NORMALIZED TRAINING FOR HMM-BASED VISUAL SPEECH RECOGNITION

Yoshihiko Nankaku†, Keiichi Tokuda†, Tadashi Kitamura† and Takao Kobayashi‡

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

‡Interdisciplinary Graduate School of Science and Engineering
Tokyo Institute of Technology, Yokohama, 226-8502 Japan

ABSTRACT

This paper presents 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 presented a normalized training method in which the normalization process is integrated in the model training. This paper extends it for contrast normalization in addition to average-intensity and location normalization. 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. Experiments on M2VTS database show that the recognition performance can be significantly improved by the normalized training.

1. INTRODUCTION

In visual speech recognition, one of the difficulties 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 [1]–[3] and image- or pixel-based approach[3]–[5]. In the model-based approach, a contour model of lips is first constructed, and it is represented by a small number of parameters. Although the advantage of this approach is that the parameters have less influence of variability of lighting condition, lip location, rotation and scaling, it has a difficulty in the construction of a robust and efficient lip contour model which can locate and track lips. On the other hand, 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, contrast and lip location 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, contrast and location simultaneously within an ML formulation. In our previous work [6], we proposed the average-intensity and location normalized training. In this paper, the normalized training algorithm is extended for contrast normalization, and the M2VTS database [7] was used instead of the Tulips1 database [8] for larger task experiments.

This paper is organized as follows. The next section defined the normalization transformation used in this work. An ML-based normalization framework and the re-estimation algorithm for the normalized training are described in section 3 and section 4, respectively. Section 5 presents experimental results. Concluding remarks and our plans for future work are presented in the final section.

2. NORMALIZATION TRANSFORMATION

2.1. Geometric Transformation

In image-based visual speech recognition, we extract mouth part from the original image by a linear transformation:

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

where $\mathbf{o}^{(r)}(t)$ and $\hat{\mathbf{o}}^{(r)}(t)$ are the original image vector and the extracted 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. Although (1) can normalize geometric transformations such as lip location, rotation and scaling, we considered only location normalization in this work. 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.

2.2. Intensity Transformation

In our method, HMM parameters are transformed by a linear transformation for Intensity normalization:

$$\hat{\boldsymbol{\mu}}_m^{(r)} = \check{\mathbf{C}}^{(r)}\boldsymbol{\mu}_m + \check{\mathbf{b}}^{(r)} \tag{2}$$

$$\hat{\boldsymbol{\Sigma}}_m^{(r)} = \check{\mathbf{C}}^{(r)}\boldsymbol{\Sigma}_m\check{\mathbf{C}}^{(r)T} \tag{3}$$

where $\boldsymbol{\mu}_m$ and $\boldsymbol{\Sigma}_m$ are the mean vector and the covariance matrix of the m-th Gaussian component, respectively. This transformation framework can represent various normalizations, e.g., color normalization. In this paper, we assume that the transformation consists of additive coefficients for average-intensity normalization and multiplicative coefficients for contrast normalization:

$$\check{\mathbf{b}}^{(r)} = \check{b}^{(r)}\mathbf{1} \quad (4)$$

$$\check{\mathbf{C}}^{(r)} = \check{c}^{(r)}\boldsymbol{I} \quad (5)$$

where $\mathbf{1} = [1 \ \dots \ 1]^T$, and $\check{\mathbf{b}}^{(r)}$ is an average-intensity normalization vector, $\check{\mathbf{C}}^{(r)}$ is a matrix for contrast normalization. We also assume that the lip location and average-intensity do not change very much during one utterance, one transformation is prepared for each utterance.

3. ML-BASED NORMALIZATION FRAMEWORK

In the conventional normalization approach, lip images must be normalized prior to estimating the parameters of HMM. In our method, the transformation $\mathcal{T}^{(r)} = (\mathbf{A}^{(r)}, \check{\mathbf{b}}^{(r)}, \check{\mathbf{C}}^{(r)})$ is determined so as to maximize the likelihood of the HMM parameters $\mathcal{M}$, that is, the optimal transformation $\mathcal{T}'^{(r)}$ is derived as

$$\mathcal{T}'^{(r)} = \underset{\mathcal{T}^{(r)}}{\operatorname{argmax}} P(\hat{\mathbf{O}}^{(r)} \mid \mathcal{M}, (\check{\mathbf{b}}^{(r)}, \check{\mathbf{C}}^{(r)})) \quad (6)$$

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

$$\lambda' = \underset{\lambda}{\operatorname{argmax}} P(\hat{\mathbf{O}} \mid \mathcal{M}, (\check{\mathbf{b}}, \check{\mathbf{C}})) \quad (7)$$

where $\hat{\mathbf{O}} = \{\hat{\mathbf{O}}^{(1)}, \hat{\mathbf{O}}^{(2)}, ..., \hat{\mathbf{O}}^{(R)}\}$ is all training utterances, and λ consists of the set of transformations for all utterances and HMM parameters:

$$\lambda = \{\mathbf{A}, (\check{\mathbf{b}}, \check{\mathbf{C}}), \mathcal{M}\} \quad (8)$$

where

$$\mathbf{A} = \{\mathbf{A}^{(r)} \mid r = 1, 2, \dots, R\} \quad (9)$$

$$(\check{\mathbf{b}}, \check{\mathbf{C}}) = \{(\check{\mathbf{b}}^{(r)}, \check{\mathbf{C}}^{(r)}) \mid r = 1, 2, \dots, R\} \quad (10)$$

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

4. RE-ESTIMATION ALGORITHM

4.1. 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:

$$Q(\lambda, \lambda') \geq Q(\lambda, \lambda) \ \Rightarrow$$
$$P(\hat{\mathbf{O}}' \mid \mathcal{M}', (\check{\mathbf{b}}, \check{\mathbf{C}})') \geq P(\hat{\mathbf{O}} \mid \mathcal{M}, (\check{\mathbf{b}}, \check{\mathbf{C}})) \quad (11)$$

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

$$Q(\lambda, \lambda') = K - \frac{1}{2} \sum_{r,m,t}^{R,M,T_r} \gamma_m^{(r)}(t) \times \big[K_m + \log(|\Sigma_m|) - 2\log(|\mathbf{C}^{(r)}|) + (\tilde{\mathbf{o}}^{(r)}(t) - \boldsymbol{\mu}_m)^T \boldsymbol{\Sigma}_m^{-1} (\tilde{\mathbf{o}}^{(r)}(t) - \boldsymbol{\mu}_m)\big] \quad (12)$$

where 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:

$$\gamma_m^{(r)}(t) = P(q_t = m \mid \mathbf{O}^{(r)}, \mathcal{T}^{(r)}, \mathcal{M}) \quad (13)$$

and $\tilde{\mathbf{o}}^{(r)}(t)$ is the normalized training data defined by

$$\tilde{\mathbf{o}}^{(r)}(t) = \mathbf{C}^{(r)}\hat{\mathbf{o}}^{(r)}(t) + \mathbf{b}^{(r)} \quad (14)$$

where

$$\mathbf{b}^{(r)} = -\{\check{\mathbf{C}}^{(r)}\}^{-1}\check{\mathbf{b}}^{(r)} = b^{(r)}\mathbf{1} \quad (15)$$

$$\mathbf{C}^{(r)} = \{\check{\mathbf{C}}^{(r)}\}^{-1} = c^{(r)}\boldsymbol{I} \quad (16)$$

Thus, the intensity transformation in our method can be implemented as a transformation of the feature vector and a simple addition of the term $\log(|\mathbf{C}^{(r)}|)$. Therefore the log-likelihoods are calculated as

$$\log(P(\hat{\mathbf{o}}^{(r)}(t) \mid \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m, \check{\mathbf{b}}^{(r)}, \check{\mathbf{C}}^{(r)})) = \log(\mathcal{N}(\tilde{\mathbf{o}}^{(r)}(t) \ ; \ \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m)) + \log(|\mathbf{C}^{(r)}|) \quad (17)$$

and there is no need to change the original HMM parameters.

The iterative approach using Q-function is adopted in which one of the parameter sets (transformations $\{\mathbf{A}, (\mathbf{b}, \mathbf{C})\}$ 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.

4.2. Maximization of Q-function

We first maximize the Q-function with respect to $\{\mathbf{A}, (\mathbf{b}, \mathbf{C})\}$ 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. Therefore, 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.

Then, we compute the transformation set $(\mathbf{b}^{(r)}, \mathbf{C}^{(r)})$ while keeping $\mathbf{A}^{(r)}$ fixed to the current values. The solution of

$$\frac{\partial Q(\lambda, \lambda')}{\partial c'^{(r)}} = 0, \quad r = 1, 2, \ldots, R, \tag{18}$$

gives the updated contrast normalization coefficient as

$$c'^{(r)} = \frac{\xi^{(r)} + \sqrt{\xi^{(r)2} + 4nT\psi^{(r)}}}{2\psi^{(r)}} \tag{19}$$

where

$$\xi^{(r)} = \sum_{m,t}^{M,T} \gamma_m^{(r)}(t) \hat{\mathbf{o}}^{(r)}(t) \mathbf{\Sigma}_m^{-1} \hat{\mathbf{o}}^{(r)}(t), \tag{20}$$

$$\psi^{(r)} = \sum_{m,t}^{M,T} \gamma_m^{(r)}(t) (\boldsymbol{\mu}_m - \mathbf{b}^{(r)})^T \mathbf{\Sigma}_m^{-1} \hat{\mathbf{o}}^{(r)}(t) \tag{21}$$

and n is the size of the feature vector $\hat{\mathbf{o}}^{(r)}(t)$. Similarly, using the updated $c'^{(r)}$, the updated average-intensity normalization coefficient $b'^{(r)}$ are given as

$$b'^{(r)} = \frac{\sum_{m,t}^{M,T} \gamma_m^{(r)}(t) (\boldsymbol{\mu}_m - \mathbf{C}'^{(r)} \hat{\mathbf{o}}'^{(r)}(t))^T \mathbf{\Sigma}_m^{-1} \mathbf{1}}{\sum_{m,t}^{M,T} \gamma_m^{(r)}(t) \mathbf{1}^T \mathbf{\Sigma}_m^{-1} \mathbf{1}} \tag{22}$$

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

$$\boldsymbol{\mu}'_m = \frac{\sum_{r,t}^{R,T_r} \gamma_m^{(r)}(t) \tilde{\mathbf{o}}'^{(r)}(t)}{\sum_{r,t}^{R,T_r} \gamma_m^{(r)}(t)} \tag{23}$$

$$\mathbf{\Sigma}'_m = \frac{\sum_{r,t}^{R,T_r} \gamma_m^{(r)}(t) (\tilde{\mathbf{o}}'^{(r)}(t) - \boldsymbol{\mu}'_m)(\tilde{\mathbf{o}}'^{(r)}(t) - \boldsymbol{\mu}'_m)^T}{\sum_{r,t}^{R,T_r} \gamma_m^{(r)}(t)} \tag{24}$$

By inspection, these equations are the same as the standard estimation using the training data $\tilde{\mathbf{o}}'^{(r)}(t)$ obtained by the updated normalization transformation $\{\mathbf{A}'^{(r)}, (\mathbf{b}^{(r)}, \mathbf{C}^{(r)})'\}$

The normalized training procedure is summarized as follows:

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

Step 1. Compute the values of $\gamma_m^{(r)}(t)$, and estimate the transformations $(\mathbf{b}, \mathbf{C})'$, and this step is iterated until the change of likelihood is small.

Step 2. Compute the values of $\gamma_m^{(r)}(t)$, and estimate the transformations $\mathbf{A}'$.

Step 3. Compute the values of $\gamma_m^{(r)}(t)$, and estimate the model parameters $\mathcal{M}'$.

Step 4. 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:

$$Q(\lambda, \lambda') \geq Q(\lambda, \lambda) \tag{25}$$

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

$$\begin{aligned} P(\hat{\mathbf{O}} \mid \mathcal{M}, (\check{\mathbf{b}}, \check{\mathbf{C}})) &\leq P(\hat{\mathbf{O}}' \mid \mathcal{M}, (\check{\mathbf{b}}, \check{\mathbf{C}})') \\ &\leq P(\hat{\mathbf{O}}' \mid \mathcal{M}', (\check{\mathbf{b}}, \check{\mathbf{C}})') \end{aligned} \tag{26}$$

where $\hat{\mathbf{O}}'$ is the training data normalized by the transformation set $\mathbf{A}'$ updated in Step 2.

5. EXPERIMENT

Visual word recognition experiments were performed. Each word model was represented by one HMM which is left-to-right model of 8 states with two Gaussian distributions of diagonal covariance. The size of extracted lip image was 80×40 and the block size of sub-sampling was 5×5. The image vector (static feature vector), the temporal difference vector (delta feature vector) and second order time derivatives (delta-delta feature vector) were combined to form the feature vector $\mathbf{o}^{(r)}(t)$.

For the experiment, the M2VTS database was used which is a bimodal database comprising of face image sequences and speech signals of 37 speakers. This database provides 5 shots for each speaker and during each shot speakers pronounce French numbers from "0" to "9" continuously. Shots were taken at one week intervals to account for minor face changes like beards. For each speaker, the most difficult shot to recognize is the fifth, because of some face and voice variations have been included. This study was carried out by using the first four shots. The visual frame rate is 25 frame/s and each frame is a full color 350×286 pixel image. They were converted into grey-level images for the experiments in this study. We performed speaker independent word recognition tests using the "leave-one-out method". In the method, one of 37 subjects was used for testing and the remaining 36 subjects

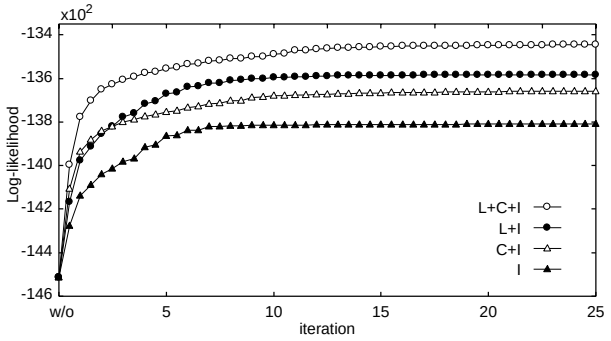

Figure 1. Log likelihoods of HMMs for training data.

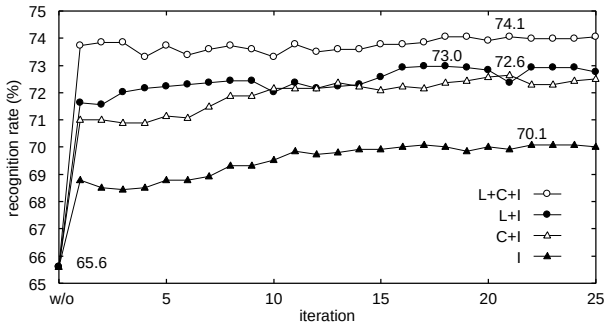

Figure 2. Recognition rate for each iteration of normalized training.

were used for training. This was repeated 37 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. We investigate the performance of the various normalized training described in this paper: the combination of the average-Intensity "I", Contrast "C"and Location "L" normalization. From 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 transformation for the testing data was obtained after three iterations of the update (Step 1-2). The normalization process 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. A recognition rate of 74.1% and error reduction of 24.7% were achieved in the case of "L+C+I"normalization. When we used the conventional normalization approach, in which the average-intensity and the contrast of the training and testing data are normalized independently of HMM, a recognition rate of 70.8% was obtained. In the case of only the intensity normalization, a recognition rate of 69.9% was achieved. These results can be regarded as those obtained by the conventional normalization approach in which the location is also normalized prior to the model training since the lip areas were extracted manually. Notice that the normalized training of "I" and "C+I" are better than the conventional normalization approach using the same location data. In addition, the normalized training which include the location normalization ("L+I" and "L+C+I") further improved the performance. 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.

6. CONCLUSION

We proposed an approach to simultaneous intensity, contrast 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 74.1% and an error-reduction of 24.7% were achieved. In the proposed algorithm, the likelihood for the training data is guaranteed to increase at each iteration of parameter re-estimation. Extension to continuous visual speech recognition and integration of the visual information to auditory information will be future works.

ACKNOWLEDGMENT

This work was partially supported by the Ministry of Education, Science, Sports and Culture Japan, Grant-in-Aid for Encouragement of Young Scientists, 10780226.

REFERENCES

[1] J. Luettin, N. Thacker, S. Beet, "Speechreading using Shape and Intensity Information," Proc. ICSLP, pp. 58–61, 1996.

[2] J. Luettin, "Towards Speaker Independent continuous Speechreading," Proc. Eurospeech, pp. 1991–1994, 1997.

[3] G. Potamianos and A. Potamianos, "Speaker Adaptation for Audio-Visual Speech Recognition" Proc. Eurospeech, pp. 1291–1294, 1999.

[4] 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.

[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] Y. Nankaku, K. Tokuda and T. Kitamura, "Intensity- and Location-Normalized Training for HMM-Based Visual Speech Recognition," Proc. Eurospeech, pp. 1287–1290, 1999.

[7] http://www.uk.infowin.org/ACTS/RUS/PROJECTS /ac102.htm

[8] 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.