EURASIP Journal on Applied Signal Processing 2005:9, 1382–1399

A Two-Channel Training Algorithm for Hidden Markov Model and Its Application to Lip Reading

Liang Dong
Department of Electrical and Computer Engineering, National University of Singapore, Singapore 119260
Email: engp0564@nus.edu.sg

Say Wei Foo
School of Electrical and Electronic Engineering, Nanyang Technological University, 50 Nanyang Avenue, Singapore 639798
Email: eswfoo@ntu.edu.sg

Yong Lian
Department of Electrical and Computer Engineering, National University of Singapore, Singapore 119260
Email: eleliany@nus.edu.sg

Received 1 November 2003; Revised 12 May 2004

Hidden Markov model (HMM) has been a popular mathematical approach for sequence classification such as speech recognition since 1980s. In this paper, a novel two-channel training strategy is proposed for discriminative training of HMM. For the proposed training strategy, a novel separable-distance function that measures the difference between a pair of training samples is adopted as the criterion function. The symbol emission matrix of an HMM is split into two channels: a static channel to maintain the validity of the HMM and a dynamic channel that is modified to maximize the separable distance. The parameters of the two-channel HMM are estimated by iterative application of expectation-maximization (EM) operations. As an example of the application of the novel approach, a hierarchical speaker-dependent visual speech recognition system is trained using the two-channel HMMs. Results of experiments on identifying a group of confusable visemes indicate that the proposed approach is able to increase the recognition accuracy by an average of 20% compared with the conventional HMMs that are trained with the Baum-Welch estimation.

Keywords and phrases: viseme recognition, two-channel hidden Markov model, discriminative training, separable-distance function.

1. INTRODUCTION

The focus of most automatic speech recognition techniques is on the spoken sounds alone. If the speaking environment is noise free and the recognition engine is well configured, high recognition rate is attainable for most speakers. However, in real-world environments such as office, bus station, shop, and factory, the speech captured may be greatly polluted by background noise and cross-speaker noise. Presenting such a signal to a sound-based speech recognition system, the recognition accuracy may drop dramatically. One solution to enhance the speech recognition accuracy under noisy conditions is to jointly process information from multiple modalities of speech. Automatic lip reading is one mode in which the visual aspect of speech is considered for speech recognition.

It has long been observed that the presence of visual cues such as the movement of lips, facial muscles, teeth, and tongue may enhance human speech perception. Systematic studies on lip reading have been carried out since 1950s [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17]. Sumby and Pollack [1] showed that the incorporation of visual information added an equivalent 12 dB gain in signal-to-noise ratio.

Among the various techniques for visual speech recognition, hidden Markov model (HMM) holds the greatest promise due to its capabilities in modeling and analyzing temporal processes as reported in [9, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29]. Most of the HMM-based visual speech processing systems reported take an individual word as the basic recognition unit and an HMM is trained to model it. Such an approach works well with limited vocabulary such as digit set [15, 30], a small number of *AVletters* [31], and isolated words or nonsense words [32], but it is difficult to extend the methods to large-vocabulary recognition task as a great number of word models has to be trained. One solution to this problem is to build subword models such as phoneme models. Any word that is presented to the recognition system is broken down into subwords. In this way, even if a word is not included in training the system, the system can still make a good guess on its identity.

The smallest visibly distinguishable unit of visual speech is commonly referred to as viseme [33]. Like phonemes that are the basic building blocks of sound of a language, visemes are the basic constituents for the visual representation of words. The time variation of mouth shape in speech is small compared with the corresponding variation of acoustic waveform. Some previous experiments indicate that the traditional HMM classifiers, which are trained with the Baum-Welch algorithm, are sometimes incompetent to separate mouth shapes with small difference [34]. Such small difference has prompted some researchers to regard the relationship between phonemes and visemes as many-to-one mapping. For example, although phonemes /b/, /m/, /p/ are acoustically distinguishable, the sequence of mouth shape for the three sounds are not readily distinguishable, hence the three phonemes are grouped into one viseme category. An early viseme grouping was suggested by Binnie et al. [35]. The MPEG-4 multimedia standard adopted the same viseme grouping strategy for face animation, in which fourteen viseme groups are included [36]. However, different groupings are adopted by different researchers to fulfill specific requirements [37, 38].

Motivated by the need to find an approach to differentiate visemes that are only slightly different, we propose a novel approach to improve the discriminative power of the HMM classifiers. The approach aims at amplifying the separable-distance between a pair of training samples. A two-channel HMM is developed, one channel, called the static channel, is kept fixed to maintain the validity of the probabilistic framework, and the other channel, called the dynamic channel, is modified to amplify the difference between the training pair.

A hierarchical classifier is also proposed based on the two-channel training strategy. At the top level, broad identification is performed and fine identification is subsequently carried out within the broad category identified. Experimental results indicate that the proposed classifier excels the traditional ML HMM classifier in identifying the mouth shapes.

Although the proposed method is developed for the recognition of visemes, it can also be applied to any sequence classification problem. As such, the theoretical background and the training strategy of the two-channel discriminative training method are introduced first in Sections 2, 3, and 4. This is followed by discussion of the general properties and extensions of the training strategy in Sections 5 and 6, respectively. Details of the application of the method to viseme recognition and the experimental results obtained are given in Section 7. The concluding remarks are presented in Section 8.

2. REVIEW OF HIDDEN MARKOV MODEL

Hidden Markov model is also referred to as hidden Markov process (HMP) as the latter emphasizes the stochastic process rather than the model itself. HMP was first introduced by Baum and Petrie [39] in 1966. The basic theories/properties of HMP were introduced in full generality in a series of papers by Baum and his colleagues [40, 41, 42, 43], which include the convergence of the entropy function of an HMP, the computation of the conditional probability, and the local convergence of the maximal likelihood (ML) parameter estimation of HMM. Application of HMM to speech processing took place in the mid-1970s. A phonetic speech recognition system that adopts HMM-based classifier was first developed in IBM [44, 45]. Applications of HMM for speech processing were further explored by Rabiner and Juang [46, 47].

The beauty of HMM is that it is able to reveal the underlying process of signal generation even though the properties of the signal source remain greatly unknown. Assume that $O^M = \{O_1, O_2, \ldots, O_M\}$ is the discrete set of observed symbols and $S^N = \{S_1, S_2, \ldots, S_N\}$ is the set of states; an N-state-M-symbol discrete HMM $\theta(\pi, A, B)$ consists of the following three components.

(1) The probability array of the initial state: $\pi = [\pi_i] = [P(s_1 = S_i)]_{1\times N}$, where s_1 is the first state in the state chain.
(2) The state-transition matrix: $A = [a_{ij}] = [P(s_{t+1} = S_j | s_t = S_i)]_{N\times N}$, where s_{t+1} and s_t denote the $t+1$th state and the tth state in the state chain.
(3) The symbol emission probability matrix: $B = [b_{ij}] = [P(o_t = O_j | s_t = S_i)]_{N\times M}$, where o_t is the tth observed symbol in the observation sequence.

In a K-class identification problem, assume that $x^T = (x_1, x_2, \ldots, x_T)$ is a sample of a particular class, say class d_i. The probability of occurrence of the sample x^T given the HMM $\theta(\pi, A, B)$, denoted by $P(x^T|\theta)$, is computed using either the forward or backward process and the optimal hidden-state chain is revealed using Viterbi matching [46]. Training of the HMM is the process of determining the parameters set $\theta(\pi, A, B)$ to fulfill a certain criterion function such as $P(x^T|\theta)$ or the mutual information [46, 48]. For training of the HMM, the Baum-Welch training algorithm is popularly adopted. The Baum-Welch algorithm is an ML estimation; thus the HMM so obtained, θ_{ML}, is one that maximizes the probability $P(x^T|\theta)$. Mathematically,

$$\theta_{\mathrm{ML}} = \arg\max_{\theta} \left[P(x^T|\theta)\right]. \tag{1}$$

The Baum-Welch training can be realized at a relatively high speed as the expectation-maximization (EM) estimation is adopted in the training process.

However, the parameters of the HMM are solely determined by the correct samples while the relationship between the correct samples and incorrect ones is not taken into consideration. The method, in its original form, is thus not developed for fine recognition. If another sample y^T of class d_j $(j \neq i)$ is similar to x^T, the scored probability $P(y^T|\theta)$ may be close to $P(x^T|\theta)$, and θ_{ML} may not be able to distinguish x^T and y^T. One solution to this problem is to adopt a training strategy that maximizes the mutual information $I_M(\theta, x^T)$ defined as

$$I_M(\theta, x^T) = \log P(x^T|\theta) - \log \sum_{\theta' \neq \theta} P(x^T|\theta')P(\theta'). \tag{2}$$

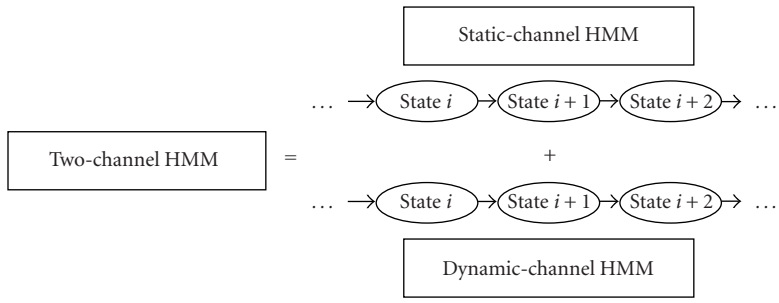


Figure 1: The block diagram of a two-channel HMM.

This method is referred to as maximum mutual information (MMI) estimation [48]. It increases the a posteriori probability of the model corresponding to the training data, and thus the overall discriminative power of the HMM obtained is guaranteed. However, analytical solutions to (2) are difficult to realize and implementation of MMI estimation is tedious. A computationally less intensive approach is desirable.

3. PRINCIPLES OF TWO-CHANNEL HMM

To improve the discriminative ability of HMM and at the same time, to facilitate the process of parameter tuning, the following two-channel training method is proposed, where the HMM is specially tailored to amplify the difference between two similar samples.

The block diagram of the two-channel HMM is given in Figure 1. It consists of a static-channel HMM and a dynamic-channel HMM. For the static channel, a normal HMM derived from a parameter-smoothed ML approach is used. A new HMM for the dynamic channel is to be derived. Details of the derivation of the dynamic-channel HMM are described in the following paragraphs.

Assume that in a two-class identification problem, $\{x^T : d_1\}$ and $\{y^T : d_2\}$ are a pair of training samples, where $x^T = (x_1^T, x_2^T, \ldots, x_T^T)$ and $y^T = (y_1^T, y_2^T, \ldots, y_T^T)$ are observation sequences of length T and d_1 and d_2 are the class labels. The observed symbols in x^T and y^T are from the symbol set O^M. $P(x^T|\theta)$ and $P(y^T|\theta)$ are the scored probabilities for x^T and y^T given HMM θ, respectively. The pair of training samples x^T and y^T must be of the same length so that their probabilities $P(x^T|\theta)$ and $P(y^T|\theta)$ can be suitably compared. Such a comparison is meaningless if the samples are of different lengths; the shorter sequence may give larger probability than the longer one even if it is not the true sample of θ.

Define a new function $I(x^T, y^T, \theta)$, called the separable-distance function, as follows:

$$I(x^T, y^T, \theta) = \log P(x^T|\theta) - \log P(y^T|\theta). \tag{3}$$

A large value of $I(x^T, y^T, \theta)$ would mean that x^T and y^T are more distinct and separable. The strategy then is to determine the HMM θ_{MSD} (MSD for maximum separable distance) that maximizes $I(x^T, y^T, \theta)$. Mathematically,

$$\theta_{\text{MSD}} = \arg\max_{\theta} \left[I(x^T, y^T, \theta) \right]. \tag{4}$$

For the proposed training strategy, the parameter set for the static-channel HMM is determined in the normal way such as the ML approach. For the dynamic-channel HMM, to maintain synchronization of the duration and transition of states, the same set of values for π and A as derived for the static-channel HMM is used; only the parameters of matrix B are adjusted.

As a first step towards the maximization of the separable-distance function $I(x^T, y^T, \theta)$, an auxiliary function $F(x^T, y^T, \theta, \lambda)$ involving $I(x^T, y^T, \theta)$ and the parameters of B is defined as

$$F(x^T, y^T, \theta, \lambda) = I(x^T, y^T, \theta) + \sum_{i=1}^{N} \lambda_i \left(1 - \sum_{j=1}^{M} b_{ij} \right), \tag{5}$$

where λ_i is the Lagrange multiplier for the ith state and $\sum_{j=1}^{M} b_{ij} = 1$ $(i = 1, 2, \ldots, N)$. By maximizing $F(x^T, y^T, \theta, \lambda)$, $I(x^T, y^T, \theta)$ is also maximized. Differentiating $F(x^T, y^T, \theta, \lambda)$ with respect to b_{ij} and setting the result to 0, we have

$$\frac{\partial \log P(x^T|\theta)}{\partial b_{ij}} - \frac{\partial \log P(y^T|\theta)}{\partial b_{ij}} = \lambda_i. \tag{6}$$

Since λ_i is positive, the optimum value obtained for $I(x^T, y^T, \theta)$ is a maximum as solutions for b_{ij} must be positive. In (6), $\log P(x^T|\theta)$ and $\log P(y^T|\theta)$ may be computed by summing up all the probabilities over time T:

$$\log P(x^T|\theta) = \sum_{\tau=1}^{T} \log \sum_{i=1}^{N} P(s_\tau^T = S_i) b_i(x_\tau^T). \tag{7}$$

Note that the state-transition coefficients a_{ij} do not appear explicitly in (7); they are included in the term $P(s_\tau^T = S_i)$.

The two partial derivatives in (6) may be evaluated separately as follows:

$$\begin{aligned}\frac{\partial \log P(x^T|\theta)}{\partial b_{ij}} &= \sum_{\substack{\tau=1\\ x_\tau^T=O_j}}^{T} P(s_\tau^T = S_i|\theta, x^T)\\ &= b_{ij}^{-1}\sum_{\tau=1}^{T} P(s_\tau^T = S_i, x_\tau^T = O_j|\theta, x^T),\\ \frac{\partial \log P(y^T|\theta)}{\partial b_{ij}} &= \sum_{\substack{\tau=1\\ y_\tau^T=O_j}}^{T} P(s_\tau^T = S_i|\theta, y^T)\\ &= b_{ij}^{-1}\sum_{\tau=1}^{T} P(s_\tau^T = S_i, y_\tau^T = O_j|\theta, y^T).\end{aligned} \tag{8}$$

By defining

$$\begin{aligned}E(S_i, O_j|\theta, x^T) &= \sum_{\tau=1}^{T} P(s_\tau^T = S_i, x_\tau^T = O_j|\theta, x^T),\\ E(S_i, O_j|\theta, y^T) &= \sum_{\tau=1}^{T} P(s_\tau^T = S_i, y_\tau^T = O_j|\theta, y^T),\\ D_{ij}(x^T, y^T, \theta) &= E(S_i, O_j|\theta, x^T) - E(S_i, O_j|\theta, y^T),\end{aligned} \tag{9}$$

equation (6) can be written as

$$\begin{aligned}&\frac{E(S_i, O_j|\theta, x^T) - E(S_i, O_j|\theta, y^T)}{b_{ij}}\\ &\quad= \frac{D_{ij}(x^T, y^T, \theta)}{b_{ij}} = \lambda_i, \quad 1 \le j \le M.\end{aligned} \tag{10}$$

By making use of the fact that $\sum_{j=1}^{M} b_{ij} = 1$, it can be shown that

$$b_{ij} = \frac{D_{ij}(x^T, y^T, \theta)}{\sum_{j=1}^{M} D_{ij}(x^T, y^T, \theta)}, \quad i = 1, 2, \ldots, N,\ j = 1, 2, \ldots, M. \tag{11}$$

The set $\{b_{ij}\}$ $(i = 1, 2, \ldots, N,\ j = 1, 2, \ldots, M)$ so obtained gives the maximum value of $I(x^T, y^T, \theta)$.

An algorithm for the computation of the values may be developed by using standard expectation-maximization (EM) technique. By considering x^T and y^T as the observed data and the state sequence $s^T = (s_1^T, s_2^T, \ldots, s_T^T)$ as the hidden or unobserved data, the estimation of $E_\theta(I) = E[I(x^T, y^T, s^T|\widetilde{\theta})|x^T, y^T, \theta]$ from incomplete data x^T and y^T is then given by [49]

$$\begin{aligned}E_\theta(I) &= \sum_{s^T\in S} I(x^T, y^T, s^T|\widetilde{\theta}) P(x^T, y^T, s^T|\theta)\\ &= \sum_{s^T\in S} [\log P(x^T, s^T|\widetilde{\theta}) - \log P(y^T, s^T|\widetilde{\theta})]\\ &\qquad \times P(x^T, y^T, s^T|\theta),\end{aligned} \tag{12}$$

where θ and $\widetilde{\theta}$ are the HMM before training and the HMM after training, respectively, and S denotes all the state combinations with length T. The purpose of the E-step of the EM estimation is to calculate $E_\theta(I)$. By using the auxiliary function $Q_x(\widetilde{\theta}, \theta)$ proposed in [48] and defined as follows:

$$Q_x(\widetilde{\theta}, \theta) = \sum_{s^T\in S} \log P(x^T, s^T|\widetilde{\theta}) P(x^T, s^T|\theta), \tag{13}$$

equation (12) can be written as

$$E_\theta(I) = Q_x(\widetilde{\theta}, \theta) P(y^T|s^T, \theta) - Q_y(\widetilde{\theta}, \theta) P(x^T|s^T, \theta). \tag{14}$$

$Q_x(\widetilde{\theta}, \theta)$ and $Q_y(\widetilde{\theta}, \theta)$ may be further analyzed by breaking up the probability $P(x^T, s^T|\widetilde{\theta})$ as follows:

$$P(x^T, s^T|\widetilde{\theta}) = \widetilde{\pi}(s_0) \prod_{\tau=1}^{T} \widetilde{a}_{s_{\tau-1}, s_\tau} \widetilde{b}_{s_\tau}(x_\tau), \tag{15}$$

where $\widetilde{\pi}$, $\widetilde{a}$, and $\widetilde{b}$ are the parameters of $\widetilde{\theta}$. Here, we assume that the initial distribution starts at $\tau = 0$ instead of $\tau = 1$ for notational convenience. The Q function then becomes

$$\begin{aligned}Q_x(\widetilde{\theta}, \theta) &= \sum_{s^T\in S} \log \widetilde{\pi}(s_0) P(x^T, s^T|\theta)\\ &\quad + \sum_{s^T\in S} \left(\sum_{\tau=1}^{T} \log \widetilde{a}_{\tau-1,\tau} \right) P(x^T, s^T|\theta)\\ &\quad + \sum_{s^T\in S} \left(\sum_{\tau=1}^{T} \log \widetilde{b}_\tau(x_\tau) \right) P(x^T, s^T|\theta).\end{aligned} \tag{16}$$

The parameters to be optimized are now separated into three independent terms.

From (14) and (16), $E_\theta(I)$ can also be divided into the following three terms:

$$E_\theta(I) = E_\theta(\widetilde{\pi}, I) + E_\theta(\widetilde{a}, I) + E_\theta(\widetilde{b}, I), \tag{17}$$

where

$$\begin{aligned}E_\theta(\widetilde{\pi}, I) &= \sum_{s^T\in S} \log \widetilde{\pi}(s_0)\\ &\qquad \times [P(x^T, y^T, s^T|\theta) - P(x^T, y^T, s^T|\theta)] = 0,\\ E_\theta(\widetilde{a}, I) &= \sum_{s^T\in S} \sum_{\tau=1}^{T} \log \widetilde{a}_{\tau-1,\tau}\\ &\qquad \times [P(x^T, y^T, s^T|\theta) - P(x^T, y^T, s^T|\theta)] = 0,\\ E_\theta(\widetilde{b}, I) &= \sum_{s^T\in S} \left[\sum_{\tau=1}^{T} \log \widetilde{b}_\tau(x_\tau) - \sum_{\tau=1}^{T} \log \widetilde{b}_\tau(y_\tau) \right]\\ &\qquad \times P(x^T, y^T, s^T|\theta).\end{aligned} \tag{18}$$

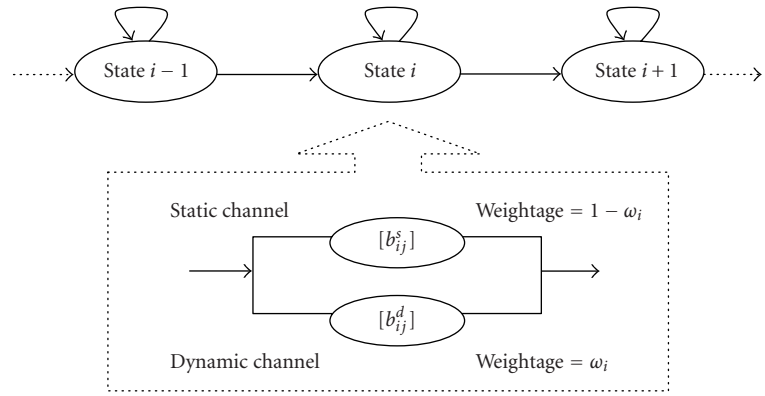


Figure 2: The two-channel structure of the ith state of a left-right HMM.

$E_\theta(\widetilde{\pi}, I)$ and $E_\theta(\widetilde{a}, I)$ are associated with the hidden-state sequence s^T. It is assumed that x^T and y^T are drawn independently and emitted from the same state sequence s^T, hence both $E_\theta(\widetilde{\pi}, I)$ and $E_\theta(\widetilde{a}, I)$ become 0. $E_\theta(\widetilde{b}, I)$, on the other hand, is related to the symbols that appear in x^T and y^T and contributes to $E_\theta(I)$. By enumerating all the state combinations, we have

$$E_\theta(\widetilde{b}, I) = \sum_{i=1}^{N} \sum_{\tau=1}^{T} [\log \widetilde{b}_i(x_\tau^T) - \log \widetilde{b}_i(y_\tau^T)] \times P(x^T, y^T, s_\tau^T = S_i|\theta). \tag{19}$$

If $\sum_{\tau=1}^{T}[\log \widetilde{b}_i(x_\tau^T) - \log \widetilde{b}_i(y_\tau^T)]$ is arranged according to the order of appearance of the symbols (O_j) within x^T and y^T, we have

$$\begin{aligned} &E_\theta(\widetilde{b}, I) \\ &= \sum_{i=1}^{N} \sum_{j=1}^{M} \log \widetilde{b}_{ij} \sum_{x_\tau^T = y_\tau^T = O_j}^{T} \\ &\times [P(x_\tau^T = O_j, s_\tau^T = S_i|\theta, x^T) - P(y_\tau^T = O_j, s_\tau^T = S_i|\theta, y^T)] \\ &\times P(x^T, y^T|\theta) \end{aligned} \tag{20}$$

or

$$\begin{aligned} &E_\theta(\widetilde{b}, I) \\ &= \sum_{i=1}^{N} \sum_{j=1}^{M} \log \widetilde{b}_{ij} [E(S_i, O_j|\theta, x^T) - E(S_i, O_j|\theta, y^T)] \\ &\quad \times P(x^T, y^T|\theta). \end{aligned} \tag{21}$$

In the M-step of the EM estimation, $\widetilde{b}_{ij}$ is adjusted to maximize $E_\theta(\widetilde{b}, I)$ or $E_\theta(I)$. Since $\sum_{j=1}^{M} \widetilde{b}_{ij} = 1$ and (21) has the form $K \sum_{j=1}^{M} w_j \log v_j$, which attains a global maximum at the point $v_j = w_j / \sum_{j=1}^{M} w_j$ $(j = 1, 2, \ldots, M)$, the re-estimated value of $\widetilde{b}_{ij}$ of $\widetilde{\theta}$ that lead to the maximum $E_\theta(I)$ is given by

$$\begin{aligned} \widetilde{b}_{ij} &= \frac{E(S_i, O_j|\theta, x^T) - E(S_i, O_j|\theta, y^T)}{\sum_{j=1}^{M} [E(S_i, O_j|\theta, x^T) - E(S_i, O_j|\theta, y^T)]} \\ &= \frac{D_{ij}(x^T, y^T, \theta)}{\sum_{j=1}^{M} D_{ij}(x^T, y^T, \theta)}. \end{aligned} \tag{22}$$

This equation, compared with (11), enables the re-estimation of the symbol emission coefficients $\widetilde{b}_{ij}$ from expectations of the existing HMM. The above derivations strictly observe the standard optimization strategy [49], where the expectation of the value of the separable-distance function, $E_\theta(I)$, is computed in the E-step and the coefficients b_{ij} are adjusted to maximize $E_\theta(I)$ in the M-step. The convergence of the method is therefore guaranteed. However, b_{ij} may not be estimated by applying (22) alone; other considerations will be taken into account such as when $D_{ij}(x^T, y^T, \theta)$ is less than or equal to 0. Further discussion on the determination of values of b_{ij} is given in the subsequent sections.

To modify the parameters according to (22) and simultaneously ensure the validity of the model, a two-channel structure as depicted in Figure 2 is proposed. The elements (b_{ij}) of matrix B of the two-channel HMM are decomposed into two parts as

$$b_{ij} = b_{ij}^s + b_{ij}^d \quad (\forall i = 1, 2, \ldots, N,\ j = 1, 2, \ldots, M), \tag{23}$$

b_{ij}^s for the static channel and b_{ij}^d for the dynamic channel. The dynamic-channel coefficients b_{ij}^d are the key source of the discriminative power. b_{ij}^s are computed using parameter-smoothed ML HMM and weighted. As long as b_{ij} computed from (22) is greater than b_{ij}^s, b_{ij}^d is determined as the difference between b_{ij} and b_{ij}^s according to (23); otherwise b_{ij}^d is set to be 0.

To avoid the occurrence of zero or negative probability, b_{ij}^s $(\forall i = 1, 2, \ldots, N,\ \forall j = 1, 2, \ldots, M)$ should be kept greater than 0 in the training procedure and at the same time,

the dynamic-channel coefficient b_{ij}^d ($\forall i = 1, 2, \ldots, N,\ \forall j = 1, 2, \ldots, M$) should be nonnegative. Thus the probability constraint $b_{ij} = b_{ij}^s + b_{ij}^d \geq b_{ij}^s > 0$ is met.

In addition, the relative weightage of the static channel and the dynamic channel may be controlled by the credibility weighing factor ω_i ($i = 1, 2, \ldots, N$) (different states may have different values). If the weightage of the dynamic channel is set to be ω_i by scaling of the coefficients

$$\sum_{j=1}^{N} b_{ij}^d = \omega_i, \quad 0 \leq \omega_i < 1\ \forall i = 1, 2, \ldots, N, \tag{24}$$

then the weightage of the static channel has to be set as follows:

$$\sum_{j=1}^{N} b_{ij}^s = 1 - \omega_i, \quad 0 \leq \omega_i < 1\ \forall i = 1, 2, \ldots, N. \tag{25}$$

4. TWO-CHANNEL TRAINING STRATEGY

4.1. Parameter initialization

The parameter-smoothed ML HMM of x^T, $\widetilde{\theta}_{\mathrm{ML}}^x$, which is trained using the Baum-Welch estimation, is referred to as the base HMM. The static-channel HMM is derived from the base HMM after applying the scaling factor. Parameter smoothing is carried out for $\widetilde{\theta}_{\mathrm{ML}}^x$ to prevent the occurrence of zero probability. Parameter smoothing is the simple management that b_{ij} is set to some minimum value, for example, $\varepsilon = 10^{-3}$, if the estimated conditional probability $\widetilde{b}_{ij} = 0$ [46]. As a result, even though symbol O_j never appears in the training set, there is still a nonzero probability of its occurrence in $\widetilde{\theta}_{\mathrm{ML}}^x$. Parameter smoothing is a posttraining adjustment to decrease error rate because the training set, which is usually limited by its size, may not cover erratic samples.

Before carrying out discriminative training, ω_i (credibility weighing factor of the ith state), b_{ij}^s (static-channel coefficients), and b_{ij}^d (dynamic-channel coefficients) are initialized.

The static-channel coefficients b_{ij}^s are given by

$$\{b_{i1}^s b_{i2}^s \cdots b_{iM}^s\} = (1 - \omega_i)\{\widetilde{b}_{i1}\widetilde{b}_{i2} \cdots \widetilde{b}_{iM}\}, \quad 1 \leq i \leq N,\ 0 \leq \omega_i < 1, \tag{26}$$

where $\widetilde{b}_{ij}$ is the symbol emission probability of $\widetilde{\theta}_{\mathrm{ML}}^x$.

As for the dynamic-channel coefficients b_{ij}^d, a random or uniform initial distribution usually works well. In the experiments conducted in this paper, uniform values equal to ω_i/M are assigned to b_{ij}^d's as initial values.

The selection of ω_i is flexible and largely problem-dependent. A large value of ω_i means large weightage is assigned to the dynamic channel and the discriminative power is enhanced. However, as we adjust b_{ij}^d toward the direction of increasing $I(x^T, y^T, \theta)$, the probability of the correct observation $P(x^T|\theta)$ will normally decrease. This situation is undesirable because the two-channel HMM obtained is unlikely to generate even the correct samples.

A guideline for the determination of the value of ω_i is as follows. If the training pairs are very similar to each other such that $P(x^T|\widetilde{\theta}_{\mathrm{ML}}^x) \approx P(y^T|\widetilde{\theta}_{\mathrm{ML}}^x)$, ω_i should be set to a large value to guarantee good discrimination; on the other hand, if $P(x^T|\widetilde{\theta}_{\mathrm{ML}}^x) \gg P(y^T|\widetilde{\theta}_{\mathrm{ML}}^x)$, ω_i should be set to a small value to make $P(x^T|\theta)$ reasonably large. In addition, different values will be used for different states because they contribute differently to the scored probabilities. However, the values of ω_i for the different states should not differ greatly.

Based on the above considerations, the following procedures are taken to determine ω_i. Given the base HMM $\widetilde{\theta}_{\mathrm{ML}}^x$ and the training pair x^T and y^T, the optimal state chains are searched using the Viterbi algorithm. If $\widetilde{\theta}_{\mathrm{ML}}^x$ is a left-right model and the expected (optimal) duration of the ith state ($i = 1, 2, \ldots, N$) of x^T is from t_i to $t_i + \tau_i$, $P(x^T|\widetilde{\theta}_{\mathrm{ML}}^x)$ is then written as follows:

$$P(x^T|\widetilde{\theta}_{\mathrm{ML}}^x) = P(x_{t_1}^T, \ldots, x_{t_1+\tau_1}^T|\widetilde{\theta}_{\mathrm{ML}}^x)P(x_{t_2}^T, \ldots, x_{t_2+\tau_2}^T|\widetilde{\theta}_{\mathrm{ML}}^x) \cdots P(x_{t_N}^T, \ldots, x_{t_N+\tau_N}^T|\widetilde{\theta}_{\mathrm{ML}}^x); \tag{27}$$

$P(y^T|\widetilde{\theta}_{\mathrm{ML}}^x)$ is decomposed in the same way.

Let $P_{\mathrm{dur}}(x^T, S_i|\widetilde{\theta}_{\mathrm{ML}}^x) = P(x_{t_i}^T, \ldots, x_{t_i+\tau_i}^T|\widetilde{\theta}_{\mathrm{ML}}^x)$. This probability may be computed as follows:

$$P_{\mathrm{dur}}(x^T, S_i|\widetilde{\theta}_{\mathrm{ML}}^x) = \prod_{t=t_i}^{t_i+\tau_i} \left[\sum_{j=1}^{N} P(s_t^T = S_j) b_j(x_t^T) \right]. \tag{28}$$

$P_{\mathrm{dur}}(x^T, S_i|\widetilde{\theta}_{\mathrm{ML}}^x)$ may also be computed using the forward variables $\alpha_t^x(i) = P(x_{t_i+1}^T, \ldots, x_{t_i+t}^T, s_{t_i+t}^T = S_i|\widetilde{\theta}_{\mathrm{ML}}^x)$ or/and the backward variables $\beta_t^x(i) = P(x_{t_i+t+1}^T, \ldots, x_{t_i+\tau_i+1}^T|s_{t_i+t}^T = S_i, \widetilde{\theta}_{\mathrm{ML}}^x)$ [46].

However, if $\widetilde{\theta}_{\mathrm{ML}}^x$ is not a left-right model but an ergodic model, the expected duration of a state will consist of a number of separated time slices, for example, k slices such as t_{i1} to $t_{i1} + \tau_{i1}$, t_{i2} to $t_{i2} + \tau_{i2}$, and t_{ik} to $t_{ik} + \tau_{ik}$. $P_{\mathrm{dur}}(x^T, S_i|\widetilde{\theta}_{\mathrm{ML}}^x)$ is then computed by multiplying them together as shown:

$$P_{\mathrm{dur}}(x^T, S_i|\widetilde{\theta}_{\mathrm{ML}}^x) = P(x_{t_{i1}}^T, \ldots, x_{t_{i1}+\tau_{i1}}^T|\widetilde{\theta}_{\mathrm{ML}}^x)P(x_{t_{i2}}^T, \ldots, x_{t_{i2}+\tau_{i2}}^T|\widetilde{\theta}_{\mathrm{ML}}^x) \cdots P(x_{t_{ik}}^T, \ldots, x_{t_{ik}+\tau_{ik}}^T|\widetilde{\theta}_{\mathrm{ML}}^x). \tag{29}$$

The value of ω_i is derived by comparing the corresponding $P_{\mathrm{dur}}(x^T, S_i|\widetilde{\theta}_{\mathrm{ML}}^x)$ and $P_{\mathrm{dur}}(y^T, S_i|\widetilde{\theta}_{\mathrm{ML}}^x)$. If $P_{\mathrm{dur}}(x^T, S_i|\widetilde{\theta}_{\mathrm{ML}}^x) \gg P_{\mathrm{dur}}(y^T, S_i|\widetilde{\theta}_{\mathrm{ML}}^x)$, this indicates that the coefficients of the ith state of the base model are good enough for discrimination, ω_i should be set to a small value to preserve the original ML configurations. If $P_{\mathrm{dur}}(x^T, S_i|\widetilde{\theta}_{\mathrm{ML}}^x) < P_{\mathrm{dur}}(y^T, S_i|\widetilde{\theta}_{\mathrm{ML}}^x)$ or $P_{\mathrm{dur}}(x^T, S_i|\widetilde{\theta}_{\mathrm{ML}}^x) \approx P_{\mathrm{dur}}(y^T, S_i|\widetilde{\theta}_{\mathrm{ML}}^x)$, this indicates that state S_i is not able to distinguish between x^T and y^T, thus ω_i must be set to a value large enough to ensure $P_{\mathrm{dur}}(x^T, S_i|\widetilde{\theta}) > P_{\mathrm{dur}}(y^T, S_i|\widetilde{\theta})$, where $\widetilde{\theta}$ is the two-channel

HMM. In practice, ω_i can be manually selected according to the conditions mentioned above (which is preferred), or they can be computed using the following expression:

$$\omega_i = \frac{1}{1 + Cv^D}, \tag{30}$$

where $v = P_{\text{dur}}(x^T, S_i|\widetilde{\theta}^x_{\text{ML}})/P_{\text{dur}}(y^T, S_i|\widetilde{\theta}^x_{\text{ML}})$. C ($C > 0$) and D are constants that jointly control the smoothness of ω_i with respect to v. Since $C > 0$ and $v > 0$, $\omega_i < 1$, by using suitable values of C and D, a set of credibility factors ω_i are computed for the states of the target HMM. For example, if the range of v is $10^{-3} \sim 10^5$, a typical setting is $C = 1.0$ and $D = 0.1$.

Once the values of ω_i ($i = 1, 2, \ldots, N$) are determined, they will not be changed in the training process.

4.2. Partition of the observation symbol set

Let θ denote the HMM with the above initial configurations. The coefficients of the dynamic channel are adjusted according to the following procedures. First, $E(S_i, O_j|\theta, x^T)$ and $E(S_i, O_j|\theta, y^T)$ are computed through the counting process. Using the forward variables $\alpha^x_\tau(i) = P(x^T_1, \ldots, x^T_\tau, s^T_\tau = S_i|\theta)$ and backward variables $\beta^x_\tau(i) = P(x^T_{\tau+1}, \ldots, x^T_T|s^T_\tau = S_i, \theta)$ [46], the following two probabilities are computed:

$$\begin{aligned}
\xi^x_\tau(i, j) &= P(s^T_\tau = S_i, s^T_{\tau+1} = S_j|x^T, \theta) \\
&= \frac{\alpha^x_\tau(i)a_{ij}b_j(x^T_{\tau+1})\beta^x_{\tau+1}(j)}{\sum_{i=1}^N \sum_{j=1}^N \alpha^x_\tau(i)a_{ij}b_j(x^T_{\tau+1})\beta^x_{\tau+1}(j)}, \\
\gamma^x_\tau(i) &= P(s^T_\tau = S_i|x^T, \theta) = \sum_{j=1}^N \xi^x_\tau(i, j);
\end{aligned} \tag{31}$$

$\xi^y_\tau(i, j)$ and $\gamma^y_\tau(i)$ are obtained in the same manner. By counting the state, we have

$$\begin{aligned}
E(S_i, O_j|\theta, x^T) &= \sum_{\substack{\tau=1 \\ x^T_\tau = O_j}}^T \gamma^x_\tau(i), \\
E(S_i, O_j|\theta, y^T) &= \sum_{\substack{\tau=1 \\ y^T_\tau = O_j}}^T \gamma^y_\tau(i).
\end{aligned} \tag{32}$$

It is shown in (22) that to maximize $I(x^T, y^T, \theta)$, b_{ij} should be set proportional to $D_{ij}(x^T, y^T, \theta)$. However, for certain symbols, for example, O_p, the expectation $D_{ip}(x^T, y^T, \theta)$ may be less than 0. Since the symbol emission coefficients cannot take negative values, these symbols have to be specially treated. For this reason, the symbol set $O^M = \{O_1, O_2, \ldots, O_M\}$ is partitioned into the subset $V = \{V_1, V_2, \ldots, V_K\}$ and its complement set $U = \{U_1, U_2, \ldots, U_{M-K}\}$ ($O^M = U \cup V$) according to the following criterion:

$$\{V_1, V_2, \ldots, V_K\} = \arg_{O_j}\left[\frac{E(S_i, O_j|\theta, x^T)}{E(S_i, O_j|\theta, y^T)} > \eta\right] \quad (\eta \geq 1), \tag{33}$$

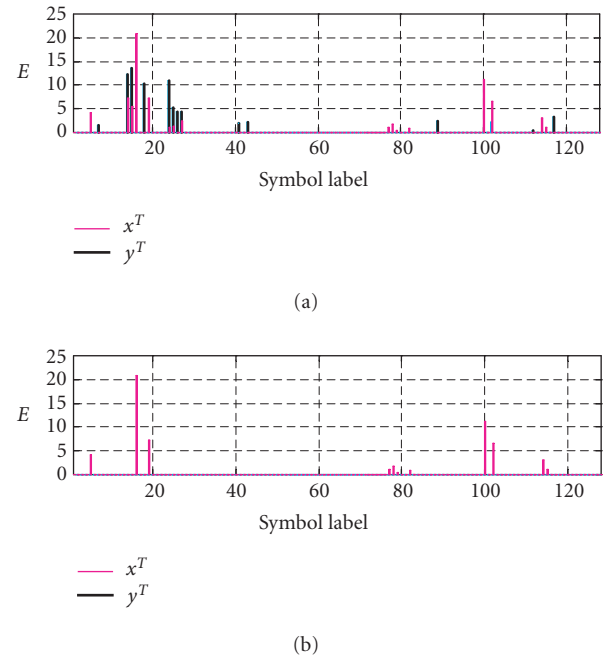


FIGURE 3: (a) Distributions of $E(S_i, O_j|\theta, x^T)$ and $E(S_i, O_j|\theta, y^T)$ for various symbols. (b) Distribution of $E(S_i, O_j|\theta, x^T)$ for the symbols in V.

where η is the threshold with a typical value of 1. η will be set to a larger value if it is required that the set V will contain fewer dominant symbols. With $\eta \geq 1$, $E(S_i, V_j|\theta, y^T) - E(S_i, V_j|\theta, x^T) > 0$. As an illustration, the distributions of the values of $E(S_i, O_j|\theta, x^T)$ and $E(S_i, O_j|\theta, y^T)$ for different symbol labels are shown in Figure 3a. The filtered symbols in set V when η is set l are shown in Figure 3b.

4.3. Modification to the dynamic channel

For each state, the symbol set is partitioned according to the procedures described in Section 4.2. As an example, consider the ith state. For symbols in the set U, the symbol emission coefficient $b_i(U_j)$ ($U_j \in U$) should be set as small as possible. Let $b^d_i(U_j) = 0$, and so $b_i(U_j) = b^s_i(U_j)$. For symbols in the set V, the corresponding dynamic-channel coefficient $b^d_i(V_k)$ is computed according to (34), which is derived from (22):

$$\begin{aligned}
b^d_i(V_k) &= P_D(S_i, V_k, x^T, y^T) \\
&\quad \times \left(\omega_i + \sum_{j=1}^K b^s_i(V_j)\right) - b^s_i(V_k), \quad k = 1, 2, \ldots, K,
\end{aligned} \tag{34}$$

where

$$\begin{aligned}
&P_D(S_i, V_k, x^T, y^T) \\
&= \frac{E(S_i, V_k|\theta, x^T) - E(S_i, V_k|\theta, y^T)}{\sum_{j=1}^K [E(S_i, V_j|\theta, x^T) - E(S_i, V_j|\theta, y^T)]}.
\end{aligned} \tag{35}$$

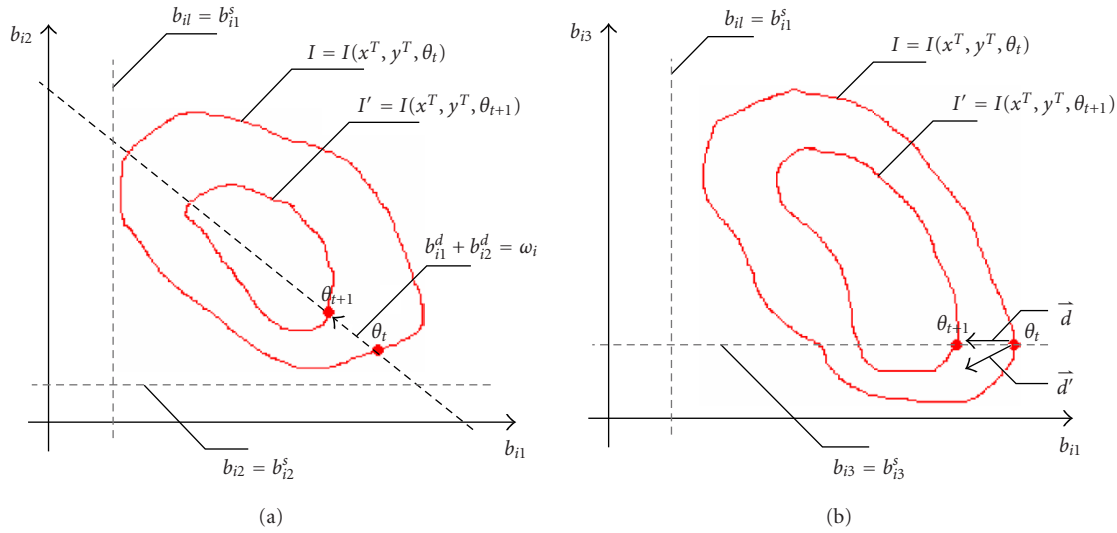


Figure 4: The surface of I and the direction of parameter adjustment.

However, some coefficients obtained may still be negative, for example, $b^d_i(V_l) < 0$ because of large value of $b^s_i(V_l)$. In which case, it indicates that $b^s_i(V_l)$ alone is large enough for separation. To prevent negative values appearing in the dynamic channel, the symbol V_l is transferred from V to U and $b^d_i(V_l)$ is set to 0. The coefficients of the remaining symbols in V are reestimated using (34) until all $b^d_i(V_k)$'s are greater than 0. This situation (some $b^d_i(V_l) < 0$) usually happens at the first few epochs of training and it is not conducive to convergence because there is a steep jump in the surface of $I(x^T, y^T, \theta)$. To relieve this problem, a larger value of η in (33) will be used.

4.4. Termination

Optimization is done through iteratively calling the training epoch described in Sections 4.2 and 4.3. After each epoch, the separable-distance $I(x^T, y^T, \widetilde{\theta})$ of the HMM $\widetilde{\theta}$ obtained, is calculated and compared with that obtained in the last epoch. If $I(x^T, y^T, \widetilde{\theta})$ does not change more than a predefined value, training is terminated and the target two-channel HMM is established.

5. PROPERTIES OF THE TWO-CHANNEL TRAINING STRATEGY

5.1. State alignment

One of the requirements for the proposed training strategy is that the state durations of the training pair, say x^T and y^T, are comparable. This is a requirement for (22). If the state durations, for example, $E(S_i|\theta, x^T)$ and $E(S_i|\theta, y^T)$, differ too much, $D_{ij}(x^T, y^T, \theta)$ will become meaningless. For example, if $E(S_i|\theta, x^T) \ll E(S_i|\theta, y^T)$, the symbol O_j takes much greater portion in $E(S_i|\theta, x^T)$ than in $E(S_i|\theta, y^T)$, the computed $D_{ij}(x^T, y^T, \theta)$ may also be less than 0. The outcome is that b_{ij} is always set to b^s_{ij} rather than adjusted to increase $I(x^T, y^T, \theta)$. Fortunately, if the corresponding state durations of the training pair are very different, the normal ML HMMs are usually adequate to distinguish the states.

The following state-duration validation procedure is added to make the training strategy complete. After each training epoch, $E(S_i|\theta, x^T)$ and $E(S_i|\theta, y^T)$ are computed and compared with each other. Using the forward variables and backward variables, the state duration of x^T is obtained as follows:

$$E(S_i|\theta, x^T) = \sum_{\tau=1}^{T} \frac{\alpha^x_\tau(i)\beta^x_\tau(i)}{\sum_{i=1}^{N} \alpha^x_\tau(i)\beta^x_\tau(i)}, \quad i = 1, 2, \ldots, N, \tag{36}$$

and $E(S_i|\theta, y^T)$ is computed in the same way. If $E(S_i|\theta, x^T) \approx E(S_i|\theta, y^T)$ (not necessary to be the same), for example, $1.2E(S_i|\theta, y^T) > E(S_i|\theta, x^T) > 0.8E(S_i|\theta, y^T)$, training continues; otherwise, training stops even if $I(x^T, y^T, \theta)$ keeps on increasing.

If the $I(x^T, y^T, \widetilde{\theta})$ of the final HMM $\widetilde{\theta}$ does not meet certain discriminative requirement, for example, $I(x^T, y^T, \widetilde{\theta})$ is less than a desired value, a new base HMM or a smaller ω_i should be used instead.

5.2. Speed of convergence

As discussed in Section 3, the convergence of the parameter-estimation strategy proposed in (22) is guaranteed according to the EM optimization principles. In the implementation of discriminative training, only some of the symbol emission coefficients in the dynamic channel are modified according to (22) while the others remain unchanged. However, the convergence is still assured because firstly the surface of $I(x^T, y^T, \theta)$ with respect to b_{ij} is continuous, and also adjusting the dynamic-channel elements according to the two-channel training strategy leads to increased $E_\theta(I)$. A conceptual illustration is given in Figure 4 on how b_{ij} is modified when the symbol set is divided into subsets V and U. For ease of explanation, we assume that the symbol set contains only three symbols O_1, O_2, and O_3 with $O_1, O_2 \in V$ and $O_3 \in U$

for state S_i. Let θ_t denote the HMM trained at the tth round and let θ_{t+1} denote the HMM obtained at the $t+1$th round. The surface of the separable distance (I surface) is denoted as $I' = I(x^T, y^T, \theta_{t+1})$ for θ_{t+1} and $I = I(x^T, y^T, \theta_t)$ for θ_t. Clearly $I' > I$. The I surface is mapped to the b_{i1}-b_{i2} plane (Figure 4a) and the b_{i1}-b_{i3} plane (Figure 4b). In the training phase, b_{i1} and b_{i2} are modified along the line $b_{i1}^d + b_{i2}^d = \omega_i$ to reach a better estimation θ_{t+1}, which is shown in Figure 4a. In the b_{i1}-b_{i3} plane, b_{i3} is set to the constant b_{i3}^s while b_{i1} is modified along the line $b_{i3} = b_{i3}^s$ with the direction $\vec{d}$ as shown in Figure 4b. The direction of parameter adjustment given by (22) is denoted by $\vec{d}'$. In the two-channel approach, since only b_{i1} and b_{i2} are modified according to (22) while b_{i3} remains unchanged, $\vec{d}$ may lead to lower speed of convergence than $\vec{d}'$ does.

5.3. Improvement to the discriminative power

The improvement to the discriminative power is estimated as follows. Assume that $\widetilde{\theta}$ is the two-channel HMM obtained. The lower bound of the probability $P(y^T|\widetilde{\theta})$ is given by

$$P(y^T|\widetilde{\theta}) \geq (1-\omega_{\max})^T P(y^T|\widetilde{\theta}^x_{\mathrm{ML}}), \tag{37}$$

where $\omega_{\max} = \max(\omega_1, \omega_2, \ldots, \omega_N)$.

Because the base HMM is the parameter-smoothed ML HMM of x^T, it is natural to assume that $P(x^T|\widetilde{\theta}^x_{\mathrm{ML}}) \geq P(x^T|\widetilde{\theta})$. The upper bound of the separable distance is given by the following expression:

$$\begin{aligned} I(x^T, y^T, \widetilde{\theta}) &\leq \log \frac{P(x^T|\widetilde{\theta}^x_{\mathrm{ML}})}{(1-\omega_{\max})^T P(y^T|\widetilde{\theta}^x_{\mathrm{ML}})} \\ &= -T\log(1-\omega_{\max}) + I(x^T, y^T, \widetilde{\theta}^x_{\mathrm{ML}}). \end{aligned} \tag{38}$$

In practice, the gain of $I(x^T, y^T, \widetilde{\theta})$ is much smaller than the theoretical upper bound. It depends on the resemblance between x^T and y^T, and the setting of ω_i.

6. EXTENSIONS OF THE TWO-CHANNEL TRAINING ALGORITHM

6.1. Training samples with different lengths

Up to this point, the training sequences are assumed to be of equal length. This is necessary as we cannot properly compare the probability scores of two sequences of different lengths. To extend the training strategy to sequences of different lengths, linear adjustment is first carried out as follows. Given the training pair x^{T_x} of length T_x and y^{T_y} of length T_y, the objective function (10) is modified as follows:

$$\begin{aligned} \lambda_i = &\frac{\sum_{\tau=1}^{T_x} P(s_\tau^{T_x} = S_i, x_\tau^{T_x} = O_j|\theta, x^{T_x})}{b_{ij}} \\ &- \frac{(T_x/T_y)\sum_{\tau=1}^{T_y} P(s_\tau^{T_y} = S_i, y_\tau^{T_y} = O_j|\theta, y^{T_y})}{b_{ij}} \\ &\forall j = 1, 2 \ldots, M. \end{aligned} \tag{39}$$

Parameter estimation is then carried out as follows:

$$\widetilde{b}_{ij} = \frac{E(S_i, O_j|\theta, x^{T_x}) - (T_x/T_y)E(S_i, O_j|\theta, y^{T_y})}{\sum_{j=1}^{M}\left[E(S_i, O_j|\theta, x^{T_x}) - (T_x/T_y)E(S_i, O_j|\theta, y^{T_y})\right]}. \tag{40}$$

The expectations of different states of y^{T_y} are normalized using the scale factor T_x/T_y. This approach is easy to implement; however, it does not consider the nonlinear variance of signal such as local stretch or squash. If the training sequences demonstrate obvious nonlinear variance, some nonlinear processing such as sequence truncation or symbol prune may be carried out to adjust the training sequences to the same length [50].

6.2. Multiple training samples

In order to obtain a reliable model, multiple observations must be used to train the HMM. The extension of the proposed method to include multiple training samples may be carried out as follows. Consider two labeled sets $X = \{x^{(1)}, x^{(2)}, \ldots, x^{(k)} : d_1\}$ and $Y = \{y^{(1)}, y^{(2)}, \ldots, y^{(l)} : d_2\}$ of samples, where X has k number of samples and Y has l number of samples. The separable-distance function that takes care of all these samples is given by

$$I(X, Y, \theta) = \frac{1}{k}\sum_{m=1}^{k}\log P(x^{(m)}|\theta) - \frac{1}{l}\sum_{n=1}^{k}\log P(y^{(n)}|\theta). \tag{41}$$

For simplicity, if we assume that the observation sequences in X and Y have the same length T, then (10) may be rewritten as

$$\begin{aligned} &\frac{(1/k)\sum_{m=1}^{k} E(S_i, O_j|\theta, x^{(m)}) - (1/l)\sum_{n=1}^{l} E(S_i, O_j|\theta, y^{(n)})}{b_{ij}} \\ &\quad = \lambda_i, \quad 1 \leq j \leq M. \end{aligned} \tag{42}$$

The probability coefficients are then estimated using the following:

$$\widetilde{b}_{ij} = \frac{(1/k)\sum_{m=1}^{k} E(S_i, O_j|\theta, x^{(m)}) - (1/l)\sum_{n=1}^{l} E(S_i, O_j|\theta, y^{(n)})}{\sum_{j=1}^{M}\left[(1/k)\sum_{m=1}^{k} E(S_i, O_j|\theta, x^{(m)}) - (1/l)\sum_{n=1}^{l} E(S_i, O_j|\theta, y^{(n)})\right]}. \tag{43}$$

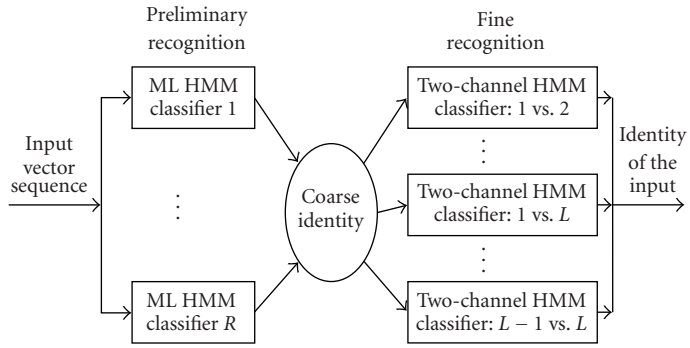


Figure 5: Block diagram of the viseme recognition system.

Table 1: The 18 visemes selected for the experiments.

/a:/, /ai/, /æ/, /ei/, /i/, /j/, /ie/, /o/, /oi/, /th/, /sh/, /tZ/, /dZ/, /eu/, /au/, /p/, /m/, /b/

7. APPLICATION TO LIP READING

The proposed two-channel HMM method is applied to speaker-dependent lip reading for modeling and recognizing the basic visual speech elements of the English language. For the experiments reported in this paper, the visemes are treated as having a one-to-one mapping with the phonemes in order to test the discriminative power of the proposed method. As there are 48 phonemes in the English language [47], 48 visemes are considered.

The block diagram of the viseme recognition system is given in Figure 5. The lip movement is captured with a video camera and the sequence of images is processed to extract the essential features relevant to the lip movement. For each frame of image, a feature vector is extracted. The sequence of feature vectors thus represents the movement of lips during viseme production. This vector sequence is then presented as input to the proposed classifier. A hierarchical structure is adopted such that for a system with K visemes to be recognized, R (usually $R < K$) ML HMM classifiers are employed for preliminary recognition. The output of the preliminary recognition is a coarse identity, which may include L (usually $1 < L < K$) viseme classes. Fine recognition is then performed using a bank of two-channel HMMs. The most probable viseme is then chosen as the identity of the input. Details of the various steps involved are given in the following sections.

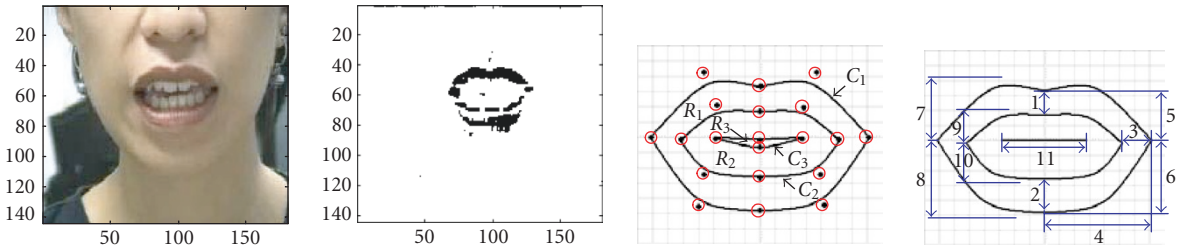
7.1. Data acquisition

For our experiments, a professional English speaker is engaged. The speaker is asked to articulate every phone me of the 18 phonemes in Table 1 one hundred times. The 18 visemes are chosen as some of them bear close similarity to others. The lip movements of the speakers are captured at 50 frames per second. Each pronunciation starts from a closed mouth and ends with a closed mouth. This type of samples is referred to as text-independent viseme samples, which is different from the type of samples extracted from various contexts, for example, from different words. The video clips that indicate the productions of context-independent visemes are normalized such that all the visemes have uniform duration of 0.5 second, or equivalently 25 frames.

7.2. Feature extraction

Each frame of the video clip reveals the lip area of the speaker during articulation (Figure 6a). To eliminate the effect caused by changes in the brightness, the RGB (red, green, blue) factors of the image are converted into HSV (hue, saturation, value) factors. The RGB to HSV conversion algorithm proposed in [51, 52] is adopted in our experiments. As illustrated in the histograms of distribution of the hue component shown in Figure 7, the hue factors of the lip region and the remaining lip-excluded image occupy different regions of the histogram. A threshold may be manually selected to segment the lip region from the entire image as shown in Figure 6b. This threshold usually corresponds to a local minimum point (valley) in the histogram as shown in Figure 7a. Note that for different speakers and lighting conditions, the threshold may be different.

The boundaries of the lips are tracked using a geometric template with dynamic contours to fit an elastic object [53, 54, 55]. As the contours of the lips are simple, the requirement on the selection of the dynamic contours that build the template is thus not stringent. Results of lip tracking experiments show that Bezier curves can well fit the shape of the lip [34]. In our experiments, the parameterized template consists of ten Bezier curves with eight of them characterizing the lip contours and two of them describing the tongue when it is visible (Figure 6c). The template is controlled by points marked as small circles in Figure 6c. Lip tracking is carried out by fitting the template to minimize a certain energy function. The energy function comprises the



(a) (b) (c) (d)

FIGURE 6: (a) Original image. (b) Segmented lip area. (c) Parameterized lip template. (d) Geometric measures extracted from the lip template. (1) Thickness of the upper bow. (2) Thickness of the lower bow. (3) Thickness of the lip corner. (4) Position of the lip corner. (5) Position of the upper lip. (6) Position of the lower bow. (7) Curvature of the upper-exterior boundary. (8) Curvature of the lower-exterior boundary. (9) Curvature of the upper-interior boundary. (10) Curvature of the lower-interior boundary. (11) Width of the tongue (when it is visible).

following four terms:

$$\begin{aligned} E_{\text{lip}} &= -\frac{1}{R_1}\int_{R_1} H(x)dx, \\ E_{\text{edge}} &= -\frac{1}{C_1+C_2}\int_{C_1+C_2} \left|H^+(x)-H(x)\right| \\ &\qquad + \left|H^-(x)-H(x)\right|dx, \\ E_{\text{hole}} &= -\frac{1}{R_2-R_3}\int_{R_2-R_3} H(x)dx, \\ E_{\text{inertia}} &= \left|\left|\Gamma_{t+1}-\Gamma_t\right|\right|^2, \end{aligned} \tag{44}$$

where R_1, R_2, R_3, C_1, and C_2 are areas and contours as illustrated in Figure 6c. $H(x)$ is a function of the hue of a given pixel; $H^+(x)$ is the hue function of the closest right-hand side pixel and $H^-(x)$ is that of the closest left-hand side pixel. Γ_{t+1} and Γ_t are the matched templates at time $t+1$ and t. $\|\Gamma_{t+1}-\Gamma_t\|$ indicates the Euclidean distance between the two templates (further details may be found in [55]). The overall energy of the template E is the linear combination of the components defined as

$$E = c_1E_{\text{lip}} + c_2E_{\text{edge}} + c_3E_{\text{hole}} + c_4E_{\text{inertia}}. \tag{45}$$

Similarly, the energy terms for the tongue template include

$$\begin{aligned} E_{\text{tongue-area}} &= -\frac{1}{R_3}\int_{R_3} H(x)dx \quad \text{if } R_3 > 0, \\ E_{\text{tongue-edge}} &= -\frac{1}{C_3}\int_{C_3} \left|H^+(x)-H(x)\right| \\ &\qquad + \left|H^-(x)-H(x)\right|dx \quad \text{if } C_3 > 0, \\ E_{\text{tongue-inertia}} &= \left|\left|T_{\text{tongue},t+1}-T_{\text{tongue},t}\right|\right|^2, \end{aligned} \tag{46}$$

and the overall energy is

$$E_{\text{tongue}} = c_5E_{\text{tongue-area}} + c_6E_{\text{tongue-edge}} + c_7E_{\text{tongue-inertia}}. \tag{47}$$

Initially, the dynamic contours are configured to provide a crude match to the lips. This can be done via comparing the enclosed region of the template and the segmented lip region as depicted in Figure 6b. Following that, the template is matched to the image sequence by adopting different values of the parameters $\{c_i\}$ $(i = 1, 2, \ldots, 7)$ in a number of searching epochs (a detailed discussion is given in [53, 54, 55]). The matched template is pictured in Figure 6d. It can be seen that the matched template is symmetric and smooth, and is therefore easy to process.

Eleven geometric parameters as shown in Figure 6d are extracted to form a feature vector from the matched template. These features indicate the thickness of various parts of the lips, the positions of some key points, and the curvatures of the bows. They are chosen as they uniquely determine the shape of the lips and they best characterize the movement of the lips.

Principal components analysis (PCA) is carried out to reduce the dimension of the feature vectors from eleven to seven. The resulting feature vectors are clustered into groups using K-means algorithm. In the experiments conducted, 128 clusters are created for the vector database. The means of the 128 clusters form the symbol set $O^{128} = (O_1, O_2, \ldots, O_{128})$ of the HMM. They are used to encode the vector sequences presented to the system.

7.3. Configuration of the viseme model

Investigation on the lip dynamics reveals that the movement of the lips can be partitioned into three phases during the production of a text-independent viseme. The initial phase begins with a closed mouth and ends with the start of sound production. The intermediate phase is the articulation phase, which is the period when sound is produced. The third phase is the end phase when the mouth restores to the relaxed state. Figure 8 illustrates the change of the lips in the three phases and the corresponding acoustic waveform when the phoneme /u/ is uttered.

To associate the HMM with the physical process of viseme production, three-state left-right HMM structure as shown in Figure 9 is adopted.

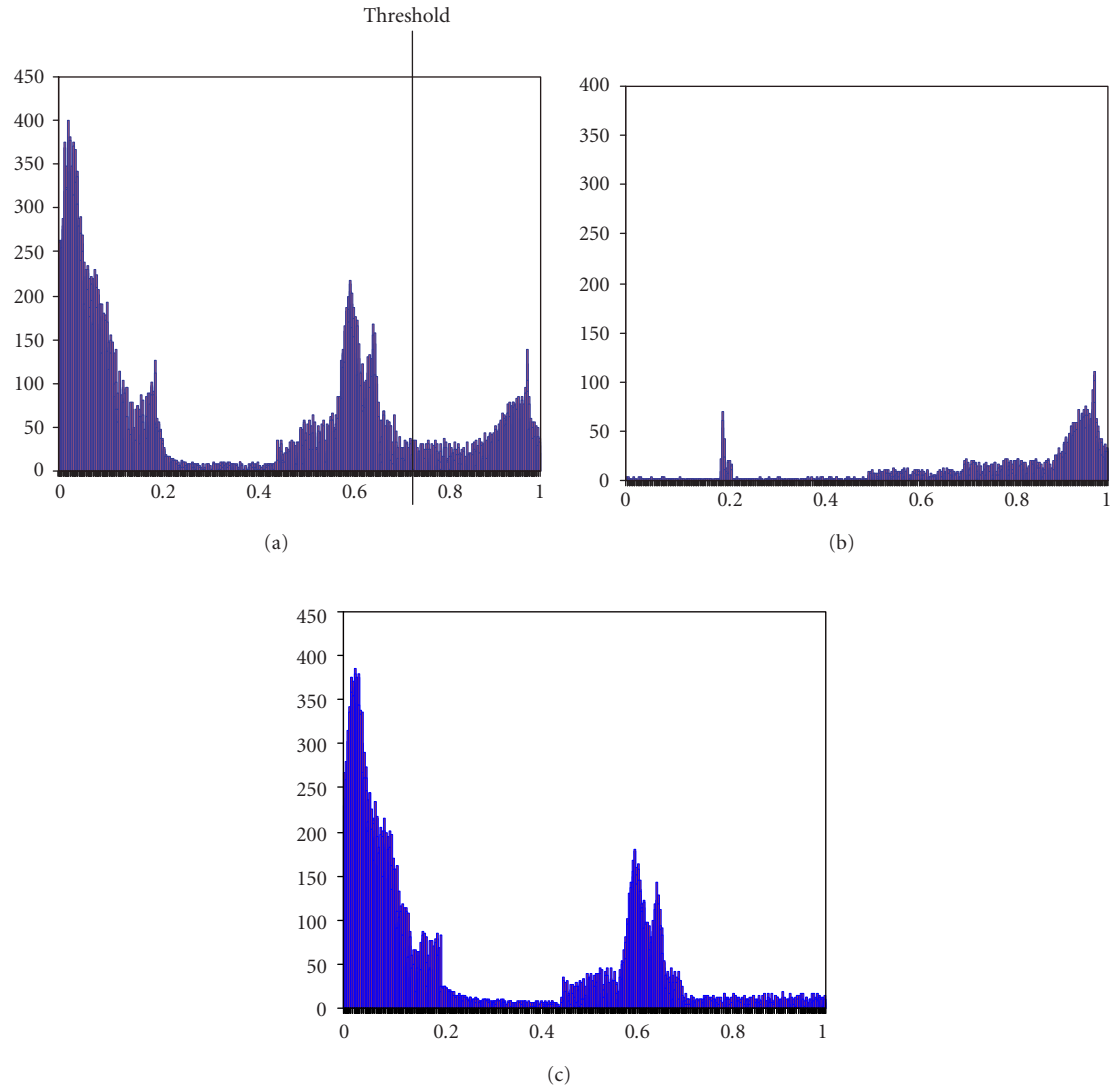


FIGURE 7: Isolation of the lip region from the entire image using hue distribution. (a) Histogram of the hue component for the entire image. (b) Histogram of the hue component for the actual lip region. (c) Histogram of the hue component for the actual lip-excluded image.

Using this structure, the state-transition matrix A has the form

$$A = \begin{bmatrix} a_{1,1} & a_{1,2} & 0 & 0 \\ 0 & a_{2,2} & a_{2,3} & 0 \\ 0 & 0 & a_{3,3} & a_{3,4} \\ 0 & 0 & 0 & 1 \end{bmatrix}, \tag{48}$$

where the 4th state is a null state that indicates the end of viseme production. The initial values of the coefficients in matrices A and B are set according to the statistics of the three phases. Given a viseme sample, the approximate initial phase, articulation phase, and end phase are segmented from the image sequence and the acoustic signal (an illustration is given in Figure 8), and the duration of each phase is counted. The coefficients $a_{i,i}$ and $a_{i,i+1}$ are initialized with these durations. For example, if the duration of state S_i is T_i, the initial value of $a_{i,i}$ is set to be $T_i/(T_i+1)$ and the initial value of $a_{i,i+1}$ is set to be $1/(T_i+1)$ as they maximize $a_{i,i}^{T_i} a_{i,i+1}$. Matrix B is initialized in a similar manner. If symbol O_j appears $T(O_j)$ times in state S_i, the initial value of b_{ij} is set to be $T(O_j)/T_i$. For such arrangement, the states of the HMM are aligned with the three phases of viseme production and hence are referred to as the initial state, articulation state, and end state.

For each of the 18 visemes in Table 1, an HMM with the above the configuration is trained using the Baum-Welch estimation. After implementing parameter smoothing, the parameter-smoothed ML HMM is ready for the subsequent two-channel discriminative training.

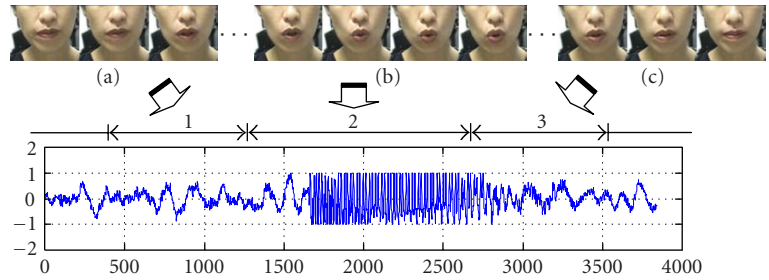


Figure 8: The three phases of viseme production. (a) Initial phase. (b) Articulation phase. (c) End phase.

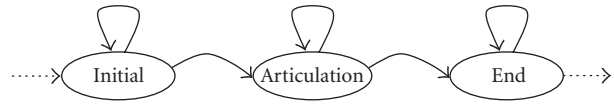


Figure 9: Three-state left-right viseme model.

7.4. Viseme classifier

The block diagram of the proposed hierarchical viseme classifier is given in Figure 10. For visemes that are too similar to be separated by the normal ML HMMs, they are clustered into one macro class. In the figure, $\theta_{\text{Mac1}}, \theta_{\text{Mac2}}, \ldots, \theta_{\text{Mac}R}$ are the R number of ML HMMs for the R macro classes. The similarity between the visemes is measured as follows.

Assume that $X_i = \{x_1^i, x_2^i, \ldots, x_{l_i}^i : d_i\}$ is the training samples of viseme d_i ($i = 1, 2, \ldots, 18$, as 18 visemes are involved), where x_j^i is the jth training sample and l_i is the number of the samples. An ML HMM is trained for each of the 18 visemes using the Baum-Welch estimation. Let $\theta_1, \theta_2, \ldots, \theta_{18}$ denote the 18 ML HMMs. For $\{x_1^i, x_2^i, \ldots, x_{l_i}^i : d_i\}$, the joint probability scored by θ_j is computed as follows:

$$P(X_i|\theta_j) = \prod_{n=1}^{l_i} P(x_n^i|\theta_j). \tag{49}$$

A viseme model θ_i is able to separate visemes d_i and d_j if the following condition applies:

$$\log P(X_i|\theta_i) - \log P(X_i|\theta_j) \geq Kl_i \quad \forall j = 1, 2, \ldots, 18,\ j \neq i, \tag{50}$$

where K is a positive constant that is set according to the length of the training samples. For long training samples, a large value of K is desired. For the 25-length samples adopted in our experiments, K is set to be equal to 2. If the condition stated in (50) is not met, visemes d_i and d_j are categorized into the same macro class. The training samples of d_i and d_j are jointly used to train the ML HMM of the macro class. $\theta_{\text{Mac1}}, \theta_{\text{Mac2}}, \ldots, \theta_{\text{Mac}R}$ are obtained in this way.

For an input viseme z^T to be identified, the probabilities $P(z^T|\theta_{\text{Mac1}}), P(z^T|\theta_{\text{Mac2}}), \ldots, P(z^T|\theta_{\text{Mac}R})$ are computed and compared with one another. The macro identity of z^T is determined by the HMM that gives the largest probability.

A macro class may consist of several similar visemes. Fine recognition within a macro class is carried out at the second layer. Assume that Macro Class i comprises L visemes: $V_1, V_2, \ldots, V_L$. A number of two-channel HMMs are trained with the proposed discriminative training strategy. For V_1, $L-1$HMMs, $\theta_{1\wedge 2}, \theta_{1\wedge 3}, \ldots, \theta_{1\wedge L}$, are trained to separate the samples of V_1 from those of $V_2, V_3, \ldots, V_L$, respectively. Take $\theta_{1\wedge 2}$ as an example, the parameter-smoothed ML HMM of V_1, $\tilde{\theta}_{\text{ML}}^1$, is adopted as the base HMM. The samples of V_1 are used as the correct samples (x^T in (3)) and the samples of V_2 are used as the incorrect samples (y^T in (3)) while training $\theta_{1\wedge 2}$. There is a total of $L(L-1)$ two-channel HMMs in Macro Class i.

For an input viseme z^T to be identified, the following hypothesis is made:

$$H_{i\wedge j} = \begin{cases} i & \text{if } \log P(z^T|\theta_{i\wedge j}) - \log P(z^T|\theta_{j\wedge i}) > K, \\ 0 & \text{otherwise,} \end{cases} \tag{51}$$

where K is the positive constant as defined in (47). For the 25-frame sequence input to the system, K is chosen to be equal to 2. $H_{i\wedge j} = i$ indicates a vote for V_i. The decision about the identity of z^T is made by a majority vote of all the two-channel HMMs. The viseme class that has the maximum number of votes is chosen as the identity of z^T, denoted by $\text{ID}(z^T)$. Mathematically,

$$\text{ID}(z^T) = \max_i \left[\text{Number of } H_{i\wedge j} = i\right] \quad \forall i, j = 1, 2, \ldots, L,\ i \neq j. \tag{52}$$

If two viseme classes, say V_i and V_j, receive the same number of votes, the decision about the identity of z^T is made

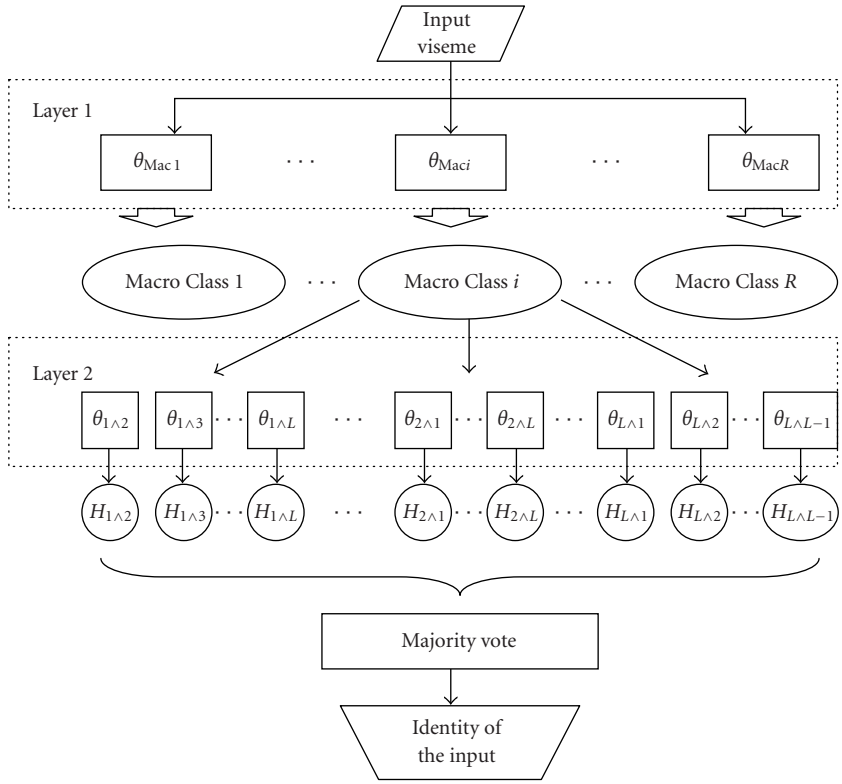


Figure 10: Flow chart of the hierarchical viseme classifier.

by comparing $P(z^T|\theta_{i\wedge j})$ and $P(z^T|\theta_{j\wedge i})$. Mathematically,

$$\text{ID}\,(z^T) = \begin{cases} i & \text{if } \log P(z^T|\theta_{i\wedge j}) > \log P(z^T|\theta_{j\wedge i}), \\ j & \text{otherwise.} \end{cases} \tag{53}$$

The decision is based on pairwise comparisons of the hypotheses. The proposed hierarchical structure greatly reduces the computational load and increases the accuracy of recognition because pairwise comparisons are carried out within each macro class, which comprises much fewer candidate classes than the entire set. If coarse identification is not performed, the number of classes increases and the number of pairwise comparisons goes up rapidly.

The two-channel HMMs act as the boundary functions for the viseme they represent. Each of them serves to separate the correct samples from the samples of another viseme. A conceptual illustration is given in Figure 11 where the macro class comprises five visemes $V_1, V_2, \ldots, V_5$. $\theta_{1\wedge2}, \theta_{1\wedge3}, \ldots, \theta_{1\wedge5}$ build the decision boundaries for V_1 to delimit it from the similar visemes.

The proposed two-channel HMM model is specially tailored for the target viseme and its "surroundings". As a result, it is more accurate than the traditional modeling method that uses single ML HMM.

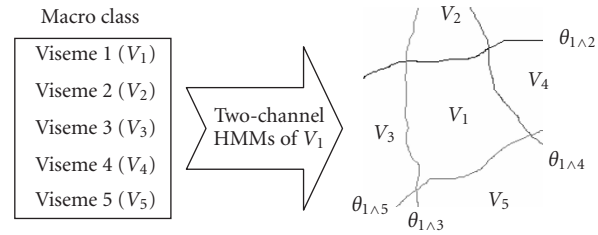


Figure 11: Viseme boundaries formed by the two-channel HMMs.

7.5. Performance of the system

Experiments are carried out to assess the performance of the proposed system. For the experiments conducted in this paper, 100 samples are drawn for each viseme with 50 for training and the remaining 50 for testing. By computing and comparing the probabilities scored by different viseme models using (49) and (50), the 18 visemes are clustered into 6 macro classes as illustrated in Table 2.

The results of fine recognition of some confusable visemes are listed in Table 3. Each row in Table 3 shows the two similar visemes that belong to the same macro class. The first viseme label (in boldface) is the target viseme and is denoted by $\boldsymbol{x}$. The second viseme is the incorrect viseme and is denoted by y. $\widetilde{\theta}_{\text{ML}}$ denotes the parameter-smoothed ML

Table 2: The macro classes for coarse identification.

Macro classes	Visemes	Macro classes	Visemes
1	/a:/, /ai/, /æ/	4	/o/, /oi/
2	/ei/, /i/, /j/, /ie/	5	/th/, /sh/, /tZ/, /dZ/
3	/eu/, /au/	6	/p/, /m/, /b/

Table 3: The average values of probability and separable-distance function of the ML HMMs and two-channel HMMs.

Viseme pair		$\widetilde{\theta}_{\mathrm{ML}}$		θ_1^*		θ_2^{**}				
$\boldsymbol{x}$	y	$\overline{P}$	$\overline{I}$	$\overline{P}$	$\overline{I}$	$\overline{P}$	$\overline{I}$	ω_1	ω_2	ω_3
/a:/	/ai/	−14.1	1.196	−17.1	5.571	−18.3	6.589	0.5	0.5	0.5
/ei/	/i/	−14.7	2.162	−19.3	5.977	−20.9	7.008	0.6	0.8	0.6
/au/	/eu/	−15.6	2.990	−18.1	5.872	−18.5	6.555	0.6	0.5	0.6
/o/	/oi/	−13.9	0.830	−17.5	2.508	−18.7	3.296	0.5	0.5	0.5
/th/	/sh/	−15.7	0.602	−19.0	2.809	−18.5	2.732	0.4	0.4	0.4
/p/	/m/	−16.3	1.144	−19.0	3.102	−17.1	2.233	0.4	0.5	0.4

Configuration of the two-channel HMMs:
*For θ_1, ω_1, ω_2, and ω_3 are set according to (30), with $C = 1.0$ and $D = 0.1$.
**For θ_2, ω_1, ω_2, and ω_3 are manually selected.

HMMs that are trained with the samples of $\boldsymbol{x}$. With $\widetilde{\theta}_{\mathrm{ML}}$ being the base HMM, two two-channel HMMs, θ_1 and θ_2, are trained with the samples of $\boldsymbol{x}$ being the target training samples and the samples of y being the incorrect training samples. Different sets of the credibility factors (ω_1, ω_2, and ω_3 for the three states) are used for θ_1 and θ_2. $\overline{P}$ is the average log probability scored for the testing samples and is computed as $\overline{P} = (1/l)\sum_{i=1}^{l} \log P(x_i|\theta)$, where x_i is the ith testing sample of viseme $\boldsymbol{x}$ and l is the number of the testing samples. $\overline{I} = (1/l^2)\sum_{i=1}^{l}\sum_{j=1}^{l} I(x_i, y_j, \theta)$ is the average separable distance. The value of $\overline{I}$ gives an indication of the discriminative power, the larger the value of $\overline{I}$, the higher the discriminative power.

For all settings of $(\omega_1, \omega_2, \omega_3)$, the two-channel HMMs give a much larger separable-distance than the ML HMMs. It shows that better discrimination capabilities are attained using the two-channel viseme classifiers than using the ML HMM classifiers. In addition, different levels of capabilities can be attained by adjusting the credibility factors. However, the two-channel HMM gives smaller average probability for the target samples than the normal ML HMM. It indicates that the two-channel HMMs perform well at discriminating confusable visemes but are not good at modeling the visemes.

The change of $I(x, y, \theta)$ with respect to the training epochs in the two-channel training is depicted in Figure 12. For the three-state left-right HMMs and 25-length training samples adopted in the experiment, the separable-distance becomes stable after ten to twenty epochs. Such speed of convergence shows that the two-channel training is not computationally intensive for viseme recognition. It is also observed that $I(x, y, \theta)$ may drop at the first few training epochs. This phenomenon can be attributed to the fact that some symbols in subset V are transferred to U while training the dynamic-channel coefficients as explained in Section 4.3. Figure 12d illustrates the situation of early termination. The training process stops even though $I(x, y, \theta)$ still shows the tendency of increasing. As explained in Section 5.1, if the state durations of the target training samples and incorrect training samples differ greatly, that is, the state alignment condition is violated, the two-channel training should terminate immediately.

The performance of the proposed hierarchical system is compared with that of the traditional recognition system where ML HMMs (parameter-smoothed) are used as the viseme classifiers. The ML HMMs and the two-channel HMMs involved are trained with the same set of training samples. The credibility factors of the two-channel HMMs are set according to (30), with $C = 0.1$ and $D = 0.1$. The decision about the identity of an input testing sample is made according to (47), (49), (50), and (51), where $K = 2$. The false rejection error rates (FRRs) or Type-II error of the two types of viseme classifiers are computed for the 50 testing samples of each of the 18 visemes. Note that as some of the 18 visemes can be accurately identified by the ML HMMs with FRRs less than 10% [34], the improvement resulting from the two-channel training approach is not prominent for these visemes. In Table 4, only the FRRs of 12 confusable visemes are listed.

Compared with the conventional ML HMM classifier, the classification error of the proposed hierarchical viseme classifier is reduced by about 20%. Thus the two-channel training algorithm is able to increase the discriminative ability of HMM significantly for identifying visemes.

8. CONCLUSION

In this paper, a novel two-channel training strategy for hidden Markov model is proposed. A separable-distance function, which measures the difference between a pair of training samples, is applied as the objective function. To maximize the separable distance and maintain the validity of the probabilistic framework of HMM at the same time, a two-channel HMM structure is used. Parameters in one channel, named the dynamic channel, are optimized in a series of expectation-maximization (EM) estimations if feasible while parameters in the other channel, the static channel, are kept fixed. The HMM trained in this way amplifies the difference between the training samples. This strategy is especially suited to increase the discriminative ability of HMM over confusable observations.

The proposed training strategy is applied to viseme recognition. A hierarchical system is developed with normal ML HMM classifier implementing coarse recognition and two-channel HMM carrying out fine recognition. To extend the classification from binary-class to multiple-class, a decision rule based on majority vote is adopted. Experimental results show that the classification error of the proposed viseme

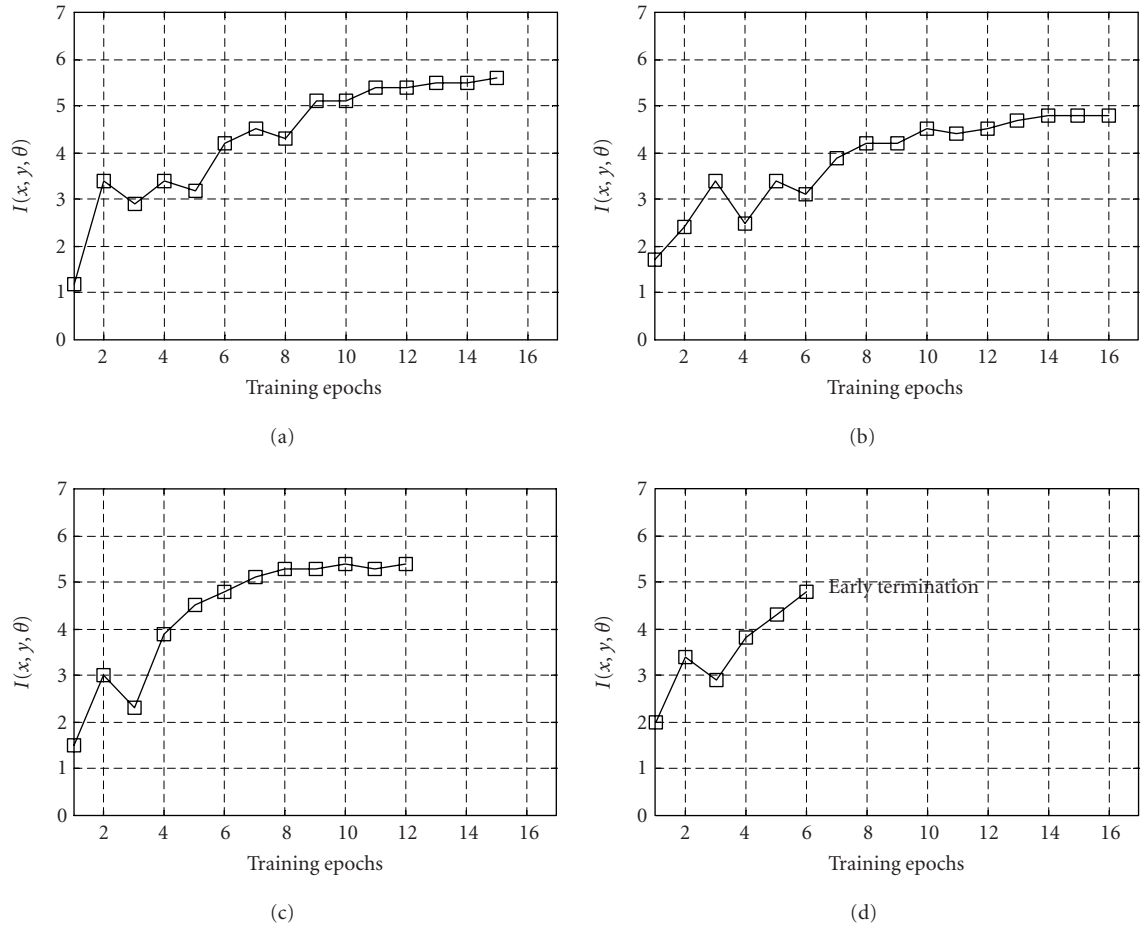


Figure 12: Change of $I(x, y, \theta)$ during the training process.

Table 4: Classification error ε_1 of the conventional classifier and classification error ε_2 of the two-channel classifier.

Viseme	ε_1	ε_2	Viseme	ε_1	ε_2
/a:/	64%	12%	/o/	46%	28%
/ai/	60%	40%	/oi/	36%	8%
/ei/	46%	22%	/th/	18%	16%
/i/	52%	32%	/sh/	20%	12%
/au/	30%	18%	/p/	36%	12%
/eu/	26%	16%	/m/	32%	32%

classifier is on the average 20% less than that of the popular ML HMM classifier while only 10 ∼ 20 training epochs are required in the training process.

The two-channel training strategy thus provides significant improvement over the traditional Baum-Welch estimation in fine recognition. However, the proposed method requires state alignment among the training samples; in other words, the samples should be of sufficient similarity such that the durations of the corresponding states are comparable.

Although the two-channel HMM is illustrated for viseme classification in this paper, the method is applicable to any sequence classification problem where the sequences to be recognized are of comparable length. Such applications include speech recognition, speaker identification, and handwriting recognition.

REFERENCES

[1] W. H. Sumby and I. Pollack, "Visual contributions to speech intelligibility in noise," *Journal of the Acoustical Society of America*, vol. 26, pp. 212–215, 1954.

[2] K. K. Neely, "Effect of visual factors on the intelligibility of speech," *Journal of the Acoustical Society of America*, vol. 28, no. 6, pp. 1275–1277, 1956.

[3] C. A. Binnie, A. A. Montgomery, and P. L. Jackson, "Auditory and visual contributions to the perception of consonants," *Journal of Speech and Hearing Research*, vol. 17, pp. 619–630, 1974.

[4] D. Reisberg, J. McLean, and A. Goldfield, "Easy to hear but hard to understand: A lipreading advantage with intact auditory stimuli," in *Hearing by Eye: The Psychology of Lipreading*, B. Dodd and R. Campbell, Eds., pp. 97–113, Lawrence Erlbaum Associates, Hillsdale, NJ, USA, 1987.

[5] H. McGurk and J. MacDonald, "Hearing lips and seeing voices," *Nature*, vol. 264, pp. 746–748, 1976.

[6] D. W. Massaro, *Speech Perception by Ear and Eye: A Paradigm for Psychological Inquiry*, Lawrence Erlbaum Associates, Hillsdale, NJ, USA, 1987.

[7] R. Campbell and B. Dodd, "Hearing by eye," *Quarterly Journal of Experimental Psychology*, vol. 32, pp. 85–99, 1980.

[8] E. D. Petajan, *Automatic lipreading to enhance speech recognition*, Ph.D. dissertation, University of Illinois at Urbana-Champaign, Urbana, Ill, USA, 1984.

[9] A. J. Goldschen, *Continuous automatic speech recognition by lipreading*, Ph.D. dissertation, George Washington University, Washington, DC, USA, 1993.

[10] B. P. Yuhas, M. H. Goldstein, and T. J. Sejnowski, "Integration of acoustic and visual speech signals using neural networks," *IEEE Commun. Mag.*, vol. 27, no. 11, pp. 65–71, 1989.

[11] D. G. Stork, G. Wolff, and E. Levine, "Neural network lipreading system for improved speech recognition," in *Proc. IEEE International Joint Conference on Neural Networks (IJCNN '92)*, pp. 285–295, Baltimore, Md, USA, June 1992.

[12] C. Bregler and S. Omohundro, "Nonlinear manifold learning for visual speech recognition," in *Proc. IEEE 5th International Conference on Computer Vision (ICCV '95)*, pp. 494–499, Cambridge, Mass, USA, June 1995.

[13] P. Silsbee and A. Bovik, "Computer lipreading for improved accuracy in automatic speech recognition," *IEEE Trans. Speech Audio Processing*, vol. 4, no. 5, pp. 337–351, 1996.

[14] D. G. Stork and H. L. Lu, "Speechreading by Boltzmann zippers," in *Machines that Learn Workshop*, Snowbird, Utah, USA, April 1996.

[15] T. Chen, "Audiovisual speech processing," *IEEE Signal Processing Mag.*, vol. 18, no. 1, pp. 9–21, 2001.

[16] D. G. Stork and M. E. Hennecke, "Speechreading: An overview of image processing, feature extraction, sensory integration and pattern recognition techniques," in *Proc. 2nd International Conference on Automatic Face and Gesture Recognition*, pp. 16–26, Killington, Vt, USA, October 1996.

[17] S. W. Foo, Y. Lian, and L. Dong, "Recognition of visual speech elements using adaptively boosted hidden Markov models," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 14, no. 5, pp. 693–705, 2004.

[18] A. J. Goldschen, O. N. Garcia, and E. D. Petajan, "Continuous optical automatic speech recognition by lipreading," in *Proc. 28th Asilomar Conference on Signals, Systems, and Computers*, vol. 1, pp. 572–577, Pacific Grove, Calif, USA, October–November 1994.

[19] W. J. Welsh, A. D. Simon, R. A. Hutchinson, and S. Searby, "A speech-driven 'talking-head' in real time," in *Proc. Picture Coding Symposium*, pp. 7.6-1–7.6-2, Cambridge, Mass, USA, March 1990.

[20] P. L. Silsbee and A. C. Bovic, "*Visual lipreading by computer to improve automatic speech recognition accuracy*," Tec. Rep. TR-93-02-90, University of Texas Computer and Vision Research Center, Austin, Tex, USA, 1993.

[21] P. L. Silsbee and A. C. Bovic, "Medium vocabulary audiovisual speech recognition," in *NATO ASI New Advances and Trends in Speech Recognition and Coding*, pp. 13–16, Bubion, Granada, Spain, June–July 1993.

[22] M. J. Tomlinson, M. J. Russell, and N. M. Brooke, "Integrating audio and visual information to provide highly robust speech recognition," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP '96)*, vol. 2, pp. 821–824, Atlanta, Ga, USA, May 1996.

[23] J. Luettin, N. A. Thacker, and S. W. Beet, "Speechreading using shape and intensity information," in *Proc. 4th International Conference on Spoken Language Processing (ICSLP '96)*, pp. 58–61, Philadelphia, Pa, USA, October 1996.

[24] X. Z. Zhang, R. M. Mersereau, and M. A. Clements, "Audio-visual speech recognition by speechreading," in *Proc. 14th International Conference on Digital Signal Processing (DSP '02)*, vol. 2, pp. 1069–1072, Santorini, Greece, July 2002.

[25] G. Gravier, G. Potamianos, and C. Neti, "Asynchrony modeling for audio-visual speech recognition," in *Proc. International Conference on Human Language Technology (HLT '02)*, San Diego, Calif, USA, March 2002, available on proceeding CD.

[26] A. V. Nefian, L. Liang, X. Pi, X. Liu, and K. Murphy, "Dynamic Bayesian networks for audio-visual speech recognition," *EURASIP Journal on Applied Signal Processing*, vol. 2002, no. 11, pp. 1274–1288, 2002.

[27] S. Dupont and J. Luettin, "Audio-visual speech modeling for continuous speech recognition," *IEEE Trans. Multimedia*, vol. 2, no. 3, pp. 141–151, 2000.

[28] S. W. Foo and L. Dong, "A boosted multi-HMM classifier for recognition of visual speech elements," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP '03)*, vol. 2, pp. 285–288, Hong Kong, China, April 2003.

[29] J. J. Williams and A. K. Katsaggelos, "An HMM-based speech-to-video synthesizer," *IEEE Trans. Neural Networks*, vol. 13, no. 4, pp. 900–915, 2002.

[30] J. Luettin, N. A. Thacker, and S. W. Beet, "Speechreading using shape and intensity information," in *Proc. 4th International Conference on Spoken Language Processing (ICSLP '96)*, vol. 1, pp. 58–61, Philadelphia, Pa, USA, October 1996.

[31] I. Matthews, T. F. Cootes, J. A. Bangham, S. Cox, and R. Harvey, "Extraction of visual features for lipreading," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 24, no. 2, pp. 198–213, 2002.

[32] A. Adjoudani and C. Benoît, "On the integration of auditory and visual parameters in an HMM-based ASR," in *Speech Reading by Humans and Machines*, D. G. Stork and M. E. Hennecke, Eds., NATO ASI Series, pp. 461–471, Springer Verlag, Berlin, Germany, 1996.

[33] E. Owens and B. Blazek, "Visemes observed by hearing impaired and normal hearing adult viewers," *Journal of Speech Hearing and Research*, vol. 28, pp. 381–393, 1985.

[34] S. W. Foo and L. Dong, "Recognition of visual speech elements using hidden Markov models," in *Proc. 3rd IEEE Pacific Rim Conference on Multimedia (PCM '02)*, pp. 607–614, December 2002.

[35] C. Binnie, A. Montgomery, and P. Jackson, "Auditory and visual contributions to the perception of consonants," *Journal of Speech Hearing and Research*, vol. 17, pp. 619–630, 1974.

[36] A. M. Tekalp and J. Ostermann, "Face and 2-D mesh animation in MPEG-4," *Signal Processing: Image Communication*, vol. 15, no. 4-5, pp. 387–421, 2000, special issue on MPEG-4.

[37] S. Morishima, S. Ogata, K. Murai, and S. Nakamura, "Audio-visual speech translation with automatic lip synchronization and face tracking based on 3-D head model," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP '02)*, vol. 2, pp. 2117–2120, Orlando, Fla, USA, May 2002.

[38] S. W. Foo, Y. Lian, and L. Dong, "A two-channel training algorithm for hidden Markov model to identify visual speech elements," in *Proc. IEEE Int. Symp. Circuits and Systems (ISCAS '03)*, vol. 2, pp. 572–575, Bangkok, Thailand, May 2003.

[39] L. E. Baum and T. Petrie, "Statistical inference for probabilistic functions of finite state Markov chains," *Annals of Mathematical Statistics*, vol. 37, pp. 1554–1563, 1966.

[40] L. E. Baum and G. R. Sell, "Growth functions for transformations on manifolds," *Pacific Journal of Mathematics*, vol. 27, no. 2, pp. 211–227, 1968.

[41] T. Petrie, "Probabilistic functions of finite state Markov chains," *Annals of Mathematical Statistics*, vol. 40, no. 1, pp. 97–115, 1969.

[42] L. E. Baum, T. Petrie, G. Soules, and N. Weiss, "A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains," *Annals of Mathematical Statistics*, vol. 41, pp. 164–171, 1970.
[43] L. E. Baum, "An inequality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes," in *Inequalities*, vol. 3, pp. 1–8, Academic Press, New York, NY, USA, 1972.
[44] J. K. Baker, "The DRAGON system—An overview," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 23, no. 1, pp. 24–29, 1975.
[45] F. Jelinek, "Continuous speech recognition by statistical methods," *Proc. IEEE*, vol. 64, no. 4, pp. 532–556, 1976.
[46] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc. IEEE*, vol. 77, no. 2, pp. 257–286, 1989.
[47] L. R. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition*, Prentice Hall, Englewood Cliffs, NJ, USA, 1993.
[48] L. R. Bahl, P. F. Brown, P. V. de Souza, and R. L. Mercer, "Maximum mutual information estimation of hidden Markov model parameters for speech recognition," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP '86)*, pp. 49–52, Tokyo, Japan, April 1986.
[49] G. J. McLanchlan and T. Krishnan, *The EM Algorithm and Extensions*, Wiley Series in Probability and Statistics, John Wiley & Sons, New York, NY, USA, 1997.
[50] R. J. Schalkoff, *Pattern Recognition: Statistical, Structural and Neural Approaches*, John Wiley & Sons, New York, NY, USA, 1992.
[51] X. Z. Zhang, C. Broun, R. M. Mersereau, and M. A. Clements, "Automatic speechreading with applications to human-computer interfaces," *EURASIP Journal on Applied Signal Processing*, vol. 2002, no. 11, pp. 1228–1247, 2002, special issue on Joint Audio-Visual Speech Processing.
[52] T. W. Lewis and D. M. W. Powers, "Lip feature extraction using red exclusion," *Selected Papers from the Pan-Sydney Workshop on Visualization*, vol. 2, pp. 61–67, Sydney, Australia, 2000.
[53] A. Yuille and P. Hallinan, "Deformable templates," in *Active Vision*, A. Blake and A. Yuille, Eds., pp. 21–38, MIT Press, Cambridge, Mass, USA, 1992.
[54] T. Coianiz, L. Torresani, and B. Caprile, "2D deformable models for visual speech analysis," in *Speech Reading by Humans and Machines*, D. G. Stork and M. E. Hennecke, Eds., NATO ASI Series, pp. 391–398, Springer Verlag, New York, NY, USA, 1996.
[55] M. E. Hennecke, K. V. Prasad, and D. G. Stork, "Using deformable templates to infer visual speech dynamics," in *Proc. 28th Asilomar Conference on Signals, Systems and Computers*, vol. 1, pp. 578–582, Pacific Grove, Calif, USA, October–November 1994.

Liang Dong received the B. Eng. degree in electronic engineering from Beijing University of Aeronautics and Astronautics, China, in 1997, and the M. Eng. degree in electrical engineering from the Second Academy of China Aerospace in 2000. Currently, he is a Ph.D. candidate in the National University of Singapore and working in the Institute for Infocomm Research, Singapore. His research interests include speech processing, image processing, and video processing.

Say Wei Foo received the B. Eng. degree in electrical engineering from the University of Newcastle, Australia, in 1972, the M.S. degree in industrial and systems engineering from the University of Singapore in 1979, and the Ph.D. degree in electrical engineering from Imperial College, University of London, in 1983. From 1972 to 1973, he was with the Electrical Branch, Lands and Estates Department, Ministry of Defense, Singapore. From 1973 to 1992, he worked in the Electronics Division of the Defense Science Organization, Singapore, where he conducted research and carried out development work on security equipment. From 1992 to 2001, he was the Associate Professor with the Department of Electrical and Computer Engineering, National University of Singapore. In 2002, he joined the School of Electrical and Electronic Engineering, Nanyang Technological University. He has authored and coauthored over one hundred published articles. His research interests include speech signal processing, speaker recognition, and musical note recognition.

Yong Lian received the B.S. degree from the School of Management, Shanghai Jiao Tong University, China, in 1984, and the Ph.D. degree from the Department of Electrical Engineering, National University of Singapore, Singapore, in 1994. He was with the Institute of Microcomputer Research, Shanghai Jiao Tong University, Brighten Information Technology Ltd., SyQuest Technology International, and Xyplex Inc. from 1984 to 1996. He joined the National University of Singapore in 1996 where he is currently an Associate Professor in the Department of Electrical and Computer Engineering. His research interests include digital filter design, VLSI implementation of high-speed digital systems, biomedical instrumentation, and RF IC design. Dr. Lian received the 1996 IEEE Circuits and Systems Society's Guillemin-Cauer Award for the best paper published in IEEE Transactions on Circuits and Systems Part II. He currently serves as an Associate Editor for the IEEE Transactions on Circuits and Systems Part II and has been an Associate Editor for Circuits, Systems and Signal Processing since 2000. Dr. Lian serves as the Secretary and Member of IEEE Circuits and Systems Society's Biomedical Circuits and Systems Technical Committee and Digital Signal Processing Technical Committee, respectively.