIT</td>
<td>GMM</td>
<td>62%</td>
</tr>
<tr>
<td>MODGDF+MFCC</td>
<td>NTIMIT</td>
<td>GMM</td>
<td>66%</td>
</tr>
</tbody>
</table>

#### 4.2. Syllable based Continuous Speech Recognition

The results listed below in Table 2 are for the Telugu (Indian language) broadcast news database [8], for one complete news bulletin comprising of 4700 syllables. Ten news bulletins of 15 minute duration are used for training HMMs (5 state 3 mixture) apriori, for a vocabulary of 265 syllables. The baseline system segments continuous speech using minimum phase group delay functions, followed by isolated style recognition using hidden markov models (HMMs) [7]. An innovative local forced Viterbi realignment scheme is used to achieve improved recognition performance, without any language models.

<table>
<thead>
<tr>
<th>Category</th>
<th>(I)</th>
<th>(II)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Whole syllable</td>
<td>39.6%</td>
<td>50.6%</td>
</tr>
<tr>
<td>Similar syllables</td>
<td>14.8%</td>
<td>22.8%</td>
</tr>
<tr>
<td>Only vowel part</td>
<td>6.1%</td>
<td>8.8%</td>
</tr>
<tr>
<td>Only consonant part</td>
<td>3.2%</td>
<td>5.5%</td>
</tr>
<tr>
<td>Total</td>
<td>63.7%</td>
<td>87.7%</td>
</tr>
</tbody>
</table>

### 5. Conclusions

In this paper the MODGDF derived from from Fourier transform phase is investigated for several objective feature evaluation criteria. It is found to perform well in all evaluations for optimality. Theoretical justifications for robustness to channel effects and white noise are provided. The results for automatic speaker identification and continuous speech recognition have been listed to prove its importance in practice. Automatic speaker recognition experiments using joint features derived from the MODGDF and MFCC at feature level gave a promising recognition of 98.5% for clean speech and 66% for telephone speech, while for continuous speech recognition they yielded a good 87.7% overall recognition. Based on the results of investigation of feature evaluation criteria and performance evaluation on actual data, we conclude that the MODGDF could be used as an alternate spectral representation of speech across all automatic speech recognition tasks.

### 6. References