A NOVEL METHOD FOR RAPID SPEAKER ADAPTATION BASED ON SUPPORT SPEAKER WEIGHTING

Tie Cai and Jie Zhu

Department of Electronic Engineering, Shanghai Jiao Tong University
1954, Huashan Road, 106# Shanghai Jiao Tong University, Shanghai, P.R.China, 200030

ABSTRACT

In this paper we propose a novel model-based speaker adaptation method called Support Speaker Weighting (SSW), which performs the adaptation scheme of model combination based on the selected speakers. These speakers, who are acoustically close to the test speaker, are selected from reference speakers using support vector machines (SVM). Compared with GMM/HMM based speaker selection method, the proposed method can quickly obtain a more optimal speaker subset because the selection is dynamically determined according to the distribution of reference speakers around the test. Experimental results for a large-vocabulary task given in this paper show that this method is both cheaper in terms of memory and more effective than Reference Speaker Weighting (RSW) for tiny amounts of adaptation data. Relative error rate reduction of 4.1% is achieved when only one adaptation sentence is available.

1. INTRODUCTION

Rapid speaker adaptation has been an interesting and challenging problem for Large Vocabulary Continuous Speech Recognition (LVCSR) for some time. Reducing the acoustic mismatches due to speaker variability between the training conditions and the testing conditions with a small amount of adaptation data is very important in many applications. Typical speaker adaptation method includes the MAP family [1] and the linear transformation family (e.g., MLLR) [2]. These two families require significant amounts of adaptation date from the new speaker in order to perform better than a SI system. Recently, a family of model combination based speaker adaptation schemes (e.g., RSW, Eigenvoice) has received much attention [3][4]. This approach utilizes the correlations among different reference speakers and performs effectively in rapid speaker adaptation even if only one adaptation sentence has been used. Reference Speaker Weighting (RSW) is a typical example of model combination based speaker adaptation [3]. It trains a speaker dependent model for each of reference speakers and assumes that the adapted model for the test speaker must be a linear combination of the reference models. In practice, the large data storage requirement impacts the application of RSW. Eigenvoice method constructs the speaker space by spanning a K-space via Principal Component Analysis (PCA) and represents the target speaker as a weighted combination of K eigenvectors [4]. However, the PCA process of Eigenvoice is particularly difficult for large scale HMM systems. In addition, the performance of above methods is very sensitive to the choice of the reference speakers. They are very efficient for implementation especially when the reference speakers in the training set are acoustically close to the test speaker.

In this paper, we formulate a novel method for rapid speaker adaptation in LVCSR using Support Speaker Weighting (SSW). This method can select a specific subset of speakers who are close to the test speaker using SVM and performs speaker adaptation based on it. The speaker selection technique, which is not included in RSW and Eigenvoice, reduces the number of parameters to be estimated and space to store the SD models during adaptation. With a very limited amount of adaptation data, SSW outperforms RSW since it uses a smart way to choose an optimal set of reference models. In the next section, the method of speaker selection is described. Our proposed method, SSW, is explained in section 3. Section 4 shows experimental results of conventional model combination and the proposed method on the Mandarin large vocabulary continuous speech recognition task.

2. SPEAKER SELECTION

As mentioned in first section, we should select a subset of speakers who are acoustically close to the test speaker from a pool of reference speakers. The appropriate speaker representation is a key issue in procedure of speaker selection. There are two representations of speaker, namely transformation matrix based and model based [5][6]. We can select the similar speakers for the target speaker based on these two representation methods.

2.1. Speaker Representation

0-7803-8874-7/05/$20.00 ©2005 IEEE

2.1.1 Representation based on MLLR matrix

As described in [5], we can use the MLLR transformation matrix (including offset) to describe the characteristics of a speaker. But when the amount of adaptation data is very limited, the MLLR transformation matrix will be poorly estimated and can't represent speakers appropriately.

2.1.2 Representation based on model

In speech technologies, GMM and HMM are widely and successfully used to represent speakers. They are able to model the main characteristics of a speaker in details. In Eigenvoice method [4], the reference speakers are represented by the supervectors which are constructed using the means of the HMM output Gaussians extracted from the SD models. In SSW, we use the same way to represent the test speaker and the reference speakers.

2.2. Support Speaker Selection

According to the methods to represent speakers, we can compute Euclidean distance based on transformation matrix or likelihood based on GMM/HMM to find the similar speakers for the target speaker [6]. Experiments in [6] show that likelihood scores from GMM is the most efficient method. But it must build one model for each reference speaker and compute the likelihood of all SD models during selection. Furthermore, the number of selected speakers in above all methods is fixed, although the optimal number of similar speakers for each test speaker is different in practice.

Support vector machine (SVM) is a promising machine learning technique developed from the theory of Structural Risk Minimization [7]. Usually, the final classification function of SVM only depends on a small part of the training samples that are called support vectors. These support vectors lie close to the decision boundary between the two classes and carry all relevant information about the classification problem [8]. In other words, they are the nearest samples between the two classes to be classified.

Regarding the reference speakers and the test speaker as two classes, we can use their feature vectors to train a SVM. The support vectors in reference speakers are approximately close to the class of test speaker, especially when these reference speakers distribute around the test in equality. Then the reference speakers corresponding to these support vectors can be selected as a subset, called support speakers. Figure 1 illustrates the principle of this method. Thus, the model combination method based on this subset for rapid adaptation is called Support Speaker Weighting (SSW).

As compared with the GMM/HMM based selection method, the support speaker selection only needs to train a SVM for all the speakers. Its computation cost is much lower since GMM/HMM has to calculate the likelihood for all SD models. In addition, the number of selected speakers in SSW is not fixed for each test speaker. It varies according to the distribution of reference speakers around the test. As the number of reference speakers grows, SSW can get more accurate support speakers who are acoustically close to the test.

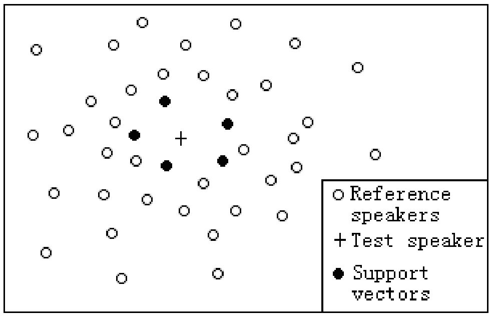

Fig.1. Speaker selection in support speaker weighting

3. SUPPORT SPEAKER WEIGHTING (SSW)

3.1. The Proposed Algorithm

Before performing the procedure of SSW, the parameters of the speaker-adapted model for each reference speaker are estimated by using the maximum likelihood linear regression technique of [2]. These adapted models can be considered as the approximation of the speaker dependent (SD) models for each of the reference speakers. The detailed procedure of SSW is summarized in Figure 2.

First, we construct a reference set of R well-trained SA models, plus an SI model. In SSW, these SA models are used as SD models for each of reference speakers. Second, we apply MAP adaptation to the test speaker and extract the mean vectors of the Gaussians updated by MAP to form a supervector $\hat{X}_{MAP}$ in an arbitrary order. To characterize the speaker more accurate with tiny adaptation data, the prior parameter τ of MAP is set to 0 in this paper. Third, we construct R supervectors X_r ($r = 1,2,\cdots,R$) from the SD models in the reference set, as long as the extracted Gaussians and the order are the same as that of $\hat{X}_{MAP}$. Using these R+1 supervectors to represent speakers of two classes (one class is the test speaker and the other is the reference speakers), a SVM can be trained with the training set $\{\ (\hat{X}_{MAP}, \hat{y})\ ,\ (X_1, y_1), \cdots, (X_r, y_r), \cdots, (X_R, y_R)\ \ \}$ where $\hat{y} = +1$ and $y_r = -1$. Then we are able to select a subset of speakers corresponding to the support vectors in SVM. Finally, we compute the combination weight coefficients through a maximum-likelihood (ML) estimator, and linearly combine the selected SD models.

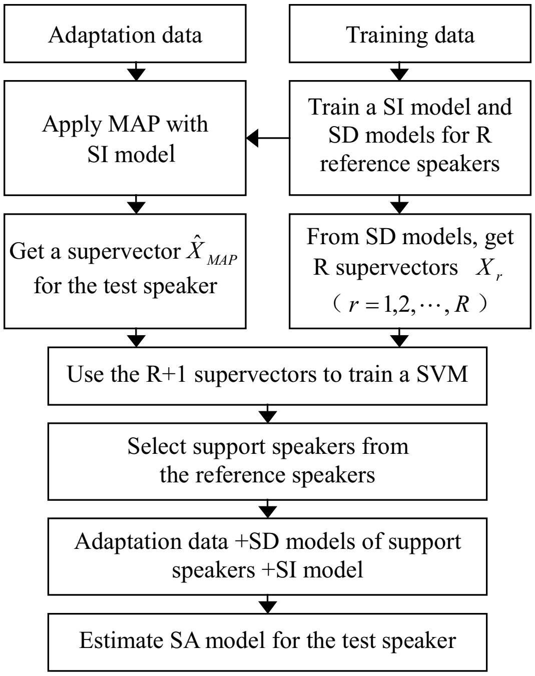
Adaptation data
Training data
Apply MAP with SI model
Train a SI model and SD models for R reference speakers
Get a supervector $\hat{X}_{MAP}$ for the test speaker
From SD models, get R supervectors X_r （$r = 1, 2, \cdots, R$）
Use the R+1 supervectors to train a SVM
Select support speakers from the reference speakers
Adaptation data +SD models of support speakers +SI model
Estimate SA model for the test speaker

Fig.2. Block diagram for support speaker weighting

3.2. Parameter Estimation

Assume that there are M speakers selected. In order to adjust the adaptation parameters with the available amount of adaptation data, we introduce the Gaussian binary tree structure of MLLR into SSW. Each node of this tree represents a regression class. A set of M weights associated with this node will be applied only to the corresponding Gaussians belonging to it.

For the k -th Gaussian component of state i in a regression class, the mean vector for test speaker, $\hat{\mu}_{ik}$, is given by

$$\hat{\mu}_{ik} = \sum_{j=1}^{M} w(j)\mu_{ik}^{SV}(j) = E_{ik}W \tag{1}$$

where E_{ik} is the matrix of mean vectors belonging to M support speakers for this Gaussian component,

$$E_{ik} = \left[\mu_{ik}^{SV}(1), \mu_{ik}^{SV}(2), \cdots, \mu_{ik}^{SV}(M)\right] \tag{2}$$

$\mu_{ik}^{SV}(j)$ is the mean vector of the k -th Gaussian component of state i associated with support speaker j .

$W = \left[w(1), w(2), \cdots, w(M)\right]^T$ is the weight vector .

To maximize the likelihood of adaptation data from test speaker, the maximum-likelihood (ML) estimator is used to find the value of W as follows:

$$W^{opt} = \arg\max_{W} \left\{\log p(O \mid \hat{\lambda})\right\} \tag{3}$$

where $\hat{\lambda}$ is the adapted model which is determined by the weight vector W . $O = \left\{o_1, o_2, \cdots, o_T\right\}$ is the observation sequence of adaptation data.

Construct the auxiliary function as follows:

$$Q(\lambda, \hat{\lambda}) = -\frac{1}{2}P(O \mid \lambda) \times \sum_i \sum_k \sum_t \gamma_{ik}(t) f(o_t, i, k) \tag{4}$$

where

$$f(o_t, i, k) = \left[-n\log(2\pi) - \log\left|\Sigma_{ik}\right| + h(o_t, i, k)\right] \tag{5}$$

and

$$\begin{aligned} h(o_t, i, k) &= (o_t - \hat{\mu}_{ik})^T\, \Sigma_{ik}^{-1}(o_t - \hat{\mu}_{ik}) \\ &= (o_t - E_{ik}W)^T\, \Sigma_{ik}^{-1}(o_t - E_{ik}W) \end{aligned} \tag{6}$$

The λ is current model and Σ_{ik} is the covariance matrix of current Gaussian component. $\gamma_{ik}(t)$ is the occupation probability at time t .

To maximize $Q(\lambda, \hat{\lambda})$, set $\partial Q / \partial W = 0$. The weight vector W can be obtained from the following equation:

$$\sum_i \sum_k \sum_t \gamma_{ik}(t) E_{ik}^T \Sigma_{ik}^{-1} o_t = \sum_i \sum_k \sum_t \gamma_{ik}(t) E_{ik}^T \Sigma_{ik}^{-1} E_{ik} W \tag{7}$$

4. EXPERIMENTAL RESULTS AND ANALYSIS

4.1. Experimental Conditions

The experiments are performed on a large-vocabulary continuous Mandarin speech recognition system based on tri-phone HMMs. The speech feature vector is MFCC with 39-dimension, which consists of 12 mel-frequency cepstral coefficients plus the energy term and their first and second derivatives. All the speech data is provided by the database of Microsoft Research Mandarin Speech Toolkit. The speaker independent (SI) model is trained with the training set, which contains 100 male speakers, each speaking approximately 200 sentences. Evaluation of the adaptation techniques is performed on an additional test set of 10male speakers with 20 sentences per speaker. For each speaker in test set, 10 sentences are taken as the adaptation data while the rest are for recognition. Before model adaptation, 100 SA models are obtained by MLLR as the SD models in reference set. It should be noted that we focus on very rapid adaptation of large-vocabulary systems in this paper. All the adaptation methods in experiments are performed with only one adaptation sentence (5s).

4.2. Experimental Results

When considering different kernels of SVM in SSW, we have experimented with three kinds of kernels for speaker selection: linear, polynomial and RBF. Table 1 shows average recognition rates of SSW (75 reference speakers in SSW) with different kernels of SVM. SV represents the number of support speakers. We can conclude from table 1 that linear kernel SVM based SSW obtains the best recognition accuracy. When we use non-linear kernel SVM to the supervectors that are linearly separable, they

Table 1: Comparison with different kernel of SVM (Only one adaptation sentence)

Recognition rates (%)	SSW		SSW		SSW	
	Linear	SV	Poly3	SV	RBF	SV
Average	55.05	10.6	54.62	12.7	54.04	21.3

are mapped into a high dimensional feature space in which the mapped data becomes non-separable. Thus, the support vectors obtained by non-linear kernel SVM are associated with classification errors that bring a big error in speaker selection.

As discussed in section 2.2, when the number of reference speakers grows, SSW can get more accurate support speakers who are acoustically close to the test. The performance of this algorithm will be improved as the amount of reference speakers ready for speaker selection increases. Figure 3 illustrates the adaptation results of the proposed algorithm varying the number of reference speakers (one adaptation sentence per speaker).

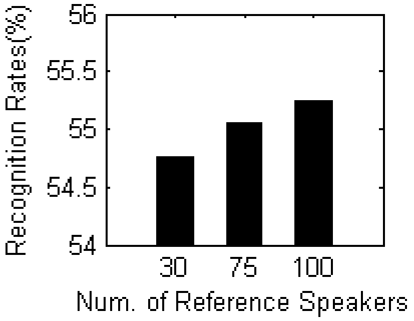

Fig. 3.Comparison with different number of reference speakers

Table 2: Average recognition rates (%) of different adaptation methods with only one adaptation sentence (75 reference speakers in SSW)

%	SI	MAP	MLLR	RSW (15)	RSW (30)	SSW
Recognition rates	53.13	53.05	52.13	53.60	54.48	55.05
Rel. Err. Reduction	--	-0.17	-2.13	1.00	2.88	4.10

To compare with the conventional methods, the MAP, MLLR and RSW are examined. Table 2 demonstrates the recognition results for adaptation on one sentence. For very rapid adaptation of large-vocabulary systems, neither MAP nor MLLR in their original form are effective. RSW applies strong constraints to the adapted model, which speeds up the adaptation process and provides improvement of performance with tiny adaptation data. To obtain a good adapted model, SSW incorporates a speaker selection procedure into adaptation scheme. Given 75 reference speakers in experiment, it selects a subset of almost 10 speakers who are acoustically close to the test speaker by training a SVM.

Table 2 shows SSW yields better performance than RSW(30) (30 SD models of reference speakers are used) by using only one-third of total memory of RSW(30) to store SD models. From the experimental results, we can conclude that SSW can improve rapid adaptation performance over that of RSW method with only a small amount of adaptation data, especially when the amount of reference speakers is enough to reliably select support speakers.

5. CONCLUSION

Model combination method has been proved effective in rapid speaker adaptation. But its performance is very sensitive to the choice of initial models, and it needs a large memory to store SD models. A novel model-based speaker adaptation method, Support Speaker Weighting (SSW), is proposed in this paper. It realizes the specific speaker selection by finding the support speaker subset from many reference speakers. This method yields major improvements in performance for tiny amounts of adaptation data. Experimental results show that relative error rate reduction of 4.1% is achieved when only one adaptation sentence is available. In comparison with RSW method, SSW achieves a further error rate reduction and needs less memory during adaptation.

6. REFERENCES

[1] J. L. Gauvain and C. H. Lee, “Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains,” IEEE Trans. Speech Audio Proc., vol. 2, pp. 291-298, 1994.

[2] C. J. Leggetter and P. C. Woodland, “Maximum likelihood linear regression for speaker adaptation of continuous-density hidden markov models,” Computer Speech Language, vol. 9, pp. 171-185, 1995.

[3] T. Hazen, “The use of speaker correlation information for automatic speech recognition,” Ph.D. diss., Mass. Inst. Technol., Combridge, Jan. 1998.

[4] R. Kuhn, J. -C. Junqua, P. Nguyen and N. Niedzielski, “Rapid speaker adaptation in eigenvoice space,” IEEE Trans. Speech Audio Proc., vol. 8,pp. 695-707, 2000.

[5] C. Huang, T. Chen, S. Li, E. Chang and J. L. Zhou, “Analysis of speaker variability,” in Proc. Eurospeech 2001, vol. 2, pp. 1377-1380, 2001.

[6] C. Huang, T. Chen and E. Chang, “Speaker selection training for large vocabulary continuous speech recognition,” in Proc. ICASSP2002, pp. 609-612, 2002.

[7] V. Vapnik, “The nature of statistical learning theory,” Springer Verlag, New York, 1995.

[8] S. R. Gunn, “Support vector machines for classification and regression,” Technical Report, Image Speech and Intelligent Systems Research Group, University of Southampton, 1997.