INCORPORATING ACOUSTIC FEATURE DIVERSITY INTO THE LINGUISTIC SEARCH SPACE FOR SYLLABLE BASED SPEECH RECOGNITION

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

[1]Indian Institute of Technology Madras, Chennai, India
Email: ee05s047@iitm.ac.in,hema@cse.iitm.ac.in

[2]Department of Electrical Engineering
Indian Institute of Technology Kanpur, Kanpur, India
Email: rajesh_m_hegde@yahoo.com

ABSTRACT

Acoustic features derived from the short time magnitude and phase spectrum provide complementary information. In this paper, we discuss the significance of incorporating this diverse information into the linguistic search space for syllable based speech recognition. The diversity of group delay acoustic features computed from the phase spectrum, and MFCC computed from the magnitude spectrum, is first illustrated in a lower dimensional feature space. Motivated by this diversity of information in the acoustic feature space, we derive syllable-feature pairs. The selection of syllable-feature pairs is based on isolated syllable recognition results, computed apriori using the two acoustic feature streams. During the recognition process, based on the syllable-feature pair information likelihoods are appropriately weighted using a weighted likelihood scheme. The syllable lattice is now rescored using these weighted syllable-feature pairs in the linguistic search space. This technique of appropriately weighting the relevant acoustic feature for each syllable during the decoding process in the linguistic search space, yields reduced word error rate (WER), for experiments conducted on the TIMIT and the DBIL databases.

1. INTRODUCTION

Fusion of heterogeneous acoustic features at the feature level and the likelihood level is an interesting approach to improving speech recognition performance [1]. Other methods of feature fusion based on discriminative model combination [2], and lattice combination [3], use different features to generate different hypotheses and ultimately combine them to arrive at a single best hypothesis. These methods are proven to increase the recognition accuracy. Fusion of features computed from the short time phase spectrum (MODGDF) and those computed from the short time spectral magnitude (MFCC), and their benefits for various speech tasks such as syllable recognition, speaker identification and language recognition was explored in [4]. In this work, we utilize the diversity in the aforementioned features to come up with a syllable-feature pair for each syllable in the vocabulary. It is illustrated that the ability of a particular acoustic feature stream to discriminate between syllables depends on the syllable pair in question. Syllable discriminability analysis of the two acoustic features, MODGDF and MFCC, is performed using non linear dimensionality reduction techniques like Sammon mapping, Isometric mapping (Isomap) and also linear discriminant analysis (LDA). We also illustrate the results of separability analysis in the high dimensional feature space using Bhattacharya distance. These results are verified using recognition experiments on the confusable syllable sets. Further, results of isolated recognition experiments on the train data are also used to generate syllable-feature pairs. Syllable-feature pair information is incorporated in the linguistic search space during the decoding phase by appropriately weighting the likelihoods. Word error rates (WER) obtained using the conventional decoding process are compared with the WER after incorporating the multi stream syllable-feature pairs into the linguistic search space. This technique gives reasonable reductions in WER, for both weighted and average likelihood combination techniques.

2. ACOUSTIC FEATURES

A wide variety of acoustic features have been used in speech recognition. The most popular among them is the Mel Frequency cepstral coefficients (MFCC) which is essentially computed from the short time magnitude spectrum [5]. In our earlier work on features derived from the short time phase spectrum, we have proposed a new feature called the MODGDF for speech recognition applications [6, 7]. Its complementarity to the MFCC has been explored in [4]. We briefly summarize the MODGDF for clarity. Group delay is defined as the negative derivative of Fourier transform phase

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

where $\theta(\omega)$ is the continuous phase spectrum of the signal. Group delay function can also be computed from the signal using

$$\begin{aligned}\tau(\omega) &= -Im\left[\frac{d(\log(X(\omega)))}{d\omega}\right] \\ &= \frac{X_R(\omega)Y_R(\omega)+Y_I(\omega)X_I(\omega)}{|X(\omega)|^2}\end{aligned} \tag{2}$$

where $X(\omega)$ and $Y(\omega)$ are the Fourier transforms of the signals $x(n)$ and $nx(n)$ respectively. The subscripts R and I denote the real and the imaginary parts of the Fourier transform. It is already been established that the group delay function is well behaved only for minimum phase signals [6]. Modified group delay function is defined as

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

where

$$\tau(\omega) = \frac{X_R(\omega)Y_R(\omega)+Y_I(\omega)X_I(\omega)}{S(\omega)^{2\gamma}} \tag{4}$$

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 [6, 7].

3. ANALYSIS OF ACOUSTIC FEATURE DIVERSITY

In this Section, the diversity of the two acoustic features namely, MFCC and MODGDF is analyzed using three different methods, Sammon mapping[8], Isomap [9] and linear discriminant analysis (LDA) [10] in the lower dimensional feature space and using Bhattacharya distance in the high dimensional feature space.

3.1 Cluster analysis

We perform cluster analysis in the two dimensional feature space to explore the relation between discriminability and the acoustic feature stream. Since it would be reasonable to assume that features extracted from the speech signal lie on a non linear manifold in the

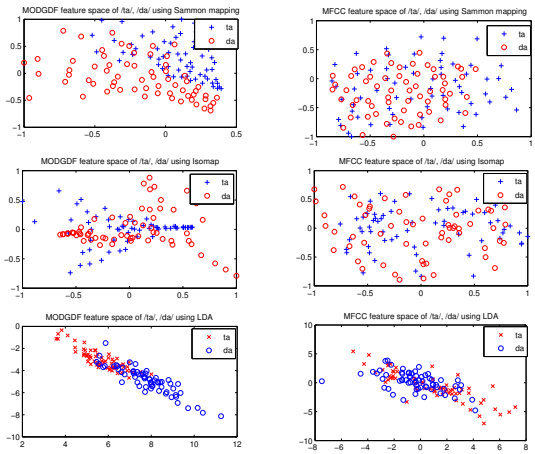

Figure 1: Cluster structures of syllables /ta/ and /da/ for MODGDF and MFCC using Sammon mapping, Isomap and LDA

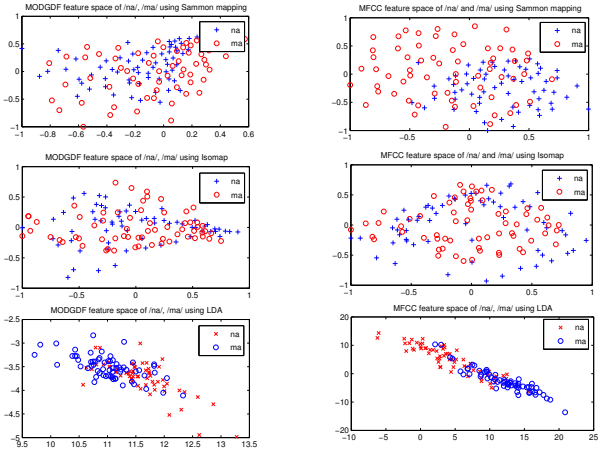

Figure 2: Cluster structures of syllables /ma/ and /na/ for MODGDF and MFCC using Sammon mapping, Isomap and LDA

high dimensional feature space, we use three different non linear dimensionality reduction techniques for cluster analysis in the two dimensional feature space.

(i) The Sammon mapping technique [8], is used to map the high dimensional feature onto the low-dimensional feature space such that the inherent structure of the feature space is preserved for visualization. This technique works on the principle of preserving interpoint distances between the points in high dimensional space to that of the low-dimensional space by minimizing the error function given by

$$E_{\text{sam}} = \frac{1}{\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} D_{ij}} \sum_{i=1}^{N-1}\sum_{j=i+1}^{N} \frac{(d_{ij}-D_{ij})^2}{D_{ij}} \qquad (5)$$

where d_{ij} and D_{ij} are the distances between two points i and j in d-dimensional output space and D-dimensional input space respectively; N is the number of points in input or output space.

(ii) Isometric mapping (Isomap) [9], is a manifold learning technique, that reduces the high-dimensional data to low-dimensional data using local metric information to learn global geometry of the data to verify the cluster structures.

(iii) Linear discriminant analysis is a method used for data classification, it maximizes the ratio of between-class variance to the within-class variance thereby guaranteeing maximal separability between classes. LDA also helps to better understand the distribution of the feature spaces [10].

Figures 1 and 2, illustrates the visualization of cluster structures computed using Sammon mapping, Isomap and LDA, for two sets of confusable syllables, namely {/ta/,/da/} and {/ma/,/na/}. It can be inferred from these cluster plots that discriminability of a particular feature (MODGDF and MFCC) is different for different units. Figure 1, shows that units {/ta/,/da/} are separated reasonably better using the MODGDF, when compared to the MFCC. On the other hand, for units {/ma/,/na/}, MFCC exhibits better separability than MODGDF as shown in Figure 2.

3.2 Separability analysis

The Bhattacharya distance [11] is a mathematical measure of separability between classes. Since it is primarily defined for a two class problem it suits well for the pairwise syllable separability analysis. For the two sets of confusable syllables namely {/ta/,/da/} and {/ma/,/na/}, analyzed earlier, we compute the Bhattacharya distance and the cumulative Bhattacharya distance for both MODGDF and MFCC at all feature dimensions. From the illustrations shown in Figures 3 and 4, it can be inferred that the measure of separability is higher for syllable pair {/ta/,/da/}, while it is lower for syllable pair {/ma/,/na/} using MODGDF when compared to MFCC carrying forward our conjecture from Figures 1 & 2.

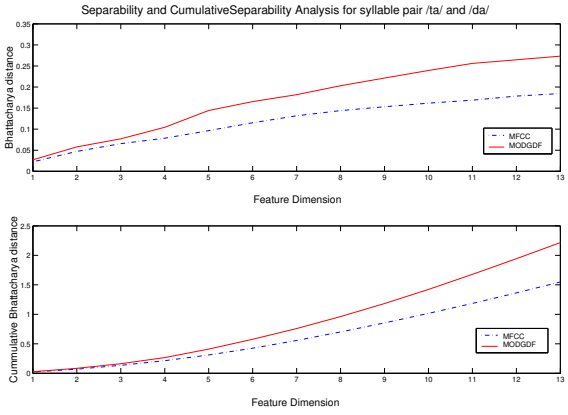

Figure 3: Separability of syllables /ta/ and /da/ with MODGDF and MFCC using Bhattacharya distance

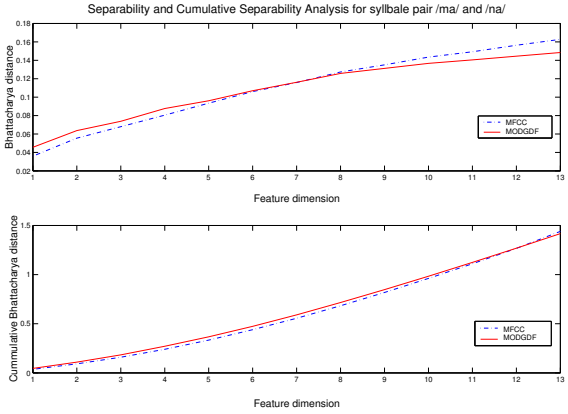

Figure 4: Separability of syllables /ma/ and /na/ with MODGDF and MFCC using Bhattacharya distance

3.3 Recognition of confusable syllable sets

To ensure the consistency of the inferences of separability analysis from Sections 3.1 and 3.2, isolated style recognition experiments

are carried out on each pair of confusable syllables for each separate feature stream. The results of recognition performed on the selected confusable sets are tabulated in Table 1. For syllable pair {/ta/,/da/}, the performance of MODGDF is reasonably better, while for the syllable pair {/ma/,/na/}, the performance is poor when compared to MFCC. These results are consistent with the results from cluster and separability analysis performed earlier.

Table 1: Syllable recognition performance for confusable syllables

Syllables	MFCC	MODGDF
/ta/ and /da/	78.9%	85.3%
/la/ and /La/	85.3%	92.6%
/tu/ and /du/	67.7%	71.7%
/a/ and /A/	79.7%	81.8%
/sa/ and /sA/	52.5%	57.6%
/tam/ and /kam/	70.7%	70.7%
/pat/ and /kat/	97.1%	91.5%
/ma/ and /na/	98.9%	97.1%

3.4 Rescoring the syllable lattice using feature switching

In an effort to utilize the diversity of the acoustic feature streams in the linguistic search space we propose a method of rescoring the syllable lattice based on feature switching. For the sake of clarity, we illustrate 5-Best syllable lattice structures for the phrase "nALAI maRunaL" in Figures 5-8. In Figures 5 and 6, it can be noted at the first syllable segment position, both MFCC and MODGDF are able to find the correct syllable in the lattice structure. At the second and fourth syllable segment positions, MODGDF is able to place the correct syllable /LAI/ and /Ru/ in the 5-best list, where as MFCC does not. A technique that can switch to MODGDF at this position in the lattice structure will output the correct syllable transcription. Similarly at the fifth segment position MFCC is able to place the syllable /nA/ where as MODGDF does not. In Figures 7 and 8 it is illustrated that the syllable transcription can be corrected by switching the feature either to MODGDF at the second and fourth positions or to MFCC at the fifth position.

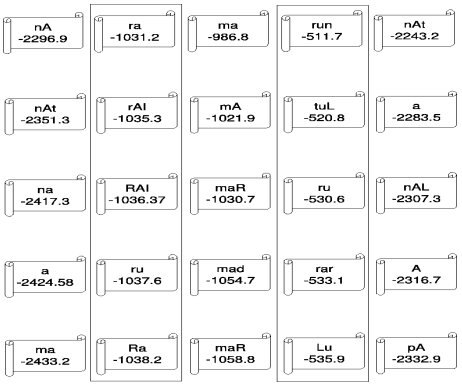

Figure 5: 5-Best syllable lattice for the phrase "nALAI maRunaL" generated using MFCC

4. INCORPORATING MODGDF AND MFCC IN THE LINGUISTIC SEARCH SPACE

In this Section, we discuss the methodology to incorporate the acoustic feature diversity in the MODGDF and the MFCC into the linguistic search space using an Automatically annotated recognizer

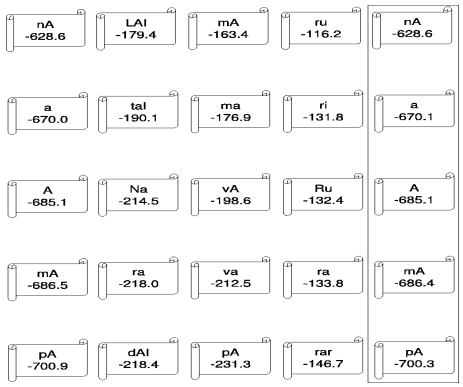

Figure 6: 5-Best syllable lattice for the phrase "nALAI maRunaL" generated using MODGDF

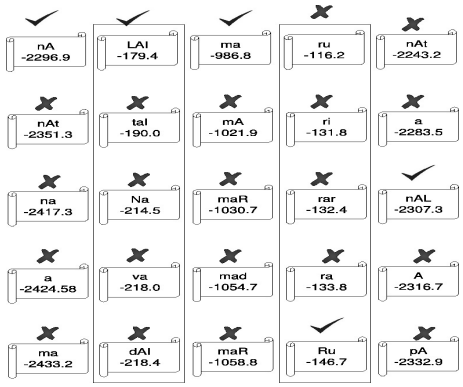

Figure 7: Re-Alignment of sentence "nALAI maRunaL" using MFCC with MODGDF

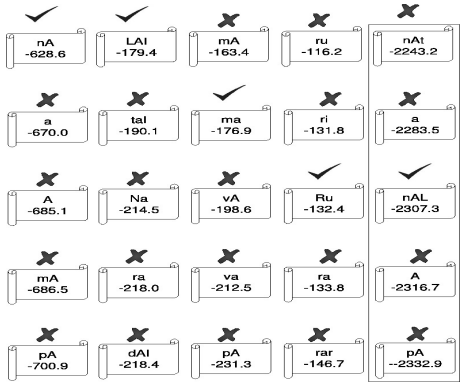

Figure 8: Re-Alignment of sentence "nALAI maRunaL" using MODGDF with MFCC

(AAR) [12]. 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. The block diagram of the training phase in an AAR system is illustrated in Figure 9.

During the testing phase, the utterance is segmented into syllable like units using group delay based segmentation algorithm. Word outputs from the AAR system are generated using syllable

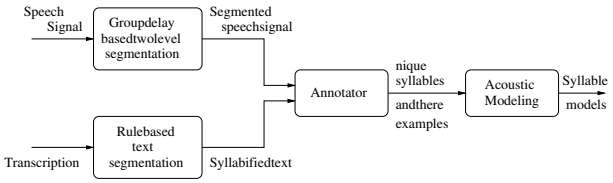

Figure 9: Block diagram of the training phase in an AAR system

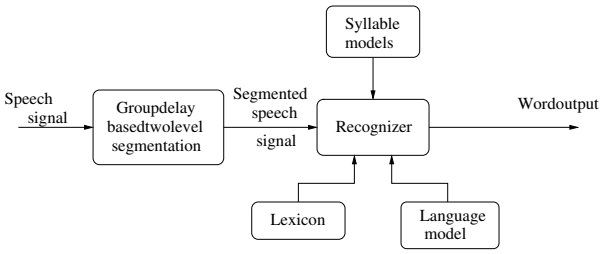

Figure 10: Block diagram of the testing phase in an AAR system

models, segmented test signal, lexicon and language models(LM) [13]. Block diagram of AAR system test phase is shown in Figure 10. 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 bigram word probabilities. The algorithm also takes care of insertions and deletions using the multipath 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. Apriori 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.

4.1 Comparison with HDA+MLLT

Temporal information in the speech signal is utilized by adding the first and second time derivatives and also by concatenating the current frame of the acoustic feature vector with preceding and succeeding frame of acoustic feature vectors. This very high dimensional feature vector is then projected to a low dimensional space using linear discriminant analysis (LDA)[10]. LDA while projecting onto the lower dimension feature space tries to preserve the discriminative capability of the acoustic features. But LDA is based on the assumption that covariance of all the classes is same. Heteroscedastic discriminant analysis (HDA) relaxes this assumption, but it yields a feature space which is highly correlated. In order to overcome this, maximum likelihood linear transformation (MLLT) is used which diagonalizes the HDA feature space. The combination of HDA and MLLT is shown to perform better in [14, 15]. However this work is not related to finding optimal feature spaces for computing the covariance matrices without loss of acoustic information. It deals with incorporating multiple acoustic feature streams into the linguistic search space during the decoding process, using the syllable-feature pair information generated apriori.

5. PERFORMANCE EVALUATION

In this Section, we discuss the performance improvements in speech recognition in terms of word error rate (WER), by incorporating the acoustic diversity of the MODGDF and MFCC into the linguistic search space. Experiments are performed using AAR system on the TIMIT database [16] and the DBIL database [17]. The phonemes in the TIMIT databases are syllabified prior to the performance evaluation, while the DBIL database inherently consists of syllable level transcriptions.

5.1 Syllable based recognition using the TIMIT Database

The TIMIT database consists of phoneme level transcriptions. The TIMIT dictionary is first syllabified using NIST syllabification software [18] available from NIST. The NIST syllabification software [18], is based on rules that define possible syllable-initial and syllable-final consonant clusters, as well as prohibited syllable-initial consonant clusters. TIMIT database contains 2 SA sentences per speaker which are same across all the 630 speakers. SA sentences are removed from both train and test databases as they introduce unfair bias. A total of 3570 unique syllables are present in the training data and 986 unique syllables in the core test set. Test syllables which are not in the training data are replaced with corresponding phonemes. For each syllable model, number of states is decided based on the number of phonemes it contains and number of mixtures in a Gaussian is decided based on the number of syllable examples in the training data. Standard result for TIMIT WER is reported as 6 % in [19], where the syllable unit system was intialized from the corresponding context dependent triphone system [19].

5.2 Syllable based recognition using the DBIL database

The DBIL Tamil (Indian Language) database, has 19 female bulletins for training and 4 female bulletins for testing, each of 20 minutes duration. The training data has 2550 unique syllables. For frequently occurring syllables, features are extracted using single frame size where only one frame size and frame shift is used. For infrequently occurring syllables, features are extracted using multiple frame size and multiple frame rate technique which generates multiple instances from each utterance by varying frame size and frame shift. Five state HMMs with three mixtures per state are built for all syllables. Syllable recognition accuracy (SRA) of 47.1% has been reported earlier on the DBIL Tamil database in [12].

5.3 Experimental results

The SRA and WER using MODGDF, MFCC and joint features fused at the feature level for the TIMIT database is listed in Table 2. Table 3 lists similar results for the DBIL database. In order to

Table 2: Recognition results on the TIMIT database for individual and joint features with D(delta), A(acceleration), E(energy) parameters

Feature	% SRA	%WER
MFCC+D+A+E	78.04%	6.5%
MODGDF+D+A+E	76.87%	10.1%
MFCC+MODGDF+D+A+E	80.42%	4.3%

Table 3: Recognition results on the DBIL database for individual and joint features with D(delta), A(acceleration), E(energy) parameters

Feature	% SRA	%WER
MFCC+D+A+E	48.2%	30.4%
MODGDF+D+A+E	47.0%	35.7%
MFCC+MODGDF+D+A+E	53.0%	23.5%

study the effects of acoustic modeling alone, language models (LM) and lexicon for the TIMIT and the DBIL database are generated using all the words and word combinations in both the train and test data sets. Joint features derived by fusing MFCC and MODGDF at the feature level show reasonable improvements in SRA both on the TIMIT and the DBIL databases. It must be noted here that the group delay based automatic segmentation algorithm is used for experiments on the DBIL database. For experiments on the TIMIT data, we have utilized the explicit boundary information given in the database. Further 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 4. The relative improvement in performance using the syllable-feature pair information in weighted likelihood method seems promising for future experiments.

Table 4: 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%

6. CONCLUSION

In this work we have proposed techniques by which the diversity of information present in acoustic features derived from the short time phase spectrum (MODGDF) and magnitude spectrum (MFCC) can be integrated into the linguistic search space for syllable based speech recognition. The proposed technique contrasts conventional feature fusion methods and HDA+MLLT techniques conceptually and in terms of computational complexity at the decoding phase. The integration is based on apriori generation of syllable-feature pairs generated from cluster structure, separability analysis and recognition experiments of confusable syllable sets. The synergistic use of a feature constrained linguistic search space and feature switching, where the syllable in question is recognized based on the appropriate weighting of the relevant feature, can reduce recognition complexity when compared to conventional early and late fusion methods used in speech recognition. The appropriate normalization of the likelihood scores computed from each feature stream is an issue that needs to be addressed given the large variation in their dynamic range. Other possible refinements in the language model directed AAR system for continuous speech recognition need to be investigated in the future.

REFERENCES

[1] Andrew K. Halberstadt and James R. Glass, "Heterogeneous acoustic measurements and multiple classifiers for speech recognition," *ICSLP*, 1998.

[2] P Beyerlein, "Discriminative model combination," *ICASSP*, vol. 1, pp. 481–484, 1998.

[3] Xiang Li, Rita Singh, and Richard Stern, "Combining search spaces of heterogeneous recognizers for improved speech recognition," *ICSLP*, September 2002.

[4] 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.

[5] 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.

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

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

[10] S. Balakrishnama and A. Ganapathiraju, "Linear Discriminant Analysis - A brief Tutorial," `http://www.isip.msstate.edu/publications/reports/`, March 1998.

[11] Fukunaga.K, *Introduction to Statistical Pattern Recognition*, Accademic Press, Boston, 1990.

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

[13] Lakshmi A. and Hema A. Murthy, "A new approach to continuous speech recognition in Indian languages," *NCC*, 2008.

[14] G. Saon, M. Padmanabhan, R. Gopinath, and S. Chen, "Maximum likelihood discriminant feature spaces," *ICASSP*, vol. 2, pp. 1129–1132, 2000.

[15] S. Deligne, S. Dharanipragada, R. Gopinath, B. Maison, P. Olsen, and H. Printz, "A robust high accuracy speech recognition system for mobile applications," *IEEE Trans. Speech Audio Processing*, vol. 10, pp. 551561, November 2002.

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

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

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

[19] A. Hämäläinen, J. de Veth, and L. Boves, "Longer-Length acoustic units for continuous speech recognition," *Proceedings of EUSIPCO*, 2005.