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 $\{y_1, d_1\}$, $\{y_2, d_2\}$ … $\{y_M, d_M\}$, 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 $\pi = [\pi_i]$, 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 $X_p = \{x_1, x_2, \cdots x_{R_p}\}$ and $x_i = \{x_1^i, x_2^i, \cdots x_{T_i}^i\}$, the parameters of θ are estimated as follows.

$$\bar{a}_{ij} = \frac{\sum_{k=1}^{R_p} \frac{1}{P_k} \sum_{t=1}^{T_k-1} \alpha_t^k(i) a_{ij} b_j(x_{t+1}^k) \beta_{t+1}^k(j)}{\sum_{k=1}^{R_p} \frac{1}{P_k} \sum_{t=1}^{T_k-1} \alpha_t^k(i) \beta_{t+1}^k(j)} \quad (1.a)$$

$$\bar{b}_j(O_l) = \frac{\sum_{k=1}^{R_p} \frac{1}{P_k} \sum_{s.t.x_t^k = O_l}^{T_k-1} \alpha_t^k(j) \beta_t^k(j)}{\sum_{k=1}^{R_p} \frac{1}{P_k} \sum_{t=1}^{T_k-1} \alpha_t^k(j) \beta_t^k(j)} \quad (1.b)$$

where $P_k = P(x_k \mid \theta)$ is the likelihood scored by θ, T_k is the length of the k-th training sample, $\alpha_t^k(i) = P(x_1^k, x_2^k, \cdots x_t^k, s_t = S_i \mid \theta)$ is the forward variable and $\beta_t^k(i) = P(x_{t+1}^k, x_{t+2}^k, \cdots x_{T_k}^k \mid s_t = S_i, \theta)$ is the backward variable. Arslan and Hansen prove that if weights are added to the training samples, the convergency of the ML estimation still holds [7]. Equations (1.a) and (1.b) then become

$$\bar{a}_{ij} = \frac{\sum_{k=1}^{K} \frac{D_k}{P_k} \sum_{t=1}^{T_k-1} \alpha_t^k(i) a_{ij} b_j(x_{t+1}^k) \beta_{t+1}^k(j)}{\sum_{k=1}^{K} \frac{D_k}{P_k} \sum_{t=1}^{T_k-1} \alpha_t^k(i) \beta_{t+1}^k(j)} \quad (2.a)$$

$$\bar{b}_j(O_l) = \frac{\sum_{k=1}^{K} \frac{D_k}{P_k} \sum_{s.t.x_t^k = O_l}^{T_k-1} \alpha_t^k(j) \beta_t^k(j)}{\sum_{k=1}^{K} \frac{D_k}{P_k} \sum_{t=1}^{T_k-1} \alpha_t^k(j) \beta_t^k(j)} \quad (2.b)$$

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 $\{y_1, d_1\}$ $\{y_2, d_2\}$ … $\{y_M, d_M\}$ 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 $D_1(i) = 1/M$ is assigned to the M training samples $\{y_1, d_1\}$ $\{y_2, d_2\}$ … $\{y_M, d_M\}$.

2. A base training algorithm is called to train the classifier using the distribution D_t.

3. Obtain the hypothesis $h_t : y \to \{-1, +1\}$ and calculate the classification error $\varepsilon_t = P(h_t(x_i) \neq d_i)$.

4. Compute the weight of the t-th classifier $\alpha_t = \frac{1}{2} \ln(\frac{1-\varepsilon_t}{\varepsilon_t})$.

5. Update the distribution using (3),

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

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:

$$H(x) = sign(\sum_{t=1}^{T} \alpha_t h_t(x)) \quad (4)$$

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,...K) has l_i HMM classifiers– $\{\theta_1^i, \theta_2^i, \cdots, \theta_{l_i}^i\}$ and R_i samples– $X_i = \{x_1, x_2, \cdots x_{R_i}\}$. 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.

$$\theta^* = \arg\max_{\theta} P(x \mid \theta_j^i) \quad \forall i = 1,2,\cdots K,\ j = 1,2,\cdots l_i \quad (5)$$

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 θ_p^i, the p-th classifier of Class i:

$$h_p^i(x_q) = \begin{cases} 1 & if\ P(x_q \mid \theta_p^i) > P(x_q \mid \theta_k^j) \\ -1 & otherwise \end{cases} \quad (6)$$

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

$$\varepsilon_p^i = \frac{D(x_q)E(h_p^i(x_q) = -1)}{R_i} = \sum_{\substack{q=1 \\ s.t.h_p^i(x_q)=-1}}^{R_i} \frac{D(x_q)}{R_i} \quad (7)$$

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,…R_k) is assigned to the R_k examples – $x_1^k, x_2^k, \cdots x_{R_k}^k$ of Class d_k (k=1,2,…K).

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

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

4. Calculate $\alpha_t^k = \frac{1}{2}\ln(\frac{1-\varepsilon_t^k}{\varepsilon_t^k})$ for θ_t^k, where ε_t^k is the error rate. If the error rate of some existing HMM changes, the corresponding α is also modified.

5. Update the distribution using (8),

$$D_{t+1}(k, j) = \frac{D_t(k, j)}{Z_t} e^{\alpha_t^k h_t^k(x_j)} \quad (8)$$

where $h_t^k(x_j) = 1$ or -1 and Z_t 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 Θ_k is the boosted (synthesized) classifier of class d_k, which is comprised of L_k HMMs - $\{\theta_1^k, \theta_2^k, \cdots \theta_{L_k}^k\}$, the normalized log likelihood of an observed sequence x given Θ_k is defined as follows.

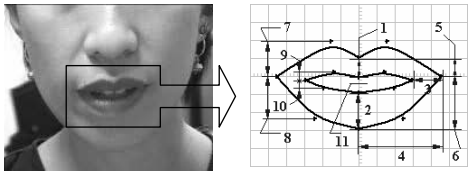
$$\overline{P}(x \mid \Theta_k) = \frac{1}{L_k} \sum_{t=1}^{L_k} \alpha_t^k \log P(x \mid \theta_t^k) \quad (9)$$

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

$$ID(x) = \arg\max_k \overline{P}(x \mid \Theta_k) \quad (1 \le k \le K) \quad (10)$$

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.



(a) (b)

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

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

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

[1] H. McGurk and J. MacDonald, "Hearing lips and seeing voices," *Nature*, Vol. 264, no. 5588, pp. 756-748, 1976

[2] D. Reisberg, J. McLean and A. Goldfield, "Easy to hear, but hard to understand: A lipreading advantage with intact auditory stimuli," *Hearing by Eye: The Psychology of Lipreading*, London: Lawrence Erlbaum, pp. 97-113, 1987

[3] W. J. Welsh, A. D. Simons, R. A. Hutchinson and S. Searby, "A speech-driven 'talking – head' in real time," *Proc. of Picture Coding Symposium*, pp. 7.6-1-7.2-2 1990

[4] M. Tomlinson, M. Russell and N. Brooke, "Integrating audio and visual information to provide highly robust speech recognition," *ICASSP*, Vol. 2, pp. 821-824, 1996

[5] J. Luettin, N. A. Thacker and S. W. Beet, "Speechreading using shape and intensity information," *Int. Conf. on Spoken Language Processing*, pp. 58-61, 1996

[6] L. R. Rabiner, "A tutorial on Hidden Markov Models and selected applications in speech recognition", *Proc. IEEE*, Vol. 77, No. 2, pp 257-286, Feb. 1989

[7] Levent M. Arslan and John H. L. Hansen, "Selective training in Hidden Markov Model recognition," *IEEE Trans. On Speech and Audio Processing*, Vol. 7 No. 1, pp. 46-54, Jan. 1999

[8] Robert E. Schapire, "A brief introduction to boosting," *Proc. of the Sixteenth International Joint Conference on Artificial Intelligence*, pp. 1401-1405, 1999

[9] Y. Freund and Robert E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *Journal of Computer and system Science*, 55(1). pp. 119-139, Aug. 1997

[10] M. Tekalp and J. Ostermann, "Face and 2-D mesh animation in MPEG-4," *Image Communication J*. Aug. 1999

[11] L. R. Rabiner and B. H. Juang, "Fundamentals of speech recognition," *Prentice Hall Signal Processing Series*, 1993

[12] Tsuhan Chen and Ram R. Rao, "Audio-visual Integration in Multimodal Communication," *Proc. IEEE*, vol. 86, No.5 pp. 837-852, May, 1998