AN ONLINE MODEL ADAPTATION METHOD FOR COMPENSATING SPEECH MODELS FOR NOISE IN CONTINUOUS SPEECH RECOGNITION

Raymond Lee, Eric H. C. Choi

Motorola Labs, Motorola Australian Research Centre
{Raymond.H.Lee, Eric.Choi}@motorola.com

ABSTRACT

This paper presents a method for online model adaptation based on parallel model combination (PMC) method. The proposed method makes use of the concept of Gaussian model clustering to reduce the computation load required by PMC. This model clustering in combination with a set of transformation equations derived provide a potential framework for online model adaptation in noisy speech recognition. The proposed method reduces the computation in adaptation by about 45% with only a slight degradation in improvements of an average 18% for a connected digit task and 9% for a large vocabulary Mandarin task when comparing with standard PMC method.

1. INTRODUCTION

Large vocabulary speech recognition systems are deployed in ever increasing numbers and situations. They need to be sufficiently flexible to cope with a wide range of environments in which there may be different levels of background noise. This so-called mismatch problem is causing degradation of recognition accuracy due to the differences between the operating environment and the environment in which the training data are collected.

Numerous approaches have been proposed to tackle the problems of noisy speech recognition by compensating acoustic models. These include the use of robust distance measure [1][2], adaptive model compensation [3][4][5] and predictive model compensation [6][7]. The adaptive approach tries to compensate a set of baseline models based on speech data acquired in operating environment. Examples of this type of approach include maximum a posteriori (MAP) adaptation, stochastic matching and maximum likelihood linear regression (MLLR). These techniques normally require a reasonable amount of field speech data to be effective. Contrast with the adaptive approach, the predictive approach does not need any speech data from actual operating environment. It only needs models of noise sources to compensate speech models. Commonly used predictive techniques include vector Taylor series (VTS) and parallel model combination (PMC).

PMC involves the creation of "noisy" acoustic models by combining original speech models with a noise model estimated from the operating environment. Although PMC can improve recognition performance under mismatch conditions, it has one major drawback: it is computationally expensive. This is because the transformations and model combinations needed in PMC are complex. The computational cost is linearly proportional to the number of models in the speech recognition system, since each model needs to be compensated individually for each incoming utterance. Because it is necessary to compensate the models dynamically to adapt to the present environment, a considerable delay is introduced to the speech recognition system. This means that it is impractical to apply PMC dynamically for medium-to-large vocabulary speech recognition systems, as the time delay would be prohibitively high.

To this end, we have developed a fast method for doing model adaptation based on the same mismatch functions used by PMC. The basic concept is to reduce the computation need for the online model adaptation task by first dividing the speech model into clusters. Standard PMC is applied only to the cluster centers and approximate transformations are applied for the rest of the non-center models. The approximation transformations are derived from the mismatch functions and are 2^{nd} order polynomial functions of the distance between the model and its cluster center in the log spectral domain.

2. SYSTEM OVERVIEW

Figure 1 shows a block diagram of a speech recognition system incorporating the proposed acoustic model adaptation module.

The front-end features are applied to the noise model estimator where a noise model is created from the non-speech portion of the input speech utterance. This noise model is estimated online (dynamically) for each input utterance to capture the current noise condition in which the speech recognition system is operating.

The model clustering module computes a cluster map from the acoustic models in which a small subset of the mean vectors from the acoustic models is assigned as cluster centers. Each of the rest of the mean vectors is then mapped to one of the assigned cluster centers. The cluster map is then applied to the model adaptation module.

The model adaptation module takes in the noise model from the noise model estimator and the cluster map from the model clustering module and generates the adapted models. The adapted models are then applied to the back end decoder of the speech recognizer.

3. MODEL ADAPTATION ALGORITHM

The model adaptation algorithm proposed here is designed to reduce the computational load of applying PMC to speech models. It consists of two major components:

(1) A model clustering module that divides the speech models into a smaller number of clusters using a novel clustering algorithm and

(2) A model adaptation module in which standard PMC is applied only to the cluster centers and approximate transformations are applied for the rest of the non-center models.

3.1 Model clustering module

The model clustering module divides the original speech models into a small number of clusters. The models within each cluster are considered close to one another according to a distance measure used by the module. Only the mean vectors of the models are used in the clustering algorithm. The mean vectors are first divided into N non-overlapping bins, using the first dimension in the MFCC vector, that is c0. The c0 value provides a measure of the energy of the underlying front-end features represented by the model mean vector. Standard minMax clustering algorithm is then applied to each bin to divide the models. A distance measure based on the model values in the log spectral domain is used. That is, the distance D_{ij} between two MFCC mean vectors $\underline{S}_i^c$ and $\underline{S}_j^c$ is the norm of the multiplication of an inverse discrete cosine transform matrix and the difference between the two vectors is as shown in the equation below:

$$D_{ij} = | IDCT * (S_i^c - S_j^c) | \qquad (1)$$

where IDCT is the inverse discrete cosine transform. The IDCT is used here to generate distance measures matching those used by the model adaptation module (see next sub-section). The operation "|.|" denotes the computation of square root of the sum of elements of its vector argument.

The clustering algorithm terminates when if all the samples in the cluster have distances to its center smaller than a pre-defined threshold.

3.2 Model adaptation module

The model adaptation module takes in the noise model and the cluster map and transforms the input acoustic models into the adapted models that match the acoustic condition in which the speech recognizer is operating. The module applies PMC only to the cluster centers in the cluster map. For each of the rest of the acoustic models, an approximate transformation is constructed and the adapted model is then computed by using the transformation and the PMC-adapted model of the cluster center in which the current model is mapped to.

The approximate transformations for the MFCCs and the first order temporal version of the MFCCs are described here. It is assumed here that the speech signal in the operating environment is corrupted by additive noise. The mismatch function in log-spectral domain is given by [8]

$$Y^l = \log(\exp(S^l) + \exp(N^l)) \qquad (2)$$

where Y^l is the noisy model parameter. As the MFCCs are generated by multiplying the S^l etc. by the DCT, it is necessary to perform an inverse discrete cosine transform (IDCT) on the original model parameters to get S^l and N^l.

For another speech model parameter S_1 "close to" S:

$$S^l{}_1 = S^l + \Delta S^l \qquad (3)$$

So that the mismatch function is given as:

$$Y_1^l = \log(\exp(S_1^l) + \exp(N^l)) \qquad (4)$$

Assuming that the relationship between the two compensated models is given by:

$$Y_1^l = Y^l + \Delta Y^l \qquad (5)$$

This gives:

$$\Delta Y^l \approx k\Delta S^l + \frac{k(1-k)}{2}(\Delta S^l)^2 \qquad (6)$$

The above approximation is valid when ΔS^l is smaller than 1.0, where k is a constant given by:

$$k = \frac{\exp(S^l)}{\exp(S^l) + \exp(N^l)} \qquad (7)$$

Hence,

$$Y_1^l \approx Y^l + k\Delta S^l + \frac{k(1-k)}{2}(\Delta S^l)^2 \qquad (8)$$

The above procedures are applied to the first order temporal version of MFCCs using the mismatch function:

$$dY^l = \log(\exp(S^l + dS^l) + \exp(N^l + dN^l)) - Y^l \qquad (9)$$

And the corresponding approximation equation is:

$$dY_1^l \approx dY^l + k_1(\Delta S^l + \Delta dS^l) + \frac{k_1(1-k_1)}{2}(\Delta S^l + \Delta dS^l)^2 \qquad (10)$$

Where k_1 is given by the equation:

$$k_1 = \frac{\exp(S^l + dS^l)}{\exp(S^l + dS^l) + \exp(N^l + dN^l)} \qquad (11)$$

Equation (6), (7), (10) and (11) are used for compensating the non-center model parameters instead of the standard PMC method.

4. EXPERIMENTS

The proposed method are tested on two tasks:

(1) A speaker independent connected digit recognition task for speech recorded in car noise environments

(2) A speaker independent large vocabulary recognition for Mandarin speech corrupted by adding car noises.

The front-end consists of 13 MFCCs, their delta and delta-delta parameters. The Motorola Lexicus speech recognition engine [9] is used in this work. State tied multiple mixture Gaussian continuous density hidden Markov models are used by the engine. The HMMs are trained using the original "clean" data. For the first task, that corresponds to data collected when the car is stationary.

Databases for both tasks are collected within Motorola. The in-car digit database consists of real noisy speech signals recorded in a car at four driving conditions. It contains utterances from 102 speakers recorded in 14 different cars.

Table 1 shows the recognition results for the connected digit task. It also compares the computational requirement between the standard PMC and the proposed method on a Sun UltraSparc II 250 MHz workstation. It can be seen that both adaptation methods provide considerable improvement over the baseline performance. Also, the proposed method only requires only about half of the computation as compared with PMC method with slight degradations in word error rates.

Table 2 shows the results for the Mandarin large vocabulary task. The database consists of speech of 100 speakers uttering sentences from various read text materials recorded in an office environment. Data from 90 speakers is used for training while that from the remaining 10 speakers are used for testing. The recognizer has a vocabulary of 30K words. A statistical tri-gram model is used to provide the grammar constrain for the search engine. The noisy test data is generated by adding car noise to clean speech data at various signal to noise ratios.

Consistent results are obtained for both tests. In particular, the proposed method reduces the computation by about 45% with a slight degradation in improvements of an average 18% for the connected digit task and 9% for the Mandarin task.

5. DISCUSSIONS

MLLR has become the de-facto standard online model adaptation method because of its low computation requirement and the gain in accuracy performance for noisy speech recognition. We have performed some preliminary experiment with MLLR on the tasks discussed here. We note that there is little or no improvement in recognition accuracy. In some cases, especially when the signal-to-noise ratio is low, the performance actually degrades. We attribute this to the limited amount of training data used to find the MLLR transformation. Also, the accuracy of method highly relies on the availability of accurately aligned data. Such data may not be available in noisy environment.

The computational cost of the proposed method depends on (1) the number of model clusters (and hence the number of PMC calculations) and (2) the complexity of the approximation transformations. We show that effective clustering can be achieved by performing the clustering in the log domain rather than the cepstral domain. We derived the transformation equations in Section 3.2 in the linear and log domains so as to match the mismatch function assumption used by PMC. Different transformations will need to be derived for other PMC methods that use different mismatch functions. The approximation transformation equations can be computed effectively as most of the terms are either computed once for each cluster or can be pre-computed. 2nd order polynomial is used here to reduce the number of clusters required in the cluster map.

6. CONCLUSIONS

This paper presents a method for online model adaptation based on the parallel model combination method. The proposed method makes use of the concept of Gaussian model clustering to reduce the computation load required by PMC. Experimental results show that this in combination with the transformation equations derived provide a potential framework for online model adaptation in noisy speech recognition.

7. REFERENCES

[1] Chien J.T. et. al., “A Novel Projection-based Likelihood Measure for Noisy Speech Recognition”, Speech Communication, vol. 24, pp. 287-297, July 1998.

[2] Carlson A. and Clements M.A., “A Projection-based Likelihood Measure for Speech Recognition in Noise”, IEEE Trans. Acoustics, Speech and Signal Processing, vol. 2 (1) part 1, pp. 97-102, 1994.

[3] Gauvain J.L. and Lee C.H., “Maximum A Posteriori Estimation for Multivariate Gaussian Mixture Observations of Markov Chains”, IEEE Trans. Speech and Audio Processing, vol. 2, pp. 291-298, Feb 1994.

[4] Huang C.S. and Wang H.C., “An SNR-incremental Stochastic Matching (SISM) Algorithm for Noisy Speech Recognition”, Proc. Automatic Speech Recognition and Understanding Workshop, pp. 39-43, Dec 1999.

[5] Leggetter C.J. and Woodland P.C., “Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density HMMs”, Computer Speech and Language, vol. 9, pp. 171-186, 1995.

[6] Sagayama S., “Differential Approach to Acoustic Model Adaptation”, Proc. Robust Methods for Speech Recognition in Adverse Conditions, pp. 61-66, May 1999.

[7] Gales M.J.F, “Predictive Model-based Compensation Schemes for Robust Speech Recognition”, Speech Communication, vol. 25, pp. 49-74, Aug. 1998.

[8] Gales, M.J.F. and Young S.J., “PMC For Speech Recognition In Additive and Convolutional Noise”, Technical Report TR154, Cambridge University Engineering Department.

[9] Sreeram V. Balakrishnan, “Effect of Task Complexity on Search Strategies for the Motorola Lexicus Continuous

Speech Recognition System", Proceedings of International Conference on Spoken Language Processing, 1998.

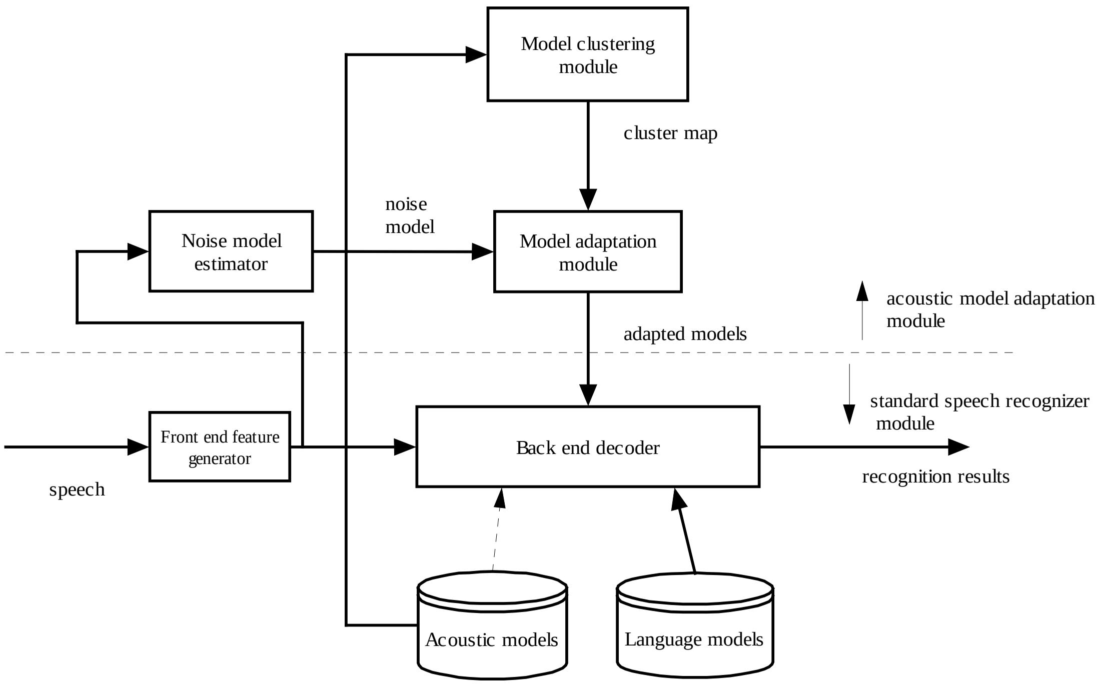

Figure 1: A block diagram for the online model adaptation algorithm in augmentation with an automatic speech recognizer.

Table 1: Recognition results for real car noisy data.

	Word error rate (%)		
Condition	Baseline	Standard PMC	Proposed method
Engine idle	1.8	1.9	2.0
30-45 MPH	5.7	4.2	4.6
55-65 MPH	32.2	12.0	14.4
Mixed	7.21	4.6	5.0
	Computation load for adaptation (% PMC)		
All conditions	0	100	55

Table 2: Recognition results for large vocabulary task.

	Character error rate (%)		
SNR (db)	Baseline	Standard PMC	Proposed method
12	78.9	30.8	35.2
18	45.8	16.3	18.1
24	20.1	11.6	12.7
	Computation load for adaptation (% PMC)		
All conditions	0	100	57