A New Approach to Lipreading Using Time-Varying Signal Analysis

Keren Yu, Xiaoyi Jiang, and Horst Bunke
Institut für Informatik und angewandte Mathemathik
Universität Bern, NeuBrückstrasse 10, CH-3012 Bern, Switzerland
E-mail: yu,jiang,bunke@iam.unibe.ch

Abstract

This paper describes a novel approach to visual speech recognition. The intensity of each pixel in an image sequence is considered as a function of time. One-dimensional wavelet and Fourier transform is applied to this intensity-versus-time function to model the lip movements. We present experimental results performed on two databases of English digits and letters, respectively.

CR Categories and Subject Descriptors: I.4.10 [Image Processing]: Image Representation; I.5.2 [Pattern Recognition]: Design Methodology; I.5.4 [Pattern Recognition]: Applications.
General Terms: Algorithms.
Additional Key Words: Fourier transform, wavelet transform, signal processing.

1 Introduction

For decades, most research efforts in automatic speech recognition have focused on the acoustic signal only. One problem in those efforts, however, is that the acoustic recognition rate often decreases significantly in noisy environments such as offices, airports, train station, factory floors, automobile and airplane cockpits, and others. One of the approaches to solving this problem is using the visual signal, which is not affected by acoustic noise. The lip movements represented in the visual signal often contain enough information for a categorization of speech. In addition, a combination of both the acoustic and visual signal possesses the potential of capturing the information from two independent sources and thus improving the overall speech recognition performance. As a matter of fact, studies in human perception have shown that visual information allows people to tolerate an extra 4-dB of noise in the acoustic signal [15]. Also in computer speech recognition, fusion of the acoustic and visual signal has been investigated to improve recognition performance [4].

There have been several works [4, 6, 9, 10, 12, 13] on lipreading based on hidden Markov models (HMM), neural networks, principal component analysis, a.s.o; see [8] for an overview and [20] for a collection of recent work in visual speech recognition. In these approaches features are usually extracted from individual images, and lip movements are modeled by HMM, neural networks, and similar methods.

In this paper, we present a novel method in which the intensity curve of pixels along the time axis is considered as the primary signal. Two transforms, one-dimensional wavelet and Fourier transform, are applied to this signal, alternatively. The wavelet and Fourier coefficients of an intensity curve encode motion information in a compact manner and are used as features for recognition. A similar idea of processing intensity-versus-time curves using Fourier transform has been successfully applied to medical images [1, 3, 16]. The wavelet transform has found other interesting applications in signal processing [7, 11, 14, 17].

In the next section we describe the wavelet and Fourier transform over time and their use for the representation of an image sequence. After that, we give a description of our model construction and recognition method in Section 3. Then, in Section 4 the topic of illumination invariance is discussed. Experimental results are reported in Section 5. Finally, in Section 6 some conclusions are given.

2 Transforms over time

In visual speech recognition, the lip movements are analyzed in a sequence of digitized images of the mouth. The intensity $I(n)$, where n is a frame number, of each pixel of an image sequence can be considered as a function of time. Clearly, the complete information of the image sequence is contained in the intensity-versus-time curves if we consider the curve in every pixel. But eventually, we are looking for features that represent the intensity-versus-time curves in a compact way. Two standard approaches to feature extraction from one-dimensional curves are wavelet and Fourier transform.

Under wavelet transform, an original signal $f(n)$ is decomposed onto an orthonormal

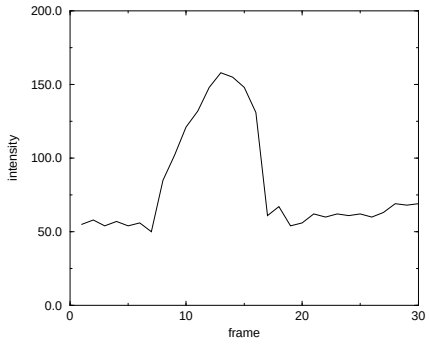

Figure 1: Intensity-versus-time curve.

basis built by dilating and translating a particular function $\psi(x)$, called an orthogonal wavelet, or alternatively, a mother wavelet. The orthonormal basis is

$$(\sqrt{2^j}\psi(2^j x - k))_{(j,k)\in Z^2}. \tag{1}$$

Notice that if the mother wavelet $\psi(x)$ is given, then the other wavelet functions can be calculated from (1) by dilation and translation. Since different mother wavelets produce different classes of wavelets, the behavior of wavelet coefficients can be quite different. In order to obtain good localization properties in both the frequency and spatial domain, the mother wavelets should be chosen carefully. In our work, we tried several wavelet bases, such as the Battle-Lemarie, the Burt-Adelson, the Daubechies, the Haar, the Spline [5], and the Coiflet [2] wavelets. Finally, we found that the Coiflet wavelets gave the best recognition results.

Since the orthonormal basis is defined by (1), it is useful to consider only signals $f(n)$ of length 2^N. This is not a real restriction as any signal with a length different from 2^N can be appropriately scaled. It is possible to use the first few wavelet coefficients to approximately represent the signal $f(n)$.

An alternative to wavelet transform is Fourier transform of the intensity function $I(n)$, which yields

$$c(k) = \frac{1}{N}\sum_{n=0}^{N-1} I(n)e^{\frac{-j2\pi kn}{N}}, \qquad k = 0, 1, \ldots, N-1,$$

where N is the number of frames of the sequence, and $c(k)$ is a complex coefficient. Using the first few Fourier coefficients we are also able to approximately describe the time evolution of a curve.

In order to demonstrate the claim that the first few coefficients represent a signal approximately, let us choose the intensity-versus-time curve of a pixel in the center of the mouth from a real image sequence. Figure 1 shows the original intensity curve of this particular pixel over time. In Figures 2 and 3 the wavelet coefficients of the curve and the real and imaginary parts of the corresponding Fourier coefficients are given, respectively. From Figures 2 and 3 we observe that the amplitudes of the wavelet and Fourier coefficients decrease as the wavelet coefficient number and the Fourier transform sequence number

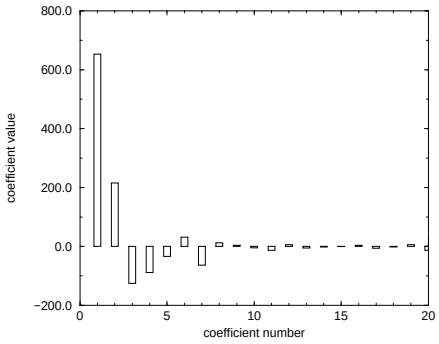

Figure 2: Wavelet coefficients of the original intensity-versus-time curve in Figure 1.

increase. Therefore, a compact representation of an intensity-versus-time curve $I(n)$ is given by the first k wavelet or Fourier coefficients. As a matter of fact, instead of Fourier coefficients, the magnitudes of them are used.

The effectiveness of this representation is demonstrated for the intensity curve in Figure 1. Its reconstruction through the inverse wavelet and Fourier transform, using the first five wavelet and Fourier coefficients, respectively, is shown in Figure 4. Obviously, the reconstructed curves have a high similarity to the original curve. Therefore, both wavelet and Fourier coefficients are suitable to approximately describe the original signal.

In our approach, an image sequence is compactly represented by means of an $h \times w$ matrix C of k-dimensional vectors

$$C = [c_{ij}]_{hw},$$

where h and w are the height and width of the images, respectively, and c_{ij} is a vector of dimension k, containing the first k wavelet coefficients or the magnitude of the first k Fourier coefficients for pixel (i, j).

As an example, Figure 5 shows the matrix C for an image sequence representing the word *zero*, where the first five wavelet coefficients and the magnitudes of the first five

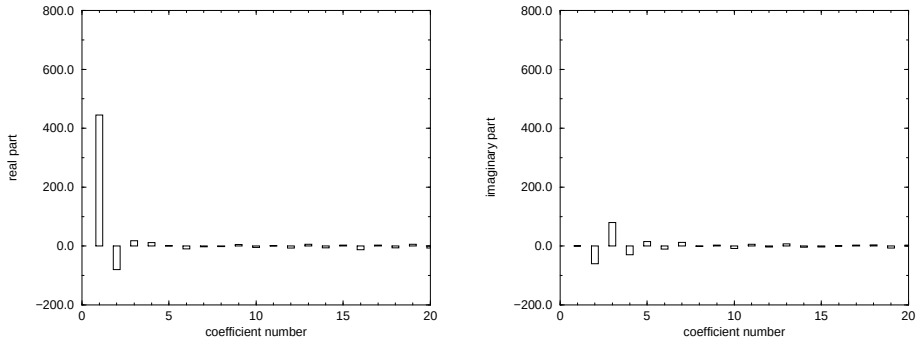

Figure 3: Fourier coefficients of the original intensity-versus-time curve in Figure 1.

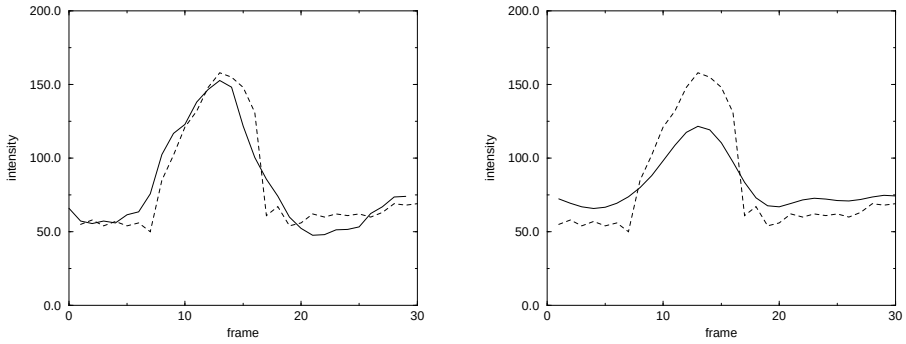

Figure 4: The reconstructions of the curve (dashed line) in Figure 1 using wavelet coefficients (left); using Fourier coefficients (right).

Fourier coefficients are represented from left to right, respectively. In this visualization a logarithmic transformation has been applied to enhance the visibility of low values.

3 Model construction and matching

After the wavelet or Fourier transform has been computed, either the matrix $C = [c_{ij}]_{hw}$ or the matrix $C' = [c'_{ij}]_{hw}$, where c'_{ij} is a logarithmic transformation of c_{ij},

$$c'_{ij} = \begin{cases} \log(c_{ij} + 1) & c_{ij} \geq 0 \\ -\log(-c_{ij} + 1) & c_{ij} < 0, \end{cases}$$

can be used as a compact representation of the original image sequence. The logarithmic transformation is motivated by great differences among the coefficients. In Section 5.2, we will see that results using the logarithmic transformation are better than those using the original coefficients.

Next we describe five different methods of model construction and recognition, based on mean distribution, Gaussian distribution, Mahalanobis distance, nearest neighbor, and individual nearest neighbor, respectively.

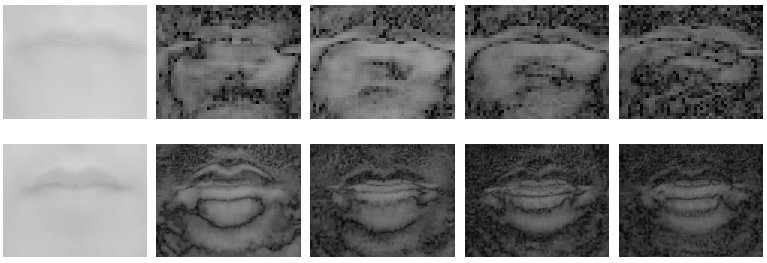

Figure 5: The coefficients of an image sequence: wavelet coefficients (upper row); Fourier coefficients (lower row).

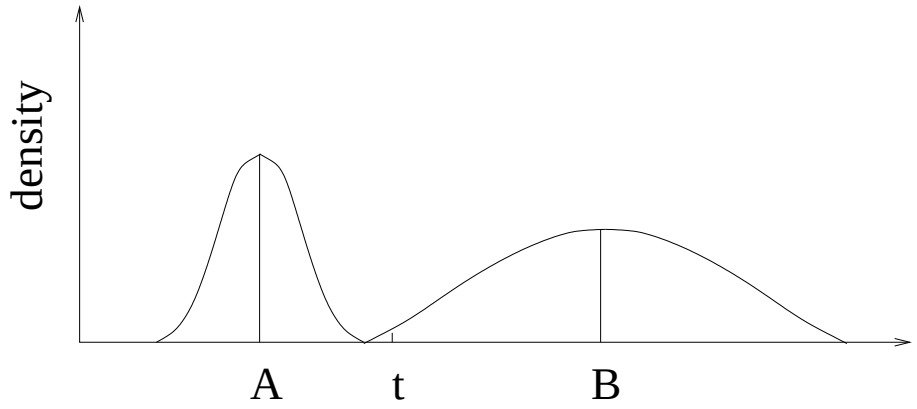

Figure 6: Distributions of two classes.

3.1 Mean method

Given L training image sequences of a class d and their corresponding wavelet or Fourier coefficient matrices, $C_l = [c_{ij}^l], l = 1, \ldots, L$[1], we construct a model matrix $M_d = [\mu_{ij}^d]$ for the class by averaging the coefficient matrices,

$$M_d = \frac{1}{L} \sum_{l=1}^{L} C_l.$$

Given a test image sequence, its coefficient matrix $T = [t_{ij}]_{hw}$ is calculated. Then, we match T against all model matrices

$$M_d = [\mu_{ij}^d], \qquad d = 1, \ldots, D,$$

where D is the total number of classes, computing the distance

$$\sum_{i=1}^{h} \sum_{j=1}^{w} ||t_{ij} - \mu_{ij}^d||.$$

The best match r is simply given by

$$r = \arg\min_d \sum_{i=1}^{h} \sum_{j=1}^{w} ||t_{ij} - \mu_{ij}^d||.$$

3.2 Gaussian distribution method

In the mean method, we only consider the mean coefficient matrix from the training sequences of a class. It might lead to wrong classification because of the different distributions of the training sequences of different classes. For example, suppose that the

[1] To keep our notation simple, C_l can either denote the matrix of original or logarithmically transformed coefficients.

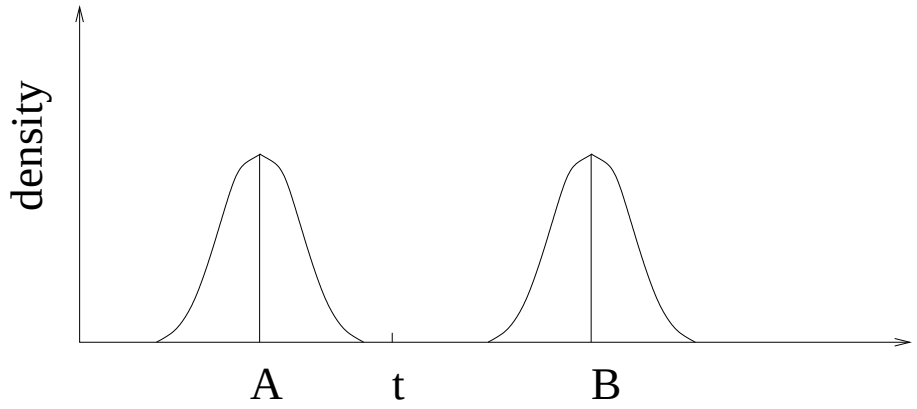

Figure 7: Distributions of two classes with small σ's.

variance of the training sequences of class A is smaller than that of class B and the distance of test t to the mean of class A is smaller than that to class B, see Figure 6. Using the mean method, t is classified as A. However, we expect that t is classified as B taking the variances of A and B into account. Therefore, we consider the Gaussian distribution of the training sequences of a class.

Given L training image sequences of a class d and their corresponding coefficient matrices, $C_l = [c_{ij}^l], l = 1, \ldots, L$, we construct a model matrix $M_d = [m_{ij}^d]$ for the class, where m_{ij}^d is a vector $(\mu_{ij}^d, \sigma_{ij}^d)$,

$$\mu_{ij}^d = \frac{1}{L} \sum_{l=1}^{L} c_{ij}^l,$$

$$\sigma_{ij}^d = \sqrt{\frac{1}{L} \sum_{l=1}^{L} (c_{ij}^l - \mu_{ij}^d)^2}.$$

Given a test image sequence and its coefficient matrix $T = [t_{ij}]_{hw}$, we match T against all model matrices

$$M_d = [m_{ij}^d], \qquad d = 1, \ldots, D,$$

by computing the probability

$$p^d = \sum_{i=1}^{h} \sum_{j=1}^{w} \frac{1}{\sqrt{2\pi}\sigma_{ij}^d} \exp \frac{-(t_{ij} - \mu_{ij}^d)^2}{2(\sigma_{ij}^d)^2}.$$

The best match r is given by

$$r = \arg \max_d p^d.$$

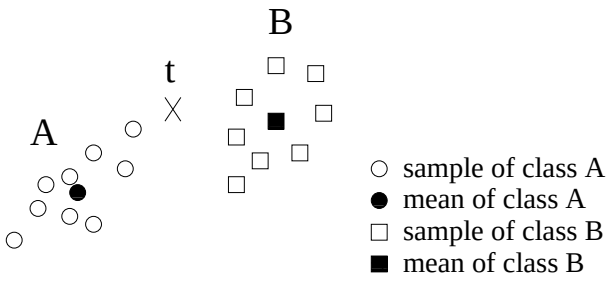

Figure 8: Nearest neighbor in two dimensional space.

3.3 Mahalanobis distance method

In the Gaussian distribution method, the calculation of function exp is costly. In addition, if the standard deviation σ is very small for two classes, the probability of a given test for those classes might be very small as well. This makes it hard to decide which class the test belongs to. See Figure 7 as an example. (Such cases actually happened in our experiments described in Section 5.3.) Therefore, we consider Mahalanobis distance as an alternative to Gaussian distribution.

We use the Gaussian distribution method to construct models. Then given a test coefficient matrix $T = [t_{ij}]_{hw}$, we match T against all model matrices by computing the Mahalanobis distance

$$Mah^d = \sum_{i=1}^{h} \sum_{j=1}^{w} \frac{|t_{ij} - \mu_{ij}^d|}{\sigma_{ij}^d}.$$

Finally the best match r is calculated by

$$r = \arg\min_d Mah^d.$$

3.4 Nearest neighbor method

Let us consider the situation shown in Figure 8. The distance of test t to the mean of class A is larger than that to the mean of class B. Therefore, t is classified as B using the mean method. However, since the nearest neighbor to t is a sample of A, t is classified as A under the nearest neighbor method.

Given L training image sequences of a class d and their corresponding coefficient matrices, $C_l = [c_{ij}^l], l = 1, \ldots, L$, we construct L model matrices $M_d^l = [m_{ij}^{dl}] = [c_{ij}^l]$ for the class.

Given a test image sequence and its coefficient matrix $T = [t_{ij}]_{hw}$, we match T against all model matrices

$$M_d^l = [m_{ij}^{dl}], \qquad d = 1, \ldots, D, \qquad l = 1, \ldots, L,$$

by computing the distance

$$\sum_{i=1}^{h} \sum_{j=1}^{w} ||t_{ij} - m_{ij}^{dl}||.$$

Through all d and l, we choose d with the smallest distance as the best match.

3.5 Individual nearest neighbor method

In the nearest neighbor method, a test pattern has to be matched to all training samples. This is computationally very costly. If there are several persons in our database, we can make a compromise between the mean and the nearest neighbor method.

Assume there are P persons in the database. Given $L_1 + \ldots + L_P$ training image sequences of a class d and their corresponding coefficient matrices, $C_l^p = [c_{ij}^{pl}], p = 1, \ldots, P, l = 1, \ldots L_p$, we construct P model matrices $M_d^p = [m_{ij}^{dp}]$ for the class by averaging the coefficient matrices,

$$M_d^p = \frac{1}{L_p} \sum_{l=1}^{L_p} C_l^p, \qquad p = 1, \ldots, P.$$

Given a test image sequence and its coefficient matrix $T = [t_{ij}]_{hw}$, we match T against all model matrices

$$M_d^p = [m_{ij}^{dp}], \qquad d = 1, \ldots, D, \qquad p = 1, \ldots, P,$$

by computing the distance

$$\sum_{i=1}^{h} \sum_{j=1}^{w} ||t_{ij} - m_{ij}^{dp}||.$$

We obtain the best match r by choosing d with the smallest distance through all d and p.

4 Illumination invariance

It is important to notice that the first wavelet and the first Fourier coefficient are proportional to the average of the intensity function $I(n)$ along the time axis. Therefore, the first dimension of the model matrices corresponding to the first wavelet and Fourier coefficient will be approximately identical for all models, as long as the illumination during the training phase is kept constant (but not necessarily uniform). That is, only coefficients other than the first one significantly contribute to the distinction of different classes. Thus, we can ignore the first wavelet and Fourier coefficients in the matching process.

If the illumination in the test phase differs from that in the training phase, it can be expected that the intensity-versus-time curve of a pixel has a shape similar to that of the same pixel in the corresponding model, although its height may be different. This implies that by ignoring the first wavelet and Fourier coefficient our lipreading method possesses the potential of illumination invariance.

5 Experimental results

The methods discussed in the previous sections have been implemented in C and run on a Sun Sparcstation. Tests on two image databases were performed.

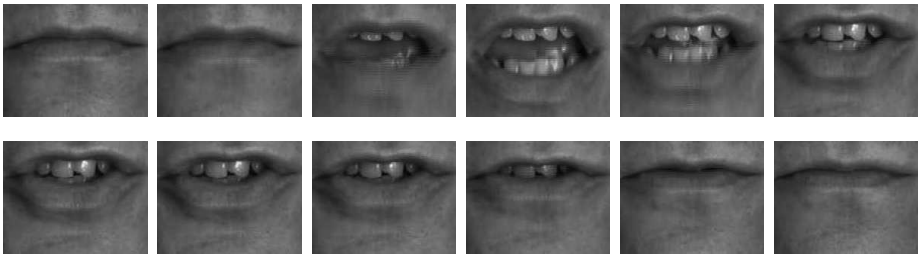

Figure 9: An image sequence of letter H.

5.1 Database acquisition

The first database was acquired in our laboratory. It based on a vocabulary of 10 English digits (from zero to nine) spoken by two speakers. From each of the two speakers 19 blocks of image sequences were collected, each block containing the ten digits. Images of 48×38 pixels and 8 bits per pixel were collected at a rate of 12 frames per second, centered around the lips, under normal lighting conditions. The second image database comes from the University of Central Florida, Orlando, and has been used in earlier work [10] on visual speech recognition. It consists of 18 blocks of lip image sequences, each block containing 10 English letters from A to J. Images of 240×200 pixels and 8 bits per pixel were collected at a rate of 15 frames per second under similar conditions as for the first image database. All sequences were supplied by a single speaker. As an example of a lip image sequence, Figure 9 shows one of letter H from the second database.

On the first image database three experiments using different data sets have been carried out. The first two data sets consisted of the image sequences acquired by each of the two speakers separately, while the whole image database was used as the third data set. In addition, a fourth experiment was conducted on the whole second image database. In each experiment the jackknife, or so-called "leave-one-out", procedure was applied. That is, leaving one image sequence out for testing, all other sequences were used for training. Therefore, there were 190 tests in the first two experiments, respectively, 380 tests in the third, and 180 tests in the fourth experiment.

In the following, we present experimental results with different model construction and matching methods as described in Section 3, including illumination invariance as presented in Section 4.

5.2 Mean method

First, we tested the approach described in section 3.1 using the first five coefficients ($k = 5$). In Figure 10, the recognition rates on our four data sets are given, where rank r means that the correct class is in the first r ranks. From Figure 10, we can see that on the four data sets without logarithmic transformation the recognition rates of wavelet transform are higher than those of Fourier transform. The recognition rates of both transforms with logarithmic transformation increase significantly. Also, with logarithmic transformation, Fourier transform become superior to wavelet transform. In addition, we notice that the recognition rate on the first data set is better than that on the second one. The reason is that the two speakers have made different head movements during

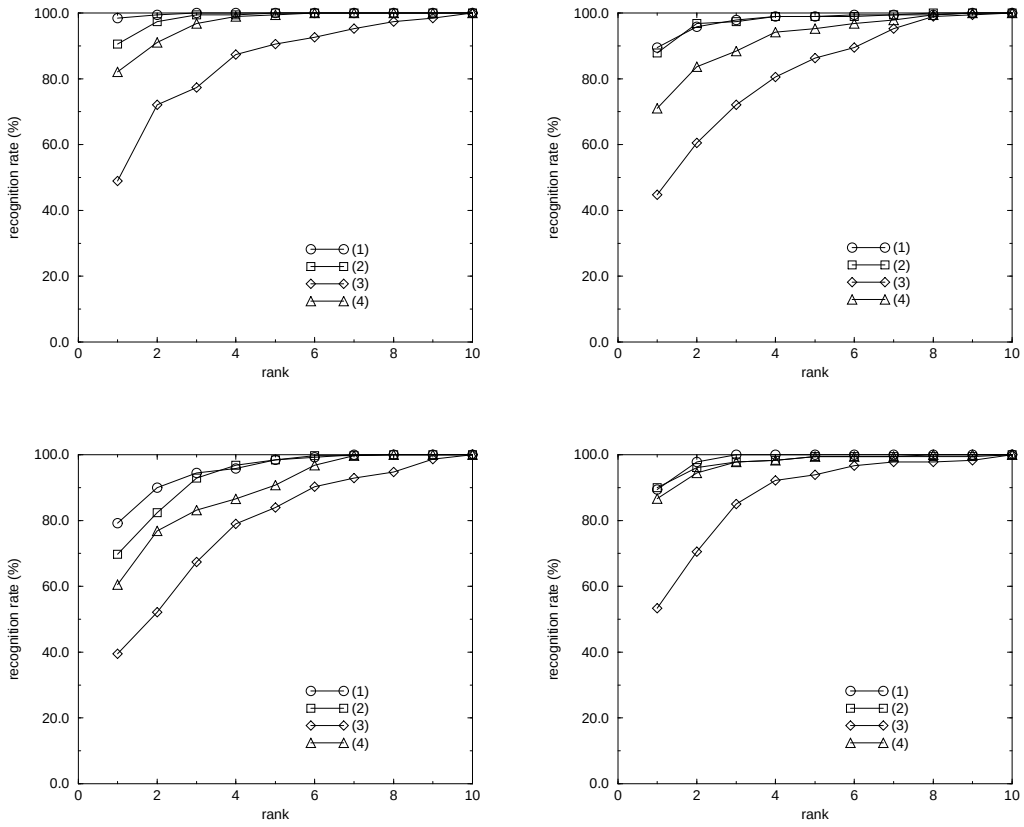

Figure 10: Recognition rates with the first five coefficients using the mean method: first data set (upper left), second data set (upper right), third data set (lower left) and fourth data set (lower right). In the figure: (1) Fourier transform with logarithmic transformation, (2) wavelet transform with logarithmic transformation, (3) Fourier transform, (4) wavelet transform.

image acquisition. While the first speaker moved his head slightly from right to left when he pronounced the digits, the second speaker moved it slightly downward. Due to the shape of the mouth, a vertical head movement changes the intensities of a larger area than the same amount of head movement in the horizontal direction. Thus, our lipreading method is more sensitive to vertical head movements. The third data set, where the image sequences provided by both speakers were put together, resulted in the lowest recognition rate.

The only parameter in our lipreading method is the number of coefficients k used to represent the intensity-versus-time curve. Therefore, we were also interested in the relationship between the recognition rate and the number k. On the four data sets, this relationship is shown in Figure 11. Similarly to Figure 10, we can observe in Figure 11 that the recognition rates using logarithmic transformation are better and the Fourier transform with logarithmic transformation achieved the best results. In addition, the recognition rates are very low when only the first coefficient is used. The reason is, as mentioned in Section 3, that the first wavelet and the first Fourier coefficient are

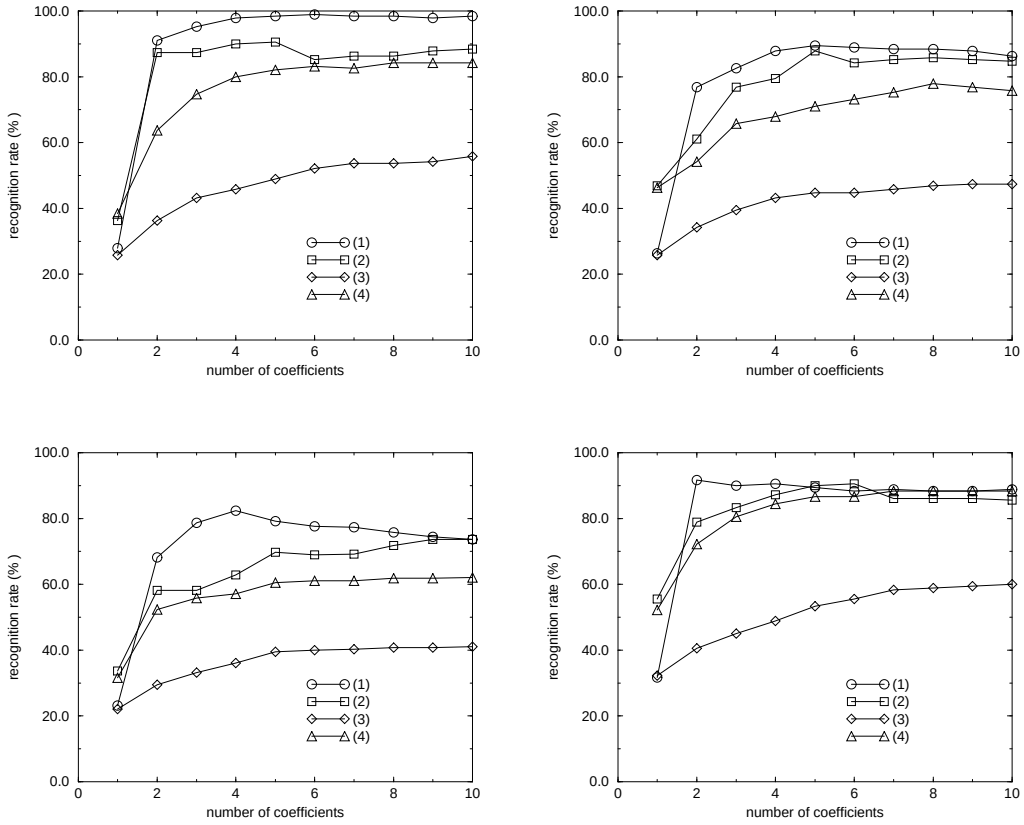

Figure 11: Recognition rates versus number of coefficients using the mean method: first data set (upper left), second data set (upper right), third data set (lower left) and fourth data set (lower right). In the figure: (1) Fourier transform with logarithmic transformation, (2) wavelet transform with logarithmic transformation, (3) Fourier transform, (4) wavelet transform.

proportional to the average of the intensity function $I(n)$ and don't significantly contribute to the distinction of different classes. Interestingly, the first two Fourier coefficients with logarithmic transformation only give already a quite reasonable recognition rate. The recognition rate becomes approximately stable after about five coefficients.

5.3 Gaussian distribution method

Figures 12 and 13 show the experimental results obtained with this method. First, we discuss the recognition rates with the first five coefficients. Comparing Figures 10 and 12, we can see that only the rates of Fourier and wavelet transform increase while those of Fourier and wavelet transform with logarithmic transformation decrease. The reason is that logarithmic transformation has changed the distribution of training classes and made the variance of the distribution smaller. In addition, the best rates of the Gaussian distribution method for each data set are lower than those of the mean method. The best rate, for example, of the Gaussian distribution method for second data set reaches about

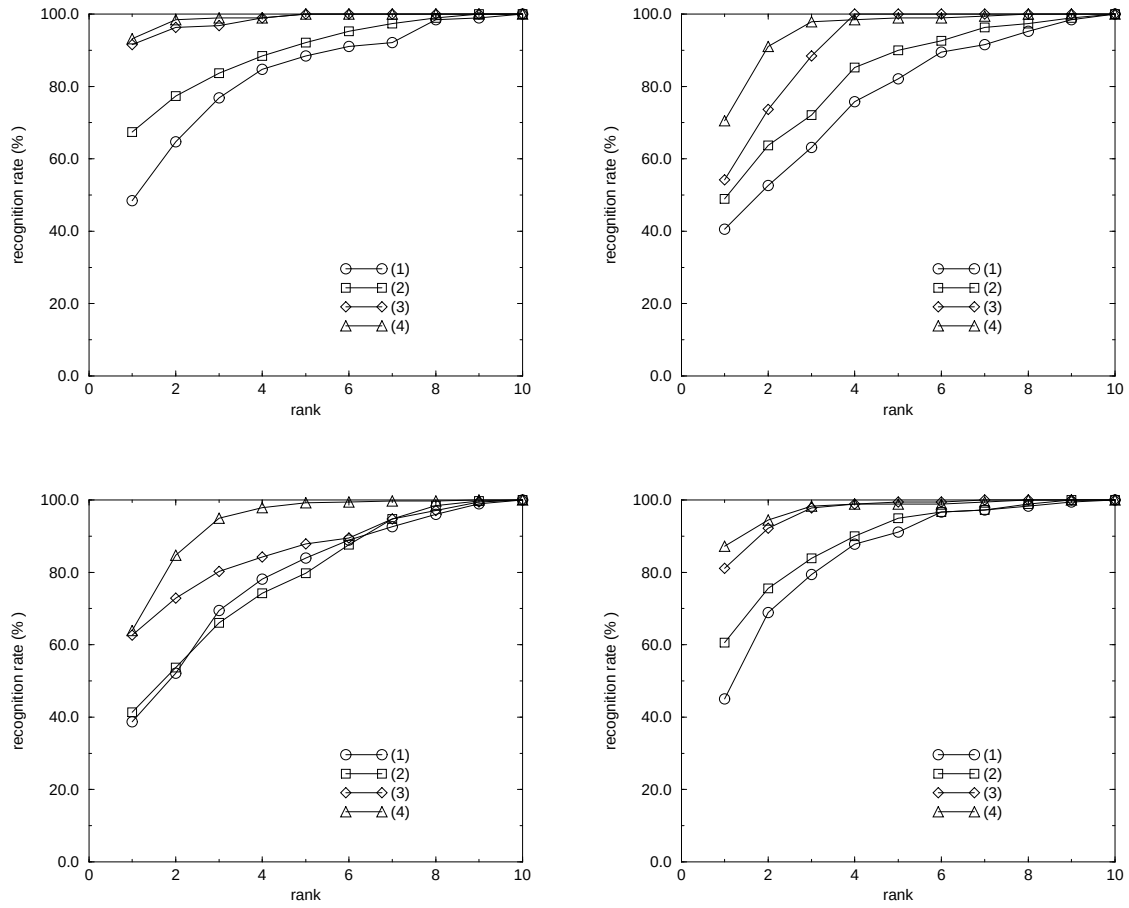

Figure 12: Recognition rates with the first five coefficients using Gaussian distribution method: first data set (upper left), second data set (upper right), third data set (lower left) and fourth data set (lower right). In the figure: (1) Fourier transform with logarithmic transformation, (2) wavelet transform with logarithmic transformation, (3) Fourier transform, (4) wavelet transform.

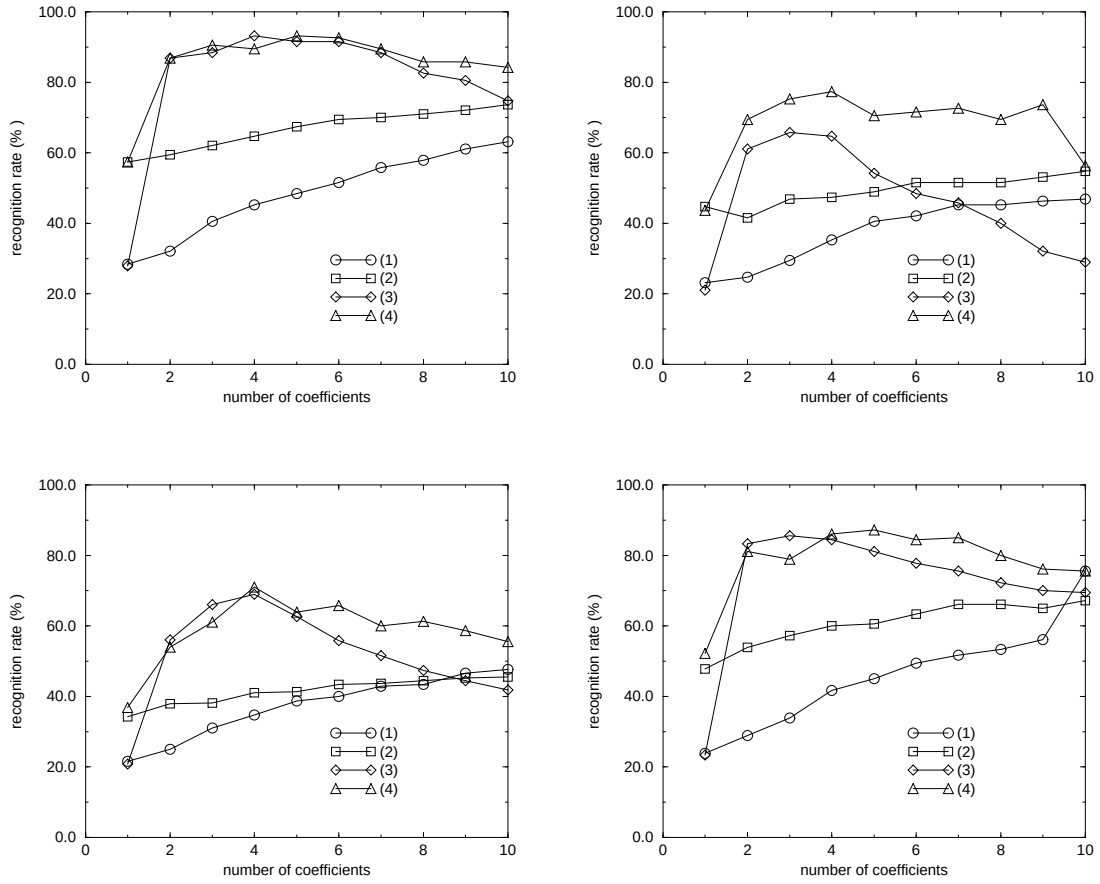

Figure 13: Recognition rates versus number of coefficients using Gaussian distribution method: first data set (upper left), second data set (upper right), third data set (lower left) and fourth data set (lower right). In the figure: (1) Fourier transform with logarithmic transformation, (2) wavelet transform with logarithmic transformation, (3) Fourier transform, (4) wavelet transform.

70% while that of the mean method is about 95%. The reason for that might be that we have not enough data to calculate the variance of Gaussian distribution. In this method, the wavelet transform on four data sets is the best of all four transform approaches.

Next, we discuss the recognition rates versus number of coefficients. Compared with the corresponding results (Figure 11) of the mean method, the rates of Fourier transform go up dramatically and the rates of wavelet transform increase slightly. On the other hand, the rates of the two transforms with logarithmic transformation decrease significantly. This means that we could not gain the benefit from the logarithmic transformation as we could in the mean method. One interesting effect that happens in the approach of Fourier transform is that rates on the third data set are higher than those on the second although the third data set was produced by two persons while the second was produced only by one.

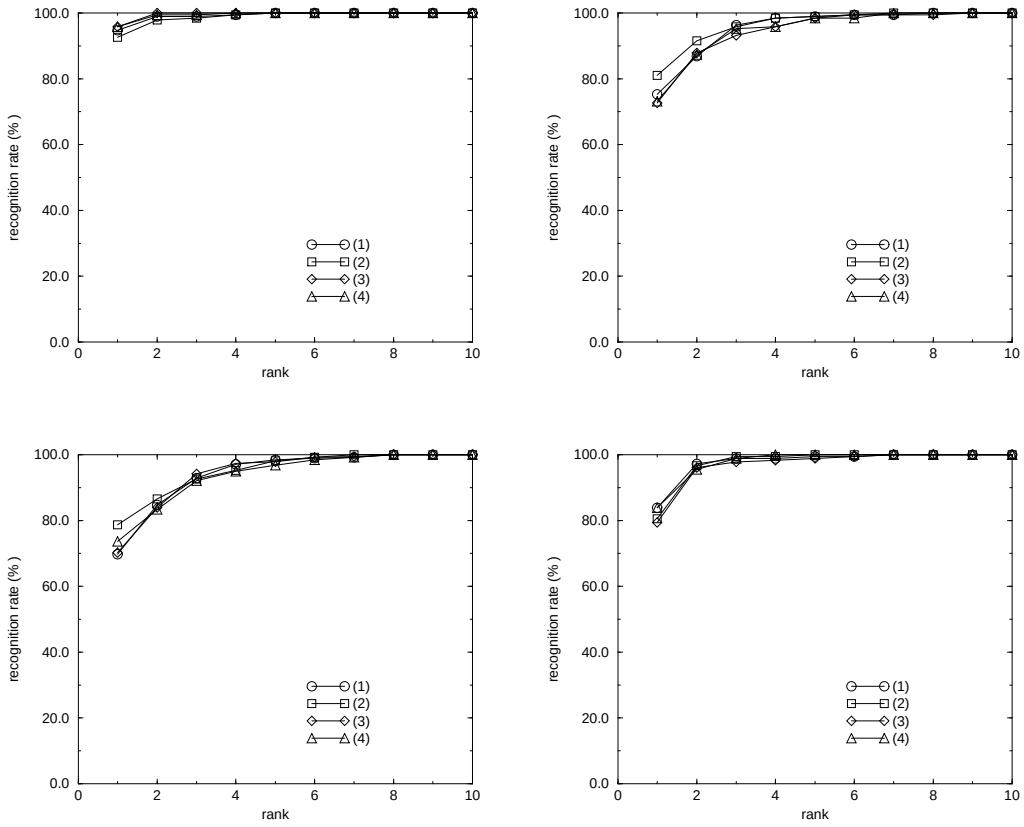

Figure 14: Recognition rates with the first five coefficients using the Mahalanobis distance method: first data set (upper left), second data set (upper right), third data set (lower left) and fourth data set (lower right). In the figure: (1) Fourier transform with logarithmic transformation, (2) wavelet transform with logarithmic transformation, (3) Fourier transform, (4) wavelet transform.

5.4 Mahalanobis distance method

In order to reduce the computation time of the Gaussian distribution method, we employed the Mahalanobis distance instead of the Gaussian distribution. The results are given in Figures 14 and 15.

First, we discuss the recognition rates with the first five coefficients. Comparing Figures 12 and 14, we observe that the rates of all four transforms are improved except the Fourier and wavelet transform with logarithmic transformation on the fourth data set. And on each data set, the rates of all four transform approaches are almost same. It means that the Mahalanobis distance method is fairly robust under different transforms.

Secondly, we discuss the recognition rates versus number of coefficients. As we can see from Figures 13 and 15, the the rates of Fourier and wavelet transform with logarithmic transformation increase significantly, while Fourier and wavelet transforms remain nearly same.

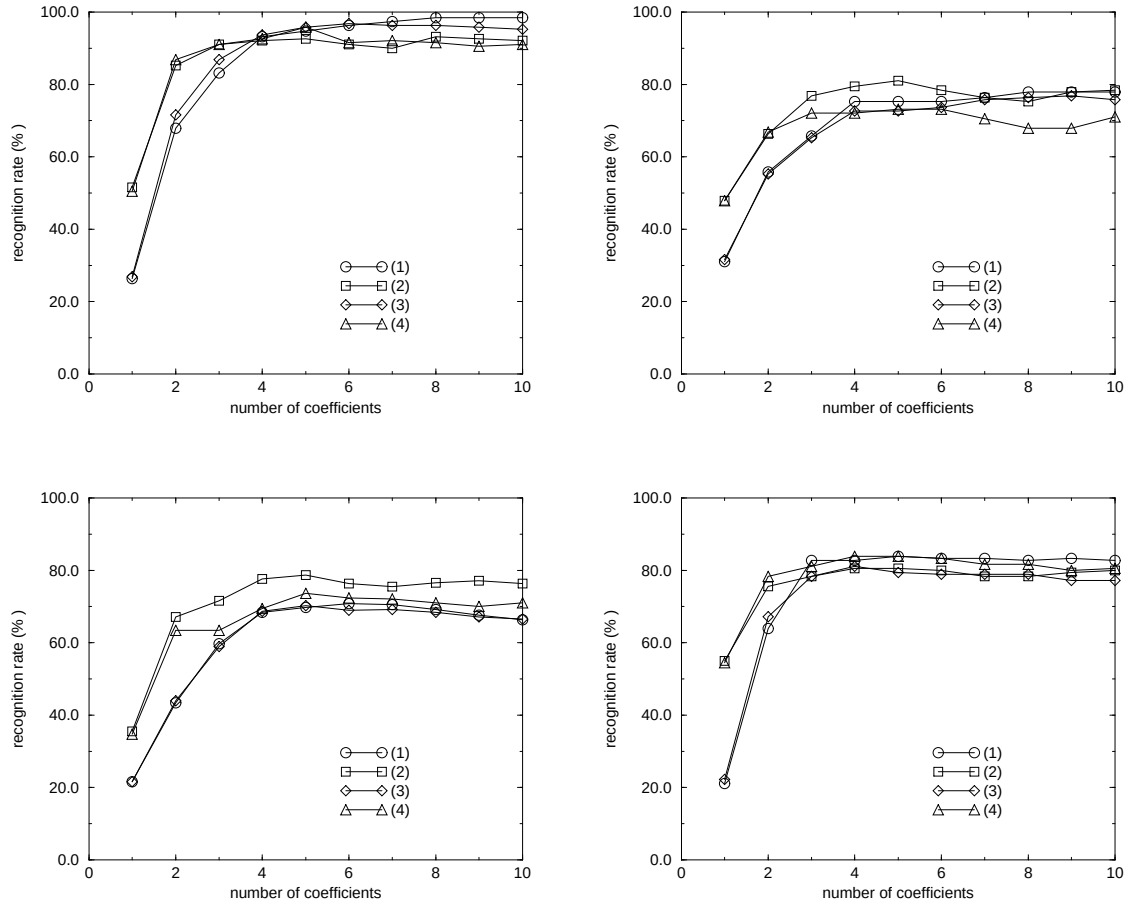

Figure 15: Recognition rates versus number of coefficients using the Mahalanobis distance method: first data set (upper left), second data set (upper right), third data set (lower left) and fourth data set (lower right). In the figure: (1) Fourier transform with logarithmic transformation, (2) wavelet transform with logarithmic transformation, (3) Fourier transform, (4) wavelet transform.

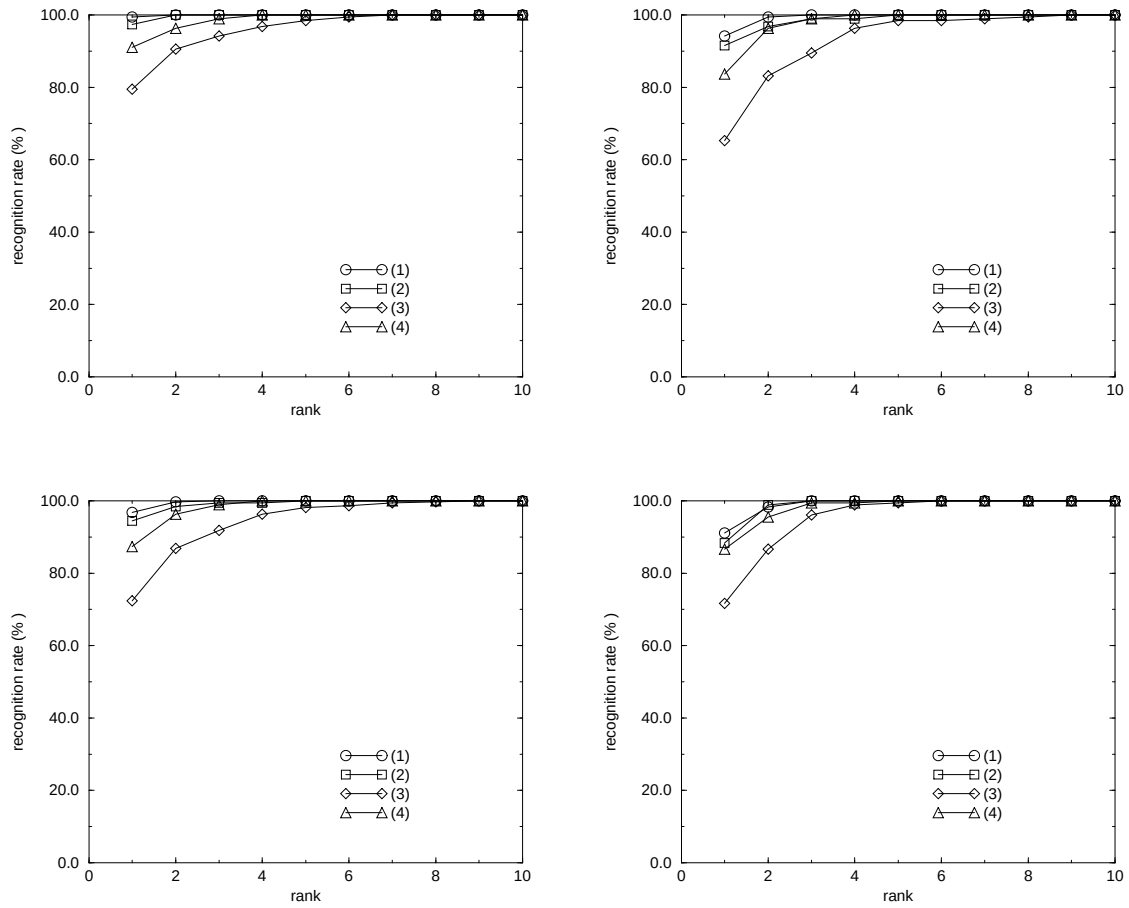

Figure 16: Recognition rates with the first five coefficients using the nearest method: first data set (upper left), second data set (upper right), third data set (lower left) and fourth data set (lower right). In the figure: (1) Fourier transform with logarithmic transformation, (2) wavelet transform with logarithmic transformation, (3) Fourier transform, (4) wavelet transform.

5.5 Nearest neighbor method

The results of this method are shown in Figures 16 and 17. This method costs much more time, but the recognition rates of the Fourier transform with logarithmic transformation are the best of all methods, especially on the second and the third data set where rates with the first two coefficients reach about 90%. Comparing Figures 14 and 16, we can observe that the rates of two transforms with logarithmic transformation of the nearest neighbor method are better than those of the Mahalanobis distance method. In addition, rates on the third data set are better than those on the second.

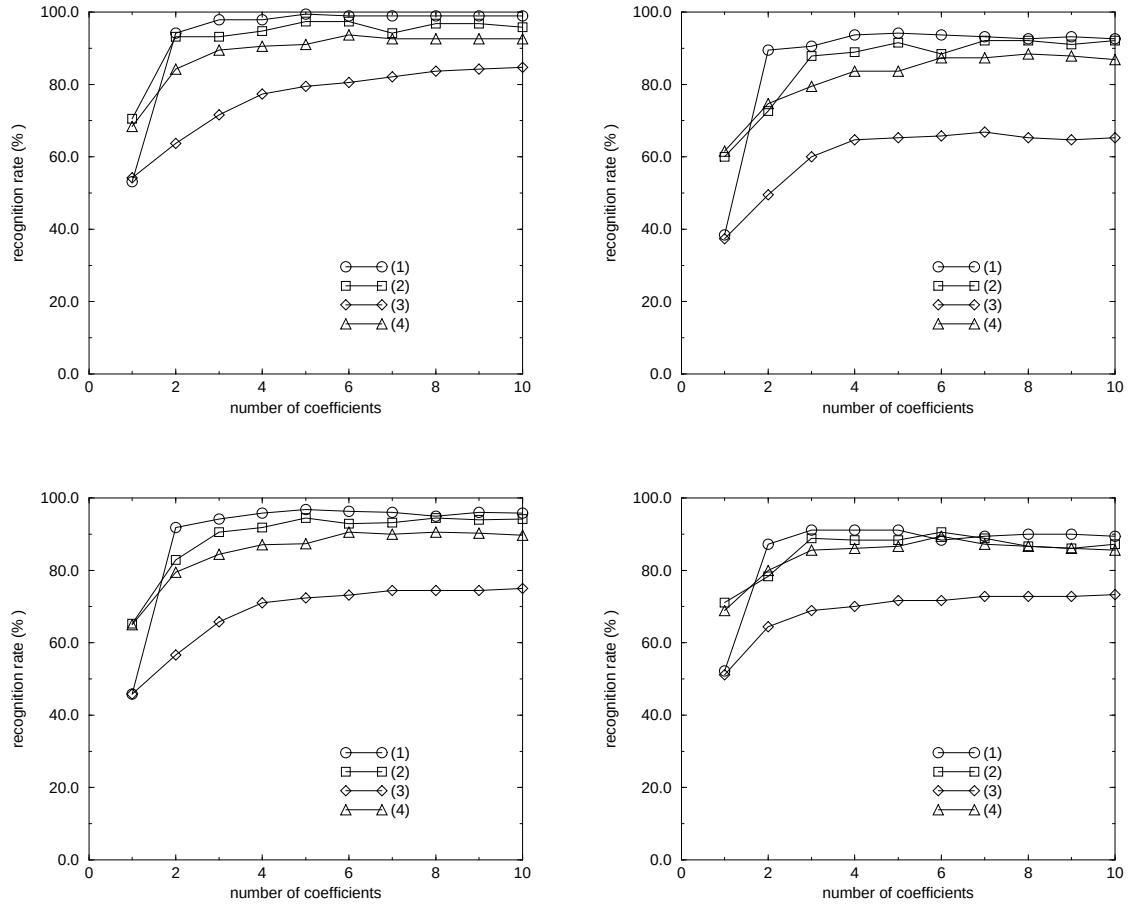

Figure 17: Recognition rates versus number of coefficients using the Nearest neighbor method: first data set (upper left), second data set (upper right), third data set (lower left) and fourth data set (lower right). In the figure: (1) Fourier transform with logarithmic transformation, (2) wavelet transform with logarithmic transformation, (3) Fourier transform, (4) wavelet transform.

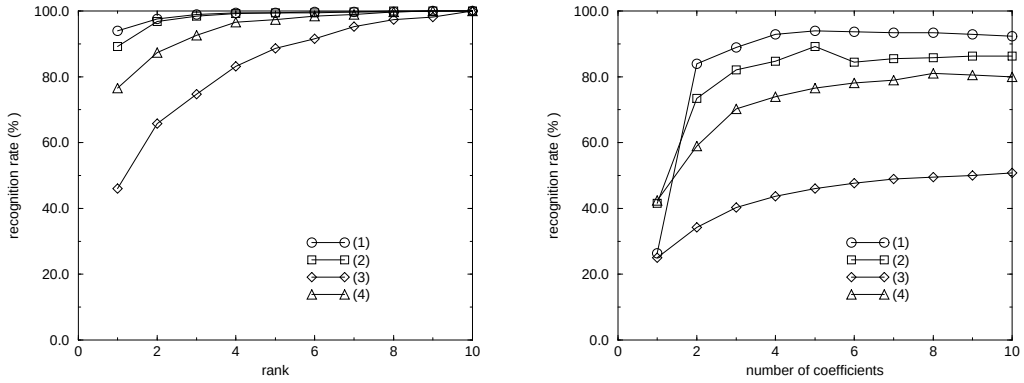

Figure 18: Recognition rate using the individual nearest neighbor method: rates with the first five coefficients (left), rates versus number of coefficients (right). In the figure: (1) Fourier transform with logarithmic transformation, (2) wavelet transform with logarithmic transformation, (3) Fourier transform, (4) wavelet transform.

5.6 Individual nearest neighbor method

We can use this method only on the third data set since the other sets were produced by just one person. The results are shown in Figure 18. From Figures 16 (lower left) and 18 (left), we can observe that all rates of this method are lower than those of the nearest neighbor method. However, we notice in Figures 14 (lower left) and 18 (left) that with this method all recognition rates are better than those with the Mahalanobis distance method, except the rates of Fourier transform. Comparing Figures 17 (lower left) and 18 (right), we can observe that using Fourier transform with logarithmic transformation the rate with the first five coefficients is about 95% although the rate decreases slightly compared with the nearest neighbor method. Using Fourier transform the rates become worse.

5.7 Illumination invariance

As mentioned in Section 4, our lipreading approach has the potential of illumination invariance. For the verification of this property with the Fourier transform, we have simulated a non-uniform illumination on the second test image database by adding position-

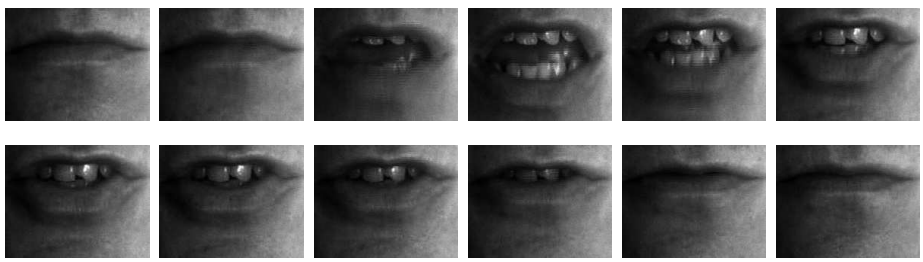

Figure 19: An image sequence of letter H under non-uniform illumination.

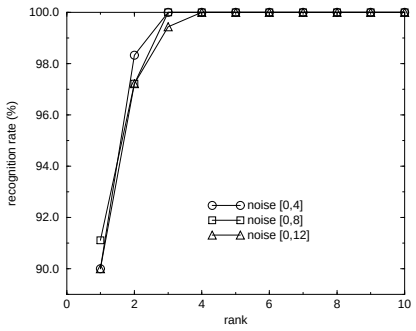

Figure 20: Recognition rates for non-uniform illumination and different noise levels.

dependent intensity value to each pixel. The intensities are further disturbed by uniformly distributed noise from the intervals [0,4], [0,8], and [0,12], respectively. As an example, Figure 19 shows the simulated image sequence of noise level 12 originating from that in Figure 9. Using the same leave-one-out procedure, a total of 180 tests has been carried out. In these tests the simulated lip sequences under non-uniform illumination are matched against models constructed from the original image data. The results are shown in Figure 20. Actually, it turned out that for all noise levels essentially the same recognition rates as shown in Figure 10 (lower right)(1) could be achieved.

6 Conclusions

In this paper we have presented new methods for lipreading. Instead of the usual approach of extracting features from the individual images, we consider the intensity-versus-time curves of the individual pixels. Lip movements are encoded by a few wavelet or Fourier coefficients of the intensity-versus-time curves. The primary experimental results on two image databases have demonstrated the usefulness of this encoding scheme. In our experiments, we used several methods of model construction and recognition. In these methods, the mean method is the simplest while the Gaussian distribution method is the most complicated. However, the recognition rates of the Gaussian distribution were not the best. Actually, the best results were achieved with the nearest neighbor method. The Mahalanobis distance method, which is an alternative to the Gaussian distribution, is fairly robust under different transforms and the individual nearest neighbor method is a compromise between the mean and the nearest neighbor method.

The proposed lipreading method also demonstrated promising results on a data set that was used by other researchers before. Our recognition rates, nevertheless, are not directly comparable with those reported in [10] on the same image database. In [10] the authors only carried out 20 tests, achieving a recognition rate of 95%. By contrast, a total of 180 tests were conducted in our experiments.

There are a number of details of our basic lipreading method to be further investigated. So far only wavelet coefficients and the magnitude of the Fourier coefficients have been considered. But we could incorporate the phase information to our image sequence

encoding scheme as well. The relative importance of the individual wavelet or Fourier coefficients to the overall lipreading performance and thus the need for a weighted matching has not been explored yet. In the two test image databases the image acquisition conditions have been kept constant. The potential of illumination invariance of our approach was only verified by simulated image data so far and has to be extended to real image data. These and other aspects will be investigated in the future.

In such complicated tasks as visual speech recognition, reliable classification is difficult to achieve for a single algorithm. Classifier combination is an effective way to improve recognition performance. In [21] we described a combination of the lipreading method proposed in the present paper with some other classifiers. It could be shown that even simple combination concepts can bring a significant improvement of classification accuracy in lipreading.

Acknowledgment

The second image database used in our experiments is supplied by Computer Vision Lab, Computer Science Department, University of Central Florida, Orlando. The Image Wavelet Package produced by Bob Lewis, the University of British Columbia, Canada, was used for our wavelet transform.

References

[1] W.E. Adam and F. Bitter, "Advances in Heart Imaging", Proc. of Int. Symposium on Medical Radionuclide Imaging, 1980.

[2] G. Beylkin, R. Coifman, and V. Rokhlin, "Fast Wavelet Transforms and Numerical Algorithms I", Comm. Pure Appl. Math., Vol. XLIV, pp. 141–183, 1991.

[3] M. Boehm, U. Obermoeller, and K.H. Hoehne, "Determination of Heart Dynamics from X-Ray and Ultrasound Image Sequences", Proc. of Int. Conf. on Pattern Recognition, pp. 403–408, 1980.

[4] C. Bregler, S. Manke, H. Hild, and A. Waibel, "Bimodal Sensor Integration on the Example of 'Speech-Reading'", Proc. of IEEE Int. Conf. on Neural Networks, pp. 667–671, 1993.

[5] I. Daubechies, "Ten Lectures on Wavelets", CBMS-NSF Regional Conference Series in Applied Mathematics, Vol. 61, Capital City Press, Montpelier, Vermont, 1992.

[6] A.J. Goldschen, O.N. Garcia, and E. Petajan, "Continuous Optical Automatic Speech Recognition by Lipreading", Proc. of 28th Annual Asilomar Conference on Signals, Systems, and Computers, pp. 572–577, 1995.

[7] N. Hazarika, J.Z. Chen, A.C. Tsoi, and A. Sergejew, "Classification of EEG signals using the wavelet transform", Signal Processing, Vol. 59, No. 1, pp. 61–72, 1997.

[8] M. Hennecke, D.G. Stork, and K.V. Prasad, "Visionary Speech: Looking Ahead to Practical Speechreading Systems", in Speechreading by Humans and Machines, D.G. Stork and M.E. Hennecke (Eds.), pp. 331–350, 1995.

[9] M. Kirby, F. Weisser, and G. Dangelmayr, "A Model Problem in the Representation of Digital Image Sequences", Pattern Recognition, Vol. 26, No. 1, pp. 63–73, 1993.

[10] N. Li, S. Dettmer, and M. Shah, "Lipreading Using Eigensequences", Proc. of Int. Workshop on Automatic Face- and Gesture-Recognition, pp. 30–34, 1995.

[11] H. Li, B.S. Manjunath, and S.K. Mitra, "Multi-Sensor Image Fusion Using the Wavelet Transform", Proc. of Int. Conf. on Image Processing, Vol. I, pp 51–55, 1994.

[12] J. Luettin and N.A. Thacker, "Speechreading Using Probabilistic Models", Computer Vision and Image Understanding, Vol. 65, No. 2, pp. 163–178, 1997.

[13] U. Meier, W. Hürst, and P. Duchnowski, "Adaptive Bimodal Sensor Fusion for Automatic Speechreading", Proc. of IEEE Int. Conf. on Acoustic, Speech and Signal Processing, 1996.

[14] J.T. Miller and C.C. Li, "Segmentation of Object Surfaces Using the Haar Wavelet at Multiple Resolutions", Proc. of Int. Conf. on Image Processing, Vol. III, pp 475-477, 1994.

[15] J.R. Movellan, "Visual Speech Recognition with Stochastic Networks", in Advances in Neural Information Processing System, G. Tesauro, D. Toruetzky, and T. Leen (Eds.), Vol. 7, MIT Press, Cambridge, 1995.

[16] C. Nastar and N. Ayache, "Time Representation of Deformations: Combining Vibration Modes and Fourier Analysis", in Object Representation in Computer Vision, M. Hebert, J. Ponce, T. Boult, and A. Gross (Eds.), pp. 263–275, 1994.

[17] P. Palisson, N. Zegadi, F. Peyrin, and R. Unterreiner, "Unsupervised Multiresolution Texture Segmentation Using Wavelet Decomposition", Proc. of Int. Conf. on Image Processing, Vol. II, pp 625–629, 1994.

[18] E.J. Stollnitz, T.D. DeRose, and D.H. Salesin, "Wavelets for Computer Graphics: A Primer, Part1", IEEE Computer Graphics and Application, pp. 76–84, May 1995.

[19] E.J. Stollnitz, T.D. DeRose, and D.H. Salesin, "Wavelets for Computer Graphics: A Primer, Part2", IEEE Computer Graphics and Application, pp. 75–85, July 1995.

[20] D.G. Stork and M.E. Hennecke (Eds.), Speechreading by Humans and Machines, Springer-Verlag, 1996.

[21] K. Yu, X. Jiang, and H. Bunke, "Lipreading: A Classifier Combination Approach", Pattern Recognition Letters, 1997 (to appear).