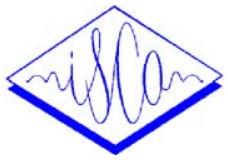

6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

INTENSITY- AND LOCATION-NORMALIZED TRAINING FOR HMM-BASED VISUAL SPEECH RECOGNITION

Yoshihiko Nankaku, Keiichi Tokuda and Tadashi Kitamura

Department of Computer Science
Nagoya Institute of Technology
Nagoya 466-8555, Japan

E-mail: {nankaku, tokuda, kitamura}@ics.nitech.ac.jp

ABSTRACT

This paper describes an approach to estimating the parameters of continuous density HMMs for visual speech recognition. One of the key issues of image-based visual speech recognition is normalization of lip location and lighting condition prior to estimating the parameters of HMMs. We present an average-intensity and location normalized training method, in which the normalization process is integrated in the model training. The proposed method provides a theoretically-well-defined algorithm based on a maximum likelihood formulation, hence the likelihood for the training data is guaranteed to increase at each iteration of the normalized training. Experimental results show that the recognition performance can be significantly improved by the normalized training.

1. INTRODUCTION

One of the difficulties in visual speech recognition is the extraction of feature parameters from the image sequence of lips. Methods to extract speech information from image sequences are largely categorized into two approaches: model-based approach and image- or pixel-based approach. In the image-based approach, pixel values of the image are preprocessed and then used as the feature vector. However, this process must take account of the variety of lighting condition, lip location, rotation, and scaling. The statistical models (e.g., HMMs) are trained with such a variation, the distributions of different classes overlap each other, and the discriminatory capabilities of the statistical models may be reduced. Therefore, the training data must be normalized prior to model training.

This paper proposes an approach to estimating the parameters of continuous density HMMs for visual speech recognition, in which normalization of average-intensity and lip location is integrated in the model training. Our approach to visual speech recognition is based on the success of a normalization approach for auditory speech recognition: SAT (speaker adaptive training) [1]–[3]. The SAT is a normalized training technique, in which speaker normalization based on the MLLR (Maximum Likelihood Linear Regression) [4] is integrated in the model training. Although the idea of [5] is similar to that of this paper, the proposed method provides a theoretically-well-defined algorithm based on the ML (maximum likelihood) criterion, and can normalize average-intensity and location simultaneously within an ML formulation.

This paper is organized as follows. An ML-based normalization framework is described in the next section. The section 3 describes the re-estimation algorithm for the normalized training. The section 4 presents experimental results. Concluding remarks and our plans for future work are presented in the final section.

2. ML-BASED NORMALIZATION FRAMEWORK

In image-based visual speech recognition, we normalize average-intensity and location of lips by an affine transformation:

$$\hat{\mathbf{o}}^{(r)}(t) = \mathbf{A}^{(r)}\mathbf{o}^{(r)}(t) + \mathbf{b}^{(r)} \tag{1}$$

where $\mathbf{o}^{(r)}(t)$ and $\hat{\mathbf{o}}^{(r)}(t)$ are the original image vector and the normalized lip image vector, respectively, associated with utterance r at time t. Note that the intensity values of all pixels in the image are collected in a long one-dimensional vector. A rectangular matrix $\mathbf{A}^{(r)}$ extracts mouth part from the original image and sub-samples the extracted image. If the size of $\mathbf{o}^{(r)}(t)$ and the block size of sub-sampling are L and k, respectively, each row of $\mathbf{A}^{(r)}$ consists of k elements of $1/k$ and $L-k$ elements of 0, and the arrangement of the non-zero elements depends on the location of extracted mouth part. The vector $\mathbf{b}^{(r)}$ consists of additive coefficients for average intensity normalization:

$$\mathbf{b}^{(r)} = [b^{(r)} \ldots b^{(r)}]^T \tag{2}$$

We assume that the lip location and average-intensity do not change very much during one utterance, one transformation is prepared for each utterance.

In the conventional normalization approach, the transformation $(\mathbf{A}^{(r)}, \mathbf{b}^{(r)})$ is determined prior to estimating the parameters of HMM. In our method, $(\mathbf{A}^{(r)}, \mathbf{b}^{(r)})$ is determined so as to maximize the likelihood of the HMM parameters $\mathcal{M}$, that is, the optimal transformation $(\mathbf{A}^{(r)}, \mathbf{b}^{(r)})$ is derived as

$$(\mathbf{A}^{(r)}, \mathbf{b}^{(r)})' = \underset{(\mathbf{A}^{(r)}, \mathbf{b}^{(r)})}{\operatorname{argmax}} P(\hat{\mathbf{O}}^{(r)} \mid \mathcal{M}) \tag{3}$$

where $\hat{\mathbf{O}}^{(r)} = [\hat{\mathbf{o}}^{(r)}(1), \hat{\mathbf{o}}^{(r)}(2), \ldots, \hat{\mathbf{o}}^{(r)}(T_r)]$ is the lip image sequence associated with an utterance r. However, to determine $(\mathbf{A}^{(r)}, \mathbf{b}^{(r)})'$, we need the HMM $\mathcal{M}$ which cannot be trained unless transformations for all the training utterances, $(\mathbf{A}^{(r)}, \mathbf{b}^{(r)}), r = 1, 2, \ldots, R$ are determined. Therefore the optimum set of transformations $(\mathbf{A}^{(r)}, \mathbf{b}^{(r)}), r = 1, 2, \ldots, R$ and the set of HMM parameters $\mathcal{M}$ are jointly estimated so as to maximize the likelihood:

$$\lambda' = \underset{\lambda}{\operatorname{argmax}} \; P(\hat{\mathbf{O}} \mid \mathcal{M}) \tag{4}$$

where

$$\hat{\mathbf{O}} = \{\hat{\mathbf{O}}^{(1)}, \hat{\mathbf{O}}^{(2)}, ..., \hat{\mathbf{O}}^{(R)}\} \quad (5)$$

is the normalized training data, and λ consists of the set of transformations for all utterances and the model parameters:

$$\lambda = \{(\mathbf{A}, \mathbf{b}), \mathcal{M}\} \quad (6)$$

$$(\mathbf{A}, \mathbf{b}) = \{(\mathbf{A}^{(r)}, \mathbf{b}^{(r)}) \mid r = 1, 2, ..., R\} \quad (7)$$

The fundamental idea of the proposed method is to determine the transformations $(\mathbf{A}, \mathbf{b})$ which normalizes the training utterances and the HMM parameters $\mathcal{M}$ simultaneously.

3. RE-ESTIMATION ALGORITHM

3.1 Q-Function

To solve the above optimization problem, we adopt the EM (Expectation-Maximization) algorithm, which is the iterative procedure of approximating ML estimates. The procedure consists of maximizing at each iteration the auxiliary function so called Q-function. The likelihood for the training data is guaranteed to increase by increasing the value of the Q-function. Hence the maximization of the Q-function value at each iteration maximizes the likelihood for the training data:

$$Q(\lambda, \lambda') \geq Q(\lambda, \lambda) \;\Rightarrow\; P(\hat{\mathbf{O}}' \mid \mathcal{M}') \geq P(\hat{\mathbf{O}} \mid \mathcal{M}) \quad (8)$$

where $\hat{\mathbf{O}}'$ is the training data normalized by the updated transformation set $(\mathbf{A}, \mathbf{b})'$, which is included in the updated parameters λ'. The Q-function with respect to the HMM parameters and the transformations can be written as

$$Q(\lambda, \lambda') = K - \frac{1}{2} \sum_{r,m,t}^{R,M,T_r} \gamma_m^{(r)}(t) \big[K_m + \log(|\mathbf{\Sigma}_m|) + (\hat{\mathbf{o}}^{(r)}(t) - \boldsymbol{\mu}_m)^T \mathbf{\Sigma}_m^{-1} (\hat{\mathbf{o}}^{(r)}(t) - \boldsymbol{\mu}_m)\big] \quad (9)$$

where $\boldsymbol{\mu}_m$ and $\mathbf{\Sigma}_m$ are the mean vector and the covariance matrix of the m-th Gaussian component, respectively, K is a constant dependent only on the transition probabilities, K_m is the normalization constant associated with Gaussian m, and $\gamma_m^{(r)}(t)$ is the posterior probability of Gaussian m at time t, that can be computed through the forward-backward algorithm:

$$\gamma_m^{(r)}(t) = P(q_t = m \mid \mathbf{O}^{(r)}, \mathcal{M}) \quad (10)$$

The iterative approach using Q-function is adopted in which one of the parameter sets (transformations $(\mathbf{A}, \mathbf{b})$ and HMM parameters $\mathcal{M}$) is estimated at each stage and the maximum likelihood re-estimation is used individually for each of the parameter sets keeping the other parameters are fixed.

3.2 Maximization of Q-function

We first maximize the Q-function with respect to $(\mathbf{A}, \mathbf{b})$ while keeping model parameters $\mathcal{M}$ fixed to current values. We cannot derive $\mathbf{A}^{(r)}$ which maximize the value of Q-function in closed form since $\mathbf{A}^{(r)}$ must satisfy the constraints described in section 2 (i.e., lip area extraction and sub-sampling). By giving a location, however, $\mathbf{A}^{(r)}$ is completely determined, and by setting

$$\frac{\partial Q(\lambda, \lambda')}{\partial b'^{(r)}} = 0, \quad r = 1, 2, ..., R, \quad (11)$$

the optimum $b'^{(r)}$ for a given $\mathbf{A}'^{(r)}$ is determined as

$$b'^{(r)} = \frac{\sum_{m,t}^{M,T_r} \gamma_m^{(r)}(t) (\boldsymbol{\mu}_m - \mathbf{A}'^{(r)} \mathbf{o}^{(r)}(t))^T \mathbf{\Sigma}_m^{-1} [1 \; ... \; 1]^T}{\sum_{m,t}^{M,T_r} \gamma_m^{(r)}(t) [1 \; ... \; 1] \mathbf{\Sigma}_m^{-1} [1 \; ... \; 1]^T} \quad (12)$$

Thus, we adopt direct search for the optimum location. To avoid a large amount of computation required for the exhaustive search, we adopted a gradient search of Q-function value for the optimal location in the experiment.

The estimation of the means of the Gaussian densities conditioned on the updated transformation set $(\mathbf{A}, \mathbf{b})'$ is expressed as

$$\boldsymbol{\mu}'_m = \frac{\sum_{r,t}^{R,T_r} \gamma_m^{(r)}(t) \hat{\mathbf{o}}'^{(r)}(t)}{\sum_{r,t}^{R,T_r} \gamma_m^{(r)}(t)} \quad (13)$$

Similarly, the estimation of covariance matrices of the Gaussian densities is expressed as

$$\mathbf{\Sigma}'_m = \frac{\sum_{r,t}^{R,T_r} \gamma_m^{(r)}(t) (\hat{\mathbf{o}}'^{(r)}(t) - \boldsymbol{\mu}'_m)(\hat{\mathbf{o}}'^{(r)}(t) - \boldsymbol{\mu}'_m)^T}{\sum_{r,t}^{R,T_r} \gamma_m^{(r)}(t)} \quad (14)$$

By inspection, these equations are the same as the standard estimation using the training data $\hat{\mathbf{o}}'^{(r)}(t)$ normalized by the transformation $(\mathbf{A}^{(r)}, \mathbf{b}^{(r)})'$.

The normalized training procedure is summarized as follows:

Step 0. Give an initial transformation set $(\mathbf{A}, \mathbf{b})$ and construct an initial model $\mathcal{M}$.

Step 1. Compute the values of $\gamma_m^{(r)}(t)$, and estimate the transformation set $(\mathbf{A}, \mathbf{b})'$.

Step 2. Compute the values of $\gamma_m^{(r)}(t)$, and estimate the model parameters $\mathcal{M}'$.

Step 3. If the change of the likelihood after the re-estimation is small, Stop. Otherwise go to Step 1.

It is easily verified that at each stage of the update process the value of the Q-function is guaranteed to increase:

$$Q(\lambda, \lambda') \geq Q(\lambda, \lambda) \quad (15)$$

Hence the likelihood for the training data is also guaranteed to increase, based on the properties of the Q-function stated earlier:

$$P(\hat{\mathbf{O}} \mid \mathcal{M}) \leq P(\hat{\mathbf{O}}' \mid \mathcal{M}) \leq P(\hat{\mathbf{O}}' \mid \mathcal{M}') \quad (16)$$

where $\hat{\mathbf{O}}'$ is the training data normalized by the transformation set $(\mathbf{A}, \mathbf{b})'$ updated in Step 1, and $\mathcal{M}'$ is the model parameters updated in Step 2. Typically, two or three iterations of Steps 1 and 2 are sufficient to ensure convergence to an optimal point.

It was explained in [4] that the feature space transformation without any constraints is not an appropriate transformation for ML-based estimation. However, in our method, transformations have constraints for extraction of lip image, sub-sampling and normalization of average intensity. These constraints avoid the problem.

3.3 Normalization in Recognition Phase

In the testing, the likelihood of each model with respect to a testing utterance is measured by using the transformation which maximizes the likelihood of the model: the transformations for models $\mathcal{M}_i = \{\mathcal{M}_i \mid i = 1, 2, ..., I\}$ are derived as

$$(\mathbf{A}_i, \mathbf{b}_i)' = \underset{(\mathbf{A}_i, \mathbf{b}_i)}{\operatorname{argmax}} P(\hat{\mathbf{O}} \mid \mathcal{M}_i) \tag{17}$$

In a manner similar to the normalized training procedure, that is, by iterating Step 1, an optimal transformation can be obtained for each model. In practice such an iteration makes small difference after the second iteration. Therefore only one iteration was applied in the experiment. The model which gives the highest likelihood is chosen as the recognition result:

$$\text{result} = \underset{i}{\operatorname{argmax}} P(\hat{\mathbf{O}}_i' \mid \mathcal{M}_i) \tag{18}$$

where $\hat{\mathbf{O}}_i'$ is the testing data normalized by the transformation $(\mathbf{A}_i, \mathbf{b}_i)'$ which is optimized for model $\mathcal{M}_i$. In the experiment, the likelihood for testing data was approximated by that calculated by the Viterbi algorithm.

4. EXPERIMENT

Visual word recognition experiments were performed. Each word model was represented by one HMM which is left-to-right model with 5 states, single Gaussian distribution of diagonal covariance. The image vector (static feature vector) and the difference between two successive frames (delta feature vector) were combined to form the feature vector $\mathbf{o}^{(r)}(t)$. The block size of sub-sampling was 5×5. For the experiment, the Tulips1 database [6] was used, which is a bimodal database comprising of lip image sequences and speech signals of 9 males and 3 females, in total 12 speakers. Each speaker pronounces the English numbers, one, two, three, and four, each twice. The visual frame rate is 30 frame/s and each frame is a 100×75 pixel image. We performed speaker independent word recognition tests using the "leave-one-out method". In the method, one of 12 subjects was used for testing and the remaining 11 subjects were used for training. This was repeated 12 times, leaving out a different subject each time.

Figure 1 shows the change of the likelihood with respect to the iteration of the normalized training. In the figure, "w/o" means that HMMs were trained without both the prior normalization and the normalized training, and they were used as the initial models for normalized training. The likelihood plotted at each iteration number is that obtained after the update of models (Step 2) and the likelihood plotted between two iteration numbers is that obtained after the update of transformations (Step 1). From

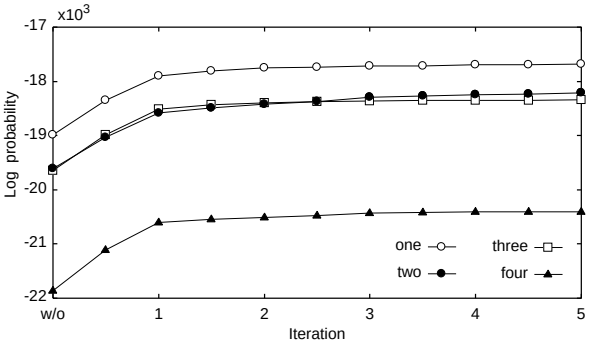

Figure 1. Log likelihoods of HMMs for training data.

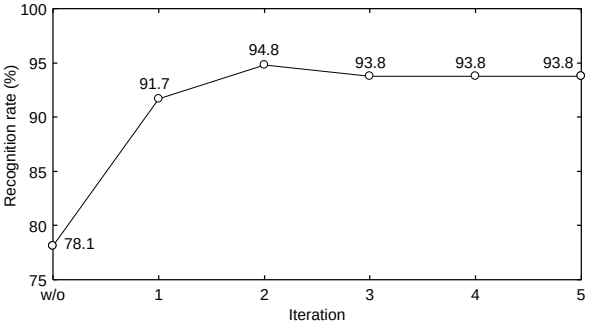

Figure 2. Recognition rate for each iteration of normalized training.

the figure, it can be confirmed that the normalized training monotonically increases the likelihood for the training data.

The word recognition rate for each iteration of the normalized training is shown in Figure 2. The normalization process described in Section 3.3 was applied to all the testing data except for "w/o". It can be seen that a significant error reduction is achieved by the proposed technique: for the second iteration, a recognition rate of 94.8% and error reduction of 76% were achieved. When we used the conventional normalization approach, in which the average-intensity of the training and testing data are normalized independently of HMM, a recognition rate of 86.5% was obtained. The result can be regarded as one in the case where the average-intensity and location are normalized prior to the model training since the lip areas were extracted manually in the Tulips1 database. These results suggest that the normalization process should be integrated in the model training to reduce the overlap among the distributions of different HMMs, and it improves the recognition performance significantly.

Figure 3 compares the obtained models for /one/ with and without normalized training. In the figure, the values of the means and variances (i.e., diagonal covariances) of static parameter were represented by gray levels. As seen in Figure 3, the images representing the mean vectors become sharp after the normalized training. Figure 4 shows the average values of variances per pixel for each iteration. The values of the variances after the normalized training process are smaller than those before it. This coincides with that the images representing the variances in Figure

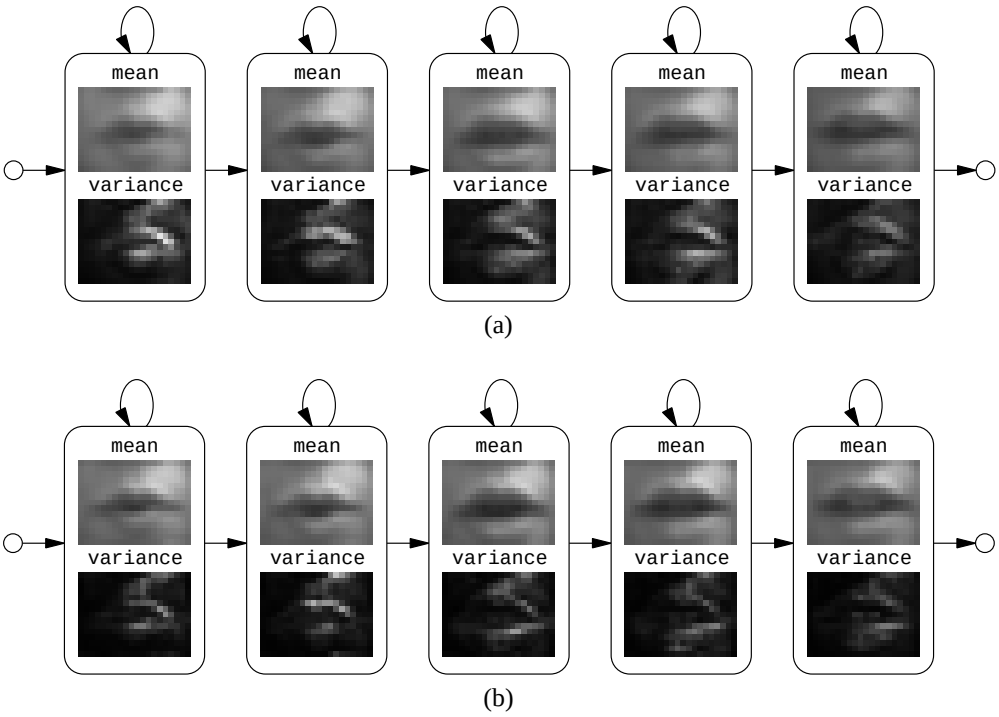

Figure 3. HMM /one/ without normalized training (a) and with normalized training (2nd iteration) (b).

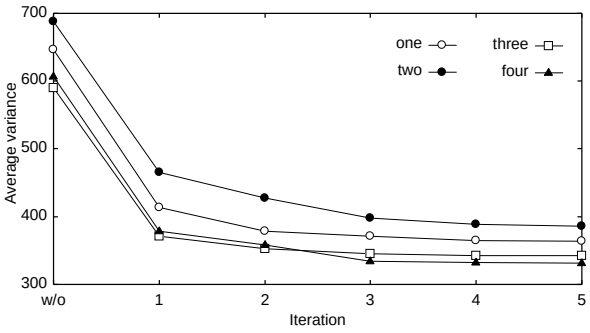

Figure 4. Average variance of HMMs.

3 become dark after the normalized training. Therefore it means that a better class separation could be obtained.

5. CONCLUSION

We proposed an approach to simultaneous intensity and location normalized training of HMMs for image-based visual speech recognition. Experimental results show that by integrating the normalization process into the model training the recognition performance is significantly improved: a word recognition rate of 94.8% and an error-reduction of 76% were achieved. In the proposed algorithm, the likelihood for the training data is guaranteed to increase at each iteration of parameter re-estimation.

In the future, the normalized training algorithm will be extended for normalizing contrast, color, lip rotation and scaling. Integration of the visual information to auditory information will also be a future work.

ACKNOWLEDGMENT

The authors would like to thank Prof. T. Kobayashi, Tokyo Institute of Technology for his comments and suggestions. This work was partially supported by the Ministry of Education, Science, Sports and Culture Japan, Grant-in-Aid for Scientific Research (c) (2), 09680394, 1997, and Encouragement of Young Scientists, 0780226, 1998.

REFERENCES

[1] T. Anastasakos, J. McDonough and J. Makhoul, “Speaker Adaptive Training: A Maximum Likelihood Approach to Speaker Normalization,” Proc. ICASSP, pp. 1043–1046, 1997.

[2] T. Anastasakos, J. McDonough, R. Schwartz and J. Makhoul, “A Compact Model for Speaker-Adaptive Training,” Proc. ICSLP, 1996.

[3] D. Pye and P. C. Woodland, “Experiments in Speaker Normalisation and Adaptation for Large Vocabulary Speech Recognition,” Proc. ICASSP, pp. 1047–1050, 1997.

[4] M. J. F. Gales, “Maximum Likelihood Linear Transformations for HMM-based Speech Recognition,” Computer Speech and Language, pp. 75–98, Dec, 1998.

[5] O. Vanegas, A. Tanaka, K. Tokuda and T. Kitamura, “HMM-based Visual Speech Recognition using Intensity and Location Normalization,” Proc. ICSLP, pp. 289–292, 1998.

[6] J. R. Movellan, “Visual Speech Recognition with Stochastic Networks,” G. Tesauro, D. Touretzky, T. Leen (eds.) , Advances in Neural Information Processing Systems 7, MIT Press Cambridge, 1995.