Significance of Group Delay based Acoustic Features in the Linguistic Search Space for Robust Speech Recognition

Ramya R[1]*, Rajesh M Hegde*[2]*, Hema A Murthy*[1]

[1]Indian Institute of Technology Madras,
Chennai, India.
[2]Department of Electrical Engineering,
Indian Institute of Technology Kanpur,
Kanpur, India
ramya,hema@lantana.tenet.res.in, rhegde@iitk.ac.in

Abstract

In this paper we discuss the complementarity of the group delay features with respect to other conventional acoustic features and also propose the use of such diverse information in the linguistic search space for robust speech recognition. A discriminability analysis is carried out on various classes of phonetic units. A class based phonetic unit analysis is conducted to compare the suitability of using different acoustic feature streams for recognition of different phonetic unit classes. The results of recognition for isolated phonemic or syllabic units, give the appropriate feature for each unit. We then turn to describe the significance of this diversity of information present in the various acoustic features and their integration into the linguistic search space for syllable based continuous speech recognition. A weighted average likelihood method is used here, which appropriately weights the relevant acoustic feature for each syllable in question during the viterbi decoding process. This technique of integrating the complementarity of acoustic features into the linguistic search space gives reasonably reduced word error rates (WER) compared to conventional single or multi-stream acoustic features for experiments conducted on the TIMIT and the DBIL databases.

Index Terms: Group Delay features, Joint Features

1. Introduction

The notion of complementarity between acoustic features derived from the short time Fourier transform magnitude and phase spectrum has been analyzed in [1, 2]. However speech recognizers in general do not incorporate the acoustic complementarity in the linguistic search space. Towards this objective we first perform discriminability analysis in the two dimensional acoustic feature space using non-linear dimensionality reduction techniques like Sammon mapping and Isomap [3, 4] and draw support vector machine (SVM) boundaries using LIBSVM [5]. The variegation in boundary information obtained in the lower dimensional feature space is further substantiated by recognition experiments performed on various classes of phonetic units (consonants, phonemic, and syllabic units) in clean and noisy environments. We also utilize the modified group delay features MODGDF [6, 7], the product spectrum group delay features [1] and also a mel warped version of these features to compare performances. Joint feature stream is used in order to incorporate the properties and advantages of multiple features. It was also observed that joint features formed by concatenating spectral magnitude based MFCC and spectral phase based MODGDF [7] gave maximum improvement in performance compared to other combinations. Other features like product spectrum [1] and mel filter bank applied to modified group delay function(MGDMEL) are also explored for the task of consonant recognition.

We then turn to the integration of the diversity of information present in MFCC and MODGDF into the linguistic search space for syllable based continuous speech recognition. A weighted average likelihood method is used here which appropriately weighs the relevant acoustic feature for each syllable during the decoding process in the linguistic search space. The relevance of the acoustic feature with respect to the syllable in question is computed using what we call the syllable-feature pair derived using a priori recognition on a development data set. This technique gives reasonable reductions in the word error rate (WER), for experiments conducted on the TIMIT and the DBIL databases.

2. Group Delay Features

Acoustic features derived from the short time Fourier transform magnitude and phase spectrum have been used in speech recognition. The most popular among the magnitude-based features are the Mel Frequency cepstral coefficients (MFCC) and Perceptual linear prediction cepstral coefficients. MFCC is essentially computed from the short time magnitude spectrum [8] and PLP modifies the short time magnitude spectrum of the speech by several psychophysically based transformations. In our earlier work on features derived from the short time phase spectrum, we have discussed the MODGDF for speech recognition applications [7]. Its complementarity to the MFCC has been explored in [2]. Feature which combines the information from both the short time magnitude and phase spectrum is the product of power spectrum and group delay as in [1]. We briefly summarize the MODGDF and Product spectrum for clarity. The modified group delay function is defined as

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

where

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

The parameters α and γ can be determined over a database and also from a signal processing perspective [7]. The modified group delay function is converted to cepstral features by using the discrete cosine transform. Related information on the MODGDF can be found in [7]. In this paper we also apply a Mel filter bank to $\tau_m(\omega)$, to compute filter bank energies. We then compute the Modified group delay mel warped feature (MGDMEL) by discrete cosine transforming the logarithm of filter bank energies. The other well studied group delay based feature is the product spectrum which is the product of the power spectrum and the GDF [1]. The product spectrum is defined as follows

$$\begin{aligned} Q(\omega) &= |X(\omega)|^2 \tau(\omega) \\ &= X_R(\omega)Y_R(\omega) + Y_I(\omega)X_I(\omega) \end{aligned} \tag{3}$$

The Product spectrum mel warped feature (PSMEL) is obtained in a similar way as MGDMEL, by applying mel filter bank on the product spectrum and then computing DCT of logarithm of filter bank energies.

3. Discriminability Analysis of Acoustic Features

We use two different non-linear dimensionality reduction techniques for discriminability analysis in the two dimensional feature space namely Sammon mapping [3] and Isometric mapping (ISOMAP) [4]. These techniques reduces the high-dimensional data to low-dimensional data by preserving inter point distances and using local metric information to learn global geometry of the data respectively. Figures 1 and 2, illustrate the visualization of cluster structures computed using Sammon mapping and Isomap, for two sets of confusable phonemes {/k/,/g/} and {/m/,/n/} in white noise at a SNR of 5 dB using MFCC, MODGDF, MGDMEL and PSMEL. Also shown in Figures, is the SVM boundary obtained using LIBSVM [5] in the two dimensional feature space. We do not delve into the detail of using an SVM to derive class boundaries in this paper but feel, it is very relevant, given the cluster structures in the lower dimensional feature space. It can be inferred from these cluster plots that discriminability of a particular feature is different for different units. Figure 1, shows that units {/k/,/g/} are separated reasonably better using the MODGDF and MGDMEL, when compared to the MFCC and PSMEL. On the other hand, for units {/m/,/n/}, MFCC and PSMEL exhibits better separability than MODGDF and MGDMEL as shown in Figure 2. MGDMEL in general behaves similar to the MODGDF but does well in case of {/m/,/n/} and worse in case of {/k/,/g/}. As seen in Figures 1 and 2 in case of {/k/,/g/}, PSMEL is not as good as MODGDF, where as in case of {/m/,/n/} it not better than MFCC.

4. Class Based Performance Analysis of Single and Multi-stream Acoustic Features

To ensure the consistency of the inferences of discriminability analysis from Section 3, a class based analysis was carried out for each acoustic feature in order to infer the suitability of a particular feature for recognizing a particular phonetic unit class. The experiments are conducted on the consonant database available as part of the consonant challenge at Interspeech 2008 [9]. The unit classes considered were semi vowels, nasals, fricatives and plosives. Table 1, summarizes the results of class based recognition averaged over SNRs from 25db to -5db. Note that all features were appended with their respective velocity and acceleration co-efficients. It is observed that the results obtained

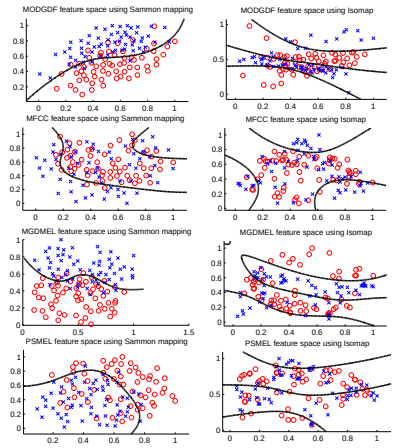

Figure 1: Cluster structures of phonemes /k/ and /g/ for MODGDF, MFCC, MGDMEL and PSMEL obtained using Sammon mapping and Isomap along with SVM boundaries

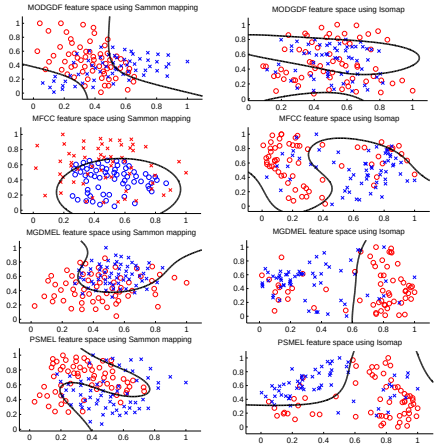

Figure 2: Cluster structures of phonemes /m/ and /n/ for MODGDF, MFCC, MGDMEL and PSMEL obtained using Sammon mapping and Isomap along with SVM boundaries

using group delay features are complementary to that of conventional features, albeit on an individual basis one can be inferior to the other as seen in Table 1. To carry forward this conjecture

Table 1: Class based performance analysis for single-stream features

Feature/Class	Semi-vowels	Nasals	Fricatives	Plosives
MFCC	96.88%	93.75%	83.52%	97.02%
PLP	96.88%	81.25%	88.07%	94.79%
MODGDF	96.88%	89.58%	81.25%	91.71%
MGDMEL	93.75%	93.75%	80.11%	91.67%
PSMEL	98.44%	93.75%	85.23%	94.79%

we derive joint feature streams, such that it has one magnitude based feature and once phase based feature. It is observed in

Table 2 that joint feature stream helps in improving the recognition performance of various classes. The MFCC+MODGDF combination gave best result for fricatives and plosives and PLP + PSMEL gave best result for semi vowels. It is also seen that the combination MFCC+MODGDF gave maximum improvement of 4.2% performance than the baseline performance using MFCC alone [9].

Table 2: Class based performance analysis for multi-stream features

Feature/Class	Semi-vowels	Nasals	Fricatives	Plosives
MFCC + MODGDF	98.44%	93.75%	**91.67**%	**97.92**%
MFCC + MGDMEL	95.31%	91.67%	85.23%	95.83%
MFCC + PSMEL	98.44%	93.75%	89.20%	95.8%
PLP + MODGDF	96.88%	93.75%	89.2%	95.83%
PLP + MGDMEL	98.44%	91.67%	88.07%	97.77%
PLP + PSMEL	**100**%	93.75%	88.07%	94.79%

5. Recognition of Consonants in Mismatched Noisy Conditions

In order to further study the utilization of complementarity of the various acoustic features for robust speech recognition applications, we use a statistical re-estimation (STAR) algorithm as in [10], which is a successful model compensation method for recognition in mismatched conditions. The experiments are conducted on the consonant database available as part of the consonant challenge at Interspeech 2008 [9]. The baseline system uses context dependent phoneme models. Separate HMMs for initial and final vowels have been used. Although we performed extensive experiments using different noise sources we illustrate the performance of the acoustic features namely MFCC, MODGDF, MGDMEL, and PSMEL and their joint counterparts in babble noise in Figures 3 and 4.

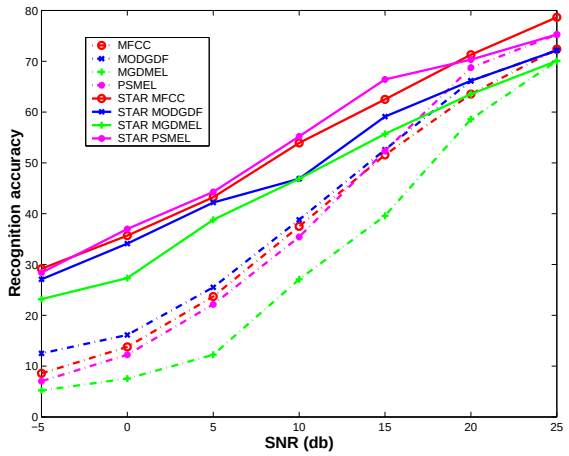

Figure 3: Recognition results using MFCC, MODGDF, MGDMEL and PSMEL with and with out STAR adaptation with test set corrupted in babble noise

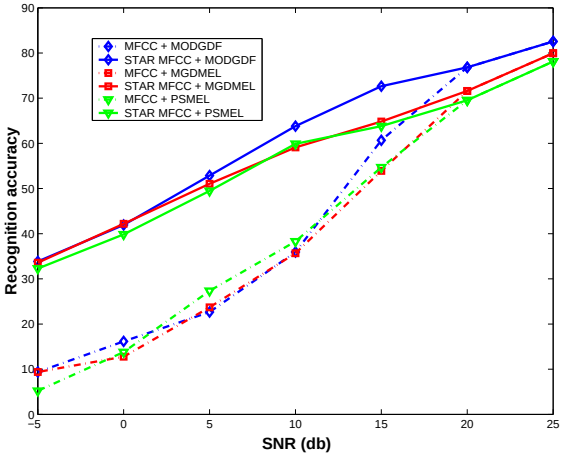

Figure 4: Recognition results using various joint features with and with out STAR adaptation with test set corrupted in babble noise

Joint features using MFCC and MODGDF gave reasonable improvement in performance when compared to the other joint features, even though the performance using MODGDF was individually less than other features. This may possibly be due to the complementary information between MODGDF and MFCC as seen in Section 3. Model adaptation along with relevant joint features does improve the recognition accuracy which can be very important at very low SNRs. In the following section we incorporated the acoustic feature diversity in MFCC and MODGDF in a more intuitive way into a syllable based continuous speech recognizer.

6. Integrating Group Delay Features into the Linguistic Search Space

The complementary information of the group delay feature streams is incorporated into the linguistic search space [1], using an automatically annotated recognizer (AAR) [11]. In the AAR system speech is segmented into syllable like units using group delay based two level segmentation algorithm. The corresponding text is segmented using rule based text segmentation. By mapping the segmented speech and text, syllable level annotations are automatically obtained for the training data. Annotated data is used to extract examples for all the unique syllables. HMMS are trained for all the unique syllables in a database. During the testing phase, the utterance is segmented into syllable like units using group delay based segmentation algorithm and each segment is tested against all the models. Word outputs from the AAR system are generated using lexicon and language models (LM). A simple directed search algorithm is implemented in an AAR system, that checks only against the set of possible syllable models at each time depending on the words that are active. The syllable HMMS form the basic nodes of the network and words are loaded as the syllable HMMS that constitute the word. The syllables at the word boundary are connected to each other using bi gram word probabilities. The algorithm also takes care of insertions and deletions using the multi path modification in the search space. As the current segment of the speech signal is recognized only against the list of units that are active, appropriate features can be used for each unit.

[1] Parts of this Section are under review for EUSIPCO 2008

A priori analysis is performed on the training data to find which feature performs better for each syllable. Syllable-feature pair information is generated using the results of this analysis for all the syllabic units in the training data. The likelihood scores given by MODGDF and MFCC are not on a similar scale making the incorporation of the syllable-feature pairs into the linguistic search space difficult. Pruning the syllable models and updating the overall likelihood may not be consistent for a reliable integration. We therefore incorporate the syllable-feature pair information into the linguistic search space using a method of weighted average likelihoods. In weighted average likelihood method, the likelihood score from the appropriate acoustic feature model is given more weight than the likelihood score from the other feature model for each syllable that is active in the network. This ensures that even if one acoustic feature fails to give a correct recognition, the most likely unit is given more priority due to the appropriate weighting. Syllable-feature pair information incorporated in this fashion ensures that the correct syllable is not pruned. The weighted average likelihood method is also compared to the average likelihood method, where the likelihood scores from different acoustic feature models are just averaged without incorporating the information in the syllable-feature pair.

6.1. Performance Evaluation

In order to study the effects of acoustic modeling alone, language models (LM) and lexicon for the TIMIT [12] and the DBIL database [13] are generated using all the words and word combinations in both the train and test data sets. It must be noted here that the group delay based automatic segmentation algorithm is used for experiments on the DBIL database, which segments speech data at boundaries of syllabic units. For experiments on the TIMIT data, we have utilized the explicit boundary information given in the database. It is important to note that the TIMIT dictionary is first syllabified using NIST syllabification software [14] available from NIST for syllable based experiments. Analysis on the DBIL Tamil database indicate that 425 syllables are recognized better with the MFCC, while 375 syllables are recognized better with the MODGDF. Similarly in experiments on the TIMIT database, 484 syllables are recognized better with the MFCC, while 300 syllables are recognized better with the MODGDF. The performance using both the features for the remaining syllables was very similar. The WER for the TIMIT and the DBIL databases using average likelihood and weighted likelihood methods is shown in Table 3. The relative improvement in performance using the syllable-feature pair information in weighted likelihood method seems promising for future experiments.

Table 3: Recognition results on the TIMIT and DBIL database for joint features with average likelihood and weighted likelihood with syllable-feature (S-F) pair information

Features	Database	
	TIMIT	DBIL
	WER	WER
MODGDF+D+A+E	10.1%	35.7%
MFCC+D+A+E	6.5%	30.4%
Average likelihood with no S-F pair information	4.2%	28.0%
Weighted likelihood with S-F pair information	3.2%	27.2%

7. Conclusion

This work analyzes the notion of complementarity of group delay features with respect to that of conventional features in terms of cluster structures and recognition of confusable phonetic units both in clean and noisy conditions. It is also noted that the variegation of information in the different features is more prominent in low SNRs based on the unit class based recognition experiments conducted in mismatched conditions. We also use a hitherto untried technique of integrating this complementary information into the linguistic search space during the decoding process for improved speech recognition. The use of a feature constrained linguistic search space, where the syllable in question is recognized based on the appropriate weighting of the relevant feature can reduce recognition complexity when compared to decoding using joint feature streams or using multiple classifiers. The appropriate normalization of the likelihood scores computed from each feature stream is an issue that needs to be addressed in the future given the large variation in their dynamic range.

8. References

[1] Donglai Zhu and Kuldip K. Paliwal, "Product of power spectrum and group delay function for speech recognition," *ICASSP*, vol. I, pp. 125–128, May 2004.

[2] Rajesh M. Hegde, Hema A. Murthy, and Venkata Ramana Rao Gadde, "Significance of joint features derived from the modified group delay function in speech processing," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2007, Jan. 2007.

[3] J.W. Sammon Jr., "A nonlinear mapping for data structure analysis," *IEEE Transactions on Computers*, pp. 401–409, 1969.

[4] Joshua B. Tenenbaum, Vin de Silva, and John C. Langford, "A global geometric framework for nonlinear dimensionality reduction," `www.science.org`, 2000.

[5] Chih-Chung Chang and Chih-Jen Lin, *LIBSVM: a library for support vector machines*, 2001, Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[6] Hema A. Murthy and Venkata Gadde, "The modified group delay function and its application to phoneme recognition," *in Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, pp. 68–71, 2003.

[7] Rajesh M. Hegde, Hema A. Murthy, and Venkata Ramana Rao Gadde, "Significance of the modified group delay feature in speech recognition," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, pp. 190–202, Jan. 2007.

[8] P. Mermelstein and S. B. Davis, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," vol. 28, pp. 357–366, August 1980.

[9] Martin Cooke and Odette Scharenborg, "The Interspeech 2008 Consonant Challenge," *Interspeech*, 2008.

[10] P.J. Moreno and B. Raj and R.M. Stern, "Data-driven environmental compensation for speech recognition: A unified approach," *Speech Communications*, vol. 24, pp. 267–285, 1998.

[11] Lakshmi A. and Hema A. Murthy, "A syllable based continuous speech recognizer for tamil," *Interspeech*, pp. 1878–1881, 2006.

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

[13] The Database of Indian Languages, *Speech and Vision Lab*, IIT Madras, 2001.

[14] W. M. Fisher, "tsylb2-1.1 syllabification software," `http://www.nist.gov/speech/tools/index.htm`, August 1996.