World Academy of Science, Engineering and Technology
International Journal of Computer and Information Engineering
Vol:3, No:11, 2009

An Adaptive Hand-Talking System for the Hearing Impaired

Zhou Yu, and Jiang Feng

Abstract—An adaptive Chinese hand-talking system is presented in this paper. By analyzing the 3 data collecting strategies for new users, the adaptation framework including supervised and unsupervised adaptation methods is proposed. For supervised adaptation, affinity propagation (AP) is used to extract exemplar subsets, and enhanced maximum a posteriori / vector field smoothing (eMAP/VFS) is proposed to pool the adaptation data among different models. For unsupervised adaptation, polynomial segment models (PSMs) are used to help hidden Markov models (HMMs) to accurately label the unlabeled data, then the ”labeled” data together with signer-independent models are inputted to MAP algorithm to generate signer-adapted models. Experimental results show that the proposed framework can execute both supervised adaptation with small amount of labeled data and unsupervised adaptation with large amount of unlabeled data to tailor the original models, and both achieve improvements on the performance of recognition rate.

Keywords—sign language recognition, signer adaptation, eMAP/VFS, polynomial segment model.

I. INTRODUCTION

THERE are about 500 million people who suffer from hearing loss worldwide [1], and there are over 27 million hearing impaired people in China [2]. The most important way for the hearing impaired to communicate with the hearing society is by sign language interaction. Whereas, most people are not familiar with sign language, which prevents them from communicating with the hearing impaired. With the development of artificial intelligence techniques in computer science, automatic recognizing and synthesizing sign language have come into reality. Together with the mature techniques of automatic speech recognition and synthesis, we can help the hearing impaired to communicate with hearing society by the aid of personal computers.

Our previous work [3] proposed a Chinese sign language / spoken language dialog system. With this system a hearing impaired person can ”talk” with a hearing person through the internet. The system consists of mainly four parts: namely sign language recognition (SLR), sign language synthesis (SLS), speech recognition, and speech synthesis. The SLR module can accurately translate sentences composed of over 5000 Chinese sign language words to texts for a specific person. Whereas the performance decreases drastically when users are unregistered in the systems' training set due to the fact that different persons sign the same sign word differently. Collecting enough signers' data to train signer-independent (SI) models can solve this problem to some extent. Even though, the models are still ”one-size-fits-all”, which means that the SI models can perform acceptably, but not well enough. One alternative to solve this problem is the adaptation techniques, by which the SI models can be tailored to new users with its labeled and unlabeled data. In this paper we propose two adaptation methods to modify the SI models' parameters so that an adaptive Chinese hand-talking system can be achieved.

The remainder of this paper is organized as follows. In Section II, we review the hand-talking system presented in [3] and propose the adaptation framework. Then we describe the supervised adaptation method and the unsupervised adaptation method in Section III and Section IV respectively. The proposed adaptation methods are evaluated in Section V. Finally the conclusions are given.

II. HAND-TALKING SYSTEM AND ADAPTATION FRAMEWORK

The system mainly consists of two terminals: the hearing impaired terminal and the hearing terminal. The hearing impaired terminal has two modules: the SLR module translates sign language signals to texts and sends them to hearing terminal; the SLS module receives texts from hearing terminal and synthesizes them to sign language image sequences. The hearing terminal also consists of two modules: the speech recognition module transcribes speech signals to texts and sends them to the hearing impaired terminal; the speech synthesis module receives the texts from the hearing impaired terminal and synthesizes them to speech signals.

SLR takes sign language signals as input and outputs corresponding texts. Hidden Markov Models (HMMs, [4]) are used as the statistical models in the hand-talking system. Each word is modeled by one HMM. For the signer-dependent (SD) case the recognition accuracy of over 90% can be achieved at the isolated word level. If we add language models to the system, real-time continuous large-vocabulary SLR can be implemented. However training the SD models for new users needs lots of data to be collected. To alleviate this problem we could train SI models using enough data from large number of signers. Nevertheless collecting training data is time consuming and very costly. Even the training data have been collected, two problems still remain:

1) The models are difficult to converge because the data of different signers vary significantly. Sometimes the distinctions among the samples of the same sign from different signers are even bigger than the distinctions among the samples of different signs from the same signer.

Zhou Yu and Jiang Feng are with the School of Computer Science and Technology, Harbin Institute of Technology, Harbin, Heilongjiang, 150001 China. e-mail: siniczhou@gmail.com.

2) The generalization ability is another problem. Well-trained SI models may achieve acceptable performance on new signers, but not perfect performance as achieved by SD models, because the distributions of the SI models are smooth.

The problems encountered above originate from that the model parameters are fixed, that is, the models are one-size-fits-all. If the models can be retrained using data of a new user, problems can be solved. The explicit data collecting process should be short considering the system's usability. As a result, the adaptation data can be collected by 3 ways:

1) Explicitly collecting small amount of labeled data before users first use the system. These data must be representative since they are not only used to adapt their corresponding models but also used to adapt other models.
2) "Implicitly" collecting labeled data when users manipulate the system. The system gives the candidate result list so that users could select the correct result when the first candidate result is wrong. By this way large amount of labeled data can be collected "implicitly".
3) Implicitly accumulating unlabeled data when users use the system. All effective data can be stored in hard disks, but these data are unlabeled.

Corresponding to the 3 data collecting strategies, adaptation methods in supervised and unsupervised settings are proposed in this paper, as shown in Fig. 1. Supervised adaptation includes two methods: adaptation using large amount of labeled data that are collected by strategy 2 and adaptation using small amount of data collected using by strategy 1. Unsupervised adaptation methods use the unlabeled data collected by strategy 3. In next two Sections we describe supervised and unsupervised adaptation respectively.

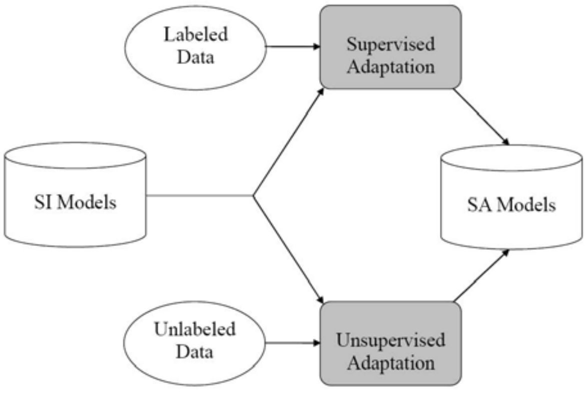


Fig. 1. The adaptation framework. If labeled data are supplied, supervised adaptation is executed, and if unlabeled data are supplied, unsupervised adaptation is executed.

III. Supervised Adaptation

A. Supervised Adaptation with MAP

The adaptation data collected by strategy 2 are labeled manually, which could assure that each HMM has at least one sample for adaptation. Consequently, Maximum a posteriori (MAP, [5]) can be used to adapt SI models. MAP, which is also called Bayesian adaptation, utilizes the prior distribution of parameters and the limited adaptation data to estimate the adapted parameters. If θ is the parameter to be adapted using observation x, and its prior probability is $p(\theta)$, then the MAP estimate of θ is:

$$\theta_{MAP} = \arg\max_{\theta}(p(\theta|x)) = \arg\max_{\theta}(p(x|\theta)\cdot p(\theta)) \qquad (1)$$

The prior distribution is generally extracted from the SI models. Previous experimental results show that for SLR the means of HMMs' mixture components are most important so we only tailor the mean parameters. If conjugate priors are used, a simple formula for MAP adaptation of the mean parameter is obtained (2):

$$\hat{m}_j = \frac{L_j}{L_j+\lambda}\bar{m}_j + \frac{\lambda}{L_j+\lambda}m_j \qquad (2)$$

where $\hat{m}_j$, $\bar{m}_j$ and m_j are the j^{th} mean adapted, the j^{th} mean of the observed adaptation data and the j^{th} mean of SI models; λ is the weight of the prior and L is the occupation likelihood of the adaptation data. If the amount of adaptation data is small, that is, the value of L_j is small compared to λ, $\hat{m}_j$ will be close to SI mean m_j. If the amount of adaptation data is large, that is, the value of L_j is big compared to λ, $\hat{m}_j$ will be close to SI mean m_j. If the amount of adaptation data is large, that is, the value of L_j is big compared to λ, $\hat{m}_j$ will be close to $\bar{m}_j$. As a result, if the user collects enough data using strategy 2, tailoring the SI models with MAP will obtain SA models that are close to SD models in performance.

Though MAP can adapt SI models very well, it needs every model has at least one adaptation sample. There are over 5000 words totally in Chinese sign language (CSL), so collecting adaptation data is a tedious job. Therefore, we must adapt SI models utilizing small amount of data. The data are collected by strategy 1.

B. Combining AP and eMAP-VFS for Adaptation

Since we need reduce the number of labeled adaptation data, we must mine the correlation among the words and select some exemplars for them. If we have not enough data to adapt all the means, we can first adapt exemplars of them, then the un-adapted means can be tailored according to the priors (from SI models) and the changes (from the adapted means). We propose the enhanced maximum a posteriori and vector field smoothing (eMAP-VFS) to adapt SI models with small amount of exemplar data, of which the exemplar data are extracted by affinity propagation (AP, [6]). The adaptation method is illustrated in Fig. 2.

We use AP to cluster the HMMs' means and detect patterns of them, and each pattern is an exemplar. A preference value and the similarity matrix whose elements are the similarity measure between pairs of two means are inputted. Then two real-valued messages between every two means are computed iteratively until a set of exemplar means with high quality emerges. This method can find exemplars with much lower error and high speed compared with other methods.

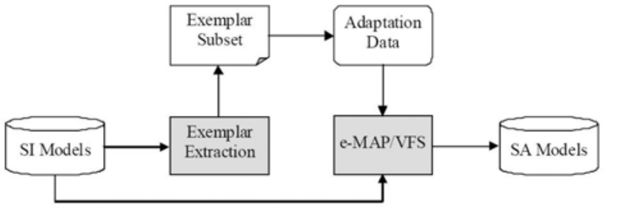


Fig. 2. Combining AP and eMAP/VFS for adaptation. First, exemplar means are extracted, and then corresponding subset is formed. Based on this subset, adaptation data are collected. SA models are generated using e-MAP/VFS.

Sign language data can be separated into 3 data streams: position and orientation of two hands (*P&O*), left hand shape (*LHS*) and right hand shape (*RHS*). These 3 data streams are almost independent from each other, and contribute differently to the recognition task. Before we compute the similarity between two means we should first weigh the 3 data streams. According to experiments implemented previously, the experiential values of $1/4 : 1/4 : 1/2$ are obtained corresponding to *P&O*, *LHS* and *RHS* respectively. As a result, the similarity $s(i, j)$ between two means m_i and m_j can be obtained.

AP clusters the means through 3 steps, depicted in (3)-(5):

$$res\,(i,p) = s\,(i,p) - \max_{q \neq p} \{ava\,(i,q) + s\,(i,q)\} \tag{3}$$

$$ava\,(i,p) = \min\left\{0, res\,(p,p) + \sum\nolimits_{q \neq i,p} \max\{0, res\,(q,p)\}\right\} \tag{4}$$

$$ava\,(p,p) = \sum\nolimits_{q \neq p} \max\{0, res\,(q,p)\} \tag{5}$$

where *res*(i, p), denoting responsibility, is the message sent from m_i to m_p; *ava*(i, p), denoting availability, is the message sent from m_p to m_i; *ava*(p, p), denoting self-availability. AP begins with all availabilities initialized to zero, and $s(p,$ $p)$ is set to the preference value that m_p be chosen as an exemplar. Then the messages are passed between all mean-pairs. Algorithm will terminate after the fixed number of iterations that the messages' changes fall below a threshold, or after the local decisions stay constant for some number of iterations [6].

The outputs of AP are the exemplar means. These exemplar means can represent a specified signer's characters to some extent. Each exemplar mean belongs to an HMM, that is, corresponds to a word. We collect one sample of an word for adaptation if the word's HMM includes one exemplar mean at least. Considering that the different exemplar means may belong to the same HMM, the number of word included in the adaptation data may be smaller than the number of exemplar means. This can also reduce the amount of adaptation data.

The adaptation data do not cover all HMM models. Supposing there are Ns samples in the adaptation data set, the corresponding Ns HMMs can be adapted using MAP. The unadapted HMMs must be estimated by utilizing the prior information available (from the SI models), the adapted HMMs, and the correlation among HMMs.

Supposing $\Omega_s$denotes the HMM means set that each of them has adaptation data available and Ω_u denotes the HMM means set that each of them has no adaptation data, the transfer vectors of means in Ω_s can be obtained by subtracting SI means from their corresponding adapted means. We assume that all the transfer vectors of $\Omega_s \cup \Omega_u$ form a smoothing vector field. So the transfer vectors of means in Ω_u can be estimated using the transfer vectors of their neighbors in Ω_s.

For each m_i in Ω_u, the estimation using MAP/VFS [7] is obtained by (6):

$$\tilde{m}_i = m_i + \frac{\sum_{m_j \in N_i^K} \omega_{ij} \nabla m_j}{\sum_{m_j \in N_i^K} \omega_{ij}} \tag{6}$$

where N_i^K is a subset ofΩ_s, represents m_i's K nearest neighbors in Ω_s; ∇m_j represents the transfer vector of m_j which equals to $(\hat{m}_j - m_j)$, where $\hat{m}_j$is defined the same as in (2); ω_{ij} is the weight of m_j to m_i, and equals to $s(i,\ j)$, which indicates that the more similar m_j is with m_i, the more information it supplies to m_i.

The estimated mean $\tilde{m}_i$ is equal to the sum of its initial value m_i and the estimated transfer vector, which is obtained by interpolation by weighing its nearest neighbors' transfer vectors. Selecting the number of K is a problem that must be solved. The neighborhood relationship may change after MAP/VFS adaptation. Previously the neighbors are obtained in the SI means space only, not considering the SA means space. Although the SA means space is not known completely, it still can supply some information. We therefore propose the e-MAP/VFS to solve this problem.

Fig. 3 shows the process of e-MAP/VFS. In (a), the target mean m_i selects 3 neighbors in the SI means space. Then (b) shows that the estimated value for m_i can be obtained with its neighbors using MAP/VFS. At the SA means space, the estimated $\tilde{m}_i$ finds the same number of neighbors. As can be seen in (c), a new neighbor appears, which means that it is still informational to m_i though it is not very close to m_i in the SI means space. (d) shows that since a new exemplar mean appears, it should be added to the neighbors at the SI means space also. The procedure iterates until the number of maximum iteration arrives or the neighbors do not change for some number of iterations.

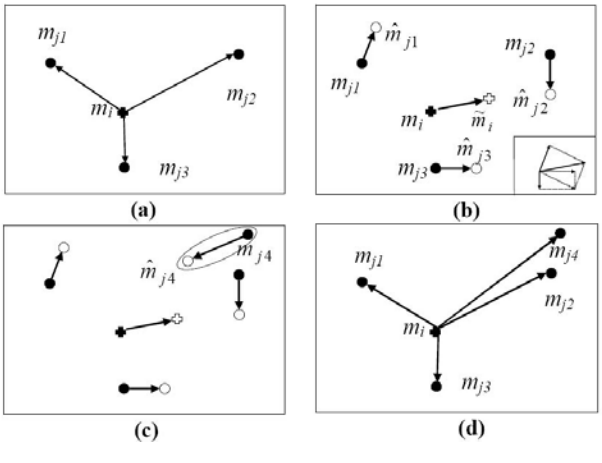


Fig. 3. The e-MAP/VFS, nearest neighbors are iteratively refined.

Though the two supervised adaptation methods can tailor the SI models to SA models, they both need adaptation data to be labeled. On the other side huge amount of unlabeled data can be collected when the user manipulate the hand-talking system. Next Section we describe the unsupervised adaptation method.

IV. Unsupervised Adaptation

Large amount of unlabeled data can be accumulated when users manipulate the hand-talking system. If these data can be used for model adaptation, the adaptation can be executed implicitly at the time users do not use the system.

The class of the unlabeled data are not known, so we can not directly use them to adapt the models. Labeling the unlabeled data accurately must be done first. Self-teaching is an alternative, that is, labeling the unlabeled data with SI models, retraining the models with the ”labeled” data. By self-teaching the noise rate of the adaptation data set may be very high, which will affect the subsequent adaptation process. We must decrease the labeling error rate further.

By analyzing the principle of HMMs we find that HMMs has the conditional independence assumption [8], that is, it assumes that the observations belonging to the same state are independent and identical distributed (i.i.d.). This assumption may be appropriate sometimes, but not always appropriate. Fig. 4 showed two signals, one of which represents a CSL word. The left signal has 5 explicit states, thus can be modeled well using a 5-state HMM. The right signal has no explicit state, and can not be well modeled by any HMM. An alternative to HMM is the polynomial segment model (PSM) [8].

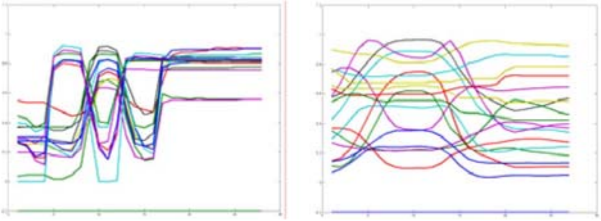


Fig. 4. Two signals representing two types of CSL words. The left signal has 5 states explicitly, the right signal has no explicit states.

A PSM can be depicted as a triplet: $\{B, \sum, N\}$, as shown in Fig. 5. Given a CSL signal $O = \{o_1, o_2, \ldots, o_N\}$, its corresponding PSM is:

$$O = Z_N B + E \tag{7}$$

where O is an $N \times D$observation matrix containing N frames of D dimensional feature vectors; B is a $(R{+}1)\times D$ parameter matrix of a R^{th} order trajectory model and E is the residual matrix the same size as O; Z_N is $N \times (R+1)$ design matrix for an R^{th} order trajectory model that normalizes the samples of different frames to [0, 1].

Since HMM and PSM can model two types of signals respectively, we can combine them to label the unlabeled data such that the labeling right rate can be improved. The process of combining HMM and PSM to label unlabeled data are shown in Fig. 6. After recognizing the unlabeled data with HMMs and PSMs, two candidate sequences are obtained. Then we compute the maximum and minimum of the two sequences respectively, and use them to normalize the two sequences to [0, 1]. After summing the two normalized sequences according to the word order, we obtained the combined likelihood. Using the combined likelihood to label the unlabeled data, the labeling right rate can be improved. The ”labeled” data together with the SI models are inputted to MAP, then SA models are obtained.

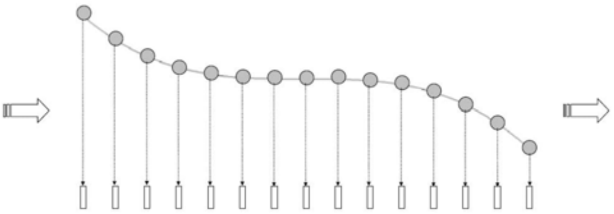


Fig. 5. Illustration for PSM. The observation sequence is assumed to be generated by a state sequence with polynomial curve type.

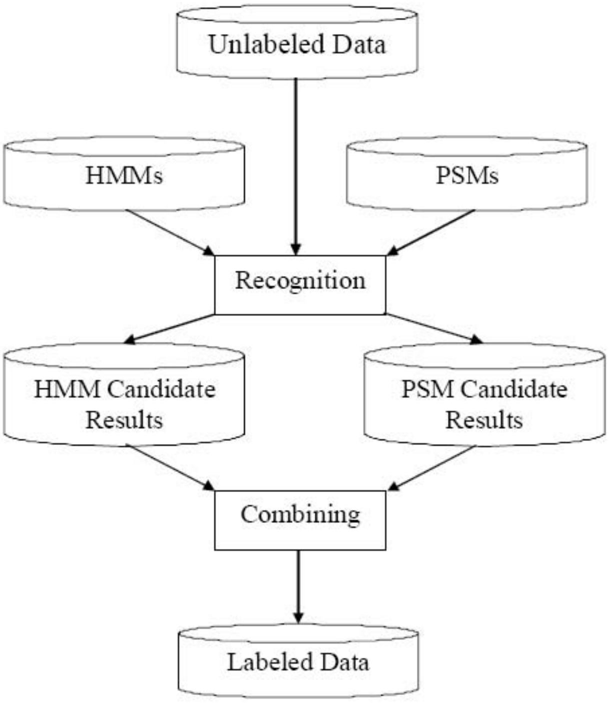


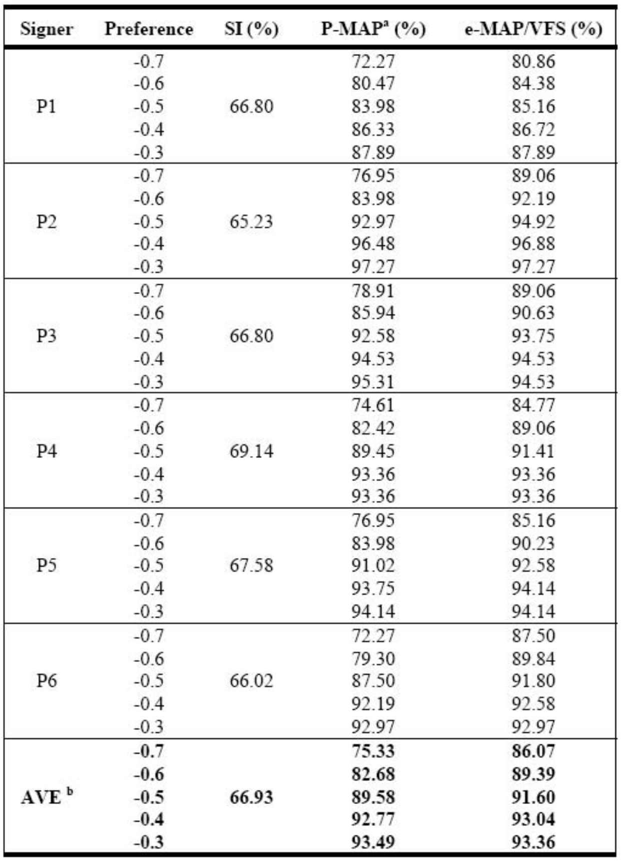
Fig. 6. Combine HMM and PSM to accurately label the unlabeled data.

V. Experiments

A. Supervised Adaptation

Two data gloves and a 3D-SPACE position trackers (with a transmitter and three receivers) were used as data input devices in the hand-talking system. The data gloves collect the information of hand shape with a 18-dimensional data for each hand, and position tracker collects the information of relative

TABLE I
EXPERIMENTAL RESULTS OF OUR METHOD.



Signer	Preference	SI (%)	P-MAP^a (%)	e-MAP/VFS (%)
P1	-0.7	66.80	72.27	80.86
	-0.6		80.47	84.38
	-0.5		83.98	85.16
	-0.4		86.33	86.72
	-0.3		87.89	87.89
P2	-0.7	65.23	76.95	89.06
	-0.6		83.98	92.19
	-0.5		92.97	94.92
	-0.4		96.48	96.88
	-0.3		97.27	97.27
P3	-0.7	66.80	78.91	89.06
	-0.6		85.94	90.63
	-0.5		92.58	93.75
	-0.4		94.53	94.53
	-0.3		95.31	94.53
P4	-0.7	69.14	74.61	84.77
	-0.6		82.42	89.06
	-0.5		89.45	91.41
	-0.4		93.36	93.36
	-0.3		93.36	93.36
P5	-0.7	67.58	76.95	85.16
	-0.6		83.98	90.23
	-0.5		91.02	92.58
	-0.4		93.75	94.14
	-0.3		94.14	94.14
P6	-0.7	66.02	72.27	87.50
	-0.6		79.30	89.84
	-0.5		87.50	91.80
	-0.4		92.19	92.58
	-0.3		92.97	92.97
AVE ^{b}	**-0.7**	**66.93**	**75.33**	**86.07**
	-0.6		**82.68**	**89.39**
	-0.5		**89.58**	**91.60**
	-0.4		**92.77**	**93.04**
	-0.3		**93.49**	**93.36**

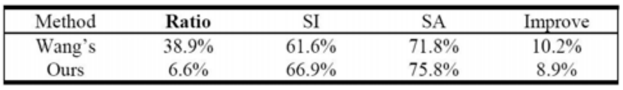
hand location and orientation with a 6-dimensional data for each hand.

Experimental data consist of 6144 samples over 256 words with each word having 24 samples (including 6 signers, each signer with 4 samples). One signer was selected as the test signer, and its first and fourth samples of all the words were used as adaptation data set and test data set respectively. The data of other 5 signers were used as the training data set. Cross validation was conducted on 6 signers. The models were 3-state Bakis HMMs, and each state's observation distribution was unimodal multivariate Gaussian.

The experimental results were summarized in TABLE I. We can see that the recognition accuracy improvements between SI and P-MAP become larger with the increase of preference value. This is because that when we increased the preference value, the number of exemplars became larger, and the corresponding number of adaptation data became larger. When the preference value was -0.3, almost every HMM had an adaptation sample, which meant that every HMM had been adapted by MAP. When the preference value was -0.7, the number of adaptation samples was about 180, which meant that about 76 HMMs had no adaptation data and their parameters were not adapted; However, if we use eMAP/VFS to estimate the un-adapted means, the recognition accuracy can be improved over 10% compared to P-MAP.

To verify the effectiveness of our method we also compared our method with Wang's method [9], and the result was illustrated in TABLE II. Using Wang's method the ratio between the size of adaptation vocabulary and that of whole vocabulary was 38.9%, and the improvement of recognition accuracy was 10.2%. By our method only 6.6% ratio was needed when the improvement was 8.9%. This meant that our method can adapt the models more quickly compared with Wang's method.

TABLE II
COMPARISON BETWEEN WANG'S AND OURS.



Method	**Ratio**	SI	SA	Improve
Wang's	38.9%	61.6%	71.8%	10.2%
Ours	6.6%	66.9%	75.8%	8.9%

B. Unsupervised Adaptation

To verify the unsupervised adaptation method, the test person's 4th group data were selected as test data, the other 3 groups data were selected as unlabeled adaptation data after we hid their labels. The experimental results were shown in TABLE III.

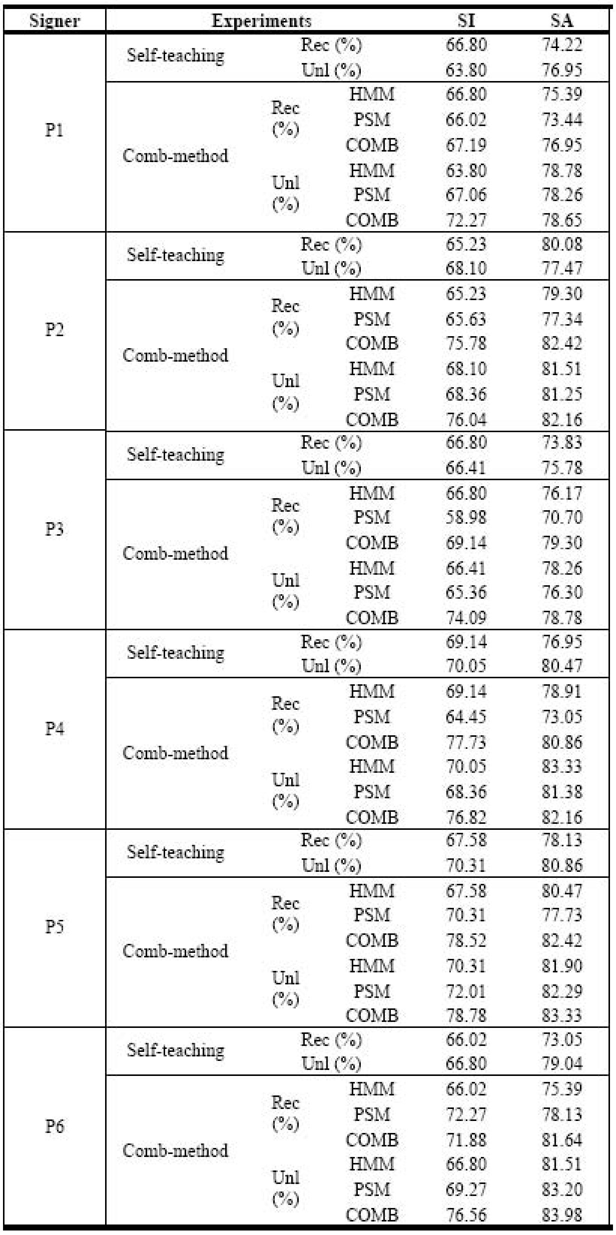
Rec (%) represented the recognition rate for test data, and Unl (%) represented the labeling right rate for unlabeled data. Self-teaching was selected as the compared method. From TABLE III, we can see that after adopting PSMs to help HMMs to label the unlabeled data, both the labeling right rate and the recognition rate were improved. This was because that by using combining method the right candidate's likelihood was taken ahead, as a result the labeling right rate was improved a lot. Thus a more 'clean' adaptation data set was obtained, and the more "clean" adaptation data set led to better adaptation performance. These results showed that introducing PSMs to the unsupervised framework was helpful.

VI. CONCLUSION

The hand-talking system suffers from the problem of drastic decrease in performance when new users are unregistered in the training set. This paper presents an adaptive framework to solve this problem. Three data collecting strategies are proposed, and supervised and unsupervised adaptation methods are implemented using these data. Experimental results show that the proposed framework can modify SI models to SA models , which can better recognize new users' data. The evaluation result shows great prospects of applying adaptation techniques to improve the hand-talking system's robustness.

Though the framework solves the problem to some extent, the amount of labeled data required are still large, which affects the applicability of the system. Utilizing unlabeled data to adapt SI models more effectively is a problem to be explored. In the adaptation problem, SI models can serve as the weak classifier, and large amount of unlabeled data exist, which supply two prerequisites for the semi-supervised learning [10] framework. Our future works will focus on utilizing the semi-supervised learning to solve the unsupervised adaptation problem.

TABLE III
UNSUPERVISED ADAPTATION RESULTS.



Signer	Experiments			SI	SA
P1	Self-teaching	Rec (%)		66.80	74.22
		Unl (%)		63.80	76.95
	Comb-method	Rec (%)	HMM	66.80	75.39
			PSM	66.02	73.44
			COMB	67.19	76.95
		Unl (%)	HMM	63.80	78.78
			PSM	67.06	78.26
			COMB	72.27	78.65
P2	Self-teaching	Rec (%)		65.23	80.08
		Unl (%)		68.10	77.47
	Comb-method	Rec (%)	HMM	65.23	79.30
			PSM	65.63	77.34
			COMB	75.78	82.42
		Unl (%)	HMM	68.10	81.51
			PSM	68.36	81.25
			COMB	76.04	82.16
P3	Self-teaching	Rec (%)		66.80	73.83
		Unl (%)		66.41	75.78
	Comb-method	Rec (%)	HMM	66.80	76.17
			PSM	58.98	70.70
			COMB	69.14	79.30
		Unl (%)	HMM	66.41	78.26
			PSM	65.36	76.30
			COMB	74.09	78.78
P4	Self-teaching	Rec (%)		69.14	76.95
		Unl (%)		70.05	80.47
	Comb-method	Rec (%)	HMM	69.14	78.91
			PSM	64.45	73.05
			COMB	77.73	80.86
		Unl (%)	HMM	70.05	83.33
			PSM	68.36	81.38
			COMB	76.82	82.16
P5	Self-teaching	Rec (%)		67.58	78.13
		Unl (%)		70.31	80.86
	Comb-method	Rec (%)	HMM	67.58	80.47
			PSM	70.31	77.73
			COMB	78.52	82.42
		Unl (%)	HMM	70.31	81.90
			PSM	72.01	82.29
			COMB	78.78	83.33
P6	Self-teaching	Rec (%)		66.02	73.05
		Unl (%)		66.80	79.04
	Comb-method	Rec (%)	HMM	66.02	75.39
			PSM	72.27	78.13
			COMB	71.88	81.64
		Unl (%)	HMM	66.80	81.51
			PSM	69.27	83.20
			COMB	76.56	83.98

REFERENCES

[1] I. C. Yoo and D. Yook, "*Automatic sound recognition for the hearing impaired*," IEEE Trans. Consumer Electron., vol. 54, no. 4, pp. 2029-2036, Nov. 2008.

[2] http://www.cdpf.com.cn/ggtz/content/2008-05/04/content_25053452.htm (in Chinese)

[3] W. Gao, Y. Chen, G. Fang, C. Yang, D. Jiang, and et al., "*HandTalker II: a Chinese sign language recognition and synthesis system*," in Proc. The 8th Int. Conf. on Control, Automation, Robotics and Vision (ICARCV 2004), pp. 759-764, 2004.

[4] L. R. Rabiner, "*A tutorial on hidden Markov models and selected applications in speech recognition*," Proceedings of the IEEE, vol. 77, no.2, pp. 257-286, 1989.

[5] J. L. Gauvain and C. H. Lee, "*Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains*," IEEE Trans. Speech Audio Process., vol. 2, no. 2, pp. 291-298, Apr. 1994.

[6] B. J. Frey and D. Dueck, "*Clustering by passing messages between data points*," Science, vol. 315, no. 5814, pp. 972-976, 2007.

[7] J. Takahashi and S. Sagayama, "*Vector-field-smoothed Bayesian learning for incremental speaker adaptation*," in International Conference on Acoustics, Speech, and Signal Processing, pp. 696-699, 1995.

[8] C. F. Li, M. H. Siu, and J. S. K. Au-Yeung, "*Recursive likelihood evaluation and fast search algorithm for polynomial segment model with application to speech recognition*," IEEE Trans. Audio Speech Lang. Process., vol. 14, no.5, pp. 1704-1718, 2006.

[9] C. Wang, X. Chen, and W. Gao, "*Generating data for signer adaptation*," in Proc. Gesture Workshop, pp. 114-121, 2007.

[10] X. Zhu, "*Semi-supervised learning literature survey*," Computer Sciences Technical Reports 1530, University of Wisconsin Madison, 2008.

Zhou Yu received his B.S. degree in Department of Computer Science and Technology, Harbin Institute of Technology (HIT), China, in 2001. And now, he is working for his PhD degree in Department of Computer Science and Technology, HIT. His research interests include pattern recognition, machine learning, and multi-modal human-computer interaction.

Jiang Feng received his PhD degree in Department of Computer Science and Technology, HIT, China, in 2008. And now, he is a lecture in Department of Computer Science and Technology, HIT. His research interests include pattern recognition, machine learning, and multi-modal human-computer interaction.