<table>
<thead>
<tr>
<th>Title</th>
<th>A boosted multi-HMM classifier for recognition of visual speech elements (Accepted version)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Author(s)</td>
<td>Foo, Say Wei; Dong, Liang</td>
</tr>
<tr>
<td>Date</td>
<td>2003</td>
</tr>
<tr>
<td>URL</td>
<td><a href="http://hdl.handle.net/10220/4590">http://hdl.handle.net/10220/4590</a></td>
</tr>
</tbody>
</table>

© 2003 IEEE. Personal use of this material is permitted. However, permission to reprint/republish this material for advertising or promotional purposes or for creating new collective works for resale or redistribution to servers or lists, or to reuse any copyrighted component of this work in other works must be obtained from the IEEE. This material is presented to ensure timely dissemination of scholarly and technical work. Copyright and all rights therein are retained by authors or by other copyright holders. All persons copying this information are expected to adhere to the terms and constraints invoked by each author's copyright. In most cases, these works may not be reposted without the explicit permission of the copyright holder. http://www.ieee.org/portal/site This material is presented to ensure timely dissemination of scholarly and technical work. Copyright and all rights therein are retained by authors or by other copyright holders. All persons copying this information are expected to adhere to the terms and constraints invoked by each author's copyright. In most cases, these works may not be reposted without the explicit permission of the copyright holder.
A BOOSTED MULTI-HMM CLASSIFIER FOR RECOGNITION OF 
VISUAL SPEECH ELEMENTS

Say Wei Foo  
School of Electrical & Electronic Engineering  
Nanyang Technological University  
Singapore 639798  
eswfoo@ntu.edu.sg

Liang Dong  
Dept. of Electrical & Computer Engineering  
National University of Singapore  
Singapore 119260  
engp0564@nus.edu.sg

ABSTRACT

A novel boosted classifier using multiple Hidden Markov Models (HMMs) is reported in this paper. The composite HMMs are specially trained to highlight certain group of training samples with the application of adaptive boosting technique. Experiments were carried out to identify the basic visual speech elements in English using the proposed boosted classifier. Comparing the results obtained using the proposed classifier and those obtained using the traditional single HMM classifier, it may be said that the proposed system is significantly better in terms of accuracy and robustness.

1. INTRODUCTION

Incorporation of visual clues into an acoustic speech recognition system becomes an interested research area of multimedia technique in recent years. However, the concept of interpreting speech content through lip movement can be dated to early years. In 1970’s, researchers had shown interest in the bimodal aspects of human speech. The well-known “McGurk effect” indicates that the human perception of the speech exists in both audio signal and visual signal [1]. The advantage brought by the visual speech analysis is that it is not affected by acoustic noise and cross talk among speakers. Many experiments proved that the incorporation of visual information could lead to significant improvement in speech recognition, especially under unfavorable environment [2].

Most previous studies on visual speech signal are based on the Hidden Markov Models (HMMs). In 1990, Welsh et al. studied audio-to-visual mapping using HMMs [3]. In 1996, Tomlinson et al. developed cross-product HMM topology for visual speech analysis [4]. And later Luettin et al. used HMMs with an early integration strategy for both speaker-independent digits recognition and speaker-independent connected-digit recognition [5]. Their work achieved some success in the respective applications of lip reading. However, the modeling of the basic visual speech elements still remains a problem due to the high elasticity of the human lip. How to effectively characterize them is the key point of constructing a word-level or connected-word visual speech recognizer.

The visual speech elements corresponding to English phonemes, namely visemes, are easily distorted by their context. As a result, single HMM classifier, e.g. the maximum likelihood (ML) model trained with the Baum-Welch method, is sometimes incompetent to cover such erratic changes. To discriminate the visemes with enough accuracy and robustness, the adoption of multiple HMMs becomes necessary. In this paper, the adaptive boosting (AdaBoost) approach is applied to the HMMs to build multiple-HMM classifiers. Such classifiers are applied to classify the visemes. Experimental results show that the boosted classifier excels the single HMM classifier in discriminating the visemes.

2. THE BAUM-WELCH TRAINING ALGORITHM AND ADAPTIVE BOOSTING TECHNIQUE

Hidden Markov Model is a powerful stochastic tool of modeling and identifying sequential signals. In applications using HMM, the Baum-Welch training algorithm is the most popular due to its fast rate of convergence and ease of implementation. Given training samples \(\{y_1,d_1\}, \{y_2,d_2\} \ldots \{y_M,d_M\}\), where \(y_i\) is the training sequence and \(d_i\) is the corresponding category label, the HMM for each class is obtained with the expectation-maximization (EM) recursions [6]. To get a reliable model, multiple observations are usually adopted to estimate the parameters. Let \(\{O_1,O_2,\cdots,O_M\}\) be the symbol set and \(\{S_1,S_2,\cdots,S_M\}\) be the state set, an \(N\)-state-M-symbol HMM \(\theta(\pi,A,B)\) is determined by the initial
state probability matrix $\pi_\theta = \{\pi_i\}$, state transition matrix $A = [a_{ij}]$ and symbol emission matrix $B = [b_{ij}]$. Assume that the training set of $d_p$ consists of $R_p$ sequences $X_p = \{x_{1p}, x_{2p}, \ldots x_{R_p}\}$ and $x_i = \{x_{1i}, x_{2i}, \ldots x_{T_i}\}$, the parameters of $\theta$ are estimated as follows.

\[
\tilde{a}_{ij} = \frac{\sum_{k=1}^{K} \sum_{i=1}^{T_k} a_{ij}^k(i)a_{ij}^k(x_{T_i})\beta_{T_i}^k(j)}{P_k \cdot \sum_{k=1}^{T_k} \sum_{i=1}^{T_k} a_{ij}^k(i)\beta_{T_i}^k(j)} \tag{1.a}
\]

\[
\tilde{b}_j(O_t) = \frac{\sum_{k=1}^{K} \sum_{i=1}^{T_k} a_{ij}^k(j)\beta_{T_i}^k(j)}{P_k \cdot \sum_{k=1}^{T_k} \sum_{i=1}^{T_k} a_{ij}^k(j)\beta_{T_i}^k(j)} \tag{1.b}
\]

where $P_k = P(x_k | \theta)$ is the likelihood scored by $\theta$, $T_k$ is the length of the $k$-th training sample, $a_{ij}^k(i) = P(x_{1i}^k, x_{2i}^k, \ldots x_{T_i}^k | s_i = S_j | \theta)$ is the forward variable and $\beta_{T_i}^k(i) = P(x_{T_i}, x_{T_i-1}, \ldots x_{i}^k | s_i = S_j, \theta)$ is the backward variable. Arslan and Hansen prove that if weights are added to the training samples, the convergence of the ML estimation still holds [7]. Equations (1.a) and (1.b) then become

\[
\tilde{a}_{ij} = \sum_{k=1}^{K} \frac{D_k \cdot \sum_{i=1}^{T_k} a_{ij}^k(i)a_{ij}^k(x_{T_i})\beta_{T_i}^k(j)}{P_k \cdot \sum_{k=1}^{T_k} \sum_{i=1}^{T_k} a_{ij}^k(i)\beta_{T_i}^k(j)} \tag{2.a}
\]

\[
\tilde{b}_j(O_t) = \sum_{k=1}^{K} \frac{D_k \cdot \sum_{i=1}^{T_k} a_{ij}^k(j)\beta_{T_i}^k(j)}{P_k \cdot \sum_{k=1}^{T_k} \sum_{i=1}^{T_k} a_{ij}^k(j)\beta_{T_i}^k(j)} \tag{2.b}
\]

where $D_k$ is the weight assigned to the $k$-th sample. A “biased” HMM is obtained by calculating the above estimates iteratively until a local maximum point is attained. In this paper, we shall refer to this training strategy as the modified Baum-Welch algorithm.

Adaptive boosting (AdaBoost) is a popular technique of “boosting” a weak classifier into a strong one. Assume that $\{y_1, d_1\}, \{y_2, d_2\} \ldots \{y_M, d_M\}$ are $M$ training samples in a two-class identification problem, where $d_i = \{-1, +1\}$, the principles of the method are generalized as follows [8]:

1. A uniform distribution $D_0(i) = 1/M$ is assigned to the $M$ training samples $\{y_1, d_1\}, \{y_2, d_2\} \ldots \{y_M, d_M\}$.
2. A base training algorithm is called to train the classifier using the distribution $D_0$.
3. Obtain the hypothesis $h_i : y \rightarrow \{-1, +1\}$ and calculate the classification error $e_i = P(h_i(x_i) \neq y_i)$.
4. Compute the weight of the $t$-th classifier $\alpha_t = \frac{1}{2} \ln(\frac{1-e_t}{e_t})$.
5. Update the distribution using (3),

\[
D_{t+1}(i) = \frac{D_t(i)}{Z_t} \times \begin{cases} e^{-\alpha_t} & \text{if } h_i(x_i) = y_i \\ e^{\alpha_t} & \text{if } h_i(x_i) \neq y_i \end{cases} \tag{3}
\]

where $Z_t$ is a normalization factor to make $D_{t+1}$ a distribution.

The above steps constitute a boosting epoch. A boosted classifier is obtained by iteratively implementing the epoch until the error rate of the classifier obtained exceeds 0.5. The final decision is made by summarizing the hypotheses of all the classifiers:

\[
H(x) = \text{sign} \left( \sum_{t=1}^{T} \alpha_t h_t(x) \right) \tag{4}
\]

For the above strategy, Freund and Schapire prove that if only each classifier has the classification error less than 0.5, the overall error rate will decrease exponentially [9]. The boosted (synthesized) classifier may have arbitrary low error rate given sufficient data.

3. PROCEDURES FOR BOOSTING HMM

To apply AdaBoost to HMMs, the following three issues need to be resolved.

1. Which base training algorithm is used to train the HMMs? The modified Baum-Welch estimation as illustrated in (2) gives a good solution. This method maintains weights over the training set while estimating the parameters of the HMM.
2. How to obtain the classification error? To solve this problem, let’s first recall the classification using the HMMs. Assume that in a $K$-class identification problem, Class $d_i$ $(i=1, 2, \ldots, K)$ has $l_i$ HMM classifiers $\{\theta_1^i, \theta_2^i, \ldots, \theta_{l_i}^i\}$ and $R_i$ samples $X_i = \{x_{1i}, x_{2i}, \ldots x_{R_i}\}$. During recognition, the probabilities of the input sequence under all the HMMs are computed and compared with one another. The one that gives the maximum likelihood is chosen as the source.

\[
\theta^* = \arg \max_{\theta} P(x | \theta^i) \quad \forall i = 1, 2, \ldots, K, \quad j = 1, 2, \ldots, l_i \tag{5}
\]

A Boolean decision (one vs. the rest) is made in this way. If the correct model gives greater likelihood value than the others, the result is correct; otherwise, an error occurs. As a result, the following hypothesis is made upon an HMM $\theta^i_p$, the $p$-th classifier of Class $i$:
\[ h_p^j(x_q) = \begin{cases} 1 & \text{if } P(x_q | \theta_p^j) > P(x_q | \theta_p^i) \\ -1 & \text{otherwise} \end{cases} \quad (6) \]

where \( x_q \in X_i \) and \( j \neq i \). Consider the weights assigned to the samples, the error rate of \( \theta_p^j \) is estimated by summarizing the hypotheses over all the training samples of \( d_i \).

\[ e_p^j = \frac{D(x_q)E(h_p^j(x_q) = -1)}{R_i} = \sum_{q=1}^{s_i} \frac{D(x_q)}{R_i} \quad (7) \]

It can be observed that the classification error of an HMM is associated with the other HMMs. As a result, with the propagation of the HMMs in boosting, the error rate of the existing HMMs will change. To make the new HMM valid for boosting (error rate < 0.5), not only its own error rate should be computed, but also all the existing HMMs should be verified.

3. How to make the final decision? The decision of the boosted HMM classifier is formulated by summarizing those of the individual HMMs. This problem will be discussed in the subsequent sections.

Take the above \( K \)-class identification problem as an example, HMM boosting takes the following steps:

1. A uniform distribution \( D_1(k,j) = 1/R_k \) \((j=1,2,\ldots R_k)\) is assigned to the \( R_k \) examples \(-x_1^k, x_2^k, \ldots, x_{R_k}^k\) of Class \( d_i \) \((k=1,2,\ldots K)\).

2. The modified Baum-Welch algorithm (2) is called to train the \( t \)-th HMM of Class \( d_i - \theta_t^i \) using the distribution \( D_t(k,j) \).

3. The binary hypothesis \( h^k : y \rightarrow \{-1,1\} \) for \( \theta_k^i \) is formulated using (6). The error rates of \( \theta_k^i \) and all the existing HMMs are computed using (7). If any of them exceeds 0.5, it manifests that the new model \( \theta_k^i \) is invalid. In this situation, the boosting is passed to the next class \( k = \begin{cases} k+1 & \text{if } k < K \\ 1 & \text{if } k = K \end{cases} \) and Step 2 is repeated.

4. Calculate \( \alpha_k^i = \frac{1}{2} \ln\left(\frac{1-e_k^i}{e_k^i}\right) \) for \( \theta_k^i \), where \( e_k^i \) is the error rate. If the error rate of some existing HMM changes, the corresponding \( \alpha \) is also modified.

5. Update the distribution using (8),

\[ D_{t+1}(k,j) = D_t(k,j) e^{\alpha_k^i h^k(x_i)} \quad (8) \]

where \( h^k(x_i) = 1 \) or -1 and \( Z_i \) is the normalization factor.

In the above procedure, the boosting of certain class may terminate earlier than that of the others. The outcome is that different classes may have different numbers of HMMs. As a result, the final decision should be normalized with the number of the HMMs. Assume \( \Theta_k \) is the boosted (synthesized) classifier of class \( d_k \), which is comprised of \( L_k \) HMMs \(-\{\theta_k^1, \theta_k^2, \ldots, \theta_k^{L_k}\}\), the normalized log likelihood of an observed sequence \( x \) given \( \Theta_k \) is defined as follows.

\[ \overline{P}(x | \Theta_k) = \frac{1}{L_k} \sum_{j=1}^{L_k} \alpha_k^j \log P(x | \theta_k^j) \quad (9) \]

The final decision is made by comparing the \( \overline{P}(x | \Theta_k) \) for all the \( K \) classes. The one that gives the maximum value is chosen as the identity of the input.

\[ ID(x) = \arg \max_k \overline{P}(x | \Theta_k) \quad (1 \leq k \leq K) \quad (10) \]

4. APPLICATION TO VISEME RECOGNITION

The above boosted HMM classifier is applied to classify the visemes in English. In our system, the raw data describing a viseme is the image sequence sampled at 25Hz. For each frame, which reveals the lip area of the speaker as shown in Fig.1(a), eleven geometric measures as indicated in Fig.1(b) are extracted to form a feature vector. These geometric measures give the thickness, position and curvature of the lip. They are chosen as they uniquely determine the lip shape and best characterize the dynamics of lip movement.

Fig. 1  (a) original image  (b) extracted image
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: length of the tongue (if visible)

The collected feature vectors are put through normalization and principal component analysis (PCA). They are finally clustered into groups using K-means algorithm. For the experiments conducted in this paper, 128 clusters (code words) are used in the vector database (code book). They are taken to encode the input
observation sequences and build the symbol set of the discrete HMMs.

The visemes to be identified are those defined in MPEG-4 multimedia standards [10]. As mentioned above, visemes are liable to be distorted by their context. For example, the visual representation of the phoneme /ai/ is different in the words like and right. To cover the morphological variation of the viseme, 200 samples are carefully chosen from different word context. Among them, 100 samples are used to train the HMMs and the remaining 100 samples are used to test the performance of the classifiers. In our experiment, all the HMMs are three-state left-right models and each viseme classifier is comprised of 15-20 HMMs.

To assess the performance enhancement of the boosted HMM classifier, the recognition results of single HMM classifier, trained using the Baum-Welch method, are also obtained for comparison. The classification errors (False Reject Rate) of the two types of classifiers in identifying the fourteen visemes are listed in Table 1.

Table 1. Classification errors (FRR) of the single HMM classifier and the boosted HMM classifier

<table>
<thead>
<tr>
<th>Viseme Categories</th>
<th>Single HMM Classifier</th>
<th>Boosted HMM Classifier</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 p, b, m</td>
<td>21%</td>
<td>15%</td>
</tr>
<tr>
<td>2 f, v</td>
<td>26%</td>
<td>22%</td>
</tr>
<tr>
<td>3 T, D</td>
<td>33%</td>
<td>18%</td>
</tr>
<tr>
<td>4 t, d</td>
<td>42%</td>
<td>16%</td>
</tr>
<tr>
<td>5 k, g</td>
<td>17%</td>
<td>16%</td>
</tr>
<tr>
<td>6 tS, dZ, S</td>
<td>21%</td>
<td>9%</td>
</tr>
<tr>
<td>7 s, z</td>
<td>44%</td>
<td>17%</td>
</tr>
<tr>
<td>8 n, l</td>
<td>79%</td>
<td>33%</td>
</tr>
<tr>
<td>9 r</td>
<td>54%</td>
<td>37%</td>
</tr>
<tr>
<td>10 A:</td>
<td>18%</td>
<td>5%</td>
</tr>
<tr>
<td>11 e</td>
<td>33%</td>
<td>7%</td>
</tr>
<tr>
<td>12 I</td>
<td>10%</td>
<td>2%</td>
</tr>
<tr>
<td>13 Q</td>
<td>35%</td>
<td>11%</td>
</tr>
<tr>
<td>14 U</td>
<td>9%</td>
<td>7%</td>
</tr>
</tbody>
</table>

Results indicate that for the set of visemes tested, the boosted HMM classifier gives better accuracy to identify the visemes than single HMM classifier.

5. CONCLUSION

In this paper, a novel system which applies adaptive boosting algorithm to the conventional Hidden Markov Model is investigated. Results of experiments using the boosted classifier in lip reading, i.e., the recognition of visemes, show that significant improvements in accuracy and robustness are achieved compared with the results obtained using the single HMM classifier. The proposed method can readily be extended to many other recognition problems especially when the observed data have erratic distribution, for examples, speech recognition, handwriting recognition and speaker identification.

6. REFERENCES