ENTROPY BASED MEASURES FOR INCORPORATING FEATURE STREAM DIVERSITY IN THE LINGUISTIC SEARCH SPACE FOR SYLLABLE BASED AUTOMATIC ANNOTATED RECOGNIZER

Chaitanya Kumar J, Hema A. Murthy

Department of Computer Science and Engineering
Indian Institute of Technology Madras
jchaitu@cse.iitm.ac.in,hema@cse.iitm.ac.in

ABSTRACT

In this paper, we propose a method to integrate acoustic features, Mel Frequency cepstral coefficients(MFCC) and Modified Group Delay Features(MODGDF), which are derived from short time magnitude and phase spectrum respectively and provide complementary information for speech recognition. Our method uses two information theoretic measures, namely, the entropy of the HMM model for merged feature stream, the entropy of the model for individual feature streams for a given syllable. During the training phase, the weights for the models of two feature streams, MFCC and MODGDF are computed. During recognition these weights are retrieved from the syllable-feature stream weight information and the likelihoods are appropriately weighed. The syllable lattice is now re-scored using these weighted likelihood scores in the linguistic search space. This technique yields reduced word error rate(WER) for experiments conducted on DBIL database [1].

1. INTRODUCTION

Integrating heterogeneous acoustic features at the feature level and the likelihood level is an interesting approach to improve speech recognition performance [2]. Fusion of MFCC and MODGDF features at feature level, and their benefits for various speech tasks such as syllable recognition, speaker identification and language recognition were explored in [3]. Incorporating syllable-feature pair information in the linguistic search space during recognition by empirically weighing the likelihoods was explored in [4]. This work is an extension of the work done in [5]. Instead of weighing the feature streams empirically, in this work, we utilize the diversity in the acoustic features MFCC and MODGDF, to assign weights to likelihood values of feature streams by emphasizing on reliability of the feature stream which gives optimum likelihood values. Ideally we would like to perform feature switching in the linguistic search space.

Our proposed method calculates the weights of models of the feature stream during the training phase. The entropy of state sequence of a HMM model λ for a feature stream X of a given syllable is, given by $H(S^T|X,\lambda)$. $H(S^T|X,\lambda)$ is a measure of the usefulness of a feature stream for a given model [6]. The smaller the entropy, greater the usefulness of a particular feature stream. In this paper, the focus is to extract the effectiveness of MFCC and MODGDF for recognition. Instead of simply merging the feature streams, an attempt is made to appropriately weight the feature streams. To obtain the weights, the delta entropy between $H(S^T|Z,\lambda_{a+b})$ and $H(S^T|X,\lambda_a)$ is compared with delta entropy between $H(S^T|Z,\lambda_{a+b})$ and $H(S^T|Y,\lambda_b)$. $\lambda_{(a+b)}$ is the model obtained with the merged features. λ_a, λ_b are the models obtained using the individual feature streams. X,Y are the features that are extracted corresponding to λ_a, λ_b. Z is the joint feature stream. S^T is the set of all possible state sequences. The weight values for each feature stream of a syllable are incorporated in the linguistic search space during the decoding phase by appropriately weighing the likelihoods[5]. Word error rates(WER) obtained using conventional decoding process are compared with the WER after incorporating the heterogeneous feature stream and weights associated with them into the linguistic search space. The performance improves upon the results obtained in [5].

This paper is organized as follows. In section 2, we briefly review modified group delay features. In section 3, we present feature stream weighs estimation using measures based on entropy. The weight obtained in section 3 are then incorporated in the linguistic search space during recognition in an Automatically Annotated Recognizer [7]. The performance of the proposed technique is tabulated in section 5.

2. ACOUSTIC FEATURES

A wide variety of acoustic features have been used in speech recognition. The most popular among them is the MFCC which is essentially computed from the short time magnitude spectrum [8]. 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 [4]

[9]. It's complementarity to the MFCC has been explored in [3]. 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 computed from the signal using

$$\begin{aligned} \tau(\omega) &= -Im\left[\frac{d(log(X(\omega)))}{d\omega}\right] &(2)\\ &= \frac{X_R(\omega)Y_R(\omega) + Y_I(\omega)X_I(\omega)}{|X(\omega)|^2} &(3)\end{aligned}$$

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 imaginary parts of the Fourier transform. It has already been established that the group delay function is well behaved only for minimum phase signal [9]. The modified group delay function is defined as

$$\tau_m(\omega) = \left(\frac{\tau(\omega)}{|\tau(\omega)|}\right)(|\tau(\omega)|)^{\alpha} \tag{4}$$

where

$$\tau(\omega) = \frac{X_R(\omega)Y_R(\omega) + Y_I(\omega)X_I(\omega)}{S(\omega)^{2\gamma}} \tag{5}$$

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 [4] [9].

3. FEATURE STREAM WEIGHT ESTIMATION

To determine weights for MFCC and MODGDF feature streams, consider HMM models λ_a and λ_b for MFCC and MODGDF respectively. Their corresponding observation sequences are O_a and O_b respectively. Let us also consider λ_{a+b} with the joint observation sequence O_{a+b}. The entropy of an individual model (λ_a) producing state sequence, for observations O_a, is given by

$$H(\mathbf{S^T}|\mathbf{O_a}, \lambda_a) = \sum_{\forall O_k \in \mathbf{O_a}} H(\mathbf{S^T}|O_k, \lambda_a) \tag{6}$$

where

$$H(\mathbf{S^T}|O_k, \lambda_a) = -\sum_{\forall s^T}\left[p(s^T|O_k, \lambda_a).log(p(s^T|O_k, \lambda_a))\right]. \tag{7}$$

where s^T is a state sequence and S^T is set of all state sequences. Algorithm to calculate $H(\mathbf{S^T}|O_k)$ explained in [6]

Similarly, the entropy of λ_b producing state sequence for given an observation sequence O_b is given by Eqn.(8). The entropy of joint feature stream λ_{a+b}, producing the state sequence for observations O_{a+b} is given by Eqn.(10)

$$H(\mathbf{S^T}|\mathbf{O_b}, \lambda_b) = \sum_{\forall O_k \in \mathbf{O_b}} H(\mathbf{S^T}|O_k, \lambda_b) \tag{8}$$

where

$$H(\mathbf{S^T}|O_k, \lambda_b) = -\sum_{\forall s^T}\left[p(s^T|O_k, \lambda_b).log(p(s^T|O_k, \lambda_b))\right]. \tag{9}$$

where s^T is a state sequence.

$$H(\mathbf{S^T}|\mathbf{O_{a+b}}, \lambda_{a+b}) = \sum_{\forall O_k \in \mathbf{O_{a+b}}} H(\mathbf{S^T}|O_k, \lambda_{a+b}) \tag{10}$$

The delta entropy of a feature stream defined as the difference between the entropy of state sequence of the merged model and entropy of state sequence of model for that feature stream. The delta entropy of model λ_a is given by Eqn.(11). Similarly, delta entropy of model λ_b is given by Eqn.(12).

$$\Delta H_a = |H(\mathbf{S^T}|\mathbf{O_{a+b}}, \lambda_{a+b}) - H(\mathbf{S^T}|\mathbf{O_a}, \lambda_a)| \tag{11}$$

$$\Delta H_b = |H(\mathbf{S^T}|\mathbf{O_{a+b}}, \lambda_{a+b}) - H(\mathbf{S^T}|\mathbf{O_b}, \lambda_b)| \tag{12}$$

For a given syllable, if the entropy of a feature stream relative to the other feature stream is less, i.e., reliability of that feature stream is more and delta entropy is less. Therefore weight can be assigned to a feature stream is inversely proportional to, delta entropy of a feature stream. Therefore,

$$W_a = \frac{1}{\Delta H_a} \tag{13}$$

$$W_b = \frac{1}{\Delta H_b} \tag{14}$$

4. INCORPORATING MODGDF AND MFCC IN THE LINGUISTIC SEARCH SPACE

In [5], it was shown that MODGDF and MFCC a) contain complementary information for a syllable. b) MFCC/MODGDF (one of the features) performs much better than the other for representation of a specific syllable unit indicating that one feature alone is sufficient for recognizing the syllable.

The goal of this work is to incorporate two acoustic feature scores into the linguistic search space. The original idea was to perform feature switching when a given segment of speech was tested against a specific syllable model [5]. But as the range of likelihoods obtained using MFCC and MODGDF were found to be significantly different, the two features were appropriately weighed based on the syllable being recognized [5]. A line search was performed on the training data to obtain the appropriate weights. In this paper we suggest an alternative to this linear search. We obtain the weights using entropy based measures discussed in section 3.

An identical experiment as in [5] was performed on DBIL data to evaluate the proposed scheme. The details of the experiment is discussed in the next section.

5. PERFORMANCE EVALUATION

In this Section, we discuss the performance improvements in speech recognition in terms of word error rate(WER), by incorporating the reliability of the MODGDF and MFCC features into the linguistic search space using the criteria discussed in section 3. Experiments are performed using AAR system on the DBIL database. The DBIL database consists of syllable level transcriptions.

5.1. Syllable based continuous speech recognition: A review[10]

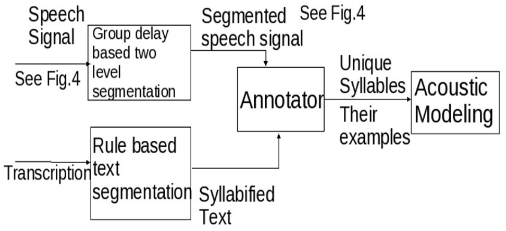

Fig. 1. Block diagram of the training phase in an AAR system

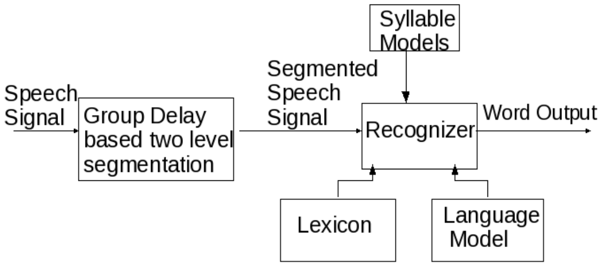

Fig. 2. Block diagram of the testing phase in an AAR system

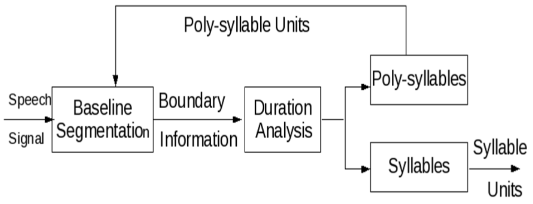

Fig. 3. Block diagram of two level segmentation algorithm

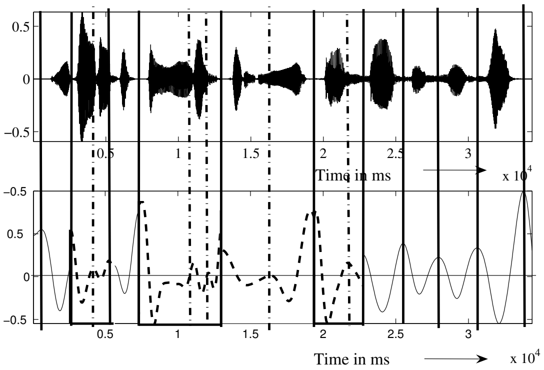

Fig. 4. The speech signal and group delay function for a simple sentence segmented using the two level segmentation algorithm

Automatically Annotated Recognizer(AAR) is a syllable based continuous speech recognition system. In AAR system, speech is segmented into syllable like units using group delay based two level segmentation algorithm. Group delay based two level segmentation algorithm is explained in [10]. 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 then used to extract examples for all unique syllables in the database. The block diagram of the training phase in an AAR system is shown in Fig 1.

During testing phase, the utterance is segmented into syllable like units using the group delay based segmentation algorithm. Word outputs from the AAR system are generated using the syllable models, the segmented test signal, the lexicons and the language models. Block diagram of AAR system test phase is shown in Fig 2. A simple directed search algorithm is implemented in 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 from 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 multipath modification in the search space.

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 as discussed in [11]. Five state HMMs with three mixtures per state are built for all syllables.

5.2. Experimental results

Group delay based automatic segmentation algorithm is used for experiments on the DBIL database. Analysis on DBIL tamil database indicate that 325 syllables are recognized better with the MODGDF, while 425 syllables are recognized better with the MFCC. The performance using both the features for the remaining syllables was very similar. The WER for the DBIL database using weighted likelihood method shown in Table 1.

Table 1. Recognition results on the DBIL database for weighted likelihood with syllable-feature pair information

Features	%WER
MFCC+D+A+E	35.7%
MODGDF+D+A+E	30.4%
Weighted likelihood (For empirically chosen weights)	27.2%
Weighted likelihood(For weights obtained in section 3)	23.9%

From Table1 we see that we have been able to reproduce the results in [5] with a significant reduction in WER.

6. CONCLUSION

In this paper we have proposed an entropy based technique by which we quantified the reliability of a feature stream for the acoustic features derived from the short time magnitude spectrum(MFCC), phase spectrum(MODGDF) and integrating them into the linguistic search space by appropriately weighing the likelihoods of feature streams for syllable based speech recognition. The use of a feature stream to constrain the linguistic search space, can not only reduce recogniser complexity but also improves recognition accuracy. Further, feature switching is perhaps more akin to human speech recognition. To incorporate complementary feature streams in a recognizer, the scores from different feature streams are appropriately normalized. We are currently exploring different approaches to normalize scores obtained from different feature streams.

7. REFERENCES

[1] *The Database of Indian Languages*, Speech and Vision Lab, 2001.

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

[3] 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*, vol. 2007, January 2007.

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

[5] Ramya R, Rajesh M Hegde, and Hema A Murthy, "Incorporating acoustic feature diversity into the linguistic search space for syllable based speech recognition," *EUSIPCO*, 2008.

[6] Diego Hernando, Valentino Crespi, and George Cybenko, "Efficient computation of the hidden markov model entropy for a given observation sequence," *IEEE Transactions on Information Theory*, vol. 51, pp. 2681–2685, July 2005.

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

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

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

[10] Lakshmi A, "A syllable based continuous speech recognizer for indian languages," *MS thesis, Indian Institute of Technology Madras, Department of Computer Science and Engg,Chennai,India*, 2007.

[11] G. Lakshmi Sarada, "Automatic transcription of continuous speech for indian languages," *MS thesis, Indian Institute of Technology Madras, Department of Computer Science and Engg,Chennai,India*, 2005.