Hindawi Publishing Corporation
EURASIP Journal on Audio, Speech, and Music Processing
Volume 2007, Article ID 79032, 13 pages
doi:10.1155/2007/79032

Research Article

Significance of Joint Features Derived from the Modified Group Delay Function in Speech Processing

Rajesh M. Hegde,[1] Hema A. Murthy,[2] and V. R. R. Gadde[3]

[1] *Department of Electrical and Computer Engineering, University of California San Diego, San Diego, CA 92122, USA*
[2] *Department of Computer Science and Engineering, Indian Institute of Technology Madras, Chennai 600 036, India*
[3] *STAR Lab, SRI International, 333 Ravenswood Avenue, Menlo Park, CA 94025, USA*

Received 1 April 2006; Revised 20 September 2006; Accepted 10 October 2006

Recommended by Climent Nadeu

This paper investigates the significance of combining cepstral features derived from the modified group delay function and from the short-time spectral magnitude like the MFCC. The conventional group delay function fails to capture the resonant structure and the dynamic range of the speech spectrum primarily due to pitch periodicity effects. The group delay function is modified to suppress these spikes and to restore the dynamic range of the speech spectrum. Cepstral features are derived from the modified group delay function, which are called the modified group delay feature (MODGDF). The complementarity and robustness of the MODGDF when compared to the MFCC are also analyzed using spectral reconstruction techniques. Combination of several spectral magnitude-based features and the MODGDF using feature fusion and likelihood combination is described. These features are then used for three speech processing tasks, namely, syllable, speaker, and language recognition. Results indicate that combining MODGDF with MFCC at the feature level gives significant improvements for speech recognition tasks in noise. Combining the MODGDF and the spectral magnitude-based features gives a significant increase in recognition performance of 11% at best, while combining any two features derived from the spectral magnitude does not give any significant improvement.

1. INTRODUCTION

Various types of features have been used in speech processing [1]. Variations on the basic spectral computation, such as the inclusion of time and frequency masking, have been used in [2–4]. The use of auditory models as the basis of feature extraction has been beneficial in many systems [5–9], especially in noisy environments [10]. Perhaps the most popular features used in speech recognition today are the Mel frequency cepstral coefficients (MFCCs) [11]. In conventional speech recognition systems, features are usually computed from the short-time power spectrum while the short-term phase spectrum is not used. This is primarily because early experiments on human speech perception have indicated that the human ear is not sensitive to short-time phase. But recent experiments described in [12, 13] have indicated the usefulness of the short-time phase spectrum in human listening tests. In this context, the short-time phase spectrum estimated via the group delay domain has been used to parameterize speech in our earlier efforts [14–17]. Cepstral features were derived from the modified group delay function and were called the modified group delay feature (MODGDF) in these efforts [18]. In this paper, we focus on the significance of the representation of speech using joint features derived from the modified group delay function and from the short-time power spectrum like the MFCC. Previous work on combining the MODGDF with MFCC also appears in [19]. The focus of this paper is on combining features before the acoustic model [20, 21], as well as after the acoustic model [22–26]. In this context, we start with a discussion on group delay functions and their significance in formant estimation of speech. The modified group delay function and extraction of cepstral features are discussed next. The significance of combining spectral magnitude and phase-based feature is illustrated next, using spectral reconstructions. Both the individual and the joint features derived from the modified group delay function and the short-time power spectrum are used for the tasks of syllable [27], speaker [28–31], and language recognition [32]. The paper concludes with

a discussion on the significance of joint features in speech processing.

2. SIGNIFICANCE OF FEATURE COMBINATIONS

The technique of combination is widely used in statistics. The simplest method of combination involves averaging the various estimates of the underlying information. This idea is based on the hypothesis that if different estimates are subject to different sources of noise, then combining them will cancel some of the errors when an averaging is done. Good examples of combining features are the works of Christensen [22] and Janin et al. [25] who have combined different features before and after the acoustic model. Other significant works on feature and likelihood combination can be found in [33–36]. A combination system works on the principle that if some characteristics of the speech signal that is deemphasized by a particular feature are emphasized by another feature, then the combined feature stream captures complementary information present in individual features.

2.1. Feature combination before the acoustic model

The combination of features before the acoustic model has been used by Christensen [22], Okawa et al. [20], and Ellis [21], where efforts have been made to capitalize on the differences between various feature streams using all of them at once. The joint feature stream is derived in such an approach by concatenating all the individual feature streams into a single feature stream.

2.2. Likelihood combination after the acoustic model

This approach uses the technique of combining the outputs of the acoustic models. Complex techniques of combining the posteriors [22–26, 33–36] have evolved. In this context, it is also worthwhile to note that if the intent is to capitalize on the complementary information in different features, the posteriors of the same classifier for individual features can be combined to achieve improved speech recognition performance.

3. THE GROUP DELAY FUNCTION AND ITS PROPERTIES

The resonances of the speech signal present themselves as the peaks of the envelope of the short-time magnitude spectrum. These resonances, often called formants, appear as transitions in the short-time phase spectrum. The problem with identifying these transitions is the masking of these transitions due to the wrapping around of the phase spectrum at multiples of 2π. The group delay function, defined as the negative derivative of phase, can be computed directly from the speech signal and has been used to extract source and system parameters [37] when the signal under consideration is a minimum phase signal. This is primarily because the magnitude spectrum of a minimum phase signal [37] and its group delay function resemble each other.

3.1. The group delay function

Group delay is defined as the negative derivative of the Fourier transform phase

$$\tau(\omega) = -\frac{d(\theta(\omega))}{d\omega}, \tag{1}$$

where the phase spectrum $(\theta(\omega))$ of a signal is defined as a continuous function of ω. The Fourier transform phase and the Fourier transform magnitude are related as in [38]. The group delay function can also be computed from the signal as in [14] using

$$\tau_x(\omega) = -\operatorname{Im}\frac{d(\log(X(\omega)))}{d\omega} \tag{2}$$

$$= \frac{X_R(\omega)Y_R(\omega) + Y_I(\omega)X_I(\omega)}{|X(\omega)|^2}, \tag{3}$$

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 $\tau(\omega)$ can also be viewed as the Fourier transform of the weighted cepstrum [37].

3.2. Relationship between spectral magnitude and phase

The relation between spectral magnitude and phase has been discussed extensively in [38]. In [38], it has been shown that the unwrapped phase function for a minimum phase signal is given by

$$\theta(\omega) = \theta_v(\omega) + 2\pi\lambda(\omega) = -\sum_{n=1}^{\infty} c(n)\sin(n\omega), \tag{4}$$

where $c(n)$ are the cepstral coefficients. Differentiating (4) with respect to ω, we have

$$\tau(\omega) = -\theta'(\omega) = \sum_{n=1}^{\infty} nc(n)\cos(n\omega), \tag{5}$$

where $\tau(\omega)$ is the group delay function. The log-magnitude spectrum for a minimum phase signal $v(n)$ [38] is given by

$$\ln|V(\omega)| = \frac{c(0)}{2} + \sum_{n=1}^{\infty} c(n)\cos(n\omega). \tag{6}$$

The relation between spectral magnitude and phase for a minimum phase signal [38], through cepstral coefficients, is given by (4) and (6). For a maximum phase signal equation

(4) holds, while the unwrapped phase is given by

$$\theta(\omega) = \theta_v(\omega) + 2\pi\lambda(\omega) = \sum_{n=1}^{\infty} c(n)\sin(n\omega) \tag{7}$$

and the group delay function $\tau(\omega)$ is given by

$$\tau(\omega) = -\theta'(\omega) = -\sum_{n=1}^{\infty} nc(n)\cos(n\omega). \tag{8}$$

Hence the relation between spectral magnitude and phase for a maximum phase signal [38], through cepstral coefficients, is given by (4) and (7). For mixed phase signals, the relation between spectral magnitude and phase is given by two sets of cepstral coefficients $\{c_1(n)\}$ and $\{c_2(n)\}$, as

$$\ln|X(\omega)| = \frac{c_1(0)}{2} + \sum_{n=1}^{\infty} c_1(n)\cos(n\omega), \tag{9}$$

where $\ln|X(\omega)|$ is the log-magnitude spectrum for a mixed phase signal and $\{c_1(n)\}$ is the set of cepstral coefficients computed from the minimum phase equivalent signal derived from the spectral magnitude. Similarly, the unwrapped phase is given by

$$\theta_x(\omega) + 2\pi\lambda(\omega) = -\sum_{n=1}^{\infty} c_2(n)\sin(n\omega), \tag{10}$$

where $\theta_x(\omega) + 2\pi\lambda(\omega)$ is the unwrapped phase spectrum for a mixed phase signal and $\{c_2(n)\}$ is the set of cepstral coefficients computed from the minimum phase equivalent signal derived from the spectral phase. Therefore, the relation between spectral magnitude and phase for a mixed phase signal [38], through cepstral coefficients, is given by (9) and (10).

3.3. Issues in group delay processing of speech

The group delay functions and their properties have been discussed in [37, 39]. The two main properties of the group delay functions [39] of relevance to this work are

(i) additive property;
(ii) high-resolution property.

3.3.1. Additive property

The group delay function exhibits an additive property. Let

$$\mathcal{H}(e^{j\omega}) = \mathcal{H}_1(e^{j\omega}) \cdot \mathcal{H}_2(e^{j\omega}), \tag{11}$$

where $\mathcal{H}_1(e^{j\omega})$ and $\mathcal{H}_2(e^{j\omega})$ are the responses of the two resonators whose product gives the overall system response. In the group delay domain, equation (11) translates to

$$\tau_h(e^{j\omega}) = \tau_{h1}(e^{j\omega}) + \tau_{h2}(e^{j\omega}), \tag{12}$$

where, $\tau_{h1}(e^{j\omega})$ and $\tau_{h2}(e^{j\omega})$ correspond to the group delay functions of $\mathcal{H}_1(e^{j\omega})$ and $\mathcal{H}_2(e^{j\omega})$, respectively. From (11) and (12), we see that multiplication in the spectral domain becomes an addition in the group delay domain.

3.3.2. High-resolution property

The group delay function has a higher resolving power when compared to both the magnitude and the LP spectrum. This property is a manifestation of the spectral additive property of the group delay function. The high-resolution property of the group delay function over both the magnitude and the linear prediction spectrum has been illustrated in [39].

3.4. Significance of pitch periodicity effects

When the short-time Fourier transform power spectrum is used to extract the formants, the focus is on capturing the spectral envelope of the spectrum and not the fine structure. Similarly, the fine structure has to be deemphasized when extracting the vocal tract characteristics from the group delay function. But the group delay function becomes very spiky in nature due to pitch periodicity effects. To illustrate this, a three-formant system is simulated whose pole-zero plot is shown in Figure 1(a). The formant locations are at 500 Hz, 1570 Hz, and 2240 Hz. The corresponding impulse response of the system is shown in Figure 1(b) and its group delay function is shown in Figure 1(c). The group delay function is able to resolve all the three formants. The system shown in Figure 1(a) is now excited with 5 impulses and the system response is shown in Figure 1(d). The group delay function of the signal in Figure 1(d) is shown in Figure 1(e). It is evident from Figure 1(e) that the group delay function becomes spiky and distorted due to pitch periodicity effects. The spikes introduced into the group delay function due to zeros close to the unit circle and also due to the pitch periodicity effects form a significant part of the fine structure and cannot be removed by normal smoothing techniques. Hence the group delay function has to be modified to suppress the effects of these spikes. These considerations form the basis for modifying the group delay function.

4. THE MODIFIED GROUP DELAY FUNCTION

As mentioned in the earlier sections, for the group delay function to be a meaningful representation, it is necessary that the roots of the transfer function are not too close to the unit circle in the z plane. Normally, in the context of speech, the poles of the transfer function are well within the unit circle. The zeros of the slowly varying envelope of speech correspond to that of nasals. The zeros in speech are either within or outside the unit circle since the zeros also have nonzero bandwidth. In this section, we modify the computation of the group delay function to suppress these effects. A similar approach was taken in an earlier paper by one of the authors [40] for spectrum estimation. Let us reconsider the group delay function derived directly from the speech signal. It is important to note that the denominator term $|X(\omega)|^2$ in (3) becomes very small at zeros that are located close to the unit circle. This makes the group delay function

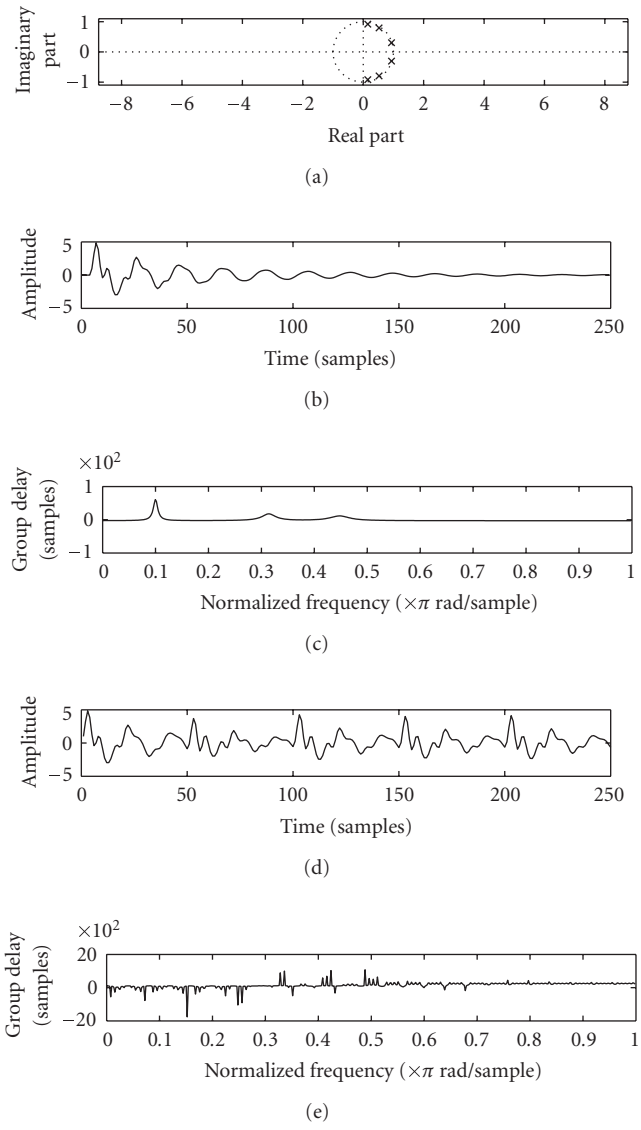

FIGURE 1: Significance of pitch periodicity effects on the group delay functions (a) the z-plane with three complex poles and their complex conjugate pairs inside the unit circle, (b) the impulse response of the system shown in (a), (c) the group delay spectrum of the signal shown in (b), (d) the response of the system shown in (a) to 5 impulses, and (e) the group delay spectrum of the signal shown in (d).

very spiky in nature and also alters the dynamic range of the group delay spectrum. 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 as in (3) with its cepstrally smoothed version, $S(\omega)$.[1] Two new parameters γ and α are further introduced to reduce the amplitude of these spikes and to restore the dynamic range of the group delay spectrum. The new modified group delay function is defined as

$$\tau_m(\omega) = \left(\frac{\tau(\omega)}{|\tau(\omega)|}\right)(|\tau(\omega)|)^{\alpha}, \tag{13}$$

where

$$\tau(\omega) = \left(\frac{X_R(\omega)Y_R(\omega) + Y_I(\omega)X_I(\omega)}{S(\omega)^{2\gamma}}\right), \tag{14}$$

where $S(\omega)$ is the smoothed version of $|X(\omega)|$. The parameters α and γ introduced vary from 0 to 1 where $(0 < \alpha \leq 1.0)$ and $(0 < \gamma \leq 1.0)$.

Figure 2(a) shows a z plane plot of a system with three resonances at 530 Hz, 1840 Hz, and 2480 Hz. In Figures 2(b) and 2(c), respectively, are shown the impulse response and the group delay function of such a system. The response of the same system excited with 5 impulses and the corresponding group delay function are shown in Figures 2(d) and 2(e), respectively. The modified group delay function (lifter$_w$ = 6, $\alpha = 0.4$, and $\gamma = 0.9$) for the signal in Figure 2(d) is shown in Figure 2(f). It is clear from Figures 2(e) and 2(f) that while the group delay function fails to capture the formant structure of the signal in Figure 2(d), the modified group delay function is able to do so.

5. PARAMETERIZING THE MODIFIED GROUP DELAY FUNCTION

Since the modified group delay function exhibits a squared magnitude behavior at the location of the roots, we refer to the modified group delay function as the modified group delay spectra henceforth. Homomorphic processing is the most commonly used approach to convert spectrum derived from the speech signal to meaningful features. This is primarily because this approach yields features that are linearly decorrelated which allows the use of diagonal covariances in modeling the speech vector distribution. In this context, the discrete cosine transform (DCT I,II,III) [41] is the most commonly used transformation that can be used to convert the modified group delay spectra to cepstral features. Hence the group delay function is converted to cepstra using the discrete cosine transform (DCT II) as

$$c(n) = \sum_{k=0}^{k=N_f} \tau_m(k) \cos\left(\frac{n(2k+1)\pi}{N_f}\right), \tag{15}$$

where N_f is the DFT order and $\tau_m(k)$ is the modified group delay spectrum. The discrete cosine transform (DCT) can also be used in the reconstruction of the modified group delay spectra from the modified group delay cepstra (MODGDF). Velocity and acceleration parameters for the new group delay function are defined in the cepstral domain, in a manner similar to that of the velocity and acceleration parameters for MFCC.

[1] A lower-order cepstral window lifter$_w$ whose length can vary from 4 to 9 is used.

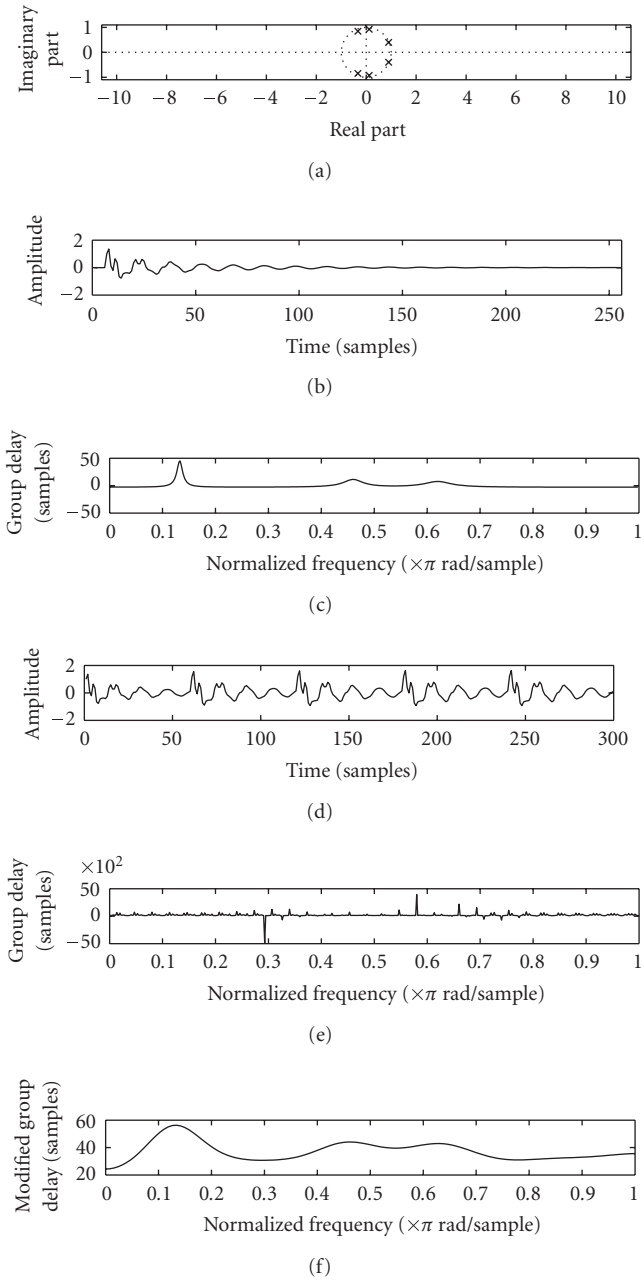

Figure 2: Comparison of the group delay and the modified group delay function to handle pitch periodicity effects. (a) The z-plane plot of a system with three complex poles and their complex conjugate pairs, (b) the impulse response of the system shown in (a), (c) the group delay function of the signal shown in (b), (d) the response of the system shown in (a) to 5 impulses, (e) the group delay function of the signal shown in (d), and (f) the modified group delay function of the signal shown in (d).

6. SIGNIFICANCE OF SPECTRAL RECONSTRUCTIONS IN COMBINING MAGNITUDE AND PHASE-BASED FEATURES

In this section, we reconstruct the formant structures or the respective short-time spectra from the MODGDF, MFCC, and the joint features. The MODGDF is derived from the modified group delay spectra as

$$c_p(n) = \sum_{k=0}^{k=N_f} \tau_m(k) \cos\left(\frac{n(2k+1)\pi}{N_f}\right), \tag{16}$$

where N_f is the DFT order. It is emphasized here that there are no filter banks used in the computation of the MODGDF. The MFCC are derived from the short time power spectra as

$$c_m(n) = \sum_{k=1}^{k=N_{fb}} X_k \cos\left(\frac{n(k-1/2)\pi}{N_{fb}}\right), \tag{17}$$

where $n = 1, 2, 3, \ldots, M$ represents the number of cepstral coefficients and $k = 1, 2, 3, \ldots, N_{fb}$ the number of filters used. X_k represents the log-energy output of the kth filter. The joint features (MODGDF + MFCC) are derived by appending the MODGDF vectors calculated as in (16) with the MFCC vectors calculated as in (17). The number of cepstral coefficients used in both the MODGDF and the MFCC is the same. To reconstruct the formant structures or the short-time spectra from the cepstra, an inverse DCT of the original DFT order has to be performed on the cepstra. The reconstructed modified group delay spectra as derived from the MODGDF is given by

$$\tau_m(k) = \sum_{n=0}^{n=N_f} c_p(n) \cos\left(\frac{n(2k+1)\pi}{N_f}\right), \tag{18}$$

where N_f is the DFT order, while the reconstructed short-time power spectra derived from the MFCC is given by

$$X_k = \sum_{n=0}^{n=N_{fb}} c_m(n) \cos\left(\frac{n(2k+1)\pi}{N_{fb}}\right), \tag{19}$$

where $n = 1, 2, 3, \ldots, M$ represents the original number of cepstral coefficients and $k = 1, 2, 3, \ldots, N_{fb}$ the original number of filters used. X_k represents the reconstructed log-energy output of the kth filter. The smooth frequency response of the original DFT order is computed by interpolating the filter bank reconstructed energies. The short-time spectra for the joint features are reconstructed as a three-step process. First, the short-time modified group delay spectra of the original DFT order are reconstructed from the n-dimensional MODGDF as in (18). Then the short-time power spectra of the original DFT order are reconstructed from the n-dimensional MFCC using (19) and an interpolation of the resulting filter bank energies. Finally, the short-time power spectra reconstructed from the MFCC and the short-time modified group delay spectra reconstructed from the MODGDF are averaged to derive the short-time composite spectra of the original DFT order. Note that the dimensionality of the MODGDF and the MFCC is the same.

6.1. Spectral reconstruction for a synthetic vowel

Typically, a vowel spectrum is characterized by the first three formants. Assuming a source system model of speech

production, the transfer function of such a system is given by:

$$H(z) = \frac{\sum_{\forall k} b_k z^{-k}}{\sum_{\forall k} a_k z^{-k}}. \tag{20}$$

The transfer function of the same system for the production of vowel assuming an all pole model is given by

$$H(z) = \frac{1}{\sum_{k=0}^{k=p} a_k z^{-k}}, \tag{21}$$

$$H(z) = \frac{1}{1 + \sum_{k=1}^{k=p} a_k z^{-k}}. \tag{22}$$

Let the vowel be characterized by the frequencies $F1$, $F2$, $F3$. Hence the poles of the system are located at

$$p = re^{\pm j\omega_i}. \tag{23}$$

By substituting (23) in (21), the system function for production of the ith formant now becomes

$$H_i(z) = \frac{1}{1 - 2r\cos\omega_i T z^{-1} + r^2 z^{-2}}. \tag{24}$$

But from resonance theory

$$r = e^{-\pi B_i T}. \tag{25}$$

By substituting (25) in (24), the system function in (24) now becomes

$$H_i(z) = \frac{1}{1 - 2e^{-\pi B_i T}\cos\omega_i T z^{-1} + e^{-2\pi B_i T} z^{-2}}. \tag{26}$$

In the above array of equations, ω_i corresponds to the ith formant frequency, B_i to the bandwidth of the ith formant frequency, and T to the sampling period. Using (26), we generate a synthetic vowel with the following values: $F1 = 500\,\text{Hz}$, $F2 = 1500\,\text{Hz}$, $F3 = 3500\,\text{Hz}$, $B_i = 10\%$ of F_i, and $T = 0.0001$ second corresponding to a sampling rate of 10 KHz. Note that $F1$, $F2$, and $F3$ are the formant frequencies in Hz. We then extract the MODGDF, MFCC, and joint features (MODGDF + MFCC) from the synthesized vowel. To reconstruct the formants, we use the algorithm described above. The reconstructed formant structures derived from the MODGDF, MFCC, joint features (MODGDF + MFCC), and also RASTA filtered MFCC are shown in Figures 3(a), 3(b), 3(c), and 3(d), respectively. The illustrations are shown as spectrogram like plots[2] where the data along the Y-axis correspond to the DFT bins and the x-axis corresponds to the frame number. It is interesting to note that while the formants are reconstructed accurately by both the MODGDF and the MFCC as in Figures 3(a) and 3(b), respectively, joint features (MODGDF + MFCC) combine the formant information in the individual features as in Figure 3(c). It is expected that RASTA filtered MFCC shown in Figure 3(d) does not capture the formant structure for the synthetic signal. RASTA filtered MFCC reconstructions are illustrated here only to show that while the MODGDF works well on clean speech, RASTA MFCC fails, which is quite expected. Although one may hypothesize that individual features capture formant information well and the need for joint features is really not there, it would be significant to note that joint features combine information gathered from individual features as in Figure 3(c). Similar spectrogram-like plots to illustrate formant reconstructions for a synthetic speech signal with varying formant trajectories are shown in Figure 4. It is interesting to note that in Figure 4(a), all the 3 formants are visible. In Figure 4(b), while the first 2 formants are visible, the third formant is not clearly visible. In Figure 4(c), while the first 2 formants are clear, the third formant is further emphasized. Hence it is clear that joint features are able to combine the information that is available in both the MODGDF and MFCC.

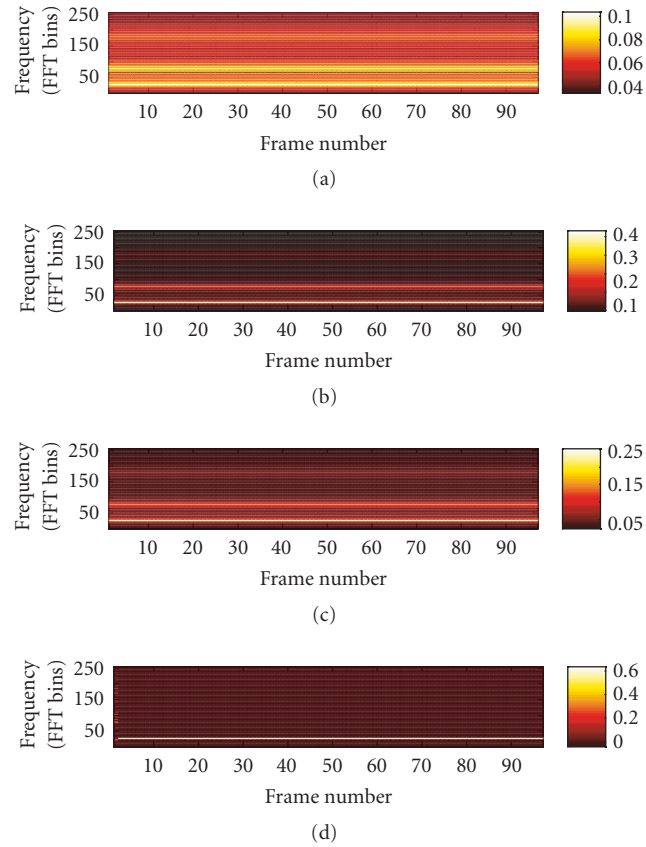

Figure 3: Spectrogram-like plots to illustrate formant reconstructions for a synthetic vowel. (a) The short-time modified group delay spectra reconstructed from MODGDF, (b) the short-time power spectra reconstructed from MFCC, (c) the short-time composite spectra reconstructed from joint features (MODGDF+MFCC), and (d) the short-time power spectra reconstructed from RASTA filtered MFCC.

[2] The differences between the subplots are better visualized in color than in gray scale.

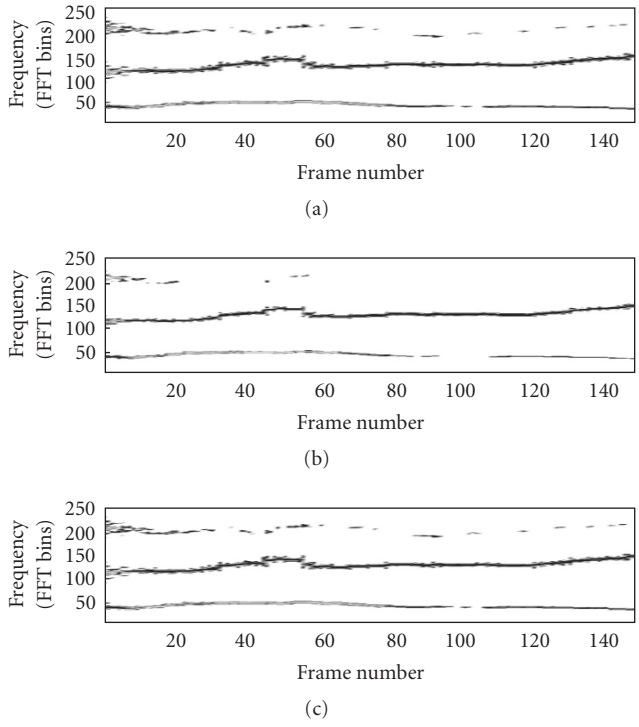

Figure 4: Spectrogram-like plots to illustrate formant reconstructions for a synthetic speech signal with varying formant trajectory. (a) The short-time modified group delay spectra reconstructed from the MODGDF, (b) the short-time power spectra reconstructed from MFCC, and (c) the short-time composite spectra reconstructed from joint features (MODGDF + MFCC).

7. DATABASES USED IN THE STUDY

There are four databases used in the study. The databases used are the database for Indian languages (DBIL) for syllable recognition [27], TIMIT [28] and NTIMIT [29] for speaker identification, and OGI_MLTS [32] for language identification.

7.1. The database for Indian languages (DBIL)

(i) DBIL Tamil database: this corpus consists of 20 news bulletins of the Tamil language transmitted by Doordarshan India, each of 15 minutes duration comprising 10 male and 10 female speakers. The total number of distinct syllables is 2184.

(ii) DBIL Telugu database: this corpus consists of 20 news bulletins of the Telugu language transmitted by Doordarshan India, each of 15 minutes duration comprising 10 male and 10 female speakers. The total number of distinct syllables is 1896.

7.2. The TIMIT database

The DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus was a joint effort among the Massachusetts Institute of Technology (MIT), Stanford Research Institute (SRI), and Texas Instruments (TI). TIMIT contains a total of 6300 sentences, 10 sentences spoken by each of 630 speakers from 8 major dialect regions of the United States.

7.3. The NTIMIT database

The NTIMIT corpus was developed by the NYNEX Science and Technology Speech Communication Group to provide a telephone bandwidth adjunct to the popular TIMIT Acoustic-Phonetic Continuous Speech Corpus. NTIMIT was collected by transmitting all the 6300 original TIMIT utterances through various channels in the NYNEX telephone network and redigitizing them. The actual telephone channels used were varied in a controlled manner, in order to sample various line conditions. The NTIMIT utterances were time-aligned with the original TIMIT utterances so that the TIMIT time-aligned transcriptions can be used with the NTIMIT corpus as well.

7.4. The OGI_MLTS database

The OGI multilanguage telephone speech corpus consists of telephone speech from 11 languages. The initial collection, included 900 calls, 90 calls each in 10 languages and was collected by Muthusamy et al. [32]. The languages are English, Farsi, French, German, Japanese, Korean, Mandarin, Spanish, Tamil, and Vietnamese. It is from this initial set that the training (50), development (20), and test (20) sets were established. The National Institute of Standards and Technology (NIST) uses the same 50-20-20 set that was established. The corpus is used by NIST for the evaluation of automatic language identification.

8. FEATURE EXTRACTION AND COMBINATION

In this section, we discuss the methods for feature extraction, tuning, and combination of various features before and after the acoustic model. The features used in this work are the MFCC, the spectral root compressed MFCC (SRMFC), the energy root compressed MFCC (ERMFC), the normalized spectral root compressed MFCC (NSRMFC), the linear frequency cepstral coefficients (LFCC), the spectral root compressed LFCC (SRLFC), and the MODGDF.

8.1. Computation and tuning of spectral magnitude-based features

The speech signal is first pre-emphasized and transformed to the frequency domain using a fast Fourier transform (FFT). The frame size used is 20 ms and the frame shift used is 10 ms. A hamming window is applied on each frame of speech prior to the computation of the FFT. The frequency scale is then warped using the bilinear transformation proposed by Acero [42]. The frequency scale is then multiplied by a bank of filters N_f whose center frequencies are uniformly distributed in the interval $[\mathrm{Min}_f, \mathrm{Max}_f]$ along the warped frequency axis. The filter shape used at the front end is trapezoidal and its width varies from one center frequency

to another. The shape of the filter is controlled by a constant which varies from 0 to 1, where 0 corresponds to triangular and 1 corresponds to rectangular. The filter bank energies are then computed by integrating the energy in each filter. A discrete cosine transform (DCT) is then used to convert the filter bank log energies to cepstral coefficients. Cepstral mean subtraction is always applied when working with noisy telephone speech. The front end parameters are tuned carefully as in [43] for computing the MFCC so that the best performance is achieved. The LFCC are computed in a similar fashion except that the frequency warping is not done as in the computation of the MFCC. The velocity, acceleration, and the energy parameters are added for both the MFCC and LFCC in a conventional manner. The spectral root compressed MFCC are computed as described in [44] and the energy root compressed MFCC as in [35]. The computation of the spectral root compressed MFCC is the same as the computation of the MFCC except that instead of taking a log of the FFT spectrum, we raise the FFT spectrum to a power γ where the value of γ ranges from 0 to 2. In the computation of the energy root compressed MFCC, instead of raising the FFT spectrum to the root value, the Mel frequency filter bank energies are compressed using the root value. In the energy root compressed case, the value of the root used for compression can range from 0 to 1. The normalized spectral root compressed MFCC is computed by normalizing the short-time power spectrum with its cepstrally smoothed version, followed by root compression as in the case of the spectral root compressed MFCC. It is emphasized here that all the free parameters involved in the computation of all these features including the root value and the cepstral window length used in spectral smoothing have been tuned carefully so that they give the best performance and are not handicapped in any way when they are compared with the MODGDF. The values of the spectral root and the energy root used in the experiments are 2/3 and 0.08, respectively. The velocity, acceleration and the energy parameters are augmented to both forms of the root compressed MFCC in a conventional manner. Note that all free parameters in all the aforementioned features have been tuned using line search on a validation data set selected from the particular database.

8.2. Computation and tuning of free parameters for the MODGDF

There are three free parameters lifter_w, α, and γ involved in the computation of the MODGDF as discussed in Section 4. From the results of initial experiments on the databases described in Section 7, we fix the length of the lifter_w to 8 although the performance remains nearly the same for lengths from 4 to 9. Any value greater than 9 brings down the performance. Having fixed the length of the lifter_w, we then fix the values of α and γ. In order to estimate the values of α and γ, an extensive optimization was carried out for the SPINE database [45] for phoneme recognition. To ensure that the optimized parameters were not specific to a particular database, we collected the sets of parameters that gave best performance on the SPINE database and tested them on other databases like the DBIL database (for syllable recognition), TIMIT, NTIMIT (for speaker identification), and the OGI_MLTS database (for language identification). The values of the parameters that gave the best performance across all databases and across all tasks were finally chosen for the experiments. The optimization technique uses successive line searches. For each iteration, α is held constant and γ is varied from 0 to 1 in increments of 0.1 (line search) and the recognition rate is noted for the three tasks on the aforementioned databases. The value of γ that maximizes the recognition rate is fixed as the optimal value. A similar line search is performed on α (varying it from 0 to 1 in increments of 0.1) keeping γ fixed. Finally, the set of values of α and γ that give the lowest error rate across the three tasks is retained. The series of experiments conducted to estimate the optimal values for lifter_w, α, and γ using line search are summarized in Table 1. Based on the results of such line searches, the best front end for the MODGDF across all tasks is listed in Table 2.

Table 1: Series of experiments conducted on various databases with the MODGDF.

Experiments conducted on the various databases
$N_c = 10, 12, 13, 16$
$\gamma = \{0.1 - 1.0\}$ in increments of 0.1
$\alpha = \{0.1 - 1.0\}$ in increments of 0.1
$\text{lifter}_w = 4, 6, 9, 10, 12$

Table 2: Best front end for the MODGDF across all databases.

γ	α	lifter_w	N_c
0.9	0.4	8	13

8.3. Extraction of joint features before the acoustic model

The following method is used to derive joint features by combining features before the acoustic model.

(i) Compute 13-dimensional MODGDF and the MFCC streams appended with velocity, acceleration, and energy parameters.

(ii) Use feature stream combination to append the 42-dimensional MODGDF stream to the 42-dimensional MFCC stream to derive an 84-dimensional joint feature stream.

Henceforth, we use the subscript bm for joint features thus derived.

8.4. Likelihood combination after the acoustic model

The following method is used to do a likelihood combination after the acoustic model.

(i) Compute 13-dimensional MODGDF and the MFCC streams appended with velocity, acceleration, and energy parameters.

(ii) Build a Gaussian mixture model (GMM) (for phoneme, speaker, and language recognition tasks) or a hidden Markov model (HMM) (for the continuous speech recognition task).

(iii) Compute the output probability of the acoustic model for different features.

(iv) The combined output log likelihood due to different feature streams is given by

$$\log P_f = \sum_{i=1}^{M} a_{(i,f)} \log P_{(i,f)} \tag{27}$$

and $a_{i,f}$ is the weight assigned to the ith feature stream, and M is the number of feature streams used. First, a rank is assigned to the log likelihood due to each individual feature stream based on its value. The higher log likelihood value gets a higher rank. The weights $a_{i,f}$ are now computed as the reciprocal of the rank assigned.

(v) Make a decision based on maximization of the combined output log likelihood.

Henceforth, we use the subscript am for likelihood combination. Table 3 summarizes the results of performance evaluation using both feature and likelihood combination techniques.

9. PERFORMANCE EVALUATION

In this section, we first discuss the significance of dimensionality of the feature vector and the size of training data. The results of performance evaluation of the MODGDF, MFCC, LFCC, NSRMFC, SRMFC, SRLFC, and joint features derived by combining these features for syllable, speaker, and language recognition are then presented. To enable a fair comparison, the results of combining any two features derived from the short-time spectral magnitude and also the MODGDF are listed. Although we experimented with all possible combinations of all features, both at feature level and using likelihood combination, we present the results of combining MODGDF with MFCC and MFCC with the LFCC. This is because combining any two features derived from spectral magnitude gave very small or no improvements, while a combination of the MODGDF with any feature derived from the spectral magnitude gave significant improvements in recognition performance. It is also noticed that new features like the NSRMFC, SRMFC, ERMFC, and SRLFC give a small improvement in recognition performance compared to the MFCC when used in isolation, but do not give any improvement when combined with each other.

9.1. Significance of dimensionality of the feature vectors and training data

In the experimental results described in the following sections, we compute a 13-dimensional vector for each feature stream appended with energy, delta, and acceleration coefficients. For feature combination before the acoustic model, the features are concatenated to compute a joint feature stream. Experimental results indicate that simply increasing the dimensionality of each individual feature stream to match the dimensionality of the joint feature stream does not improve recognition performance for any of the three tasks mentioned above. We have also experimented with increased amounts of training data for such individual high dimensional feature streams to validate these results. For syllable recognition, training data was increased in increments of one news bulletin from the DBIL database. For the speaker and language identification tasks, the training data was increased in increments of two sentences from the TIMIT, NTIMIT, and the OGI_MLTS databases. The results from these experiments indicate that combining the MODGDF (derived from the short-time spectral phase) with features computed from the short-time spectral magnitude like MFCC gives an improvement in recognition performance even though the overall feature dimension is increased. It is hypothesized from these results that the MODGDF has some complementary information when compared to features derived from the short-time spectral magnitude.

9.2. Syllable-based speech recognition

In this section, we discuss the baseline system and experimental results for recognition of syllables on the DBIL Tamil and Telugu databases [27]. The baseline recognition system uses hidden Markov models trained apriori for 320 syllables for Tamil and 265 syllables for Telugu extracted from the broadcast news corpora from the DBIL database [27] of two Indian languages, Tamil and Telugu. The number of syllables used for training is selected based on their frequency of occurrence in the respective corpora. Any syllable that occurs more than 50 times in the corpus is selected as a candidate for which HMMs are built. All the HMMs built are 5-state and 3-mixture models. A separate model is built for silence. During the test phase, the test sentence is segmented at boundaries of syllabic units using minimum phase group delay functions derived from the root cepstrum as in [39]. These segments are now checked in isolated style against all HMMs built apriori. The HMM that gives the maximum likelihood value is declared as the correct match. The segments hence recognized are concatenated in the same order as they were segmented to realize the recognized sentence. For DBIL data of Telugu language, the MODGDF (MGD) recognition performance was at 36.6%, MFCC (MFC) at 39.6%, ($\{\mathbf{MGD} + \mathbf{MFC}\}_{\mathbf{bm}}$) at 50.6%, and ($\{\mathbf{MGD} + \mathbf{MFC}\}_{\mathbf{am}}$) at 44.6% for this task. The best increase due to feature combination was 11%. For DBIL data of Tamil language, the MODGDF (MGD) recognition performance was at 35.1%, MFCC (MFC) at 37.1%, ($\{\mathbf{MGD} + \mathbf{MFC}\}_{\mathbf{bm}}$) at 48.9%, and ($\{\mathbf{MGD} + \mathbf{MFC}\}_{\mathbf{am}}$) at 41.7% for this task. The best increase due to feature combination was 11% as indicated in Table 3. The results for combining the MODGDF with the LFCC are also tabulated in Table 3 and show very small improvements. It is worthwhile to note that syllable recognition performance is improved significantly by combining features before the acoustic model when compared to combining the likelihoods.

Table 3: Results of performance evaluation for three speech processing tasks: syllable, speaker, and language recognition for MODGDF(**MGD**), MFCC(**MFC**), LFCC(**LFC**), spectral root compressed MFCC(**SRMFC**), normalized spectral root compressed MFCC(**NSRMFC**), energy root compressed MFCC(**ERMFC**), spectral root compressed LFCC(**SRLFC**), MODGDF and MFCC combined before the acoustic model($\{\mathbf{MGD} + \mathbf{MFC}\}_{\mathbf{bm}}$), likelihood combination of MODGDF and MFCC after the acoustic model($\{\mathbf{MGD} + \mathbf{MFC}\}_{\mathbf{am}}$), MFCC and LFCC combined before the acoustic model($\{\mathbf{MFC} + \mathbf{LFC}\}_{\mathbf{bm}}$), and likelihood combination of MFCC and LFCC after the acoustic model($\{\mathbf{MFC} + \mathbf{LFC}\}_{\mathbf{am}}$).

Task	Feature	Database	Train data	Test data	Classifier	Recog. (%)	Inc. in Recog. (%)
Syllable recognition	**MGD**	DBIL (TELUGU)	10 news bulletins 15 mt. duration	2 news bulletins 9400 syllables	HMM 5 states 3 mixtures	36.6%	—
	MFC					39.6%	—
	LFC					32.6%	—
	SRMFC					35.6%	—
	NSRMFC					36%	—
	ERMFC					38%	—
	SRLFC					34.2%	—
	$\{\mathbf{MGD} + \mathbf{MFC}\}_{\mathbf{bm}}$					50.6%	11%
	$\{\mathbf{MGD} + \mathbf{MFC}\}_{\mathbf{am}}$					44.6%	5%
	$\{\mathbf{MFC} + \mathbf{LFC}\}_{\mathbf{bm}}$					41.6%	2%
	$\{\mathbf{MFC} + \mathbf{LFC}\}_{\mathbf{am}}$					40.6%	1%
Syllable recognition	**MGD**	DBIL (TELUGU)	10 news bulletins 15 mt. duration	2 news bulletins 9400 syllables	HMM 5 states 3 mixtures	35.1%	—
	MFC					37.1%	—
	LFC					31.2%	—
	SRMFC					34.1%	—
	NSRMFC					34.5%	—
	ERMFC					36.5%	—
	SRLFC					32.4%	—
	$\{\mathbf{MGD} + \mathbf{MFC}\}_{\mathbf{bm}}$					48.9%	11%
	$\{\mathbf{MGD} + \mathbf{MFC}\}_{\mathbf{am}}$					41.7%	4.6%
	$\{\mathbf{MFC} + \mathbf{LFC}\}_{\mathbf{bm}}$					39.1%	2%
	$\{\mathbf{MFC} + \mathbf{LFC}\}_{\mathbf{am}}$					38%	0.9%
Speaker identification	**MGD**	TIMIT	6 sentences/speaker	4 sentences/speaker	GMM 64 mixtures	98%	—
	MFC					98%	—
	LFC					96.25%	—
	SRMFC					97.25%	—
	NSRMFC					97%	—
	ERMFC					98%	—
	SRLFC					97%	—
	$\{\mathbf{MGD} + \mathbf{MFC}\}_{\mathbf{bm}}$					99%	1%
	$\{\mathbf{MGD} + \mathbf{MFC}\}_{\mathbf{am}}$					99%	1%
	$\{\mathbf{MFC} + \mathbf{LFC}\}_{\mathbf{bm}}$					98%	0%
	$\{\mathbf{MFC} + \mathbf{LFC}\}_{\mathbf{am}}$					98%	0%
Speaker identification	**MGD**	NTIMIT	6 sentences/speaker	4 sentences/speaker	GMM 64 mixtures	41%	—
	MFC					40%	—
	LFC					30.25%	—
	SRMFC					34.25%	—
	NSRMFC					35%	—
	ERMFC					34.75%	—
	SRLFC					31.75%	—
	$\{\mathbf{MGD} + \mathbf{MFC}\}_{\mathbf{bm}}$					47%	6%
	$\{\mathbf{MGD} + \mathbf{MFC}\}_{\mathbf{am}}$					45%	4%
	$\{\mathbf{MFC} + \mathbf{LFC}\}_{\mathbf{bm}}$					40%	0%
	$\{\mathbf{MFC} + \mathbf{LFC}\}_{\mathbf{am}}$					40%	0%

Table 3: Continued.

Task	Feature	Database	Train data	Test data	Classifier	Recog. (%)	Inc. in Recog. (%)
Language identification	**MGD**	OGI_MLTS 11-language task	45 sentences 40 males and 5 females	20 sentences 18 males and 2 females	GMM 64 mixtures	53%	—
	MFC					50%	—
	LFC					47%	—
	SRMFC					50.4%	—
	NSRMFC					50.8%	—
	ERMFC					50.6%	—
	SRLFC					48%	—
	$\{\mathbf{MGD} + \mathbf{MFC}\}_{\mathbf{bm}}$					58%	5%
	$\{\mathbf{MGD} + \mathbf{MFC}\}_{\mathbf{am}}$					57%	4%
	$\{\mathbf{MFC} + \mathbf{LFC}\}_{\mathbf{bm}}$					51%	1%
	$\{\mathbf{MFC} + \mathbf{LFC}\}_{\mathbf{am}}$					50.5%	0.5%

9.3. Speaker identification

In this section, we discuss the baseline system and experimental results for speaker identification on the TIMIT database (clean speech) and the NTIMIT database (noisy telephone speech). A series of GMMs modeling the voices of speakers for whom training data is available and a classifier that evaluates the likelihoods of the unknown speakers voice data against these models make up the likelihood maximization-based baseline system used in this section. Single state, 64 mixture Gaussian mixture models (GMMs) are trained for each of the 630 speakers in the database. The number of sentences used for training each speaker's model is 6, while 4 sentences are used to test a particular speaker during the testing phase. Hence a total of 400 tests are conducted to identify 100 speakers and the number of tests goes up to 2520 for identifying 630 speakers. For the TIMIT (clean speech) data [28], the MODGDF (MGD) recognition performance was at 98%, MFCC (MFC) at 98%, ($\{\mathbf{MGD} + \mathbf{MFC}\}_{\mathbf{bm}}$) at 99%, and ($\{\mathbf{MGD} + \mathbf{MFC}\}_{\mathbf{am}}$) at 99% for this task. The best increase due to feature combination was 1%. While for the NTIMIT (noisy telephone speech) data [29], the MODGDF (MGD) recognition performance was at 41%, MFCC (MFC) at 40%, ($\{\mathbf{MGD} + \mathbf{MFC}\}_{\mathbf{bm}}$) at 47%, and ($\{\mathbf{MGD} + \mathbf{MFC}\}_{\mathbf{am}}$) at 45% for this task. The best increase due to feature combination was 6% as indicated in Table 3. The results for combining the MODGDF with the LFCC are also tabulated in Table 3 and show minor improvements.

9.4. Language identification

In this section, we discuss the baseline system and experimental results for language identification on the OGI_MLTS database (11-language task for noisy telephone speech). The baseline system used for this task is very similar to the system used for the automatic speaker identification task as in Section 9.3, except that each language is now modeled by a GMM. From the 90 phrases available for each language, 45 are used for training and 20 are used for testing. The duration of the test utterance is 45 seconds. The results of the MODGDF and the MFCC on the OGI_MLTS [32] corpora using the GMM scheme are listed in Table 3. For the 11-language task on the OGI_MLTS data, the MODGDF (MGD) recognition performance was at 53%, MFCC (MFC) at 50%, ($\{\mathbf{MGD} + \mathbf{MFC}\}_{\mathbf{bm}}$) at 58%, and ($\{\mathbf{MGD} + \mathbf{MFC}\}_{\mathbf{am}}$) at 57% for this task. The best increase due to feature combination was 5%. The results for combining the MODGDF with the LFCC are also tabulated in Table 3, and indicate that combining two spectral magnitude-based features does not give significant improvements for the language identification task. It was also noticed from the confusion matrix created from our recognition experiments that Japanese and Korean had a high degree of confusion between themselves, and that the MODGDF was able to identify Korean better, while the MFCC performed better in recognizing Japanese.

10. CONCLUSION

This paper discusses the significance of joint features derived by combining the short-time magnitude and phase spectra in speech recognition. Indeed, the MODGDF and its significance in speech processing have been investigated in earlier efforts. The idea of combining cepstral features derived from the short-time magnitude spectra and from the modified group delay function both at feature level and at likelihood level is proposed in this paper. It is illustrated that joint cepstral features derived from the modified group delay function and MFCC essentially capture complete spectral information in the speech signal. The advantage of using joint features for noisy data and related robustness issues are discussed. The joint features are used for three speech processing tasks, namely, syllable, speaker, and language recognition. The results of the performance evaluation indicate that joint features improve recognition performance up to 11% for feature combination before the acoustic model and up to 5% for likelihood combination after the acoustic model. The results of the performance evaluation presented in this work indicate that the MODGDF complements the features derived from the short-time power spectra like the MFCC. Recognition results indicate that combining features at the feature level gave significant improvements for syllable recognition and speaker identification when compared to the method of likelihood combination. Although the results of

the performance evaluation indicate the complementarity of the MODGDF to the MFCC, it is not clear how a measure of complementarity can be defined. The use of feature pruning techniques like the sequential floating forward search with appropriate distance measures to reduce the dimensionality of the joint feature stream is another issue that needs to be addressed.

ACKNOWLEDGMENT

The work of Rajesh M. Hegde was supported by the National Science Foundation under Award numbers 0331707 and 0331690—http://www.itr-rescue.org.

REFERENCES

[1] L. R. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition*, Prentice Hall, Englewood Cliffs, NJ, USA, 1993.

[2] K. Aikawa, H. Singer, H. Kawahara, and Y. Tohkura, "A dynamic cepstrum incorporating time-frequency masking and its application to continuous speech recognition," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '93)*, vol. 2, pp. 668–671, Minneapolis, Minn, USA, April 1993.

[3] M. Bacchiani and K. Aikawa, "Optimization of time-frequency masking filters using the minimum classification error criterion," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '94)*, vol. 2, pp. 197–200, Adelaide, SA, Australia, April 1994.

[4] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *Journal of the Acoustical Society of America*, vol. 87, no. 4, pp. 1738–1752, 1990.

[5] O. Ghitza, "Auditory models and human performance in tasks related to speech coding and speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 1, part 2, pp. 115–132, 1994.

[6] K. L. Payton, "Vowel processing by a model of the auditory periphery: a comparison to eighth-nerve responses," *The Journal of the Acoustical Society of America*, vol. 83, no. 1, pp. 145–162, 1988.

[7] R. Lyon, "A computational model of filtering, detection, and compression in the cochlea," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '82)*, vol. 7, pp. 1282–1285, Paris, France, May 1982.

[8] S. Seneff, "A joint synchrony/mean-rate model of auditory speech processing," *Journal of Phonetics*, vol. 16, no. 1, pp. 55–76, 1988.

[9] J. R. Cohen, "Application of an auditory model to speech recognition," *The Journal of the Acoustical Society of America*, vol. 85, no. 6, pp. 2623–2629, 1989.

[10] M. J. Hunt, S. M. Richardson, D. C. Bateman, and A. Piau, "An investigation of PLP and IMELDA acoustic representations and of their potential for combination," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '91)*, vol. 2, pp. 881–884, Toronto, Ont, Canada, May 1991.

[11] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 4, pp. 357–366, 1980.

[12] K. K. Paliwal and L. D. Alsteris, "On the usefulness of STFT phase spectrum in human listening tests," *Speech Communication*, vol. 45, no. 2, pp. 153–170, 2005.

[13] L. D. Alsteris and K. K. Paliwal, "Some experiments on iterative reconstruction of speech from STFT phase and magnitude spectra," in *Proceedings of 9th European Conference on Speech Communication and Technology (EUROSPEECH '05)*, pp. 337–340, Lisbon, Portugal, September 2005.

[14] H. A. Murthy and V. R. R. Gadde, "The modified group delay function and its application to phoneme recognition," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '03)*, vol. 1, pp. 68–71, Hong Kong, April 2003.

[15] R. M. Hegde, H. A. Murthy, and V. R. R. Gadde, "Application of the modified group delay function to speaker identification and discrimination," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '04)*, vol. 1, pp. 517–520, Montreal, Quebec, Canada, 2004.

[16] R. M. Hegde, H. A. Murthy, and V. R. R. Gadde, "Continuous speech recognition using joint features derived from the modified group delay function and MFCC," in *Proceedings of 8th International Conference on Spoken Language Processing (INTERSPEECH '04)*, vol. 2, pp. 905–908, Jeju Island, Korea, October 2004.

[17] R. M. Hegde, H. A. Murthy, and V. R. R. Gadde, "The modified group delay feature: a new spectral representation of speech," in *Proceedings of 8th International Conference on Spoken Language Processing (INTERSPEECH '04)*, vol. 2, pp. 913–916, Jeju Island, Korea, October 2004.

[18] R. M. Hegde, H. A. Murthy, and V. R. R. Gadde, "Significance of the modified group delay feature in speech recognition," to appear in *IEEE Transactions on Speech and Audio Processing*.

[19] R. M. Hegde, H. A. Murthy, and V. R. R. Gadde, "Speech processing using joint features derived from the modified group delay function," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '05)*, vol. 1, pp. 541–544, Philadelphia, Pa, USA, March 2005.

[20] S. Okawa, E. Bocchieri, and A. Potamianos, "Multi-band speech recognition in noisy environments," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '98)*, vol. 2, pp. 641–644, Seattle, Wash, USA, May 1998.

[21] D. Ellis, "Feature stream combination before and/or after the acoustic model," Tech. Rep. TR-00-007, International Computer Science Institute, Berkeley, Calif, USA, 2000.

[22] H. Christensen, "Speech recognition using heterogenous information extraction in multi-stream based systems," Ph.D. dissertation, Aalborg University, Aalborg, Denmark, 2002.

[23] B. E. D. Kingsbury and N. Morgan, "Recognizing reverberant speech with RASTA-PLP," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '97)*, vol. 2, pp. 1259–1262, Munich, Germany, April 1997.

[24] S.-L. Wu, B. E. D. Kingsbury, N. Morgan, and S. Greenberg, "Incorporating information from syllable-length time scales intoautomatic speech recognition," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '98)*, vol. 2, pp. 721–724, Seattle, Wash, USA, May 1998.

[25] A. Janin, D. Ellis, and N. Morgan, "Multi-stream speech recognition: ready for prime time?" in *Proceedings of 6th European Conference on Speech Communication and Technology (EUROSPEECH '99)*, pp. 591–594, Budapest, Hungary, September 1999.

[26] K. Kirchhoff and J. A. Bilmes, "Dynamic classifier combination in hybrid speech recognition systems using utterance-level confidence values," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '99)*, vol. 2, pp. 693–696, Phoenix, Ariz, USA, March 1999.

[27] *Database for Indian Languages*, Speech and Vision Lab, IIT Madras, Chennai, India, 2001.

[28] NTIS, *The DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus*, 1993.

[29] C. Jankowski, A. Kalyanswamy, S. Basson, and J. Spitz, "*NTIMIT*: a phonetically balanced, continuous speech, telephone bandwidth speech database," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '90)*, vol. 1, pp. 109–112, Albuquerque, NM, USA, April 1990.

[30] L. Besacier and J. F. Bonastre, "Time and frequency pruning for speaker identification," in *Proceedings of the 14th International Conference on Pattern Recognition (ICPR '98)*, vol. 2, pp. 1619–1621, Brisbane, Qld., Australia, August 1998.

[31] K. L. Brown and E. B. George, "CTIMIT: a speech corpus for the cellular environment with applications to automatic speech recognition," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '95)*, vol. 1, pp. 105–108, Detroit, Mich, USA, May 1995.

[32] Y. K. Muthusamy, R. A. Cole, and B. T. Oshika, "The OGI multi-language telephone speech corpus," in *Proceedings of the 2nd International Conference on Spoken Language Processing (ICSLP '92)*, pp. 895–898, Banff, Alberta, Canada, October 1992.

[33] K. Turner, "Linear and order statistics combiners for reliable pattern classification," Ph.D. dissertation, University of Texas at Austin, Austin, Tex, USA, May 1996.

[34] M. P. Perrone and L. N. Cooper, "When networks disagree: ensemble methods for hybrid neural networks," in *Neural Networks for Speech and Image Processing*, pp. 126–142, Chapman-Hall, London, UK, 1993.

[35] R. Sarikaya and J. H. L. Hansen, "Analysis of the root-cepstrum for acoustic modeling and fast decoding in speech recognition," in *Proceedings of the 7th European Conference on Speech Communication and Technology (EUROSPEECH '01)*, pp. 687–690, Aalborg, Denmark, September 2001.

[36] A. Krogh and J. Vedelsby, "Neural network ensembles, cross validation, and active learning," in *Advances in Neural Information Processing Systems*, vol. 7, pp. 231–238, MIT Press, Cambridge, Mass, USA, 1995.

[37] H. A. Murthy and B. Yegnanarayana, "Formant extraction from group delay function," *Speech Communication*, vol. 10, no. 3, pp. 209–221, 1991.

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

[39] V. K. Prasad, T. Nagarajan, and H. A. Murthy, "Automatic segmentation of continuous speech using minimum phase group delay functions," *Speech Communication*, vol. 42, no. 3-4, pp. 429–446, 2004.

[40] B. Yegnanarayana and H. A. Murthy, "Significance of group delay functions in spectrum estimation," *IEEE Transactions on Signal Processing*, vol. 40, no. 9, pp. 2281–2289, 1992.

[41] P. Yip and K. R. Rao, *Discrete Cosine Transform: Algorithms, Advantages, and Applications*, Academic Press, San Diego, Calif, USA, 1997.

[42] A. Acero, "Acoustical and environmental robustness in automatic speech recognition," Ph.D. dissertation, Carnegie Mellon University, Pittsburgh, Pa, USA, 1990.

[43] H. A. Murthy, F. Beaufays, L. P. Heck, and M. Weintraub, "Robust text-independent speaker identification over telephone channels," *IEEE Transactions on Speech and Audio Processing*, vol. 7, no. 5, pp. 554–568, 1999.

[44] P. Alexandre and P. Lockwood, "Root cepstral analysis: a unified view. Application to speech processing in car noise environments," *Speech Communication*, vol. 12, no. 3, pp. 277–288, 1993.

[45] V. R. R. Gadde, A. Stolcke, J. Z. D. Vergyri, K. Sonmez, and A. Venkatraman, "The SRI SPINE 2001 Evaluation System," SRI: Menlo Park, Calif, USA, 2001.