2010 12th International Conference on Computer Modelling and Simulation

New Approach in Transform-Based Speaker Adaptation Using Minimum Classification Error

Reza Sahraian[1], Behzad Zamani[2], Ahmad Akbari[2], Ahmad Ayatollahi[1] ,Babak Nasersharif[3],
[1]Electrical engineering department , Iran University of Science and Technology
[2]Computer engineering department, Iran University of Science and Technology
[3] Computer Engineering Department, Faculty of Engineering, University of Guilan
[1,2]Tehran. Iran, [3]Guilan. Iran
rsahraian@ee.iust.ac.ir, {bzamani, Nasser_s, akbari, ayatollahi}@iust.ac.ir

Abstract— **Automatic speech recognition (ASR) systems work well when trained for a number of specific speakers. However, in most applications there are multiple speakers and they are unknown to the system; performance of ASR system may be degraded because of such speaker variations. This paper examines the use of minimum classification error (MCE) as a preprocessing operation to improve the performance of conventional MLLR (Maximum Likelihood Linear Regression) adaptation. MCE applies its effect by providing better classified components for regression tree in the case of making regression tree on the basis of acoustic space. In this case, distribution of Gaussians will be more smoothing in regression classes. Experimental results on TIMIT database show that 0.42%-0.58% relative improvement is achieved in phoneme recognition rate using our proposed method.**

Keywords- speaker adaptation; regression class trees; minimum classification error.

I. INTRODUCTION

In speaker-independent speech recognition systems, recognition accuracy varies considerably from speaker to speaker, and performance may be significantly degraded for outlier speakers such as nonnative talkers. Maximum likelihood linear regression (MLLR) adaptation has proven to be an effective speaker adaptation technique in the presence of limited adaptation data [1]. A set of linear transformations for the mean (and, possibly, variance) parameters of a mixture Gaussian HMM system is estimated such that the likelihood of the adaptation data is maximized.

Furthermore, the MLLR transforms themselves may be used as features for speaker modeling purposes, e.g., in speaker recognition [2]. It is shown that speaker-clustering procedure that models speaker variability by partitioning a large corpus of speakers in the eigenspace of their MLLR transformations and learning cluster-specific regression class tree structures leads to reduction in overall word error rates in automatic speech recognition systems [3].

To improve the performance of MLLR adaptation many efforts has been done in optimizing and developing regression trees. In [4] it has been attempted to arrive at regression class trees that are closer to the maximum likelihood solution by employing an iterative procedure starting from an initial acoustic clustering. However, unfortunately no improvements in error rate have been observed; it is shown that correlation is a good criterion for making regression class trees[5]. Two approaches to the design of regression class trees are common practice [5]:

1) Phonetic knowledge: Here, expert knowledge is used to decide which components are to be transformed together. The components are split according to broad phonetic classes (e.g., nasals, glides) or, at a lower level, into phones.

2) Acoustic space: Components are clustered according to how close they are in acoustic space, irrespective of which phone they belong to. This has the advantage of being a "data-driven" approach with no need for expert knowledge. However, the resulting classes usually cannot be assigned a phonetic identity.

In this paper we are about to explain the improvement achieved by using minimum classification error in MLLR adaptation. Minimum classification error is a well-known discriminative method used for both feature transformations and classifier training [6]. Using MCE in adaptation of parameters of Gaussian mixture continuous density HMM was first reported in [7]. In [7], the gradient probabilistic descent (GPD) algorithm was directly applied to adapt the linear regression matrices and MCE is used as an alternative for maximum likelihood to estimate transformation matrix.

In this study MCE is used to improve classification of Gaussian mixture components in regression class trees in the case of classifying according to acoustic space. When the regression class tree is made on the basis of acoustic space, MCE improves MLLR adaptation by smoothing the distribution of Gaussians in different classes.

The remainder of paper is organized as follows. In Section 2, MLLR is explained briefly. Section 3 describes MCE algorithm. Section 4 introduces our proposed method. Section 5 contains our experimental results. Finaly, Section 6 includes the conclusion.

II. MAXIMUM LIKELIHOOD LINEAR REGRESSION

In MLLR Adaptation it is assumed that a new mean vector $\hat{\mu}$ is related to its baseline mean value μ by linear regression. For an arbitrary regression class m, the relation is given by the multiplication matrix W_m , and shift vector b_m

978-0-7695-4016-0/10 $26.00 © 2010 IEEE
DOI 10.1109/UKSIM.2010.62

$$\hat{\mu} = W_m \mu + b_m \tag{1}$$

W_m is estimated to maximize the likelihood of adaptation data according to following equation

$$\sum_{r=1}^{R}\sum_{t=1}^{T}\gamma_{kr}(t)C_{kr}^{-1}o_t\xi_{kr}^{t} = \sum_{r=1}^{R}\sum_{t=1}^{T}\gamma_{kr}(t)C_{kr}^{-1}W_k\xi_{kr}\xi_{kr}^{t} \tag{2}$$

Where R is a number of Gaussians in class k, T is the length of observation $O_T = \{o_1,..., o_t,..., o_T\}$, γ is a prior probability and C is covariance matrix[1].

III. Minimum Classification Error

The main idea behind the MCE algorithm is to optimize an empirical error rate on the training set to improve the overall recognition rate. The MCE method has also been used for obtaining feature space transformations [8]. In this approach a cost function like equation (3) is used:

$$d_{k,Add}(O_{n_k}, F) = -g_k(O,F) + \frac{1}{\eta}\log\left[\frac{1}{I-1}\sum_{i=1,i\neq k}^{I}\exp\left(g_i(O,F)\eta\right)\right] \tag{3}$$

Where, F stands for the feature extraction parameters, I is the number of classes, η is cluster weighting which is tuned empirically. $g_i(O,F)$ generates logarithm likelihood for observation O.

In the MCE methods, our object is to minimize the cost function (3). We define a cost function that maps the misclassification measure between zero and one. For this purpose, we select sigmoid function as the cost function

$$l_k(O,F) = \frac{1}{1+\exp\left(-\alpha d_k(O,F)\right)} \tag{4}$$

Where, $d_k(O,F)$ is as in (3), and α is a tuning parameter greater than one. Obviously, when $d_k(O,F)$ is smaller than zero in equation (3), implying an accurate classification, $l_k(O,F)$ will be close to zero indicating no loss. For an observation, O, the total classification error is computed by

$$L = \sum_{k=1}^{M} l_k(O,F) \tag{5}$$

The transformation matrix can now be computed using the gradient descent method for function L

$$\mathbf{W}_{iter} = \mathbf{W}_{iter-1} - \beta\frac{\partial L}{\partial \mathbf{W}} \tag{6}$$

Where, W is the transformation matrix; *iter* denotes iteration number in gradient descent algorithm and β is the learning parameter. Equation (6) represents an iterative process. The iteration is stopped when total classification error, L, is not changed by consecutive iteration. In this paper, we use MCE algorithm to improve MLLR adaptation.

IV. Proposed Method

Equation (2) shows that the more number of Gaussians in each class, the more robust estimation we have for that class [9]. Since the number of Gaussians in the SI model is constant, as the number of Gaussians in one class increases, it decreases in other classes and as a result the estimation of transformation matrix for those classes degrades.

When the large number of Gaussians is dedicated to one class, the transformation matrix which is estimated for that class is more robust than other classes. However, most of Gaussians are transformed by the same matrix in this case in MLLR adaptation and it looks like a global adaptation. Thus, the improvement in multi-class MLLR adaptation is degraded.

In the case that Gaussians are classified according to their closeness in acoustic space, standard MLLR considers difference between means of Gaussians as a criterion to classify them. So for better classification all data can be simply multiplied by a constant greater than one. Suppose "Fe" is a sequence of feature vectors, "fe_t", of length T:

$$Fe = \{fe_1, fe_2,..., fe_T\} \tag{7}$$

We define a new feature space "Fe'"

$$Fe' = \{fe'_1, fe'_2,..., fe'_T\} \tag{8}$$

$$fe'_t = D \times fe_t = D \times \left[fe_{t1}\ fe_{t2}\ ...\ fe_{tn}\right]^T \tag{9}$$

Where n is the dimension of feature vector and D is a $n \times n$ diagonal matrix. Considering a Gaussian distribution for each dimension of feature vectors we have

$$\begin{aligned} fe_{tj} &: N(\mu;\sigma) \\ fe'_{tj} &: N(d_{jj}\mu; d_{jj}\sigma) \end{aligned} \tag{10}$$

Where d_{jj} is jth element of D. In HMM, probability distribution is considered as a weighted linear combination of normal distribution. Thus, for any probability distribution for each dimension of feature vector we have

$$\begin{aligned} P(fe_{tj}) &= \sum_i k_i N(\mu_i;\sigma_i) \\ P(fe'_{tj}) &= \sum_i k_i N(d_{jj}\mu_i; d_{jj}\sigma_i) \end{aligned} \tag{11}$$

Where k_i is the weight of ith Gaussian in probability distribution for each dimension of feature vector.

Minimum classification error algorithm estimates a diagonal transformation matrix which is applied on the

feature vectors and does the same as what *D* does in the above explanation. New models in new feature space are more discriminated. If the elements of the diagonal transformation matrix, as a multiplier of each dimension of feature vectors, are greater than one, as described, the Gaussians of even one model will be more discriminated.

In our proposed method, MCE is used as a preprocessing algorithm before clustering Gaussian mixture components in MLLR adaptation; it causes better classification and consequently better adaptation.

V. EXPERIMENTAL RESULTS

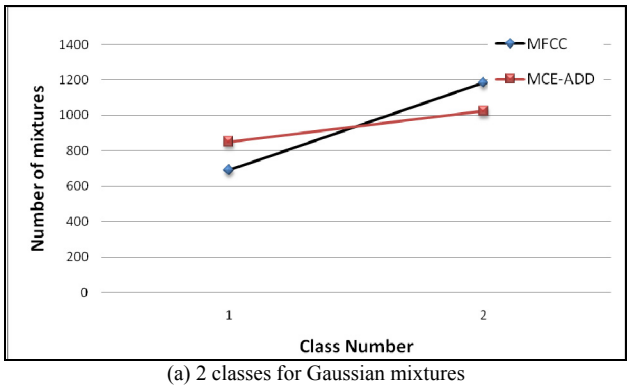
To evaluate performance of proposed method, phoneme recognition rate in speaker independent recognition system and MLLR adaptation system in two cases of with MCE and without MCE were compared on TIMIT dataset.

In all experiments, we use 39-dimension feature vectors consisting of energy, 12 MFCCs and their first and second order derivatives. The features were normalized to have mean zero and standard deviation one over TIMIT training set. We don't reduce features vector dimension in our methods. In addition, we use HMMs with 3 states and 16 Gaussian mixtures per state. Therefore, the total number of Gaussian mixtures is 1872. Table (1) contains results of phoneme recognition rate on TIMIT dataset. It shows the effect of MCE which used in MLLR adaptation based on acoustic space. We use 10 sentences for each speaker as an adaptation data. Since the amount of adaptation data for 8 and 16 classes MLLR is not available for one speaker in TIMIT database, results are reported only on 2 and 4 class MLLR adaptation in Table (1).After the text edit has been completed, the paper is ready for the template. Duplicate the template file by using the Save As command, and use the naming convention prescribed by your conference for the name of your paper. In this newly created file, highlight all of the contents and import your prepared text file. You are now ready to style your paper.

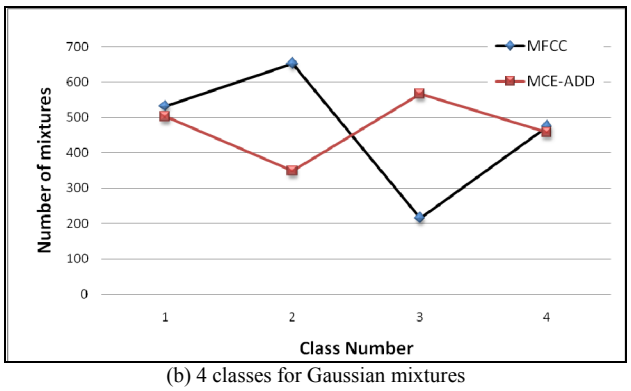
TABLE I. RESULTS OF PHONEME RECOGNITION RATE ON TIMIT DATASET IN ACOUSTIC SPACE CASE

Method	Phoneme recognition rate
SI	80.77
SI-MCE	80.91
MLLR-2	87.57
MLLR-MCE-2	88.06
MLLR-4	88.89
MLLR-MCE-4	89.51

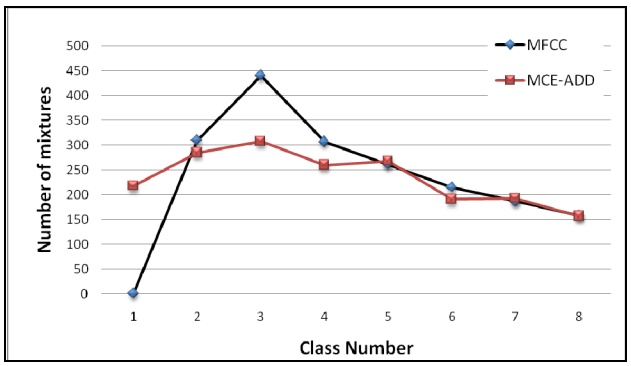
It is shown that the proposed method (MLLR-MCEE-n), where *n* is the number of classes, causes relative improvement of 0.42% for 2-class and 0.58% for 4-class MLLR adaptation in phoneme recognition rate to conventional MLLR adaptation.



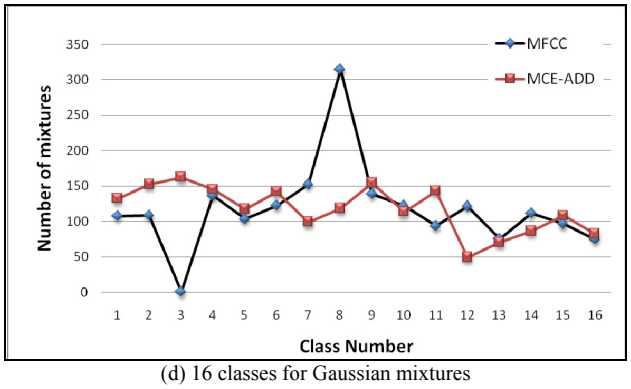
(a) 2 classes for Gaussian mixtures



(b) 4 classes for Gaussian mixtures



(c) 8 classes for Gaussian mixtures



(d) 16 classes for Gaussian mixtures

Figure 1. Distribution of Gaussian mixtures in clusters of MLLR with / without MCE

As can be seen in Figure (2) MCE has been applied before the MLLR, makes Gaussian mixtures to be distributed

more uniformly between classes. Performance of the proposed method improves as number of classes increases.

VI. CONCLUSION

In this paper we introduce a new approach to get better performance for multi-class MLLR adaptation by using minimum classification error algorithm. It is shown that our proposed method improves phoneme recognition rate in conventional MLLR adaptation in acoustic space based regression tree. When regression class is made based on acoustic space we gain 0.42% relative increase for 2-class and 0.58% relative increase for 4-class MLLR adaptation in phoneme recognition rate on TIMIT database.

REFERENCES

[1] C. J. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," J. Comput. Speech Lang., vol. 9, pp. 171–185, 1995.

[2] M. Ferr`as, C. Chi Leung, C. Barras and J. L. Gauvain, "Constrained MLLR for Speaker Recognition", Proceedings of the IEEE ICASSP, pp. 53–56, 2007.

[3] A. Mandal, M. Ostendorf, A. Stolcke, "Improving robustness of MLLR adaptation with speaker-clustered regression class trees", computer speech and language 23, 176-199, 2009.

[4] M. J. F. Gales, "The generation and use of regression class trees for MLLR adaptation," Cambridge Univ., Cambridge, U.K., Tech. Rep. CUED/F-INFENG/TR263, 1996.

[5] R. Haeb-Umbach, "Automatic generation of Phonetic regression class trees for MLLR adaptation", IEEE Trans. Speech Audio Process., vol. 9, no. 3, 2001.

[6] B. Zhang, and S., Matsoukas, Minimum phoneme error based heteroscedastic linear discriminant analysis for speech recognition. *In: Proceedings of ICASSP*, vol. 1. Philadelphia, PA, pp. 925-928, 2005.

[7] C.-H. Lin, C.-H. Wu, and P.-C. Chang, "A study on speaker adaptation for Mandarin syllable recognition with minimum error discriminative training", IEICE Trans. Inf. & Syst., Vol.E78-D, No.6, pp.712-718, 1995.

[8] Biem, and S. Katagiri. Feature extraction based on minimum classification error/generalized probabilistic descent method. *In: Proceedings of IEEE International Conference Acoustic and Speech Signal Processing*, Vol. 2, pp. 275-278, 1993.

[9] Jonathan E. Hamaker, MLLR: A Speaker Adaptation Technique for LVCSR, Lecture for a course at ISIP - Institute for Signal and Information Processing, Department of Electrical and Computer Engineering, Mississippi State University, 1999.