SPEECH PROCESSING USING JOINT FEATURES DERIVED FROM THE MODIFIED GROUP DELAY FUNCTION

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

This paper discusses the significance of joint cepstral features derived from the modified group delay function and MFCC in speech processing. We start with a definition of cepstral features derived from the modified group delay function called the modified group delay feature (MODGDF) which is derived from the Fourier transform phase. Robustness issues like similarities of the MODGDF to RASTA and cepstral mean subtraction are discussed. The efficiency with which formants can be reconstructed for noisy cellular speech using joint features derived from early fusion is illustrated. The joint features are used for four speech processing tasks phoneme, syllable, speaker, and language recognition. Based on the results of analysis and performance evaluation the significance of joint features derived from the MODGDF and MFCC are discussed.

1. INTRODUCTION

Representations of speech are almost always derived from the short-term power spectrum, and the short term phase spectrum is ignored. This is primarily because human ears are considered largely insensitive to phase effects and hence speech communication and recording equipment often do not preserve the phase structure of the original waveform. Contradictory evidence to the non usefulness of short time phase spectra for perception of speech can also be found in literature. Drucker's [1], perceptual tests have indicated that the primary cause of loss in intelligibility is the confusion among fricatives and plosives. Further this loss is primarily due to the loss of the short pauses immediately before the plosive sounds. But the results from the psychoacoustic experiments conducted by Helmholtz and Liu [2, 3] indicate that perception of voicing for plosive sounds depends strongly on phase information. It is emphasized here that all the aforementioned claims about the non usefulness of short time phase spectra of speech are based on perception or intelligibility of speech and not on the basis how well speech can be recognized by a machine. Yegnanarayana and Murthy [4] have used group delay functions for various speech processing tasks. Atal, Alsteris, Ney and Paliwal [5] have made some important contributions to the usefulness of short time phase spectra. In this context the short time phase spectra via the group delay domain has been used to parameterize speech in our earlier attempts [6, 7]. We have also proposed an alternative representation of speech which uses the modified group delay function derived from the Fourier transform phase spectra [8]. In this paper we focus on the significance of the representation of speech using joint features derived from the MODGDF and the MFCC. The focus of this paper is on combining features before the acoustic model as well as after the acoustic model [9]. The modified group delay function and extraction of joint features is discussed first. Certain robustness issues [10] are analyzed next, followed by a discussion on formant reconstruction of noisy cellular speech from joint features. The joint features derived from the MODGDF and other short time magnitude spectrum derived features are used for the tasks of automatic identification of phoneme, syllable [11], speaker, and language [12] identification. Finally we conclude with a discussion on the significance of joint features in speech processing.

2. SIGNIFICANCE OF FEATURE COMBINATIONS BEFORE AND AFTER THE ACOUSTIC MODEL

A feature combination system works on the principle that some characteristics that are de emphasized by a particular feature are emphasized by another feature, and therefore the combined feature streams capture complementary information present in individual features. The combination of features before the acoustic model have been used by Ellis [9], where an effort 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. The approach to combine features after the acoustic model uses the technique of combining the outputs of the acoustic models. Complex techniques of combining the posteriors [9] have evolved but the classic way of simple averaging of the maximum likelihoods from different estimators is the best approach to feature combination after the acoustic model. 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 MODIFIED GROUP DELAY FEATURE

The theory and significance of the modified group delay feature (MODGDF) has been discussed in detail in our previous ICASSP papers [6, 7]. The modified group delay function as in [6, 7] is defined as

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

where

$$\tau(\omega) = (\frac{X_R(\omega)Y_R(\omega) + Y_I(\omega)X_I(\omega)}{S(\omega)^{2\gamma}}) \quad (2)$$

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

$$c(n) = \sum_{k=0}^{k=N_f} \tau_x(k) \cos(n(2k+1)\pi/N_f) \quad (3)$$

where N_f is the DFT order and $\tau_x(k)$ is the group delay function.

4. ROBUSTNESS ISSUES

4.1. Similarity to RASTA

RASTA (RelAtive SpecTrA) [10] is a popular technique used to handle speech degraded with both convolutional and white noise. RASTA filters out very low temporal frequency components below 1 Hz which are primarily due to the changing auditory environment. It also filters out higher frequency temporal components greater than 13 Hz as they represent changes faster than the speech articulators can move. The basic filter used for this purpose is an IIR filter with the transfer function :

$$H(z) = 0.1z^4 * \frac{2 + z^{-1} - z^{-3} - 2z{-}4}{1 - 0.98z^{-1}} \quad (4)$$

But the issue on which we propose to compare RASTA with the MODGDF is the use of a compressing static non linear transformation (generally a logarithm operation) on the critical band spectrum. Instead of compressing the PLP spectrum logarithmically as in RASTA, the group delay spectrum is raised to the power of α. It is emphasized here that by varying the value of α a greater control over the time trajectories of spectral components can be exercised. Further by varying the value of γ, control over the zeros lying on the unit circle can be gained. We would like to therefore point out that the MODGDF has two advantages over the RASTA technique, which are :

1. The MODGDF works well for both clean and noisy speech by setting appropriate values of α and γ, while RASTA fails for clean speech.
2. The MODGDF avoids additional processing steps of RASTA, like deriving the critical band spectrum and a compressing non linear transformation while also being immune to convolutional and additive noise.

4.2. Significance of cepstral mean subtraction

Cepstral mean subtraction (CMS) is a successful technique used to filter out linear distortions in speech passed through a telephone channel. Let T(z) be the Z transform of a telephone speech signal :

$$T(z) = S(z)G(z) \quad (5)$$

where S(z) is the Z transform of clean speech and G(z) the Z transform of the channel. In the log domain :

$$logT(z) = logS(z) + logG(z) \quad (6)$$

The cepstrally mean subtracted vector is given by :

$$c_{cms}(n) = c_{mgd}(n) - E[c_{mgd}(n)] \quad (7)$$

where $E[c_{mgd}(n)]$ is the expectation of the modified group delay cepstra taken over a number of frames of channel corrupted speech. It is emphasized in [10] that CMS is capable of handling convolutive noise only and therefore RASTA with CMS always significantly improves the performance of a speech recognition system. We have already shown that the MODGDF is immune to both convolutive and additive noise in [8]. Further applying CMS on the MODGDF also gave us a good improvement in recognition performance. Indeed by taking advantage of the additive property of group delay functions, spectral subtraction in the modified group delay domain can be used to avoid the additional step of CMS on the MODGDF.

4.3. Formant reconstruction for noisy cellular speech data

In this Section we extract the MODGDF, MFCC and joint features (MODGDF + MFCC) from a phrase of noisy cellular speech data picked from the CTIMIT database. The reconstructed formant structures derived from the MODGDF, MFCC, RASTA, and joint features (MODGDF + MFCC) are shown in Figures 1, 2, 3, and 4 respectively. It is significant to note that while the MFCC captures unwanted transitions in formant structures due to noise and channel effects as in Figure 2, RASTA is able to eliminate such transitions as in Figure 3. The significance of the MODGDF is evident in the way in which the formant information is captured for noisy cellular speech as in Figure 1. It is emphasized here that while RASTA fails for clean speech and works well for noisy speech, the MODGDF works in both cases. Hence it is true that joint features derived from the MODGDF and MFCC do capture

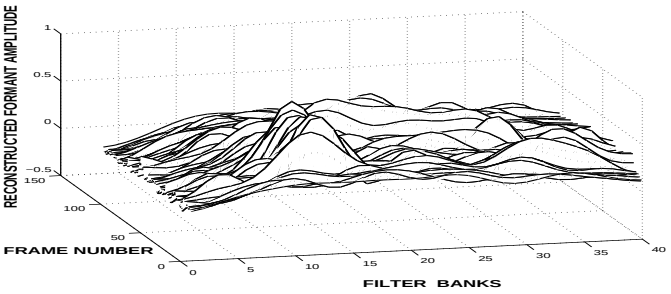

Fig. 1. Formants reconstructed from the MODGDF for noisy cellular speech

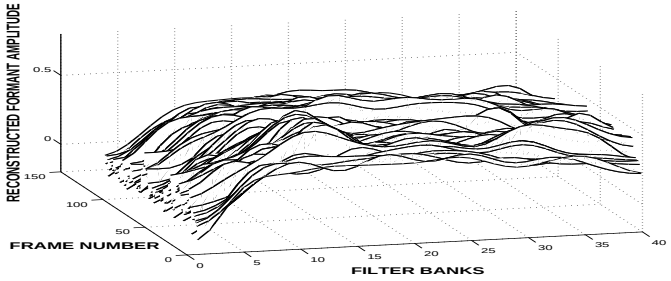

Fig. 2. Formants reconstructed from the MFCC for noisy cellular speech

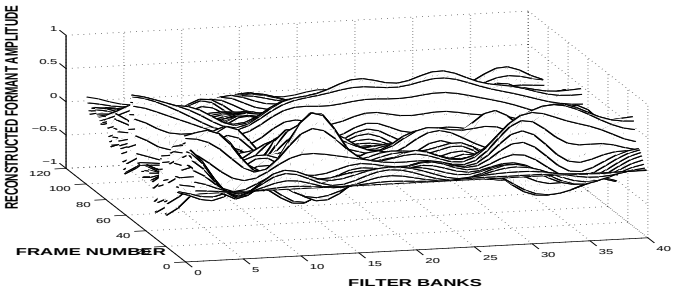

Fig. 3. Formants reconstructed from the RASTA for noisy cellular speech

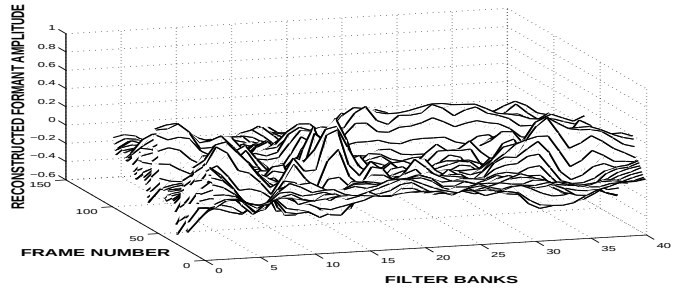

Fig. 4. Formants reconstructed from the joint features for noisy cellular speech

the complete formant information in the noisy speech signal as in Figure 4.

5. PERFORMANCE EVALUATION

In this section we present the results of performance evaluation of the MODGDF, MFCC and joint features derived by combining the MODGDF and MFCC, for phoneme, syllable, speaker, and language recognition. We have explored combining other features derived from the Fourier transform magnitude like the LFCC and model based features like the LPCC. But we present the results of of combining MFCC with the MODGDF since this combination gave the best results among all the other combinations.

5.1. Extraction of joint features before the acoustic model

The 13 dimensional MODGDF and the MFCC streams appended with velocity, acceleration and energy parameters are computed. The 42 dimensional MODGDF stream is appended to the 42 dimensional MFCC stream to derive a 84 dimensional joint feature stream. Henceforth we use the notation JF_{bm} for joint features thus derived.

5.2. Extraction of joint features after the acoustic model

The 13 dimensional MODGDF and the MFCC streams appended with velocity, acceleration and energy parameters are calculated. A GMM (for phoneme, speaker, and language tasks) or an HMM (for continuous speech recognition task) is built using these features. The posterior probability outputs derived using each model are combined by weighting the probabilities appropriately. A decision is made based on the maximization of the combined output probability. Henceforth we use the notation JF_{am} for joint features thus derived. Table 1 summarizes the results of performance evaluation using JF_{bm} and JF_{am}.

5.3. Experimental results for phoneme recognition

In this section we employ the MODGDF, MFCC, JF_{bm} and JF_{am} for recognition of speech in noisy environments (SPINE) evaluation task [6]. using the SPINE database [6]. Phonemes from 15000 utterances of the SPINE2000 database are used to train the models. Phonemes from 10000 utterances are used to test the models. Simple isolated single state GMM models are trained for each of the 48 phones (including pause) that is present in the database. The MODGDF recognition performance was at 57.3%, MFCC at 60.7%, JF_{bm} at 65.7%, and JF_{am} at 62.85% for this task. The best increase due to feature combination was 5% as indicated in Table 1.

5.4. Experimental results for syllable recognition

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 [11] of two Indian languages Tamil and Telugu. The number of syllables used for training are 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. A separate model is built for silence.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. For DBIL data of Telugu language the MODGDF recognition performance was at 36.6%, MFCC at 39.6%, JF_{bm} at 50.6%, and JF_{am} at 44.6% for this task. The best increase due to feature combination was 11%. For DBIL data of Tamil language the MODGDF recognition performance was at 35.1%, MFCC at 37.1%, JF_{bm} at 48.9%, and JF_{am} at 41.7% for this task. The best increase due to feature combination was 11% as indicated in Table 1.

5.5. Experimental results for automatic speaker identification

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. For the TIMIT (clean speech) data the MODGDF recognition performance was at 98%, MFCC at 98%, JF_{bm} at 99%, and JF_{am} at 99% for this task. The best increase due to feature combination was 1%. While for the NTIMIT (noisy telephone speech) data the MODGDF recognition performance was at 41%, MFCC at 40%, JF_{bm} at 47%, and JF_{am} at 45% for this task. The best increase due to feature combination was 6% as indicated in Table 1.

5.6. Experimental results for automatic language identification

The baseline system used for this task is very similar to the system used for the automatic speaker identification task, except that each language is now modeled by a GMM. The results of the MODGDF and the MFCC on both the DBIL and OGI_MLTS [12] corpora using the GMM scheme are listed in Table 1. For the 3 language task on the DBIL data the MODGDF recognition performance was at

Table 1. *Table summarizing the results of performance evaluation of the MODGDF (***MGD***), MFCC (***MFC***), JF_{bm}, and JF_{am} and increase in recognition (***Inc***) for four speech processing tasks phoneme, syllable, speaker, and language recognition*

Task	Feature	Database	Train Data	Test Data	Classifier	Recog	Inc.
Phoneme recognition	MGD	SPINE	15000 utterances	10000 utterances	GMM	**57.3%**	-
	MFC					60.7%	-
	JF_{bm}					65.7%	**5%**
	JF_{am}					62.85%	**2%**
Syllable recognition	MGD	DBIL (TELUGU)	10 news bulletins 15 mt. duration	2 news bulletins 9400 syllables	HMM	**36.6%**	-
	MFC					39.6%	-
	JF_{bm}					50.6%	**11%**
	JF_{am}					44.6%	**5%**
	MGD	DBIL (TAMIL)	10 news bulletins 15 mt. duration	2 news bulletins 9400 syllables	HMM	**35.1%**	-
	MFC					37.1%	-
	JF_{bm}					48.9%	**11%**
	JF_{am}					41.7%	**4.6%**
Speaker identification	MGD	TIMIT	10 sentences/speaker	4 sentences/speaker	GMM	**98%**	-
	MFC					98%	-
	JF_{bm}					99%	**1%**
	JF_{am}					99%	**1%**
	MGD	NTIMIT	10 sentences/speaker	4 sentences/speaker	GMM	**41%**	-
	MFC					40%	-
	JF_{bm}					47%	**6%**
	JF_{am}					45%	**4%**
Language identification	MGD	DBIL	45 sentences 40 Male & 5 Female	20 sentences 18 Male & 2 Female	GMM	**96%**	-
	MFC					95%	-
	JF_{bm}					98%	**2%**
	JF_{am}					97%	**1%**
	MGD	OGI_MLTS	45 sentences 40 Male & 5 Female	20 sentences 18 Male & 2 Female	GMM	**53%**	-
	MFC					50%	-
	JF_{bm}					58%	**5%**
	JF_{am}					57%	**4%**

96%, MFCC at 95%, JF_{bm} at 98%, and JF_{am} at 97% for this task. The best increase due to feature combination was 2%. For the 11 language task on the OGI_MLTS data the MODGDF recognition performance was at 53%, MFCC at 50%, JF_{bm} at 58%, and JF_{am} at 57% for this task. The best increase due to feature combination was 5%.

6. CONCLUSION

The significance of joint features derived by combining short time magnitude and phase spectra is discussed in this paper. Indeed the MODGDF and its significance in speech processing has been proved in earlier efforts. But the idea of combining the Fourier transform magnitude and phase spectra for representing speech via the group delay domain and MFCC is presented in this work. It is illustrated that joint cepstral features derived from the modified group delay function and MFCC essentially capture complete formant information in the speech signal. The joint features are used for four speech recognition tasks phoneme, syllable, speaker recognition, and language recognition. The results of performance evaluation indicate that joint features improve recognition performance up to 11% for feature combination before the acoustic model and upto 5% for feature combination after the acoustic model. This clearly indicates that combining evidences derived from different feature streams and different systems does enhance recognition performance of speech recognition systems. The use of appropriate feature combinations before the model using mutual discriminant information together with a logical combination of the classifier outputs can further improve recognition performance.

7. REFERENCES

[1] H. Drucker, " Speech processing in a high ambient noise environment,"*IEEE Transactions on Audio Electroacoustics*, Vol. AU-16, pp. 165-168, June, 1968.

[2] Helmholtz.H, " On the Sensations of Tone," *Dover*, 1954.

[3] Liu.L, He.J and Palm.G, " Effects of phase on the perception of intervocalic stop consonants," in *Speech Communication*,Vol.22, pp.403-417, 1997.

[4] B.Yegnanarayana and Hema A. Murthy,"Significance of group delay functions in spectrum estimation,"*IEEE Trans. on Signal Processing*, Vol.40, pp. 2281-2289,September 1992.

[5] K.K. Paliwal, "Usefulness of phase in speech processing," in *JNRSAS*, ISSN:1547-0407 Article 2, 01 January, 2004.

[6] Hema A. Murthy and Venkata Ramana Rao Gadde, "The Modified group delay function and its application to phoneme recognition,"*Proceedings of the ICASSP*, Vol.I, pp. 68-71, April 2003.

[7] Rajesh M.Hegde, Hema A.Murthy and Venkata Ramana Rao Gadde, " Application of the Modified Group Delay Function to Speaker Identification and Discrimination," in *Proceedings of the ICASSP*, Vol 1, pp. 517-520, May 2004.

[8] Rajesh M. Hegde, Hema A. Murthy, and VRR Gadde," Representation of speech using the modified group delay feature,"*IEEE Transactions on Speech and Audio Processing*, Communicated, July, 2004.

[9] D.Ellis,*Feature stream combination before and/or after the acoustic model*, ICSI Technical Report TR-00-007

[10] Hynek Hermansky and Nelson Morgan," RASTA processing of speech,"*IEEE Transactions on Speech and Audio Processing*, Vol. 2, No. 4, pp. 578-589, October, 1994.

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

[12] Muthusamy, Y.K., R.A. Cole, and B.T. Oshika, "The OGI multilanguage telephone speech corpus," in *Proceedings of the ICSLP*, pp. 895-898, Oct 1992.