HMM-BASED VISUAL SPEECH RECOGNITION USING INTENSITY AND LOCATION NORMALIZATION

Oscar Vanegas, Akiji Tanaka, Keiichi Tokuda, and Tadashi Kitamura

Department of Computer Science
Nagoya Institute of Technology, Nagoya, 466-8555 JAPAN
E-mail: {oscar, akiji, tokuda, kitamura}@ics.nitech.ac.jp

ABSTRACT

This paper describes intensity and location normalization techniques for improving the performance of visual speech recognizers used in audio-visual speech recognition. For auditory speech recognition, there exist many methods for dealing with channel characteristics and speaker individualities, e.g., CMN (cepstral mean normalization), SAT (speaker adaptive training). We present two techniques similar to CMN and SAT, respectively, for intensity and location normalization in visual speech recognition. Word recognition experiments based on HMM show that a significant improvement in recogniton performance is achieved by combining the two techniques.

1. INTRODUCTION

It has been shown that the image sequence of lips as well as acoustic speech signal plays an important role for improving the speech recognition performance, especially in noisy environments [1]. 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 (e.g., [2]) and image- or pixel-based approach (e.g., [3]). 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 variability of lighting condition, lip location, rotation, and sacaling.

In auditory speech recognition, there exist many techniques for dealing with variability of channels and speakers, e.g., CMN (cepstral mean normalization) [4], MAP adaptation [5], MLLR [6] and SAT (speaker adaptive training) [7]. Our approach to visual speech recognition is based on the success of the normalization approaches for auditory speech recognition. In this paper, we present a simple technique for normalizing average intensity of the image sequence and a location-normalized training technique, similar to CMN and SAT, respectively. Word recognition experiments based on HMM (hidden Markov model) show that significant error rate reduction is achieved by combining the two techniques.

2. INTENSITY NORMALIZATION

CMN (Cepstral Mean Normalization) [4] is the simplest feature-based normalization technique that is used mainly to counteract channel effects. In order to normalize the variation of the intensity of lip images, the mean intensity over the image sequence is subtracted from each pixel value in the frames in a similar manner of CMN, that is, the value of the pixel at location (x, y) in frame t after normalization is given by

$$\bar{I}_t(x, y) = I_t(x, y) - \frac{1}{TXY} \sum_{t=1}^{T} \sum_{x=1}^{X} \sum_{y=1}^{Y} I_t(x,\ y) \qquad (1)$$

where $I_t(x, y)$ is the value of the pixel at location (x, y) in frame t, T is the number of frames in the image sequence, and $X \times Y$ is the size of each lip image. Although this approach does not solve all problems of lighting variation, e.g., lighting direction, it can improve recognition performance significantly in the case where subsampled image is used as feature vector as shown in **4.1**.

3. LOCATION NORMALIZATION

An inherent difficulty of speaker independent speech recognition is that the resulting statistical models, i.e., HMMs, have to contend with a wide range of variation in the speech parameters caused by inter-speaker variability. As a result, the distributions of different classes overlap each other, and the discriminatory capabilities of the statistical model may be reduced. In order to avoid this problem, SAT (Speaker Adaptive Training) [7], a normalized training technique, in which speaker normalization was integrated in the model training, was developed for auditory speech recognition. In SAT, a set of transformations for normalizing each of training speakers and the parameters of the HMMs are jointly estimated.

We assume that the mouth part is extracted from the face image sequence by using some region extraction algorithm. However, the extracted region is considered to have some degree of variation of location. If the HMM is trained with such a variation of location, an HMM with a large variance might be obtained as in the case of speaker independent model training. Therefore, we propose a normalized training technique similar to SAT, which

integrates the location normalization for each utterance into the model training. For the location-normalized training, it is neccessary to jointly estimate the best lip location for each utterance and the parameters of the HMMs. In a similar manner of SAT, an iterative approach is adopted in which one of these set of parameters (the lip locations and the HMM parameters) is estimated at each stage and the maximum likelihood estimation is used individually for each set of parameters assuming the other parameters are fixed. Thus the training algorithm iterates the following procedure several times:

(a) **Location Normalization**
For each training utterance, find the best lip location in the sense that its likelihood is the highest for the current HMM.

(b) **Model Update**
Update the HMMs by the Baum-Welch re-estimation algorithm using all training utterances having the best location.

In the testing, to get the likelihood values of an utterance for all HMMs, procedure (a) is applied for all HMMs, and the model which gives the highest likelihood is chosen as the recognition result. We assume that the lips does not move very much during one utterance. In the procedure (a), the likelihood is measured by the Viterbi algorithm. To obtain the best location avoiding a large amount of computation required for the exhaustive search, we apply the following sub-optimum search procedure to each utterance:.

Step 0. Given an initial guess for the location of the region containing the lips.

Step 1. In total 8 kinds of lip image sequences are extracted from the original face image sequence by shifting the region to be extracted $\pm L$ pixels in x and y directions.

Step 2. From the 8 lip image sequences extracted in step 1 and the current lip image sequence, 9 lip image sequences in total, choose a lip image sequence whose likelihood is the highest for the HMM.

Step 3. If the lip image sequence chosen in step 2 is the current lip image sequence, go to step 4. Otherwise the chosen lip image sequence is used as the new current lip image sequence and go to step 1.

Step 4. If $L = 1$, stop. Otherwise set $L \leftarrow \lfloor L/2 \rfloor$ and go to step 1.

When we cannot obtain pixel values of outside area of the initial lip images, the pixel value obtained by shifting the region to be extracted is given by

$$\hat{I}_t(x, y) = I_t((x - u) \bmod X,\ (y - v) \bmod Y) \qquad (2)$$

where $I_t(x, y)$ is the value of the pixel at location (x, y) in frame t of the initial lip image sequence, (u, v) is the amount of displacement and $X \times Y$ is the size of the lip region to be extracted.

4. EXPERIMENTS

Two kinds of feature vectors, subsampled image and two-dimensional DCT (2D-DCT) coefficients, were used in these experiments. In the sub-sampling, for the dimension reduction, $m \times n$ pixels from the image were defined as one block, and the average value of the pixels in each block was considered to represent the block. In 2D-DCT, $k \times l$ coefficients were extracted from the lower order of the 2D-DCT coefficients. Both in the above subsampling and 2D-DCT, the vectors of each frame (static parameters) and the difference between succesive two frames (delta parameters or dynamic feature parameters) were combined to form the final feature vector.[1] Experiments of word recognition by using continous density HMMs were performed. Each word class was modeled by an HMM which is left-to-right model with 5 states. Each state has a single Gaussian distribution with diagonal covariance.

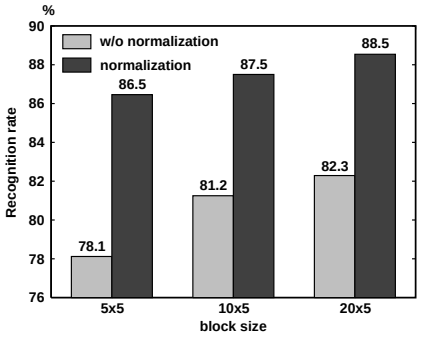

Figure 1: Effect of intensity normalization. (subsampling)

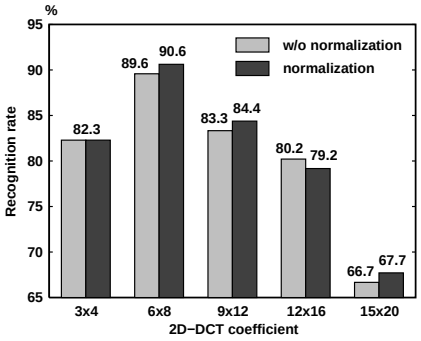

Figure 2: Effect of intensity normalization. (DCT coefficients)

For the experiments, the Tulips1 database [3] was used, which is a bimodal database consisting of lip image sequences and acoustic 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. The database reflects a large variety of lip locations and lighting conditions. 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. The initial value of "L" was set to 10.

[1] Preliminary experiments showed that the the use of delta parameters in addition to the static parameters is effective: error rate reduction of about 60% was achieved for subsampled image

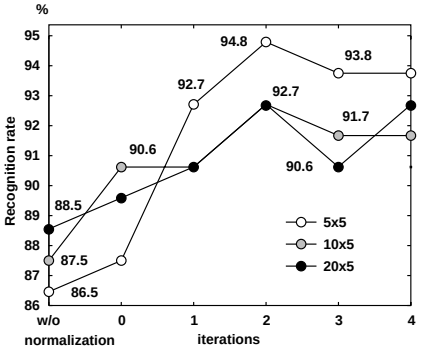

Figure 3: Effect of location-normalized traning. (subsampling)

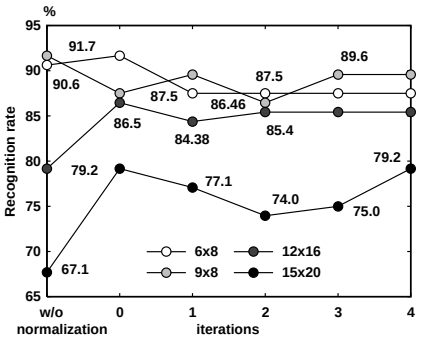

Figure 4: Effect of location-normalized traning. (DCT coefficients)

4.1. Intensity Normalization

Experimental results of intensity normalization are shown in Figs. 1 and 2. It can be seen that the intensity normalization is effective for subsampled image: recognition rate of 88.5% and the error rate reduction of 35% were achieved for the block size of 20×5. On the other hand, the effect was not significant for 2D-DCT coefficients while 2D-DCT coefficients achieved a high recognition rate of 89.6% without the intensity normalization. It is considered that 2D-DCT coefficients are robust to the variation of average intensity since the change of average intensity does not affect the values of 2D-DCT coefficients except for $(0, 0)$-th coefficient.

For subsampled image, a large block size of 20×5 gave the best recognition rate, and for 2D-DCT, the relatively small number of coefficients of 6×8 gave the best recognition rate. They are considered because the use of a large block size or a small number of 2D-DCT coefficients give a rough spatial resolution which results in less influence of variation of lip location.

4.2. Location Normalization

Figs. 3 and 4 show the results of applying location normalization in addition to intensity normalization. Iteration "0" indicates the result of recognition in which location normalization of test data was carried out whereas the location-normalized training was not applied to the HMMs.

For the block size of 5×5, a recognition rate of 94.8% and an error rate reduction of 61% were achieved. This implies 76% reduction in error rate compared with the case without intensity and location normalization. From the fact that the recognition rates for the same task by using other methods [3], [2] were about 90%, the effectiveness of the proposed normalization technique can be confirmed.

From the point of view that the location normalization was more effective for smaller block size of subsampling, it is understood that it is effective to use a feature vector with some degree of higher spatial resolution when we apply the location-normalized training. On the other hand, the location-normalization did not improve the recognition rate for 2D-DCT significantly. This coincides with the fact that 2D-DCT reduces the spatial resolution.

Fig. 5 shows the obtained models with or without location-normalized training. This figure shows the values of the mean vectors and the variances (i.e., diagonal covariances) represented by gray levels. As seen in Fig. 5, the images representing the mean vectors become sharp after location-normalized training and the values of the variances after this process are smaller than those before it. Therefore it means that a better class separation can be obtained.

5. CONCLUSIONS

In order to improve visual speech recognition performance, we proposed two techniques for normalization of lighting condition and lip location. The recognition performance is significantly improved by combining the two techniques: a recognition rate of 94.8% and an error rate reduction of 76% were achieved. It was also shown that for the location normalization it is effective to use a feature parameter with some degree of higher spatial resolution, e.g., nearly the raw pixels.

The normalized training technique will be extended for normalizing lip rotation and scaling. Integration of the visual information to auditory information will also be a future work.

6. ACKNOWLEDGEMENT

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

7. REFERENCES

1. D. G. Stork, M. E. Hennecke (eds.), *Speechreading by Humans and Machines*, Springer Verlag, 1996.
2. J. Luettin, N. A. Thacker and S. W. Beet, "Speechreading using shape and intensity information," in *Proc. ICSLP'96*, pp.58–61, 1996.
3. J. R. Movellan and B. Chadderdon, "Channel peparability in the audio-visual integration of speech: a Bayesian approach," D. G. Stork, M. E. Hennecke (eds.), *Speechreading by Humans and Machines*, Springer Verlag, 1996.

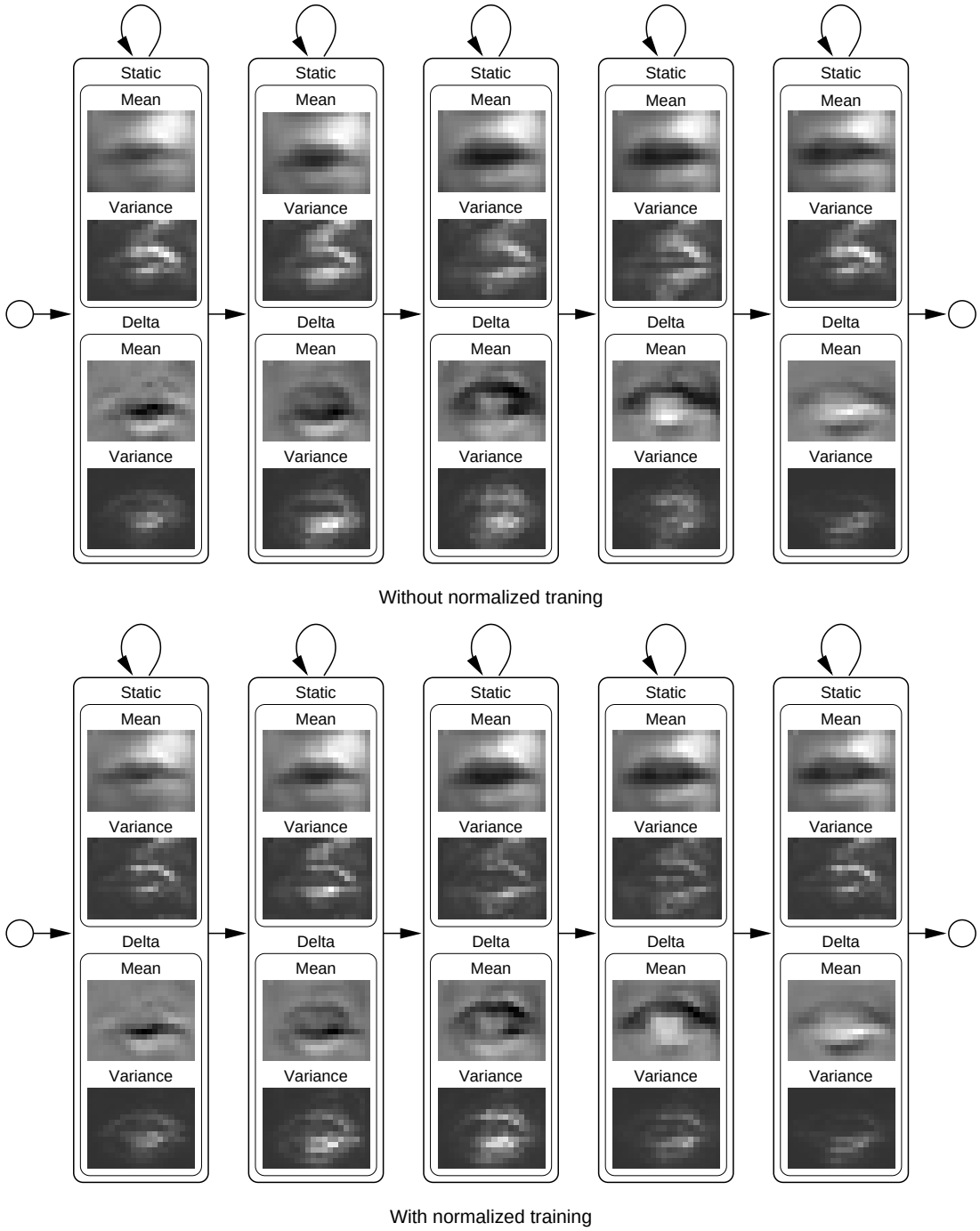

Figure 5: Location-normalized training for a model /one/.

4. B. A. Atal, "Effictiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification," *J. Acoust. Soc. Am.*, vol.68, no.4, pp.1304–1312, Apr. 1974.

5. C. H. Lee, C. H. Lin and B. H. Juang, "A study on speaker adaptation of the parameters of continuous density hidden Markov models," *IEEE Trans. Acoust., Speech, Signal Processing*, vol.39, no.4, pp.806–814, Apr. 1992.

6. M. J. F. Gales, and P. C. Woodland, "Mean and variance adaptation within the MLLR framework," *Computer Speech and Language*, vol.10, no.4, pp.249–264, Apr. 1996.

7. T. Anastasakos, J. McDonough and J. Makhoul, "Speaker adaptive training: a maximum likelihood aproach to speaker normalization," in *Proc. ICASSP'97*, pp.1043–1046, 1997.