Title	A supervised two-channel learning method for hidden Markov model and application on lip reading(Accepted version)
Author(s)	Foo, Say Wei; Dong, Liang
Citation	Foo, S. W., & Dong, L. (2002). A supervised two-channel learning method for hidden Markov model and application on lip reading. IEEE International Conference on Advanced Learning Technologies.
Date	2002
URL	http://hdl.handle.net/10220/4617
Rights	© 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.

A Supervised Two-channel Learning Method for Hidden Markov Model and Application on Lip Reading

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

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

Abstract

In this paper, a novel two-channel learning method for hidden Markov model (HMM) is proposed. This method is specially designed to train HMMs for fine recognition from similar observations. The prominent features of this method are 1.) the criterion function is based on the difference between training sequences, and 2.) a two-channel structure is adopted to maintain the validity of the HMM. This learning method has been applied on a viseme-level lip reading system. The result shows that the performance of the two channel approach is better than that of the maximum likelihood (ML) estimation.

1. Introduction

Lip reading is the technique of extracting speech information from the visual clues such as lip movement. Because of its great practical value, research on this area has attracted the attention of researchers. Time-delayed neural network [2], hidden Markov model [4] and fuzzy logic are applied to lip reading with some success. Among the techniques, HMM holds the biggest promise. Further improvement in HMM to make the individual HMM element more informative and to improve its overall performance is one of the areas of research. In this paper, an attempt in this direction is reported.

Unlike speech recognition, the ML estimation of HMM is difficult to be applied on lip reading due to some distinct features of lip motion. First, the movement of the lip is slight compared with its geometric measures, and such movement varies slowly over time. It indicates that the statistic features of lip motion chiefly concentrate around some stable states. Second, the basic lip motion elements corresponding to phonemes, namely visemes, have too many similarities with each other. Most visemes experience the same three-phase process during production: starting from closed mouth, peaking at half-opened mouth and ending with closed mouth. These features are unfavorable for recognizing with traditional HMMs such as ML model. To decide the source of an observation from similar ones with good credibility, a model with strong discriminating power is needed.

The supervised two-channel learning method presented in this paper is a possible solution to the problem. This method modifies the parameters to maximize the ratio between the probabilities of correct observation and incorrect ones. The discriminative power is therefore guaranteed. In this paper, the theoretic background and the process of the training strategy are detailed and its application on viseme recognition is introduced.

2. Two-channel modeling

Maximum likelihood model is the most popular HMM to be applied on sequential signal analysis. Given a labeled training sequence $x^T = \{x_1, x_2, \cdots x_T\}$, its ML model θ^x_{ML} satisfies the objective function (1).

$$\theta^x_{ML} = \arg\max_{\theta}[P(x^T \mid \theta)] \tag{1}$$

Take an N-state-M-observation model $\theta(\pi, A, B)$ as an example, where $\pi = [\pi_i]$ gives initial condition, $A = [a_{ij}]$ denotes state transition matrix and $B = [b_{ij}]$ denotes probability matrix, the Baum-Welch training strategy (EM method) modifies the parameters as follows:

$$\bar{a}_{ij} = \frac{E(number\ of\ state\ transition\ S_i \to S_j)}{E(number\ of\ transition\ from\ S_i)} \tag{2.a}$$

$$\bar{b}_{ij} = \frac{E(times\ of\ observing\ O_j\ in\ S_i)}{E(number\ of\ times\ in\ state\ S_i)} \tag{2.b}$$

$$\bar{\pi}_i = E(number\ of\ times\ in\ S_i\ at\ s_1) \tag{2.c}$$

where S_i identifies the i-th state and O_j is the j-th observation symbol. $P(x^T \mid \theta)$ is increased after each

expectation-maximization epoch, and finalized at a local maximum point.

From the above training strategy, it is manifested that the model is obtained without considering the relationship between the correct observation and incorrect ones. If there is another sequence, say $y^T = \{y_1, y_2, \cdots y_T\}$, which is emitted from a different source but is similar with x^T, the scored probability $P(y^T \mid \theta_{ML}^x)$ is likely to be big too.

The great likelihood of incorrect observations makes discrimination difficult. To solve this problem, a new objective function is put forward instead of (1). Given model θ, the ratio between the correct observation x^T and incorrect observation y^T is defined in (3).

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

The greater the value of $I(x^T, y^T, \theta)$, the better the model θ distinguish between x^T and y^T. As a result, the purpose of training is to find the $\widetilde{\theta}$ that maximizes $I(x^T, y^T, \theta)$,

$$\widetilde{\theta} = \arg\max_{\theta}[I(x^T, y^T, \theta)] \tag{4}$$

In most cases, the training is done to an existing HMM, e.g. an ML model. To keep the physical significance of the original model, only the coefficients in matrix B are modified while matrix A is left unchanged. If the probability constraint $\sum_{j=1}^{M} b_{ij} = 1 \quad (i = 1,2,\cdots N)$ is considered, maximizing of (3) is equivalent to maximizing the auxiliary function (5),

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

Where λ_i is the Lagrange multiplier.

Differentiate $F(x^T, y^T, \theta, \lambda)$ with respect to b_{ij} and set the result to 0, (5) gives,

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

If the solutions of b_{ij} are positive, $F(x^T, y^T, \theta, \lambda)$ will reach the maximum value. This condition is guaranteed by the training strategy discussed later. In (6), $\log P(x^T \mid \theta)$ and $\log P(y^T \mid \theta)$ are computed by summing up all the possibilities over time T.

$$\log P(x^T \mid \theta) = \sum_{\tau=1}^{T} \log \sum_{i_\tau=1}^{N} P(s_\tau = S_{i_\tau}) b_{i_\tau}(x_\tau^T) \tag{7.a}$$

and

$$\log P(y^T \mid \theta) = \sum_{\tau=1}^{T} \log \sum_{i_\tau=1}^{N} P(s_\tau = S_{i_\tau}) b_{i_\tau}(y_\tau^T) \tag{7.b}$$

If we carry out the indicated differentiation $\frac{\partial \log P(x^T \mid \theta)}{\partial b_{ij}}$ and $\frac{\partial \log P(y^T \mid \theta)}{\partial b_{ij}}$ in (6), we have, after some manipulation,

$$\frac{\partial \log P(x^T \mid \theta)}{\partial b_{ij}} = \frac{\sum_{\tau=1}^{T} P(s_\tau^T = S_i, x_\tau^T = O_j \mid \theta, x^T)}{b_{ij}} = \frac{E(S_i, O_j \mid \theta, x^T)}{b_{ij}} \tag{8.a}$$

and

$$\frac{\partial \log P(y^T \mid \theta)}{\partial b_{ij}} = \frac{E(S_i, O_j \mid \theta, y^T)}{b_{ij}} \tag{8.b}$$

The maximum point of $I(x^T, y^T, \theta)$ is obtainable by solving (9),

$$\frac{E(S_i, O_j \mid \theta, x^T) - E(S_i, O_j \mid \theta, y^T)}{b_{ij}} = \lambda_i \quad 1 \le j \le M \tag{9}$$

However, we cannot modify b_{ij} using (9) directly because **i.)** the numerator may be less than or equal to 0, and **ii.)** the unknown item b_{ij} also exists in calculating $E(S_i, O_j \mid \theta, x^T)$ and $E(S_i, O_j \mid \theta, y^T)$. Equation (9) only suggests a weighting gradient: for greater expectation of $E(S_i, O_j \mid \theta, x^T) - E(S_i, O_j \mid \theta, y^T)$, if a greater conditional probability b_{ij} is designated, the objective function $I(x^T, y^T, \theta)$ will gain.

To modify the parameters according to (9) and simultaneously maintain the validity of the model, a two channel model is devised as shown in Fig. 1.

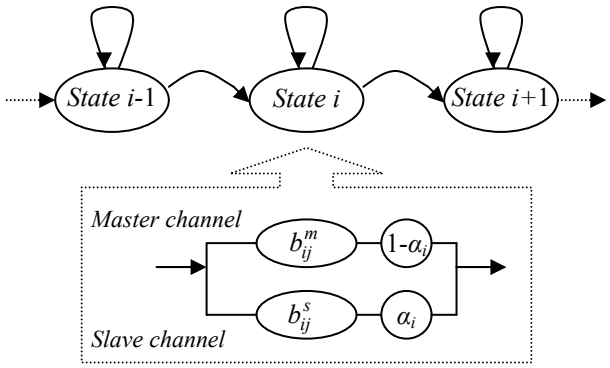

Figure 1. **Structure of a two-channel HMM**

In the above HMM, the probability parameters of each state, e.g. $\{b_{i1}\ b_{i2} \cdots b_{iM}\}$ in this instance, are decomposed into the sum of two parameter sets, which are called master channel and slave channel.

$$\{b_{i1}\ b_{i2}\cdots b_{iM}\} = \underbrace{\{b_{i1}^m\ b_{i2}^m\cdots b_{iM}^m\}}_{\text{master channel}} + \underbrace{\{b_{i1}^s\ b_{i2}^s\cdots b_{iM}^s\}}_{\text{slave channel}} \quad (10)$$

The master channel serves as the common channel for both the correct observations and incorrect ones. For the training pair x^T and y^T, the probabilities scored by this channel – $P(x^T \mid \theta_{master})$ and $P(y^T \mid \theta_{master})$ may be close to each other. The purpose of this channel is to avoid the occurrence of zero probability rather than to distinguish them. As a result, the initial values of $\{b_{i1}^m\ b_{i2}^m\cdots b_{iM}^m\}$ are always derived from a smoothed ML model such as $\widetilde{\theta}_{ML}^x$. Parameter smoothing [2] makes all the probability parameters $\widetilde{b}_{ij}$ of $\widetilde{\theta}_{ML}^x$ greater than 0. As a result, the probability of any nonzero-length observation scored under $\widetilde{\theta}_{ML}^x$ is greater than 0. Because the master channel is derived from $\widetilde{\theta}_{ML}^x$, the same characteristic also applies to it. Thus we have,

$$b_{ij}^m > 0 \quad \forall i = 1,2,\cdots N \quad j = 1,2,\cdots M \quad (11)$$

The slave channel, on the other hand, is the key source of the discriminating power. This channel aims at the difference between x^T and y^T. Its parameter set $\{b_{21}^s\ b_{22}^s\cdots b_{2M}^s\}$ is modified using equation (9). However, the parameter space of the slave channel is not complete in most cases (some may equal to 0), the probability scored by this channel alone often leads to 0. A valid model is formed by incorporating the slave channel and the master channel.

The tradeoff between the master channel and the slave channel is controlled by the credibility factor α_i (different states may have different values).

For the master channel,

$$\sum_{j=1}^{N} b_{ij}^m = 1-\alpha_i \quad 0 \le \alpha_i < 1 \quad \forall i = 1,2,\cdots N \quad (12.a)$$

And for the slave channel,

$$\sum_{j=1}^{N} b_{ij}^s = \alpha_i \quad 0 \le \alpha_i < 1 \quad \forall i = 1,2,\cdots N \quad (12.b)$$

The coefficient b_{ij}^s will be set non-negative value during learning; as a result, the probability restriction for a smoothed model $b_{ij} \ge b_{ij}^m > 0$ is automatically satisfied for the two-channel learning method.

3. Supervised learning

The initial configurations of the supervised two-channel learning method include the initial settings of α_i, b_{ij}^m of the master channel and b_{ij}^s of the slave channel. As mentioned above, b_{ij}^m is obtained from $\widetilde{\theta}_{ML}^x$ – the smoothed ML model of x^T

$$\{b_{i1}^m\ b_{i2}^m\cdots b_{iM}^m\} = (1-\alpha_i)\{\widetilde{b}_{i1}\ \widetilde{b}_{i2}\cdots\widetilde{b}_{iM}\} \quad (13)$$
$$1 \le i \le N, 0 \le \alpha_i < 1$$

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

b_{ij}^s is estimated in the learning process. For the discrete HMM discussed in this paper, a random positive initial setting usually works well.

The selection of the credibility factor α_i is very flexible and depends much on the problem itself. If α_i is set great, the slave channel plays a more important role on scoring the probability of the input sequence. Thus the discriminating power is improved. However, as we adjust b_{ij}^s toward the direction of increased $I(x^T, y^T, \theta)$, the probability of the correct observations $P(x^T \mid \theta)$ will decrease. This is undesirable because the trained model is unlikely to generate the given sequence. In addition to this, α_i of different state should also be set respectively based on its contribution to the computed probability. Considered the above requirements, choosing of α_i can be done in a tentative manner. However, after the value of α_i is settled, it cannot be modified in the learning process.

Step 1: Partition of the symbol space

In this step, we are to find from the probability matrix B which coefficients will positively affect $I(x^T, y^T, \theta)$ and which do not.

Using the forward variables $\alpha_\tau(i) = P(o_1\cdots o_\tau, s_\tau = S_i \mid \theta)$ and backward variables $\beta_\tau(i) = P(o_{\tau+1}\cdots o_T \mid s_\tau = S_i, \theta)$, it is not difficult to calculate $E(S_i, O_j \mid \theta, x^T)$ and $E(S_i, O_j \mid \theta, y^T)$. The symbol set $\{O_1, O_2\cdots O_M\}$ is partitioned by (14) into $V = \{v_1, v_2, \cdots v_K\}$ and its complement set $U = \{u_1, u_2, \cdots u_{M-K}\}$.

$$\{v_1, v_2, \cdots v_K\} = \arg_{O_j}\left[\frac{E(S_i, O_j \mid \theta, x^T)}{E(S_i, O_j \mid \theta, y^T)} > T\right] \quad (14)$$

Where T is the threshold with the typical value greater than or equal to 1.

From (9), it can be concluded that increasing the value of $b_i(v_j)$ and decreasing that of $b_i(u_j)$ (but can not be less than or equal to 0) will both lead to greater $I(x^T, y^T, \theta)$.

Step 2: Modification to the slave channel

Equation (9) illustrates that, for the symbol set U, $b_i(u_j)$ should be set as small as possible. As a result, we let $b_i^s(u_j)=0$, and so $b_i(u_j)=b_i^m(u_j)$. For the set V, the corresponding probability coefficient $b_i(v_k)$ should be distributed in proportion to the value of $E(S_i,O_j\mid\theta,x^T)-E(S_i,O_j\mid\theta,y^T)$, and simultaneously summing up to α_i.

$$b_i^s(v_k)=\frac{E(S_i,O_j\mid\theta,x^T)-E(S_i,O_j\mid\theta,y^T)}{\lambda_i}-b_i^m(v_k) \quad (15.a)$$

$$\sum_{k=1}^{K} b_i^s(v_k)=\alpha_i \quad (15.b)$$

The $b_i^s(v_k)$'s are obtained by solving (15). However, some parameter, e.g. $b_i^s(v_l)$, may be less than 0. It indicates that $b_i^m(v_l)$ alone is big enough for separation. In this case, v_l is excluded from V and $b_i^s(v_l)$ is set to 0. The left symbols in V are reevaluated until all the $b_i^s(v_k)$'s are greater than 0.

The two steps constitute a training epoch and they are implemented iteratively in the process. After each epoch, $I(x^T,y^T,\theta)$ is calculated and compared with that of the previous time. If $I(x^T,y^T,\theta)$ does not change much, e.g. less than a predefined threshold, the training stops and the model is output.

The convergence of the two-channel method is guaranteed by the Lagrange multiplier algorithm and gradient descent theory. It should be noted that for the supervised two-channel strategy, the gain of the overall discriminating power is achieved by improving that of the individual state of the HMM. The underlying assumption of the idea is that the training pair are similar enough that the durations of the corresponding state are comparable. As a result, the supervised two-channel learning method must use similar observation sequences as its training data.

4. Application on lip reading

Lip reading is a good example for the implementation of the two-channel model. As mentioned above, some visemes are statistically and dynamically close to each other, however, it does not mean that they are inseparable. Actually, there are always some distinct features exist in a viseme. For example, while the phoneme /th/ is articulated, there is short interval that the tongue can be traced. The two-channel method may make full use of these features to classify it out of other similar visemes.

In our viseme recognition system, the input visemes are image sequences sampled at 25Hz. For each frame,

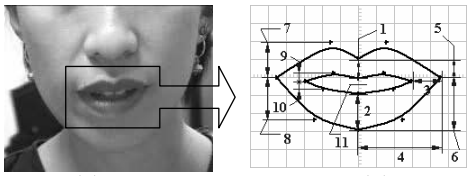

Figure 2. **Feature extraction example**
a) original image b) extracted measures

eleven geometric measures as shown in Fig. 2b are extracted to form a vector, where the 11th indicates the tongue when it is visible. These measures are chosen as they uniquely determine the lip shape and best characterize the dynamics of lip movement.

The vectors that indicate various lip poses are collected and clustered into groups using method such as *K-means* algorithm. For the experiments conducted in this paper, 128 clusters (code words) are used in the vector database (code book). The input observation sequences are encoded with them.

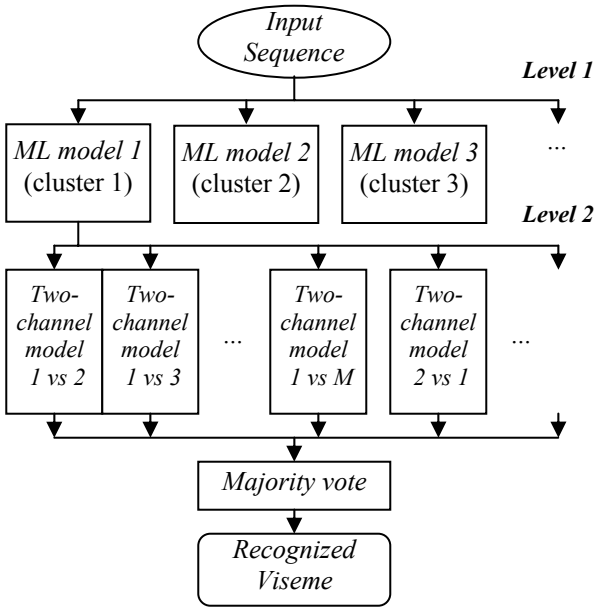

Figure 3. **Hierarchical viseme recognition system**

The viseme recognition system has a hierarchical structure as shown in Fig. 3. The ML models in level 1 carry out coarse recognition to separate visemes that are easy to be identified. For the visemes that are too close to be distinguished by the ML classifiers, such as /ai/, /a:/, they are clustered into one cluster for fine processing by two-channel classifiers. In level 2, a number of two-channel models are mounted in parallel with each of them is specially trained to distinguish between two visemes within a cluster, for example, between the visemes of /ai/ and /a:/. As a result, if a cluster contains M visemes, there

will be $M(M-1)$ models to describe all of them. Each classifier gives a conclusion on the source of the input viseme. The finial decision is made after summarized all these conclusions by means of majority vote.

5. Performance of the method

Experiments are conducted to assess the performance of the HMMs trained by the supervised two-channel method. In table I, the ratio between the probabilities of the correct and incorrect observations $I(x,y)=\frac{P(x^T \mid \theta_x)}{P(y^T \mid \theta_x)}$ is listed as the indicator of the discriminating power. The greater the $I(x,y)$, the better the distinguish accuracy is. Smoothed ML models (three-state left-right HMMs) trained with the Baum-Welch method and two-channel models, which are based on the above ML models but with different settings, are compared in the table. Here, we do not use the fourteen visemes defined in MPEG-4 multimedia standards [6] but assume each viseme is corresponding with a phoneme. As a result, a viseme is represented by its corresponding phoneme in the row of the table. The viseme in boldface is the correct observation of the model.

Table 1. I(x,y) of ML models and two-channel models

Model type / *Viseme pair*	**ML model**	**Two-channel model 1** (α_1=α_2=α_3=0.5)	**Two-channel model 2** (α_1=α_3= 0.6, α_2=0.8)
/**a:**/, /ai/	1.571e01	3.720e05	3.887e06
/**ei**/, /i/	1.452e02	9.476e05	1.019e07
/**au**/, /eu/	9.765e02	7.443e05	3.587e06
/**o**/, /oi/	6.754	3.220e02	1.978e03
/**th**/, /sh/	3.988	6.441e02	8.556e02

The results indicate that for the set of visemes tested, the two-channel models provide a much better accuracy to distinguish similar visemes than ML models. And by adjusting the credibility factors, different discriminating accuracy can be attained.

6. Conclusion

The supervised two-channel learning method excels traditional EM method in its specialty to discriminate similar observations. Although the method is applied to lip-reading as an example, the approach can be applied to many situations that require fine recognition, such as tone recognition, speaker identification and so on.

In this paper, we only introduce some fundamental concepts about the supervised two-channel learning method. In applications, the proposed method can be modified or extended to satisfy various requirements. For example, the training pair can be of different length if some linear measures are taken to rectify the objective function; and the two-channel approach can also be extended to continuous symbol distribution if some continuous pdf basis, such as Gaussian mixtures are adopted.

7. References

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

[2] C. Bregler and S. Omohundro, "Nonlinear manifold learning for visual speech recognition," *Proc. IEEE ICCV*, pp. 494-499, 1995

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

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

[5] 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," *Proc. ICASSP '86*, pp. 49-52, Apr. 1986

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

[7] Gerasinos Potamianos and Hans Peter Graf, "Linear discriminant analysis for speechreading," *Proc. IEEE Signal Processing Society Workshop on Multimedia Signal Procesing*, pp. 221-226, 1998