An Automatic Lipreading System for Spoken Digits With Limited Training Data

An Automatic Lipreading System for Spoken Digits With Limited Training Data

S. L. Wang, A. W. C. Liew, W. H. Lau, and S. H. Leung

***Abstract*—It is well known that visual cues of lip movement contain important speech relevant information. This paper presents an automatic lipreading system for small vocabulary speech recognition tasks. Using the lip segmentation and modeling techniques we developed earlier, we obtain a visual feature vector composed of outer and inner mouth features from the lip image sequence for recognition. A spline representation is employed to transform the discrete-time sampled features from the video frames into the continuous domain. The spline coefficients in the same word class are constrained to have similar expression and are estimated from the training data by the EM algorithm. For the multiple-speaker/speaker-independent recognition task, an adaptive multimodel approach is proposed to handle the variations caused by various talking styles. After building the appropriate word models from the spline coefficients, a maximum likelihood classification approach is taken for the recognition. Lip image sequences of English digits from 0 to 9 have been collected for the recognition test. Two widely used classification methods, HMM and RDA, have been adopted for comparison and the results demonstrate that the proposed algorithm deliver the best performance among these methods.**

***Index Terms*—Lipreading, visual feature extraction, visual speech recognition.**

I. Introduction

AUDIO-VISUAL speech recognition has attracted significant interest in recent years since the visual information extracted from the lip movement has been shown to improve the performance of the automatic speech recognition (ASR) system in noisy environments [1]. However, the performance of automatic lipreading, i.e., speech recognition from visual information alone, is far from satisfactory. Accurate lip segmentation and modeling, feature extraction and appropriate classifier design are the three major challenges for automatic lipreading system. Furthermore, the classifier should be effective even with limited training samples because the user will be impatient if he/she is asked to repeat the same utterance many times during training.

Manuscript received January 24, 2007; revised June 10, 2007 and September 10, 2007. First published September 16, 2008; current version published November 26, 2008. This work was supported by the Shanghai Natural Science Fund (05ZR14080) and the National Natural Science Foundation of China (No. 60702043). This paper was recommended by Associate Editor R. C. Lancini.

S. L. Wang is with the School of Information Security Engineering, Shanghai Jiaotong University, Shanghai 200030, China (e-mail: wsl@sjtu.edu.cn).

A. W. C. Liew is with the School of Information and Communication Technology, Gold Coast Campus, Griffith University, QLD 4222 Queensland, Australia (e-mail: a.liew@griffith.edu.au).

W. H. Lau and S. H. Leung are with the Department of Electronic Engineering, City University of Hong Kong, Hong Kong, China (e-mail: itwhlau@cityu.edu.hk; eeeugshl@cityu.edu.hk).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCSVT.2008.2004924

Various methods have been proposed for lip segmentation and modeling [2]–[5]. Edge-based [2] and Markov random field (MRF) based [3] approaches are the two widely used techniques. Edge-based methods can provide good results if noticeable luminance changes are present along the lip boundary. However, such requirement cannot be easily satisfied under natural condition. MRF-based methods usually incorporate local spatial information to improve the robustness of the lip segmentation. However, patches outside and holes inside the segmented lip region are generally found. With the fuzzy c-means with shape function (FCMS) algorithm proposed in [4], the lip region can be robustly segmented and accurate lip contour extraction can now be achieved in an efficient manner.

The hidden Markov model (HMM) [7] has been widely used in audio speech recognition and in audio-visual data fusion. However, HMM is a fairly complex statistical model and requires large amount of training data to optimize the model parameters. In our approach, we build a statistical word model based on the training data and perform classification using the maximum likelihood (ML) principle. A spline curve is adopted to transform each discrete feature sequence to the continuous domain. The spline curves of the same word class are assumed to have a common mean and a zero-mean random component and the EM algorithm is used to derive these quantities. The final classification is performed by selecting the word model that gives the largest ML probability. The proposed classification algorithm has the following advantages: 1) the spline representation is more appropriate in depicting the dynamic information of the continuous lip movements than the discrete observation sequence. It also makes the feature invariant to the varying length of the discrete observation sequence; 2) the proposed algorithm adopts a relatively simple classification model and is suitable for situation where limited training samples are available; 3) the spline coefficients in our algorithm are estimated from all training samples within a word class to reduce the overfitting effect.

For multispeaker/speaker-independent recognition, large within-class variations due to differences in speaking style are inevitable. In this paper, we demonstrate the use of an adaptive multimodel approach to improve the recognition performance. The training data of each digit class is divided into a number of groups and models are trained for each group. The optimum number of models is searched iteratively and the final decision is made by selecting the model with the largest probability among all the word models.

II. Visual Feature Extraction

Accurate segmentation of lip region is difficult especially for lip images with low color contrast. In our previous work [4], [6], a fuzzy clustering based algorithm is designed to differentiate lip pixels from the background ones. Then a 14-point ASM model

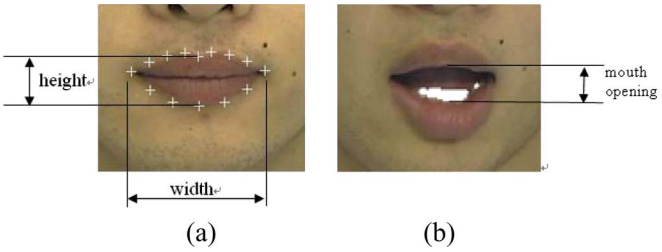

Fig. 1. (a) 14-point ASM lip model, lip width and height. (b) Inner mouth opening and teeth region (shown in white).

[5] is adopted to describe the lip contour. With the lip model, the following three kinds of visual features are extracted.

First, the width and height of the lip can be obtained directly from our 14-point lip model as shown in Fig. 1(a). In order to remove the variations caused by varying camera distance, they are normalized with respect to the values of the first image in the image sequence and are denoted as $\{w_{\text{norm}}, h_{\text{norm}}\}$.

Second, as the ASM lip model is adopted to describe the lip contour, the shape variation is contained in the eigenvalues [5]. Since the first few eigenvectors corresponding to the largest eigenvalues are sufficient for modeling the shape variation, only the first three weight values denoted by $\mathbf{b_0} = \{b_1, b_2, b_3\}$ are incorporated into the visual features vector.

Third, inner mouth features such as the teeth appearance have also been shown to contain speech relevant information [3]. Having segmented the lip region, the method described in [11] can be modified to detect the teeth pixels. The total area of the teeth region t_{area} is normalized against the entire mouth area enclosed by the lip contour. Fig. 1(b) shows an example of the teeth pixels detected. In addition, the inner mouth opening m_{open}, which is normalized against the lip height, is also taken into account. The visual feature set used for recognition is given by $\{w_{\text{norm}}, h_{\text{norm}}, \mathrm{b}_0, t_{\text{area}}, m_{\text{open}}\}$.

III. Spline-Based Classification Method

Although most of the visual features in our feature set are continuous in nature, their observations are made on discrete time points given by the sampling rate of the image sequence. Simple interpolation or curve fitting can be used to transform the discrete observations to the continuous domain but they are susceptible to overfitting due to variations caused by illumination.

A. Spline Representation of the Visual Features

In our approach, B-splines is used to model the discrete observations of each visual feature as a continuous curve. The control points of the splines are uniformly spaced covering the entire duration of the utterance. The spline coefficients of the same word class are constrained to have a common mean and covariance matrix which are estimated from all the observation sequences from the same word class. Hence, each observation sequence of a particular visual feature can be modeled by

$$Y_{ic} = S_{ic}(\mu_c + \gamma_{ic}) + \varepsilon_{ic}, \quad i = 1, 2, \ldots, m_c, c = 1, 2, \ldots, C \tag{1}$$

where Y_{ic} denotes the observation values for ith sequence in class c at time points $t_1, t_2 \ldots T_{n_{ic}}$ (n_{ic} is the number of sampling points for ith sequence in class c); S_{ic} is the spline basis functions evaluated at those time points; μ_c is the common mean of the spline coefficients in class c; m_c is the number of training sequences in class c; C is the number of word classes; γ_{ic} is a zero-mean random component representing the variation among all the observation sequences in class c and finally ε_{ic} is the noise terms. Note that the control points of the splines are uniformly spaced covering the entire duration of the utterance. If γ_{ic} and ε_{ic} are further assumed to be Gaussian distributed, i.e., $\gamma_{ic} \sim N(0, \Gamma_c)$, $\varepsilon_{ic} \sim N(0, \sigma^2 I)$, Y_{ic} is then also Gaussian distributed with $Y_{ic} \sim N(S_{ic}\mu_c, \Sigma_c)$ where $\Sigma_c = \sigma^2 I + S_{ic}\Gamma_c S_{ic}^T$. Hence, the joint distribution of the observed curves can be written as

$$\prod_{c=1}^{C}\prod_{i=1}^{m_c} \frac{1}{(2\pi)^{n_i/2}|\Sigma_c|^{1/2}} \exp \times \left(-\frac{1}{2}(Y_{ic} - S_{ic}\mu_c)^T \Sigma_c^{-1}(Y_{ic} - S_{ic}\mu_c)\right). \tag{2}$$

Directly estimating μ_c, σ and Γ_c by maximizing the joint probability in (2) is a difficult nonconvex optimization problem [9]. However, if the γ_{ic} values are available, the joint probability in (2) can be simplified to

$$\prod_{c=1}^{C}\prod_{i=1}^{m_c} \frac{1}{(2\pi)^{n_i/2}\sigma^{n_i}} \times \exp\left(-\frac{1}{2\sigma^2}\left(Y_{ic} - S_{ic}(\mu_c + \gamma_{ic})\right)^T \left(Y_{ic} - S_{ic}(\mu_c + \gamma_{ic})\right)\right) \times \frac{1}{(2\pi)^{k/2}|\Gamma_c|^{1/2}} \exp\left(-\frac{1}{2}\gamma_{ic}^T \Gamma_c^{-1}\gamma_{ic}\right) \tag{3}$$

where k is the dimension of the spline coefficients. The EM algorithm can be employed to find the optimum model parameters which maximize the likelihood in (3). Readers can refer to [9] for the derivation of the parameters by EM.

B. Speaker Dependent Classification Method

With the observation sequence of the jth feature in class c expressed in (2), with mean μ_{jc} and covariance matrix Σ_{jc}, the classification algorithm for speaker dependent tests runs as follows:

In the training step, the model parameters μ_{jc} and Γ_{jc} are estimated using the observation values obtained from the training data for each visual feature v_j and each class c by the EM algorithm. The variance of the noise term σ is also calculated.

In the classification step, a ML decision criterion is adopted. Let Y_{v_j} denote the observation sequence of the jth visual feature and f be the number of visual features. The probability of the testing sequence belonging to class c can be formulated as $P_c = \prod_{j=1}^{f} P(Y_{v_j}|\mu_{jc}, \sigma, \Gamma_{jc})^{w_j}$, where w_j is the weight factor of feature v_j and is set to be the same value ($w_j = 1/f$) in our experiments. We note that this value can be adjusted according to the significance of the individual feature. The correct word model for the testing sequence is classified to be the one with maximum likelihood probability.

C. Multiple-Speaker Classification Method

For multiple-speaker tests, large variation caused by individual's speaking style is inevitable. To solve this problem, an adaptive multimodel approach is proposed. The training samples for each digit from various speakers are divided into several groups. A model is then trained for each group that captures a certain speaking habit for that digit. Least-squared fitting is used to estimate the spline coefficients of the training sequences and the crisp k-means clustering is used for grouping. In order to determine the appropriate number of models for each digit, the Dunn's index [10] is used as the clustering validity function. Details of our approach are as follows.

1) For all sequences in the training set that correspond to a digit, estimate the spline coefficients of the visual features for each sequence by least-squared fitting. Each sequence is then represented by a spline coefficient vector. Initialize the number of groups with $M = 1$.
2) Partition the spline coefficient vectors into M groups using the k-means clustering algorithm.
3) Compute the Dunn's index $v_{\mathrm{Dunn},M}$ for M groups. Repeat step (ii) with $M = M + 1$ until $v_{\mathrm{Dunn},M} \leq v_{\mathrm{Dunn},M-1}$ or any group contains less than 3 components.
4) The appropriate number of groups is set to $M - 1$. Unlike the single speaker case, $M - 1$ models have been trained for the same digit based on the group partition.
5) The testing sequence is classified to digit c if the output probability of any model belonging to digit c is the highest among all the models.

IV. Experimental Results and Discussions

To evaluate the performance of the proposed automatic lipreading algorithm, a database consisting of 10 isolated English digits (0 to 9) uttered by eight different individuals, 6 males and 2 females (without make-ups), has been established. All speakers were asked to repeat each digit twenty times under natural lighting condition. Each digit was recorded for 1 second and contained 30 color lip images. All lip images are of frontal view and of size 110 by 90. There are about 50,000 lip images in our database. Normalization of the feature sequences to a common time frame is then performed. The starting (ending) point is assigned to the earliest (latest) frame that has significant changes with respect to the first (last) frame in the feature sequence. Using the start and end points, the normalized time values of the intermediate points can be calculated.

A. Speaker Dependent Tests

We compare the recognition accuracy of our method with HMMs and regularized discriminant analysis (RDA) [8]. For the HMM experiments, the performances for various number of Gaussian mixtures (1, 2, 3 and 4) and number of states (5, 6, 7 and 8) have been investigated. The Baum–Welch algorithm using the ML criterion has been used for training and the Viterbi algorithm for recognition. Experimental results show that the left to right HMM with six states, two continuous density Gaussian mixtures with diagonal covariance matrix associated with each state delivers the best performance. These HMM settings have been adopted for comparison.

TABLE I
Speaker Dependent Recognition Accuracy in % Obtained by Different Classification Methods With Various Parameter Settings for the Visual Feature Set $\{w_{\mathrm{norm}}, h_{\mathrm{norm}}\}$: (a) by the Proposed Method; (b) by RDA and HMM

Training samples	p	q					
		0	1	2	3	4	5
5	3	83.96	85.19	84.77	84.06	83.06	82.58
	4	85.39	**85.03**	84.08	82.52	82.27	80.12
10	3	88.44	90.77	91.11	90.82	90.82	90.99
	4	90.76	**91.49**	91.11	91.09	90.70	89.79

(a)

Training samples	Classifier	Feature set	Recognition rate
5	RDA	w_{norm}, h_{norm}	76.28
10	RDA	w_{norm}, h_{norm}	77.41
10	HMM	w_{norm}, h_{norm}	68.12
10	HMM	w_{norm}, h_{norm} and their first order derivatives	81.27

(b)

TABLE II
Speaker Dependent Recognition Accuracy in % Obtained By Different Classification Methods With Various Parameter Settings for the Visual Feature Set $\{w_{\mathrm{norm}}, h_{\mathrm{norm}}, \mathrm{b}_0\}$: (a) by the Proposed Method; (b) by RDA and HMM

Training samples	p	q					
		0	1	2	3	4	5
5	3	91.45	92.35	91.58	90.97	90.49	89.90
	4	92.34	**91.98**	90.49	89.88	89.00	87.65
10	3	94.46	96.00	95.49	95.04	94.99	94.73
	4	95.88	**95.58**	95.27	94.81	94.62	94.01

(a)

Training samples	Classifier	Feature set	Recognition rate
5	RDA	w_{norm}, h_{norm}, b_0	86.30
10	RDA	w_{norm}, h_{norm}, b_0	88.32
10	HMM	w_{norm}, h_{norm}, b_0	83.86
10	HMM	w_{norm}, h_{norm}, b_0 and their first order derivatives	91.27

(b)

Tables I, II and III show the recognition results with different visual feature set, where p denotes the order of the B-spline and $\mathrm{q} = \mathrm{number\ of\ knots} - \mathrm{p}$. For each test, n (5 and 10) sequences of each digit are randomly selected to form the training set and the rest (15 and 10) of the sequences are used for testing. For each parameter setting, 100 tests are performed and the mean recognition rate is taken as the recognition accuracy.

We observe that: First, the recognition accuracy increases when more features are used for all three methods. This demonstrates that the shape parameters b_0 and the inner mouth features $\{t_{\mathrm{area}}, m_{\mathrm{open}}\}$ are useful features that provide additional discrimination information.

Second, for each individual feature set, the proposed algorithm always outperforms the other two methods, especially for

TABLE III
SPEAKER DEPENDENT RECOGNITION ACCURACY IN % OBTAINED BY DIFFERENT CLASSIFICATION METHODS WITH VARIOUS SETTINGS FOR THE ENTIRE VISUAL FEATURE SET $\{w_{\mathrm{norm}}, h_{\mathrm{norm}}, \mathrm{b}_0, t_{\mathrm{area}}, m_{\mathrm{open}}\}$: (a) BY THE PROPOSED METHOD; (b) BY RDA AND HMM

Training samples	p	q					
		0	1	2	3	4	5
5	3	93.52	94.19	94.53	93.39	92.49	90.81
	4	94.28	**94.50**	93.17	92.63	91.05	88.91
10	3	95.89	96.11	96.54	96.35	96.59	96.18
	4	96.62	**96.69**	96.32	96.56	96.06	95.02

(a)

Training samples	Classifier	Feature set	Recognition rate
5	RDA	Entire feature set	88.77
10	RDA	Entire feature set	90.01
10	HMM	Entire feature set and its first order derivatives	93.31

(b)

TABLE IV
MULTIPLE-SPEAKER RECOGNITION ACCURACY IN % OBTAINED BY DIFFERENT CLASSIFICATION METHODS WITH VARIOUS PARAMETER SETTINGS FOR THE VISUAL FEATURE SET $\{w_{\mathrm{norm}}, h_{\mathrm{norm}}, \mathrm{b}_0, t_{\mathrm{area}}, m_{\mathrm{open}}\}$: (a) BY THE PROPOSED METHOD WITH SINGLE MODEL; (b) BY THE ADAPTIVE MULTI-MODEL APPROACH; (c) BY RDA AND HMM

p	q					
	0	1	2	3	4	5
3	60.37	68.60	73.75	77.09	78.82	80.65
4	68.61	73.30	77.10	**78.69**	80.44	81.40

(a)

p	q					
	0	1	2	3	4	5
3	86.07	87.60	86.54	87.80	88.75	87.36
4	86.47	87.07	86.55	**88.48**	88.02	88.06

(b)

Classifier	Recognition accuracy
RDA	62.37
HMM	75.87

(c)

the feature set containing the width and height only. The superior performance of the proposed algorithm can be explained as follows. 1) The speech-relevant information is contained in both the static lip shapes and the dynamic lip movements. For HMM, the input is the discrete observation sequence (and possibly the first order derivatives, i.e., $delta_i = feature_i - feature_{i-1}\, i > 1$ and $delta_1 = 0$) of the visual features. For RDA, the dynamic relationship among the observations of a sequence is described by correlation. Nevertheless, only the discrete observations and their low-order relationship are considered in these two methods. In contrast, our algorithm uses a spline curve to represent the static and dynamic information of the visual feature so that higher order relationship is embedded in the spline coefficients. Consequently, the spline representation better captures the dynamic information of the lip sequence than the discrete feature representations in the other two approaches. 2) The control points of the spline curve are estimated from all the training samples, which helps reduce overfitting. In addition, the statistical model adopted in the proposed algorithm is relatively simple compared to that of HMM and can produce reliable results with only 5 training samples. In contrast, some HMM models cannot even be initialized with such a small number of training samples.

Third, it can be seen from the results of the proposed algorithm that the recognition accuracy varies with different spline order and number of knots. Generally speaking, a small number of knots ($q = 1$ for example) achieves better performance than when the number of knots is large ($q = 5$ for example). This is because the variation of the features is relatively small and a small number of knots is sufficient to describe the observation sequence for the speaker dependent tests. Using more knots would result in overfitting the observation sequence with speech irrelevant information such as illumination variation and time-misalignment, which confuses the recognition task. The cubic spline with $p = 4$ and $q = 1$ is selected for its superior performance. It is further observed that many other choices of p and q can also provide similarly good results, and the recognition performance is insensitive to the parameter setting.

B. Multiple-Speaker/Speaker-Independent Tests

Table IV shows the multiple-speaker recognition results. For each digit, half of the total samples are used for training, and the other half for testing (80 samples each). One hundred random tests are performed for each parameter setting and the average recognition rate is taken as the accuracy. We see that the performance of HMM is comparable to that of the proposed method with single model for each digit and is better than that of RDA. This is because large variation due to different speaking style of various speakers cannot be sufficiently modeled by the simple Gaussian distribution of RDA whereas the parameters for HMM are well estimated by the adequate training samples in this case. Experimental results also show that the spline models with higher order and more knots for single word model perform better in this case because they can better handle the large variation.

Table IV(b) shows that the accuracy for multiple-speaker recognition tests can be greatly improved by using the multi-model approach. Substantial improvement has been achieved particularly for spline with lower order and a small number of knots. For this approach, one can use more models to represent the samples if large variations are present. Each model is then focused on a particular kind of speaking style and the speech-irrelevant information are therefore discarded. We see that many different combinations of p and q can deliver satisfactory results. The cubic spline with $\mathrm{p} = 4$ and $\mathrm{q} = 3$ has been selected for its high performance and relatively low complexity.

Besides the multiple-speaker recognition tests, a speaker independent evaluation is also performed. The "leave-one-out" procedure is used, i.e., seven speakers are used to train the speaker independent model and the remaining speaker is used for testing. For each of the test speaker, we perform one hundred random tests by randomly choosing half of the samples from each of the seven speakers for training and all the samples of the test speaker for testing. The average recognition rate is then taken as the recognition accuracy.

TABLE V
SPEAKER-INDEPENDENT RECOGNITION ACCURACY IN % OBTAINED BY DIFFERENT CLASSIFICATION METHODS WITH VARIOUS PARAMETER SETTINGS FOR THE VISUAL FEATURE SET $\{w_{\text{norm}}, h_{\text{norm}}, \text{b}_0, t_{\text{area}}, m_{\text{open}}\}$: (a) BY THE PROPOSED METHOD WITH SINGLE MODEL; (b) BY THE ADAPTIVE MULTI-MODEL APPROACH; (c) BY RDA AND HMM

p	q					
	0	1	2	3	4	5
3	50.57	54.24	57.79	65.10	65.63	67.68
4	58.11	62.39	64.80	**67.63**	68.71	70.12

(a)

p	q					
	0	1	2	3	4	5
3	72.19	74.62	75.98	76.27	76.71	76.20
4	74.18	74.36	76.43	**79.64**	78.73	79.97

(b)

Classifier	Recognition accuracy
RDA	55.57
HMM	67.32

(c)

From Table V, we observe that: 1) the proposed multi-model approach achieved the best performance; and 2) HMM achieved comparable accuracy as the proposed algorithm with single model and much better than the RDA. Nevertheless, all three methods perform significantly inferior than the multiple-speaker recognition tests. The multiple-speaker model is able to capture additional information about the individual test speaker, which results in better performance. In addition, the speaker-independent model is only trained using seven speakers and is not representative enough. We expect that with more speakers included in the training, the generality of the model can be improved and the recognition accuracy would increase.

C. Discussions

One of the advantages of HMM is that it provides a nice mechanism to handle the variable input sequence length. This is particularly important in speech related applications since the same word uttered twice would likely be in different speaking rate, even by the same speaker. Our spline model has this property too since it transforms the discrete input sequence into the continuous representation, thus doing an implicit re-scaling of the time axis. By using the spline coefficients as the representation, the dependency on the input sequence length is removed.

It is obvious that the accuracy of the lipreading system relies heavily on the reliability of the extracted visual features, which in turn relies on the quality of the images. Experimental results have demonstrated that our lip extraction algorithms [4], [5] perform well without the need of stringent lighting conditions. However, a major concern is the extraction of the inner mouth features $\{t_{\text{area}}, m_{\text{open}}\}$ which can be affected by the pose and lighting condition. Fortunately, the parameters γ_{ic} and ε_{ic} in our model can account for much of the variations caused by noise, detection errors, environment changes, changes among various utterances, etc. As a result, even if the extracted visual features of a single frame is erroneous or missing, the spline coefficient will not change much and the recognition result is still of high accuracy.

In automatic speech recognition, the model adaptation techniques of MLLR [12] and MAP [13] are useful for handling the problem of small training samples. Rather than directly training the speaker dependent model from the limited training samples, these methods adapt the model using only a small amount of training data from the new speaker. However, the effectiveness of this approach relies on the representativeness of the speaker independent model to start with, which may not be easily obtained from a small number of speakers. In our future work, we plan to expand the dataset and investigate appropriate model adaptation methods for visual speech recognition.

In visual lipreading, it is important that a speaker speak with clear visual articulation since lip movement is the only information source for recognition. Clearly, cooperation from the speaker is crucial. Although our system deals with only a limited vocabulary of isolated words, it is suitable for certain specialize applications such as in an intelligent navigation system, where only a small number of commands need to be recognized.

V. CONCLUSION

In this paper, an automatic lipreading system based solely on the visual features is presented. The set of visual speech features consists of the lip width and height, the ASM shape descriptors, and the inner mouth features. A spline curve has been employed to model the static and dynamic information for each visual feature in the same word class and the spline coefficients are estimated from all the training samples within the class by the EM algorithm. An adaptive multimodel approach has also been proposed to overcome the variation caused by different speaking style for the multiple speaker/speaker independent visual speech recognition tasks. Experimental results demonstrated that the proposed algorithm provides accurate recognition results even when only limited training data is available.

REFERENCES

[1] N. P. Erber, "Interaction of audition and vision in the recognition of oral speech stimuli," *J. Speech Hearing Res.*, vol. 12, pp. 423–425, 1969.
[2] I. Matthews, T. F Cootes, J. A. Bangham, S. Cox, and R. Harvey, "Extraction of visual features for lipreading," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 2, pp. 198–213, Feb. 2002.
[3] X. Zhang, R. M. Mersereau, M. Clements, and C. C. Broun, "Visual speech feature extraction for improved speech recognition," in *Proc. IEEE Int. Conf. Acoustic, Speech, . Signal Process.*, Orlando, FL, May 2002, vol. 2, pp. 1993–1996.
[4] S. H. Leung, S. L. Wang, and W. H. Lau, "Lip image segmentation using fuzzy clustering incorporating an elliptic shape function," *IEEE Trans. Image Process.*, vol. 13, no. 1, pp. 51–62, Jan. 2004.
[5] K. L. Sum, W. H. Lau, S. H. Leung, A. W. C. Liew, and K. W. Tse, "A new optimization procedure for extracting the point-based lip contour using active shape model," in *Proc. IEEE Int. Conf. Acoustics, Speech Signal Process.*, Salt Lake City, UT, May 2001, vol. 3, pp. 1485–1488.
[6] A. W. C. Liew, S. H. Leung, and W. H. Lau, "Fuzzy image clustering incorporating spatial continuity," *Proc. IEE Vis. Image Signal Process.*, vol. 147, no. 2, pp. 185–19, Apr. 2000.
[7] J. Luettin, N. A. Thacker, and S. W. Beet, "Visual speech recognition using active shape models and hidden Markov models," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Process.*, Atlanta, GA, May 1996, vol. 2, pp. 817–820.
[8] J. H. Friedman, "Regularized discriminant analysis," *J. Amer. Stat. Assoc.*, vol. 84, no. 405, pp. 165–175, Mar. 1989.
[9] G. M. James and T. J. Hastie, "Functional linear discriminant analysis for irregularly sampled curves," *J. Roy. Stat. Soc.*, ser. B JRSS B, vol. 63, pp. 533–550, Apr. 2000.

[10] J. C. Dunn, "A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters," *J. Cybern.*, vol. 3, pp. 32–57, 1973.

[11] A. W. C. Liew, S. H. Leung, and W. H. Lau, "Segmentation of color lip images by spatial fuzzy clustering," *IEEE Trans. Fuzzy Syst.*, vol. 11, no. 4, pp. 542–549, Aug. 2003.

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

[13] J. Gauvain and C. 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.