PAPER

Lip Location Normalized Training for Visual Speech Recognition

Oscar VANEGAS[†], ***Student Member***, **Keiichi TOKUDA**[†], ***and*** **Tadashi KITAMURA**[†], ***Regular Members***

SUMMARY This paper describes a method to normalize the lip position for improving the performance of a visual-information-based speech recognition system. Basically, there are two types of information useful in speech recognition processes; the first one is the speech signal itself and the second one is the visual information from the lips in motion. This paper tries to solve some problems caused by using images from the lips in motion such as the effect produced by the variation of the lip location. The proposed lip location normalization method is based on a search algorithm of the lip position in which the location normalization is integrated into the model training. Experiments of speaker-independent isolated word recognition were carried out on the Tulips1 and M2VTS databases. Experiments showed a recognition rate of 74.5% and an error reduction rate of 35.7% for the ten digits word recognition M2VTS database.

key words: *hidden Markov model, lip location normalization, lipreading, Tulips1, M2VTS*

1. Introduction

Its well known that apart from the speech signal, an important parameter on speech recognition processes is the visual characteristics of the lips in motion in order to improve the robustness and accuracy of speech recognition [1]. Basically, there are two methods to extract speech information from image sequences of the lips, the model-based [2], [4], [6] and the image-based method [7], [8]. In the model-based method, an outline of the lips is first required with a minimum amount of parameters. Models constructed by this method present less influence from the lighting conditions and the lip location. This characteristic makes it suitable for speech recognition processes but its principal disadvantage is that it is difficult to construct the model. In case of the image-based method, a larger amount of data is required. Also there is a large influence of the lighting conditions and of the location of lips, but there is not much difficulty on constructing the model.

This paper presents an approach based on the image-based method in order to solve the effect of the lip location on the speech recognition results. The proposed method is based on a search algorithm of the lip position in which a lip location normalization is integrated in the model training [10].

Manuscript received August 9, 1999.
Manuscript revised April 17, 2000.

[†]The authors are with the Nagoya Institute of Technology, Nagoya-shi, 466–8555 Japan.

In this paper, experiments based on HMMs (Hidden Markov Models) on speaker-independent isolated word recognition were carried out on two databases, the Tulips1 [7] and the M2VTS [14] database. As a result of applying the proposed method, a good recognition performance was achieved and the error rate was considerably reduced.

This paper is organized as follows. The next section describes the location normalization algorithm. Experiments on Tulips1 and M2VTS are described in Sects. 3 and 4 respectively. Finally, conclusions and future works are presented in the final section.

2. Location Normalized Training

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.

Although the mouth part is extracted from original images by some lip extraction algorithm, it has some degree of variation of location as shown in the left side of Fig. 2. If the HMM is trained with images having such a variation of location, an HMM with a large variance might be obtained. As a result, the distributions of different classes overlap each other, and the discriminatory capabilities of the statistical models may be reduced. Therefore, we propose a normalized training technique, which integrates the location normalization for each utterance into the model training. For the location-normalized training, it is necessary to jointly estimate the best lip location for each utterance and the parameters of the HMMs. 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 (a) and (b) steps several times after setting the initial model.

- **Set Initial Model**
  Get the initial HMMs$^{(0)}$ using the original training data set $\{I^{(0)}\}$.

(a) **Best Location Search**

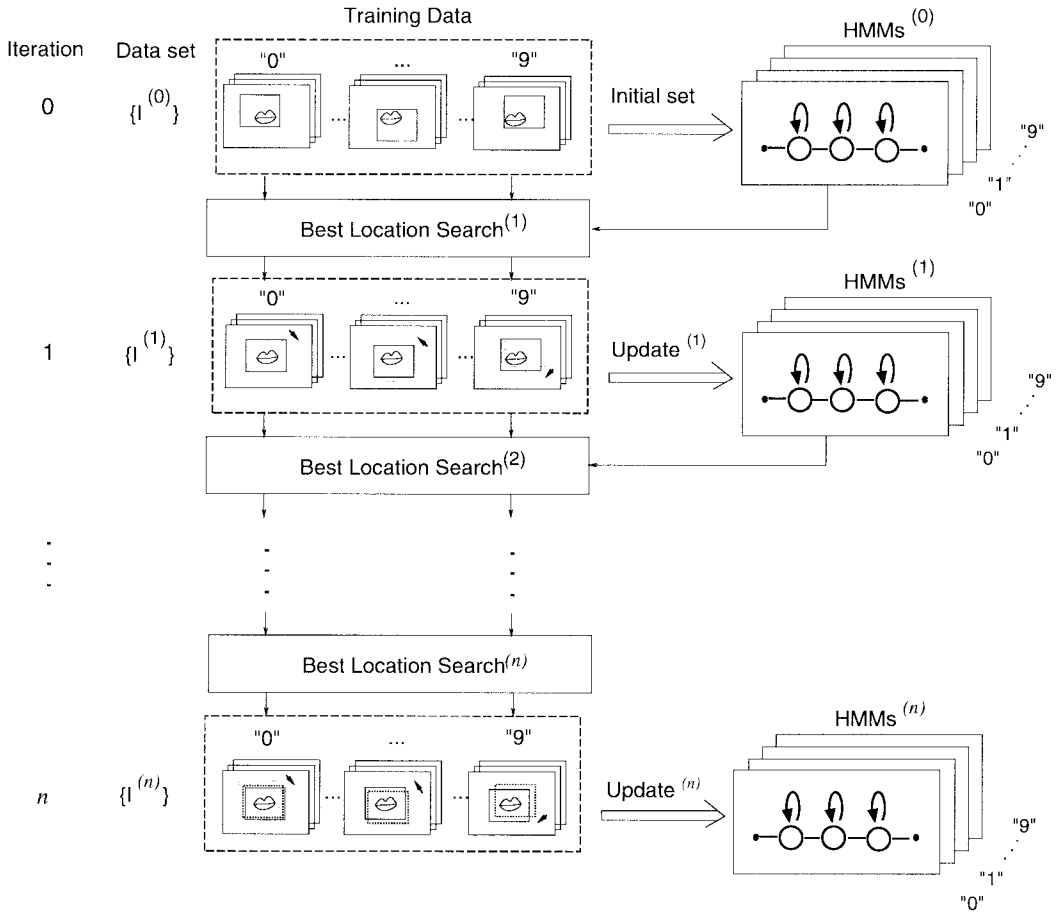


Fig. 1 Lip location normalized training method scheme.

For the training data set $\{I^{(k-1)}\}$, find the best lip location data set $\{I^{(k)}\}$ in the sense that the likelihood of each word in the data set is the highest for the corresponding HMMs$^{(k-1)}$, $(k = 1, 2, .., n-1)$.

(b) **Model Update**
Update the model HMMs$^{(k)}$ by the Baum-Welch re-estimation algorithm using all training data set $\{I^{(k)}\}$ having the best location.

Figure 1 shows the lip location normalized training process scheme. In the iteration "0," the original data set $\{I^{(0)}\}$ is taken as training data and the first HMM model HMMs$^{(0)}$ is constructed. In the iteration "1" search the lip location for the utterance data set $\{I^{(1)}\}$ having the highest likelihood for this HMM model, then, update the HMM models HMMs$^{(1)}$ by the Baum-Welch re-estimation algorithm using the utterances with the best lip location. After this process, the same procedure is repeated. That is, for each iteration, search the lip location with the highest likelihood for the current model HMMs$^{(k)}$, then update the model HMMs$^{(k+1)}$ by using the utterances with the best lip location. In the proposed method, the lip location normalization is integrated in the model training in a similar manner to SAT [10] in which the speaker normalization is integrated in the training process.

In case of the training data, the likelihood for a specific word or utterance is obtained from the corresponding HMM word model. In the procedure (a), the likelihood is measured by the Viterbi algorithm [9]. In order to obtain the best location avoiding a large amount of computation required for the exhaustive search, we apply the following best location search (Fig. 1) procedure to each utterance:.

- **Best Location Search Procedure**

  Step 0. Give 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 lip 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

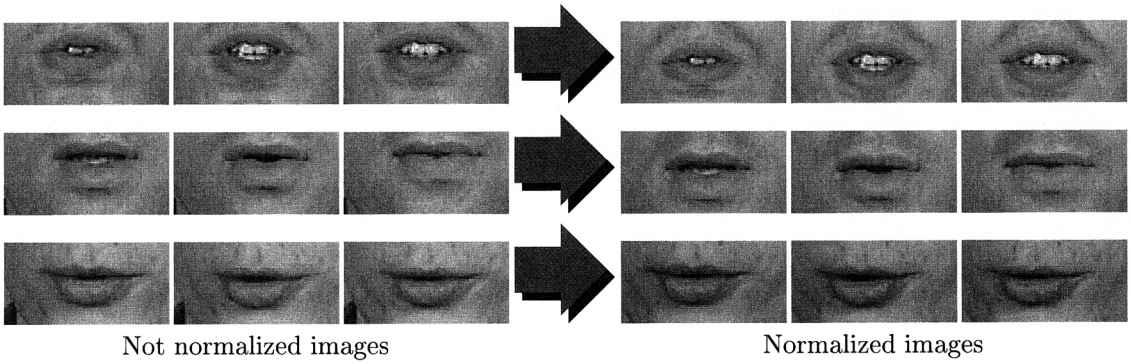


Fig. 2 Effect of location normalized training technique on M2VTS images.

is the current lip image sequence, go to step 4. Otherwise use the chosen lip image sequence 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.

Figure 2 shows M2VTS images, on the left, images without location normalized training process, on the right the same images after applying this technique. We can see that the lip location of all the images came to the center place of each frame. The initial value of L was fixed to 10 at first and reduced gradually until 1 in this study. After applying this normalization method, the error was very much reduced as shown in Sects. 3 and 4.

To normalize the lighting conditions on images, we utilized a simple method in which the average value of intensities is subtracted from all the pixels in an utterance. Intensity normalization and lip location normalized training process were applied over the complete set of utterances.

For the testing data, the best lip location for all utterances is obtained in the sense that its likelihood is the highest for all HMM word models. That is, the best lip location for the training data is determined as the maximum likelihood γ such that $\gamma \in \{l_0, l_1, \ldots, l_9\}$ and l_w means the likelihood of an utterance obtained from the HMM model of word w.

3. Experiments on the Tulips1

In order to explain the experiments carried out in this paper, we are introducing the next terminology:

"w/o N" means that no normalization method has been applied on the utterances. The utterances are in its original state.

"w I" means that only the Intensity Normalization method has been applied on utterances.

Iteration number *"k"* means the corresponding recognition results using HMMs$^{(k)}$ with intensity normalization. The lip location normalized training process is also applied over the testing data set with intensity normalization.

Experiments of speaker-independent isolated word recognition by using the Tulips1 and the M2VTS databases were carried out using frames with subsampling data in order to reduce the number of parameters, that is, original frames were divided into blocks, each block having the average intensity of the pixels inside that block. Preliminary experiments [13] showed that the use of subsampling data with small block size gives the best recognition rates, probably because a small block size provides higher spatial resolution in which the lip location can affect considerably the recognition rate. Therefore, blocks of 5×5, and 10×5 pixels per block were used in this study. The subsampled frames are constructed after making the 8 lip image sequences described in Step 1. The lip location normalized training process was applied over the complete set of utterances.

3.1 Tulips1 Database

For the first experiment, the Tulips1 database [7] was used, which is a bimodal database comprising lip image sequences and speech signals of 9 males and 3 females, in total 12 speakers. Each speaker pronounces English numbers, "one," "two," "three" and "four," each twice. Images are in grey scale with 8 bytes/pixel. Images sequences are sampled at 30 Hz and the frame size is 100×75 pixels. The database reflects a large variety of lip locations and lighting conditions. In this paper, the so-called "leave-one-out method" was applied to the experiments. In the method, one of 12 subjects was used for testing and the remaining 11 subjects were used for training, so that the number of test data for each word was 24 and it resulted in a total of 96 test data by 12 speakers.

Subsampled frames with two different block sizes were used; 5×5 and 10×5 pixels per block, that is, two-dimensional feature vectors of 300 and 150 parameters, respectively. Vectors of each frame and the difference between successive two frames (delta parameters)

were combined. Experiments of recognition by using a continuous Hidden Markov Model (HMM) [12] were carried out. Experiments were carried out varying the number of states from 3 to 5 and better results were obtained when 5 states were used. In this paper, each word model was represented by one HMM which is a left-to-right model with 5 states and a single Gaussian distribution of diagonal covariance.

As was explained in Sect. 2, in order to apply the lip location normalization method, it is necessary to shift the lip region around the current position to get the lip location with the highest likelihood for the HMM, but in case of Tulips1, images cover only the lip region and in order to shift the lip location we need to compensate intensities outside area of original images. In case of Tulips1, intensities obtained by shifting the original image are given by

$$I^1(x, y) = I^0((x - u) \bmod X, (y - v) \bmod Y) \quad (1)$$

where $I^0(x, y)$ is the intensity of the original image, (u, v) is the amount of displacement and $X \times Y$ is the size of the original image.

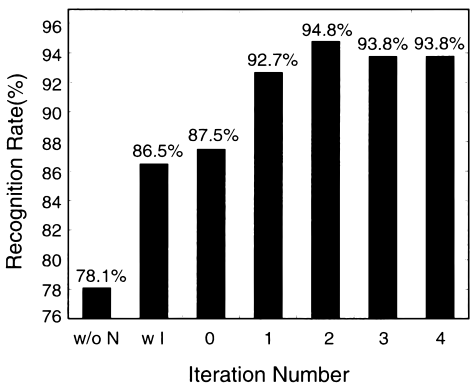


Fig. 3 Effect of lip location normalization for subsampling data with block of 5×5; Database Tulips1.

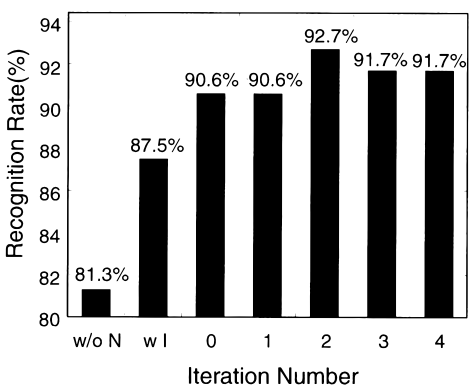


Fig. 4 Effect of lip location normalization for subsampling data with block of 10×5; Database Tulips1.

3.2 Results

Figures 3 and 4 show the corresponding results when subsampling data with block size of 5×5 and 10×5 were used. In Fig. 3, a recognition rate of 78.1% was obtained when nothing was applied on utterances, that is, neither intensity normalization nor lip location normalization. 86.5% was obtained after applying only intensity normalization on utterances, which means 38.4% of error reduction rate. In case of Iteration "0" 87.5% of recognition rate was obtained with an error reduction rate of 42.9%. For Iteration "1" the error reduction was increased up to 71.2% with a recognition rate of 92.7%. The highest obtained recognition rate was 94.8% on Iteration "2," the error was reduced up to 76.3%.

In case of subsampling data with block size of 10×5 pixels, the curve of recognition rate showed almost the same tendency. The recognition rate was increasing up to iteration "2" where the highest obtained rate of recognition was 92.7% with an error reduction of 61.1%.

From the fact that the recognition rate for the same task by using other methods [2], [5] remains at the level of 90% in case of Tulips1, the effectiveness of the proposed method can be confirmed.

4. Experiments on the M2VTS

4.1 M2VTS Database

Although the proposed method gives good results, the number of testing data in Tulips1 is not enough and original images are fixed, that is, do not offer the possibility of extracting complete lips images with different positions. Therefore, the M2VTS database was used in order to confirm the effectiveness of the proposed method. The M2VTS (Multi Modal Verification for Teleservices and Security applications) [14] is a multimodal face database recorded at UCL (Catholic University of Louvain). This database contains images and audio signal information of 37 speakers (male and female) and provides 5 shots for each person. During each shot, people have been asked to count from '0' to '9' in their native language. One shot is a sequence of the ten digits pronounced 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. Images are sampled at 25 Hz. This study was carried out by using the first four shots. The database contains full color images of 286×350 pixels and they were converted into grey-level images for the experiments in this study. For the experiments, all 37 speaker lip images and same experimental conditions as [3] were used.

For the experiments, images from the lips in motion were extracted from the database. The images

were analyzed and a central point was visually calculated trying to leave the lips in the most center place of frames. Each frame consists of 80×40 pixels. The word boundaries of the training data were found by an HMM based speech recognition system which was used to segment and label the sentences. Subsampled frames with two different block sizes were used; 5×5 and 10×5 pixels per block, that is, two-dimensional feature vectors of 128 and 64 parameters, respectively.

Experiments of recognition by using continuous HMMs were carried out varying the number of states and the highest recognition rates were obtained when 8 states were used. Each word model was represented by one HMM which is a left to right model with 8 states and two single Gaussian distributions of diagonal covariance. The feature vector consisted of the static, delta and acceleration coefficients.

For the testing process, the leave-one-out method was used. It means that 37 leave-one-out testing were carried out, each one with 1440 training words. For the testing process, 40 test words per speaker, producing 1480 testing utterances in total.

4.2 Results

Experiments on isolated word recognition were carried out by using the extracted image sequences from the M2VTS database.

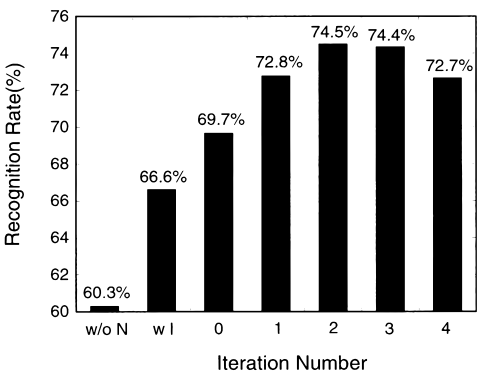


Fig. 5 Effect of lip location normalization for subsampling data with block of 5×5; Database M2VTS.

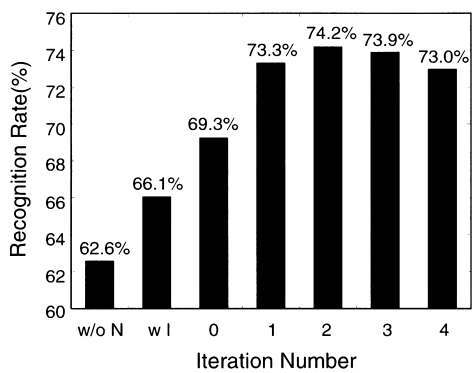


Fig. 6 Effect of lip location normalization for subsampling data with block of 10×5; Database M2VTS.

Figures 5 and 6 show the obtained results. As is shown here, in case of blocks of 5×5, a recognition rate of 60.3% was obtained without applying neither intensity normalization nor lip location normalization and 66.6% was obtained by applying only the intensity normalization over the training and the testing data; this means an error reduction rate of 15.9%. After applying the lip location normalized process, the recognition rate was increasing up to 74.5% on the second iteration, and an obtained error reduction of 35.7%. In case of blocks of 10×5, the recognition results showed the same tendency as in case of blocks of 5×5. The highest recognition rate was 74.2% obtained on the second iteration with an error reduction of 31%. The highest recognition rate was obtained with subsampling data of the smallest block size of 5×5, probably because this is the block size which provides the highest spatial resolution in which the lip location can affect considerably the recognition rate. The location normalization gave a better improvement in recognition rate than the improvement given by the intensity normalization as is shown in Figs. 5 and 6. The tendency shown in this figures for subsampling data with block size of 5×5 is similar to the tendency given by subsampling data of 10×5.

Tables 1 and 2 show the confusion matrix for "w I" and iteration "2," for block size of 5×5 respectively. As shown in the tables, better results can be appreciated in the last case. All the words increased the number

Table 1 Confusion Matrix for the case "w I" for blocks of 5×5; Database M2VTS.

	0	1	2	3	4	5	6	7	8	9	Recg.
zero	120	5	4	0	6	0	0	4	0	7	81%
un	1	99	4	0	25	1	5	3	4	2	67%
deux	7	5	106	0	2	1	9	1	4	12	72%
trois	1	6	2	107	7	1	0	1	16	3	72%
quatre	3	17	0	4	65	10	6	13	0	10	44%
cinq	4	6	3	2	10	73	30	13	0	5	49%
six	1	7	5	0	3	17	104	9	0	1	70%
sept	3	12	2	1	15	19	13	77	2	2	52%
huit	1	3	2	18	7	1	0	1	106	2	72%
neuf	12	6	15	0	8	2	0	0	2	102	69%

Table 2 Confusion Matrix for the case of iteration "2" for blocks of 5×5; Database M2VTS.

	0	1	2	3	4	5	6	7	8	9	Recg.
zero	127	5	4	0	1	4	0	1	0	5	86%
un	1	115	0	0	8	4	5	2	4	3	78%
deux	9	4	113	1	3	1	5	0	1	9	76%
trois	0	5	0	126	2	1	0	0	10	3	85%
quatre	3	19	1	4	76	9	6	5	0	6	51%
cinq	4	10	1	0	7	87	25	12	1	1	59%
six	0	7	4	0	5	19	102	10	0	0	69%
sept	3	10	2	0	4	19	10	99	0	0	67%
huit	0	3	2	23	1	2	1	1	107	5	72%
neuf	6	4	7	0	2	3	0	1	1	124	84%

of times they were recognized except word "six" which the number of recognized utterances decreased in Table 2. We can appreciate that there was a tendency of "miss-recognizing" between utterances "cinq," "six" and "sept," that is, in Table 1, 30 utterances of "cinq" were recognized as "six," and 13 as "sept." 17 words of "six" were recognized as "cinq" and 9 as "sept." 19 words of "sept" were recognized as "cinq" and 13 as "six." Probably, because these words produce almost the same lip movements. Table 2 shows almost the same tendency but between these words, the number of well recognized and bad recognized words were very much improved.

Figures 7 and 8 show the obtained recognition rate for all the 37 users at "w I" and iteration "2" in the case of blocks of 5×5 pixels. As shown in the figures, a considerable difference of the recognition rate per user can be appreciated. It is important to note that there are 3 persons with low recognition results, less than 30%. One of them corresponds to a person who does not speak the same language as the others, so that the expected recognition rate should be also low. Another speaker is a person whose mouth region is full of beard, and it makes the recognition a little hard. The third one is a person who seems to be laughing during the utterance.

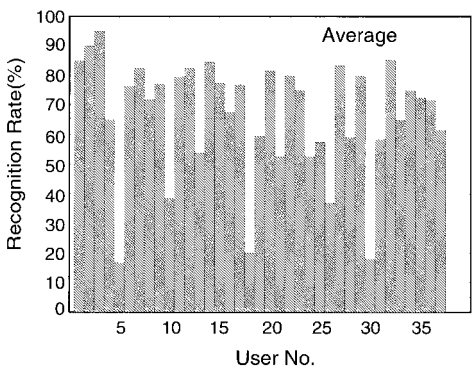


Fig. 7 Recognition rate per users at iteration "w I"; block size of 5×5.

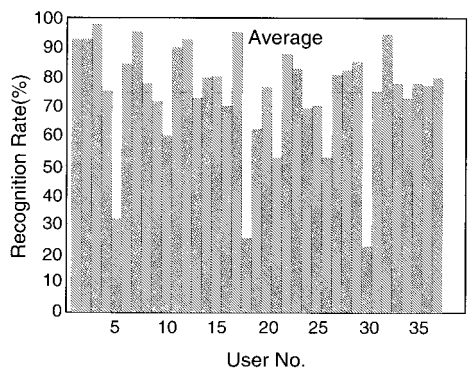


Fig. 8 Recognition rate per users at iteration "2"; block size of 5×5.

Experiments using same database and also same recordings were carried out in Ref. [3]. Those experiments were based on the model-based method, therefore images are already normalized, so that, results are not affected by the lip location. HMMs also use 8 states and 2 Gaussian distributions and for the training and testing process, experiments also are based on the leave-one-out method. Almost same conditions are given for the experiments. Experiments of isolated word recognition were carried out and a recognition rate of 60.2% was obtained by [3]. In this study, after applying the proposed normalization method for isolated word recognition, 74.5% was obtained. From this fact, the effectiveness of the proposed method can be confirmed.

4.3 Discussion

Figure 9 shows images obtained from the values of the mean vector and the square root of the variance vector of HMMs without and with lip location normalized training process for one of the mixtures on M2VTS. In the figure, values of means and variances of static and delta parameters are represented by gray levels and average values of variance vectors are also shown. We can see that normalized images are sharper than the same ones without normalization. In case of the variance vector, normalized images become darker than the ones without normalization. It means that smaller values could be obtained for the variance after the normalized training method.

Figure 10 shows the average values of the obtained variance per pixel for static and delta data on the HMM models. They were obtained by calculating the average value between variances on the two mixtures. The values of the variance are decreasing as the iteration number increases in both cases. It means that there exist a tendency of finding a better class separation after the location normalized training process.

Figure 11 shows the change of the obtained likelihood per frame for the training data through the iterations, which is obtained by subsampling data with 5×5 pixels per block for the M2VTS database. The likelihood is increasing as the iteration number increases. At first, the likelihood is increased significantly at iteration "1." It shows that the proposed method is very much effective during the first iteration.

The property of identifying positions of the lips depends on all the training data. That is, if the majority of lip images have lips located at center place and HMMs are trained with such a data, our proposed normalization method will be able to identify lips located at center place of frames. For example, assuming that the majority of the lip images have lips located at the left place of frames, the proposed method will identify lips when they are located at the left.

The identification property is different from the recognition property. The recognition depends on the

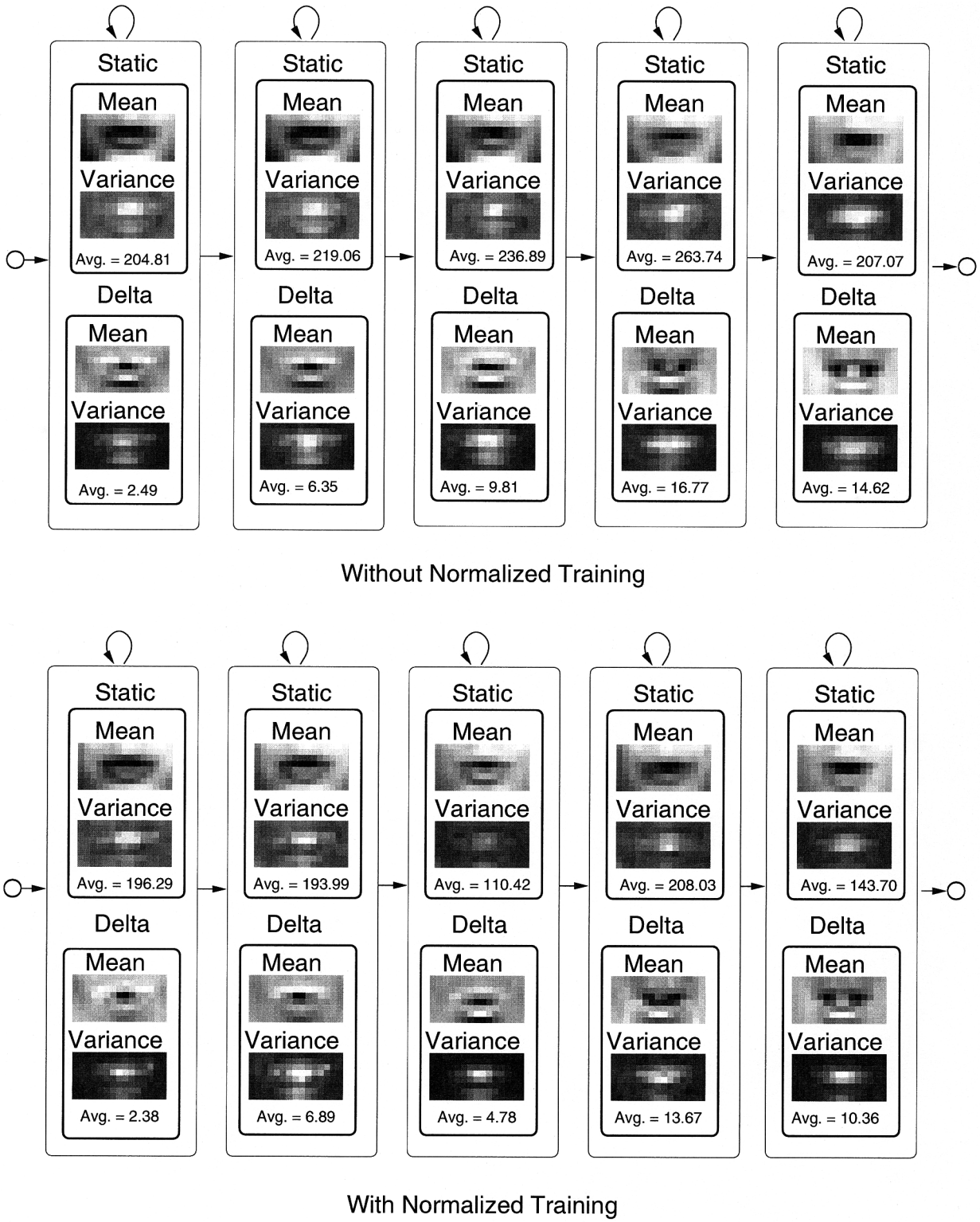


Fig. 9 Effect of lip location normalization for HMM models. Utterance "Toroa." Subsampling data with block of 5×5 (M2VTS Database). Above, data on "w I." Below, data on iteration "2."

testing data, that is, if the testing lip images correspond to difficult cases to make recognition, for example, users who have beard or people who show big teeth etc. and HMMs are not trained with this kind of data, low recognition results can be obtained. Otherwise, good recognition results can be obtained.

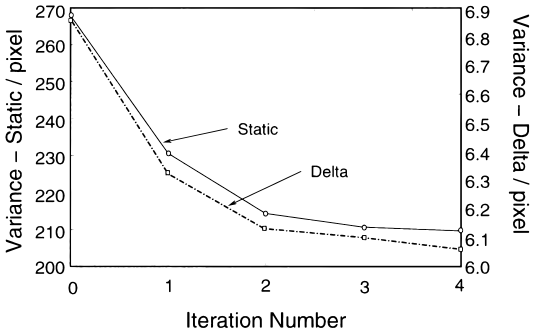


Fig. 10 Variance through the iterations; blocks of 5 × 5; M2VTS database.

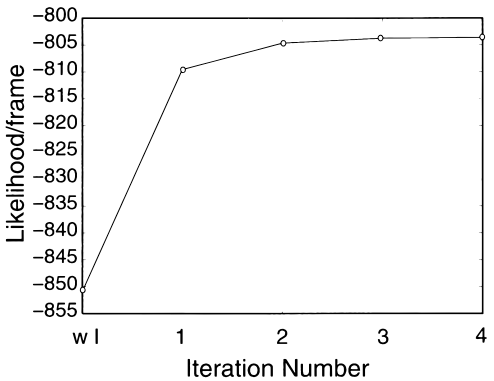


Fig. 11 Likelihood through the iterations; blocks of 5 × 5; M2VTS database.

After an analysis on Figs. 3, 4, 5 and 6 we conclude that in all cases, recognition rates are incremented until a maximun value is reached, after that, recognition rates decrease. This behavior can be because of an overtraining presented by HMMs. HMM models are trained one more time on each iteration and for all iterations same testing data is used. The model corresponding to the iteration with the highest recognition rate corresponds to the best model for the testing data. After this iteration, models are over-trained and they are not considered to be the best ones to produce better recognition rates.

Although the obtained recognition rates in this paper are considered as good results compared with other methods, in case of the M2VTS database, recognition rates still remain low. The reasons can be described as follows:

- The frame shift for visual data is very long compared with that of acoustic data. In case of the M2VTS, the frame shift is 40 ms (25 Hz) while the frame shift for acoustic data is normally set to 10–20 ms.
- The M2VTS although is one of the biggest bimodal databases available is still so small considering the big amount of cases which can be given. For example, people with big lips or small lips, people who open very much the lips to speak, people who almost do not move their mouth, people full of beard, people with big teeth, people whose tongue can be easily be seen during the speech etc. Considering this limitation, recognition rates are very much affected in the sense that the HMM models do not consider all possible cases.
- In order to improve the recognition results for speechreading using lip images, is necessary to include other kind of normalizations like the lip size normalization, the lip inclination normalization, etc.

5. Conclusions

In this study, we proposed a location normalized training technique for automatic lip-reading for image-based visual speech recognition. In order to show the effectiveness of the proposed normalization method, the Tulips1 and the M2VTS databases were used. By the proposed method the recognition rate can be significantly improved, recognition rates of 94.8% and 74.5% were obtained for Tulips1 and M2VTS respectively. The recognition error was reduced up to 76.3% in case of tulips1 and 35.7% in case of M2VTS.

In the future, we will apply the lip location normalization on continuous word recognition. Also, we will normalize the size and the slant of images and integrate visual information with audio information.

Acknowledgment

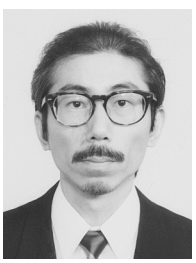
The authors would like to thank Prof. T. Kobayashi of Tokyo Institute of Technology for his comments and suggestions. Also we would like to thank Mr. Akiji Tanaka for his work on this study. This work was partially supported by the Ministry of Education, Science, Sports and Culture, Grant-in-Aid for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

References

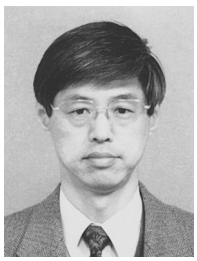
[1] J.R. Movellan and B. Chadderdon, "Channel separability in the audio-visual integration of speech: A Bayesian approach," D.G. Stork and M.E. Hennecke, eds., Speechreading by Humans and Machines, Springer Verlag, 1996.
[2] J. Luettin, N.A. Thacker, and S.W. Beet, "Visual speech recognition using active shape models and hidden Markov models," Proc. ICASSP96, vol.2, pp.817–821, 1996.
[3] J. Luettin, "Towards speaker independent continuous speechreading," Proc. Eurospeech97, pp.1991–1994, 1997.
[4] J. Luettin, N.A. Thacker, and S.W. Beet, "Active shape models for visual speech feature extraction," D.G. Stork and M.E. Hennecke, eds., Speechreading by Humans and Machines, Springer Verlag, 1996.
[5] J. Luettin, N.A. Thacker, and S.W. Beet, "Speechreading using shape and intensity information," Proc. ICSLP96, pp.58–61, 1996.

[6] M. Kass, A. Witkin, and D. Terzopoulos, "Snakes: Active contour models," Int. J. Computer Vision, pp.321–331, 1988.
[7] J.R. Movellan, "Visual speech recognition with stochastic networks," in Advances in Neural Information Processing Systems 7, eds. G. Tesauro, D. Touretzky, and T. Leen, MIT Press, Cambridge, MA, 1995.
[8] C. Bregler, H. Hild, S. Manke, and A. Waibel, "Improved connected letter recognition by lipreading," Proc. ICASSP93, pp.557–560, 1993.
[9] G.D. Forney, "The Viterbi algorithm," Proc. IEEE, vol.61, pp.268–278, March 1973.
[10] J. McDonough, T. Anastasakos, and J. Makhoul, "Speaker adaptive training: A maximum likelihood approach to speaker normalization," Proc. ICASSP97, pp.1043–1046, 1997.
[11] D.A. Reynolds, "Experimental evaluation of features for robust speaker identification," IEEE Trans. Speech & Audio Process., vol.2, pp.639–643, 1994.
[12] L. Rabiner and B. Juang, Fundamentals of Speech Recognition, Prentice-Hall, Englewood Cliffs, NJ, 1993.
[13] O. Vanegas, A. Tanaka, K. Tokuda, and T. Kitamura, "HMM-based visual speech recognition using intensity and location normalization," Proc. ICSLP98, pp.289–292, 1998.
[14] http://www.tele.ucl.ac.be/M2VTS/m2vts_db.ps.Z.

Oscar Vanegas received the B.Eng. degree in system engineering from the National University of Colombia in 1986. He studied a M.Eng. at the same university in 1992. Received a M.Eng. and Dr.Eng. degrees from Nagoya Institute of Technology, Nagoya, Japan, in 1997 and 2000, respectively. His research interests are visual speech processing, image processing and neural networks.



Keiichi Tokuda received the B.E. degree in electrical and electronic engineering from the Nagoya Institute of Technology, Nagoya, Japan, the M.E. and Dr.Eng. degrees in information processing from the Tokyo Institute of Technology, Tokyo, Japan, in 1984, 1986, 1989, respectively. From 1989 to 1996 he was a Research Associate at the Department of Electronic and Electric Engineering, Tokyo Institute of Technology. Since 1996 he has been with the Department of Computer Science, Nagoya Institute of Technology as Associate Professor. His research interests include speech spectral estimation, speech coding, speech synthesis and recognition, and adaptive signal processing. He is a member of IEEE, ASJ, JSAI and IPSJ.



Tadashi Kitamura received the B.Eng. degree from Nagoya Institute of Technology, Nagoya, Japan, 1973, M.Eng. and Dr.Eng. and Dr.Eng. degrees from Tokyo Institute of Technology, Tokyo, Japan, in 1975 and 1978, respectively. From 1978 to 1983 he was a Research Associate at the Research Laboratory of Precision Machinery and Electronics, Tokyo Institute of Technology. From 1983 to 1995 he was an Assistant Professor and an Associate Professor at Nagoya Institute of Technology. From 1993 to 1994 he was a Visiting Research Fellow at the University Wales of Swansea, UK. Since 1995 he has been a Professor at the Department of Computer Science, Nagoya Institute of Technology. His current research interests are speech processing, image processing and multi-modal person authentication. He is a member of IEEE, ASJ, IPJ and ESCA.