Correspondence

Analysis of Lip Geometric Features for Audio-Visual Speech Recognition

Mustafa N. Kaynak, Qi Zhi, Adrian David Cheok, Kuntal Sengupta, Zhang Jian, and Ko Chi Chung

***Abstract*—Audio-visual speech recognition employing both acoustic and visual speech information is a novel extension of acoustic speech recognition and it significantly improves the recognition accuracy in noisy environments. Although various audio-visual speech-recognition systems have been developed, a rigorous and detailed comparison of the potential geometric visual features from speakers' faces is essential. Thus, in this paper the geometric visual features are compared and analyzed rigorously for their importance in audio-visual speech recognition. Experimental results show that among the geometric visual features analyzed, lip vertical aperture is the most relevant; and the visual feature vector formed by vertical and horizontal lip apertures and the first-order derivative of the lip corner angle leads to the best recognition results. Speech signals are modeled by hidden Markov models (HMMs) and using the optimized HMMs and geometric visual features the accuracy of acoustic-only, visual-only, and audio-visual speech recognition methods are compared. The audio-visual speech recognition scheme has a much improved recognition accuracy compared to acoustic-only and visual-only speech recognition especially at high noise levels. The experimental results showed that a set of as few as three labial geometric features are sufficient to improve the recognition rate by as much as 20% (from 62%, with acoustic-only information, to 82%, with audio-visual information at a signal-to-noise ratio of 0 dB).**

***Index Terms*—Audio-visual speech recognition, feature fusion, lip geometric features.**

I. Introduction

Speech recognition will play an important role in future human–computer interfaces. In general, the field of speech recognition is a part of the ongoing research effort in developing computers that can hear and understand spoken information. However, current automatic speech recognition (ASR) systems' recognition rates decrease significantly in common environments where the ambient noise level is high. This presents a problem because speech recognition is desirable in realistic settings as in store fronts, offices, airports, train stations, vehicles, and mobile outdoor settings.

Humans on the other hand, often compensate for noise degradation and uncertainty in speech information by integrating multiple sources of speech information, such as visible gestures from the speaker's face and body [1]. Thus, human perception systems use visual and acoustic (audio-visual, bimodal) information to recognize speech. This is demonstrated by the "McGurk effect" [2]. When a subject is presented with contradicting acoustic and visual signals, the perceived signal can be completely different from both the acoustic and visual signals. For example, when subjects were shown a video where a speaker mouthed "gah," but the video was dubbed with a voice saying "bah," most people usually heard the sound "dah." However, when viewers turn their backs to the video, they heard the sound "bah" correctly. McGurk also discovered that viewers could not force themselves to hear the correct vocal sounds when it is told that they were being fed with the wrong visual information.

Modeling the bimodal characteristics of speech perception and production systems of human beings is one of the possible approaches to solve the noise problem of ASR systems. The most important source of additional speech information is the visual information from a speaker's face [3]. Speech perception can be improved significantly by watching the face of the speaker [1]. For example, hearing impaired people use only the visual speech information from the visible speech articulators (spread all over the face [4], [5]) to recognize the speech signal.

The primary advantage of the visual information is that it is not affected by acoustic noise and cross talk among speakers. Another advantage reported in [1] and [6], is the complementary structure of phonemes and visemes, which are the smallest acoustically and visually distinguishing units of a language, respectively. According to these studies, the acoustic and visual components of the speech signal are not purely redundant; they are complementary as well. Certain speech characteristics that are visually confusable are acoustically distinct, while those characteristics that are acoustically confusable are visually distinct. For instance, phonemes /b/ and /k/, which are produced in visibly distinct manners, have similar acoustic characteristics. In contrast, phonemes /p/ and /b/, which are visibly indistinguishable, can be acoustically differentiated by voicing [6].

Thus, the motivation of audio-visual (bimodal) speech recognition can be summarized as follows. The more speech information we have, the higher will be the potential recognition. Furthermore, studies in human perception system have shown that visual speech information allows people to tolerate an extra 4 dB of noise in the acoustic signal [7].

However, one of the fundamental problems in bimodal speech recognition is to determine which visual features are most advantageous to use. In previous research, different types of visual features were used. Petajan developed one of the first audio-visual speech recognition systems [8], [9] using the mouth area, perimeter, height, and width derived from binary mouth image as the visual features. In [1], the pixel values of a reduced area of interest in the image centered around the mouth were used as the visual features. In [10], the temporal variations of the mouth parameters were used as the visual features. In [11], both the shape and intensity information from the lip contours were used as the visual features, and active shape models were employed to extract these visual features. In [12], acoustic features were combined with either geometric visual parameters, such as the mouth height and width, or nongeometric visual parameters, such as the wavelet transform of the mouth images. In [13], gray-scale parameters associated with the mouth region of the image were considered but, first, principal-component analysis was used to reduce the dimension of the feature space. In [14], the intensity of each pixel in an image sequence was considered as a function of time and one-dimensional wavelet and Fourier transforms were applied to this intensity-versus-time function to model the lip movements. In [15], point-distribution models were used to track the lips, and the shape parameters obtained from tracking results were used as the visual features.

Manuscript received July 20, 2001; revised May 6, 2003 and November 8, 2003.

M. N. Kaynak is with the Department of Electrical Engineering, Arizona State University, Tempe, AZ 85287-5706 USA (e-mail: mustafa.kaynak@asu.edu).

Q. Zhi, A. D. Cheok, K. Sengupta, Z. Jian, and K. C. Chung are with the Department of Electrical and Computer Engineering, National University of Singapore, Singapore 117576 (e-mail: adriancheok@nus.edu.sg).

Digital Object Identifier 10.1109/TSMCA.2004.826274

It can thus be summarized that previous bimodal recognition studies have used two major types of visual speech information:

- geometric features (such as measures of height, width, area);
- image and shape features (such as pixel colors, lip outline).

The disadvantage of using image- and shape-based features is that they require complex algorithms and are computationally expensive [16]. Furthermore, in [17], it is shown that, for speech recognition, there is little point in utilizing the fine detail of the lip, as gross detail appears to provide almost the same information. Thus, even very simple geometric visual features may improve the recognition rate when used with the acoustic signal. In addition, less complex geometric features allow the development of more practical and real-time bimodal speech-recognition systems.

Nevertheless, in previous works, a rigorous comparison of suitable geometric features, both singly and in combination, has not been given. Hence, the purpose and the motivation of the work described in this paper are as follows:

- analyzing the importance of each lip geometric feature in bimodal speech recognition;
- determining the best combination of geometric visual features for both visual-only and bimodal speech recognition;
- comparing the accuracy of acoustic-only, visual-only, and bimodal speech recognition experimentally using the optimized HMMs and geometric visual features.

The experimental results are obtained for digit recognition, in order to provide a simple but rigorous example for comparison purposes.

This paper is organized as follows. First, the audio-visual database is introduced, and acoustic hidden Markov modeling is given in Section II. In Section III, geometric visual features are analyzed and the visual hidden Markov modeling is explained for both single and combined geometric visual features. In Section IV, the audio-visual speech recognition experimental setup is given and then the results are reported. Finally, the paper is concluded in Section V.

II. State of the Art

A. Audio-Visual Database and Automatic Database Processing Toolkit

To conduct research on bimodal speech recognition, a fully segmented and labeled audio-visual database is required. For this purpose, we generated an audio-visual database and developed an automatic audio-visual database-processing toolkit to process the database. In the following subsections, the audio-visual database and the toolkit are introduced briefly.

1) Audio-Visual Database: For our research, we generated an audio-visual database. Since the main focus of our research is on the determination of the relevant visual features, in order to extract the visual features accurately, blue fiducials were placed on the speakers' faces as shown in Fig. 1. Blue was used for highlighting because humans seldom have natural blue-color elements on their skin, thus, it is easy to detect blue color on the face. The markers on the nose and chin were used to normalize geometric visual features so that the extracted visual features are invariant to the distance between the camera and the speaker.

The generated English audio-visual database[1] contains over twenty hours of continuous and isolated utterances from 22 nonnative English speakers having European, Chinese, Singaporean, and Indian accents. In this research, 77 samples from nine speakers were used to train the HMM of each digit, 24 samples from ten speakers were used to test the digit models for speaker independent case, and 19 samples from nine speakers were used to test the digit models for speaker dependent case. The recordings were done in a silent isolated room using a digital camera with an IEEE 1394 (Firewire) output to produce high-quality video signal.

[1]The database and the automatic segmentation toolkit are available for research purposes. Please go to http://speech.ece.nus.edu.sg to download the database and the toolkit.

Fig. 1. Speaker with six fiducials.

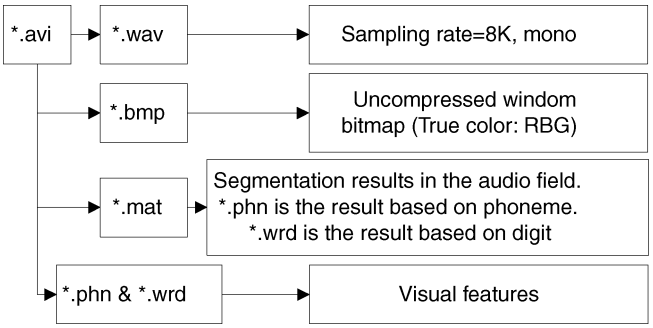

Fig. 2. Input–output file structure of the segmentation software (*.avi = video file; *.wav = audio file; *.bmp= bitmap picture file; *.phn = phoneme label text file; *.wrd = word transcription text file; and *.mat = matrix file.).

2) Automatic Audio-Video Segmentation Software Toolkit: To process the database, we developed a software package [18] in Matlab. The software is user friendly with a graphical user interface (GUI). To process the video avi files, the toolkit first extracts the bitmap and sound wave files from the avi file. Then, it calls tcl scripts from the CSLU Toolkit [19], [20] to do phoneme and word partitioning in the audio domain. Using the segmentation results in audio domain, the toolkit does the segmentation in the video domain. It then saves bitmap files corresponding to every phoneme of the digit into carefully labeled folders.

Then, the toolkit detects the markers in each frame of the corresponding bitmap files by convolving each frame with a blue mask. The local maximal points in the result of convolution correspond to the blue markers. Then, these markers are tracked automatically and results are saved into a matrix showing the movements of the markers in relation to the corresponding phonemes and digits. Finally, the tracking results of the markers are used to calculate the basic geometric visual features.

After running the software, the user obtains soundwave files segmented into phoneme level, bitmap files sampled at 25 frames/s from the original avi file, a word label text file, a phoneme label text file (which contains the phonemes of the uttered word), an audio domain segmentation file and two Matlab files containing the tracking results of the six markers for every video frame, and the basic geometric visual features. The input-output file structure of the toolkit is shown in Fig. 2. All these results are saved into separate folders, according to the folder structure based on the results of segmentation. The bitmap and wave files are saved and grouped at the phoneme level. The user

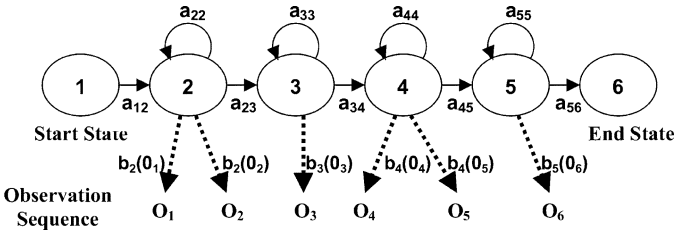

Fig. 3. Typical left-right HMM (a_{ij} is the state-transition probability from state i to state j, O_t is the observation vector at time t, and $b_i(O_t)$ is the probability that O_t is generated by state i).

can display the fully segmented audio signal and the bmp files corresponding to each phoneme using a GUI.

We note that the toolkit requires that the speakers in the database face the camera. If the speaker rotates his face, the toolkit should be modified to compensate the rotation. One possible approach is to use pose-independent visual features [21].

B. Hidden Markov Modeling for Bimodal Speech Recognition

For the speech-recognition engine, neural networks (NN), fuzzy logic, and hidden Markov models (HMMs) have been used. However, among these, HMMs have been the most widely used algorithm for speech-recognition studies for the last decade because of their success in modeling both acoustic and the visual speech signals. HMMs [22] are stochastic models and are widely used for characterizing the spectral properties of frames of patterns. The underlying assumptions of using HMMs in speech recognition are that speech signals can be well characterized as a parametric random process, and that the parameters of the stochastic process can be determined (estimation) in a precise, well-defined manner (training). Fig. 3 shows the structure of a typical left-right HMM. To model the speech signal, usually a left-right HMM is used because the underlying state sequence associated with the model has the property that as time increases, the state index increases or stays the same.

For our research we developed an HMM code in Matlab, and for comparison the Microsoft HTK[2] version 3.0 HMM code was used as a benchmark. For the developed HMM code, the Baum–Welch algorithm was used for training signal models by taking into account the practical implementation issues, such as scaling, multiple observation sequences and initial parameter estimates, which are explained in [23]. To prevent overtraining of the HMMs, when the highest accuracy is obtained for the validation data recognition, the training is terminated automatically.

C. Acoustic HMM

In this section, hidden Markov modeling of the acoustic speech signal is discussed and a comparative study is carried out between the developed and HTK HMM codes. As mentioned above, all speech signals are modeled with left-right HMMs and in order to present rigorous comparisons of the recognition results of acoustic-only, visual-only, and bimodal speech recognition, the optimum HMM is selected for each type of signal.

For the acoustic speech-recognition experiments, 13 dimensional Mel frequency cepstral coefficients (MFCC) were used as the standard audio features [23], [24]. The original dimension of the acoustic feature vector was 14, however, since zero-order coefficient is approximately equivalent to the log energy of the frame, and energy is directly computed on the time signal, we discarded the zero-order cepstral coefficient and obtained a 13 dimensional acoustic feature vector. Please note that the MFCC coefficients were extracted from 25 ms windows with 10 ms overlapping.

[2]HTK web site: http://htk.eng.cam.ac.uk/

TABLE I
ACOUSTIC SPEECH RECOGNITION FOR SPEAKER-INDEPENDENT CASE

N	4	4	4	5	5	5
M	8	16	32	8	16	32
SNR	Recognition Rate (%)					
clean	96.7	96.3	96.3	90.4	95.4	96.7
30dB	96.7	96.3	96.3	92.5	95.4	97.5
25dB	95.8	95.8	96.7	92.9	96.3	97.1
20dB	94.6	96.7	95.8	92.5	95.8	96.7
15dB	91.7	95.0	95.0	92.9	93.3	93.8
10dB	86.3	91.3	92.9	89.2	93.3	90.0
5dB	79.6	81.3	83.3	77.5	80.8	79.2
0dB	55.4	60.8	62.1	52.5	57.1	57.9

TABLE II
ACOUSTIC SPEECH RECOGNITION FOR SPEAKER-DEPENDENT CASE

N	4	4	4	5	5	5
M	8	16	32	8	16	32
SNR	Recognition Rate (%)					
clean	99.5	99.5	99.5	92.6	99.5	100
30dB	99.5	99.5	99.5	94.7	99.0	100
25dB	98.4	99.0	100	95.3	99.0	99.5
20dB	96.3	98.4	99.0	94.7	97.9	99.0
15dB	95.3	96.8	98.4	94.7	96.3	97.4
10dB	89.5	93.7	94.7	90.5	95.3	94.2
5dB	78.4	80.0	84.7	76.3	81.1	82.6
0dB	53.7	56.8	60.0	52.1	55.8	57.9

In the experiments, white noise was used to obtain noisy speech signals and the SNR was calculated as the ratio between the average powers of the speech and noise signals. To find the optimum HMM, in terms of the highest recognition accuracy, many different models were experimentally examined.

For each digit model, the state number, N, was fixed to four and five; and eight, 16, and 32 Gaussian mixtures, M, were used repeatedly in order to find the optimum model. For training each digit model, 77 samples from nine speakers were used and for evaluating the performances of the digit models for speaker independent and dependent cases, 24 samples from ten speakers and 19 samples from nine speakers were used, respectively.

Tables I and II show the average recognition results of the different types of HMMs for speaker independent and dependent cases. As expected, speaker-dependent results are better than speaker-independent results, especially at low noise levels. However, the difference between the speaker dependent and independent results is not very significant especially at high SNRs. From the tables, we also observe that as the number of Gaussian mixture components increases, the performances of the HMMs improve. Among the six different digit models tested for both speaker-dependent and -independent cases, HMMs with 32 Gaussian mixtures and four states are the best ones. Thus, to model the acoustic speech signals HMMs with four states and 32 Gaussian mixtures are used. We note that, since the difference between the recognition rates of 16 and 32 Gaussian mixture models is not statistically significant, HMM with four states and 16 Gaussian mixtures can be used as well.

An important conclusion from Tables I and II is that, even though the vocabulary used for this research is limited to ten digits, as the noise level increases, the recognition decreases to very low levels. This conclusion reveals that, in noisy environments, visual speech signal can potentially be used together with the acoustic speech signal to improve the recognition performance of the system.

For acoustic modeling, both the HTK HMM and developed Matlab HMM were used. Tables I and II show the results obtained from the developed Matlab-based HMM. Table III shows the comparison of the

TABLE III
RECOGNITION ACCURACY COMPARISON OF HTK AND MATLAB HMMS

Recognition Rate (%)								
SNR(dB)	Clean	30dB	25dB	20dB	15dB	10dB	5dB	0dB
HMM Source								
HTK HMM	100	100	97.9	94.2	84.2	75	60.4	47.1
Matlab HMM	94.2	93.8	93.8	91.7	85	77.9	60.1	47.9

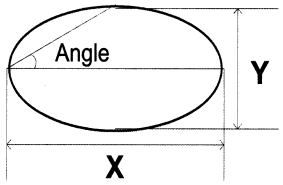

Fig. 4. Basic geometrical lip features.

recognition accuracy of the developed and HTK HMM codes using four states and four Gaussian mixtures for speaker independent case. Although HTK HMM outperforms the developed HMM by around 5% at high SNR, the performance gap diminishes as SNR decreases. For the rest of the experiments, developed HMM code is used in order that the recognition results of acoustic-only, visual-only, and audio-visual speech recognition obtained by the same code can be rigorously compared.

III. ANALYSIS OF GEOMETRIC VISUAL FEATURES AND VISUAL HIDDEN MARKOV MODELING

In this section, first single geometric visual features are analyzed and then the combinations of geometric visual features are experimented on to find the best performing visual feature combination for bimodal speech recognition.

For visual hidden Markov modeling, left-right HMMs with continuous observation densities and diagonal covariance matrices, implying that statistically independent features were used. Another reason for using diagonal covariance matrices is that more training data is required to train the HMMs for full covariance matrices. The best performing HMM was found for both the single and the combined visual features for each digit through a series of experiments on different combinations of the number of states and the number of Gaussian mixture components. In the next section, single geometric visual features are analyzed for their importance in bimodal speech recognition. This analysis gives a valuable insight into the best features that can be used for bimodal speech recognition.

A. Single Geometric Visual Feature Analysis

As explained previously, in order to have an accurate and noise-free training, in the audio-visual database six blue markers were placed on the speaker's face as shown in Fig. 1. Different topologies of the markers were reported in literature. In these studies researchers focused on finding the most information carrying part of the face and they concluded that, most of the visual speech information is carried around the lip area [16]. We also experimented different topologies, however the topology shown in Fig. 1 gave us the best performance in terms of recognition and tracking accuracy. After automatically detecting the centers of the dots using a convolution-based approach, four geometrical lip parameters were extracted. These were the outer-lip horizontal aperture (X), the outer-lip vertical aperture (Y), the outer-lip area (Δ) defined as the inside of the ellipse, and the angle of the outer-lip corner (Θ), shown in Fig. 4. The vertical distance between the points on the chin and nose was used to normalize X and Y, in order that the features became invariant to the distance between the speaker and the camera.

It should be noted that not only the shape of the outer-lip contour, but also the movement of the lip contour are important for distinguishing the digits. Therefore, the first-order derivatives of the four basic lip features are also considered for the experiments.

A systematic study was carried out to determine the most relevant visual features for speech recognition. We analyzed ten different visual features, namely, X, Y, Δ, Θ, Y/X, and the first-order derivatives of these five parameters.

The analysis was carried out in two stages. At the first stage, the best single visual features were determined and at the second stage the best visual feature combination was determined. In order to determine the optimum HMM for each digit and each single visual feature, six HMMs having four, five states with four, eight, and 16 Gaussian mixtures were trained. The recognition performances of the trained models were evaluated against the validation data using the forward backward algorithm.

Table IV shows the experiment results for the ten single visual features. In this table, "$'$" denotes the derivative over time and the average recognition rates are calculated for every single visual feature over all the digits. Note that in the table, the recognition results are shown for each single visual feature alone (with no other visual feature or audio features being used). From these results, it can be concluded that the geometric visual features along the vertical directions, such as Y and Y', are more important than the ones along the horizontal direction, such as the feature X. This is justified by the higher recognition rate for the vertical direction features. The single visual feature Θ' represents not only the coordinative lip movements along the vertical and horizontal directions but also expresses the velocities of these movements, thus it is an important feature.

From Table IV, it can also be concluded that by taking into account the best average recognition rates, X, Y, Θ, Y', Θ', and $(\mathrm{Y}/\mathrm{X})'$ performed well for almost all digits so they are the most relevant geometric visual features for bimodal speech recognition to optimize the recognition rate.

B. Combined Visual Feature Analysis

Now that the best single visual features have been determined, the question is to decide which combination of these six geometric visual features should be used for bimodal speech recognition for an acceptable recognition rate. In order to determine the best visual feature combination, further experiments were conducted on seven different combinations of the visual features, namely, X-Y, Y-Y', X-Y-Θ, X-Y-Y', X-Y-Θ', X-Y-Θ-Θ', and X-Y-Y-$(\mathrm{Y}/\mathrm{X})'$. These visual feature combinations are selected in an empirical manner among the single geometric visual features with higher average recognition rates in Table IV. For selecting the visual feature combinations, factor or principal component analysis (PCA) can be used to obtain optimal combinations from a statistical point of view.

For these visual feature combinations, for each digit six different HMMs, with four and five states and eight, 16 and 32 Gaussian mixture components were trained by using 77 samples from nine speakers. Then performances of these HMMs were evaluated against the validation data for both speaker dependent and independent cases. For speaker independent case 24 samples from ten speakers, and for speaker dependent case 19 samples from nine speakers were used as the validation date. The results are shown in Tables V and VI for speaker independent and dependent cases, respectively. According to both of these tables, the best visual feature combination is X-Y-Θ' and the best model for this combination is HMM with five states and 32 mixtures.

A very important conclusion from the results reported in Tables V and VI is that although single visual features do not perform very well when used singly (as expected due to the relatively small information

TABLE IV
VISUAL SPEECH RECOGNITION FOR THE SINGLE VISUAL FEATURE EXPERIMENT

Digit / Feature	One	Two	Three	Four	Five	Six	Seven	Eight	Nine	Zero	Average Rec. rate
X	35.4	30.6	37.5	38.9	52.8	48.6	29.9	34.7	38.2	6.9	35.4
Y	37.5	31.3	19.4	51.4	72.2	22.9	58.3	34.0	36.1	15.3	37.9
Y/X	31.3	34.7	27.8	20.8	33.3	28.5	49.3	16.7	24.3	12.5	27.9
Δ	23.6	14.6	16.0	34.0	36.8	16.7	22.2	10.4	7.6	33.3	21.5
Θ	29.2	35.4	29.2	21.5	39.6	29.9	45.8	28.5	20.8	20.8	30.1
X′	22.9	27.8	34.0	38.2	59.0	13.9	20.1	27.1	9.7	16.7	26.9
Y′	29.9	21.2	16.7	50.7	68.8	6.9	68.1	45.8	31.3	25.7	37.3
(Y/X)’	36.1	25.0	19.4	34.0	66.7	8.3	56.9	26.4	27.8	31.3	33.2
Δ'	35.4	17.4	25.7	44.4	61.8	17.4	50.7	22.2	22.9	26.4	32.4
Θ'	23.6	27.1	22.2	35.4	66.0	6.3	60.4	22.2	25.7	34.7	32.4

TABLE V
VISUAL SPEECH RECOGNITION BASED ON COMBINED VISUAL FEATURES FOR SPEAKER-INDEPENDENT CASE

N	4	4	4	5	5	5
M	8	16	32	8	16	32
Visual Feature	Average Recognition Rate (%)					
X-Y	57.9	62.5	63.3	63.3	62.9	62.9
Y-Y′	46.7	47.1	42.9	46.7	51.3	52.1
X-Y-Y′	65.0	70.8	70.8	70.4	72.5	70.8
X-Y-Θ	51.7	50.4	51.7	52.5	50.8	50.8
X-Y-Θ'	68.3	72.1	72.5	67.9	73.3	74.6
X-Y-Y′-(Y/X)′	60.8	60.0	59.2	60.0	62.5	60.8
X-Y-Θ-Θ'	60.0	60.8	62.1	59.2	60.0	60.8

TABLE VI
VISUAL SPEECH RECOGNITION BASED ON COMBINED VISUAL FEATURES FOR SPEAKER-DEPENDENT CASE

N	4	4	4	5	5	5
M	8	16	32	8	16	32
Visual Feature	Average Recognition Rate (%)					
X-Y	60.5	64.7	67.4	68.4	65.3	63.2
Y-Y′	48.4	47.9	45.3	49.5	52.1	53.7
X-Y-Y′	67.4	74.7	74.2	74.7	74.7	75.8
X-Y-Θ	62.1	61.1	60.5	64.2	62.6	61.1
X-Y-Θ'	72.6	75.3	76.3	71.6	76.8	78.4
X-Y-Y′-(Y/X)′	64.7	64.7	63.2	64.7	67.4	65.3
X-Y-Θ-Θ'	73.2	75.3	73.2	70.0	72.1	73.2

content from each single visual feature), when they are combined to form a visual feature vector, the recognition performance improves significantly; the recognition rate almost doubles and the system becomes more robust to noise. For the bimodal speech-recognition experiments, we will use X-Y-Θ' as the visual feature vector due to higher recognition accuracy of this combination.

IV. AUDIO-VISUAL HIDDEN MARKOV MODELING

Since acoustic speech signal is susceptible to acoustic noise, in noisy environments the acoustic speech signal-based recognition may not be accurate enough to identify speech. Our experimental results reported for acoustic speech recognition confirmed this argument. The visual speech recognition experiments detailed in Section IV showed that the visual signal carries relevant information for speech recognition. In this section the audio-visual speech recognition experiments and the fusion of the acoustic and visual data streams are detailed.

For the experiments, we used X-Y-Θ' and 13 dimensional MFCC as the visual and acoustic feature vectors, respectively. An important problem for bimodal speech recognition is the fusion of the acoustic and visual speech information. Some of the previous studies on audio-visual speech recognition have focused on the optimal integration of

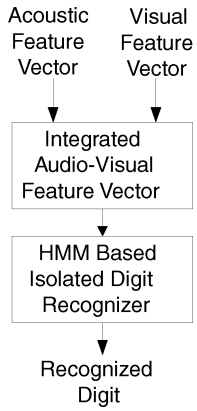

Fig. 5. Early integration, feature fusion system.

the speech information from the acoustic and visual channels. There are two widespread beliefs about how humans integrate acoustic and visual speech information [25]: 1) early integration (feature fusion or direct identification) and 2) late integration (decision fusion or separate identification).

The former approach uses one recognition engine for both the visual and acoustic features, while the latter approach uses two recognition engines; one for the visual signal, and one for the acoustic signal, and then integrates the recognition results of each recognition engine.

In this research, an early integration system through the direct identification (DI) strategy was developed (Fig. 5). By this approach, features from acoustic and visual channels should be concatenated to form a joint feature vector. However, acoustic and visual feature vectors are not synchronous, because the acoustic features are extracted from 25 ms windows with 10 ms overlapping; on the other hand, visual features are extracted from 40 ms frames without any overlapping. Hence, for the same speech signal, there are different numbers of feature sequences for visual and acoustic features. Thus, merging the acoustic and visual signals raises a problem of the nonsynchronism between the acoustic and visual channels. In order to solve this problem, first, each single visual feature signal was resampled at a higher rate using low pass interpolation. Then, from these up-sampled visual signals, new visual features were obtained from 25 ms windows with 10 ms overlapping, by averaging the samples inside each window. This way, the visual and acoustic feature sequences produced have the same number of samples.

After obtaining the audio-visual feature vector, for each digit six different HMMs having four and five states, and 8, 16, and 32 Gaussian mixtures were trained. 77 samples from nine speakers were used to train the HMM of each digit. Then, their performances were evaluated against the validation data which included the clean and noise-corrupted test data. For testing, 24 samples from ten speakers were used

TABLE VII
AUDIO-VISUAL SPEECH RECOGNITION FOR SPEAKER-INDEPENDENT CASE

N	4	4	4	5	5	5
M	8	16	32	8	16	32
SNR	Average Recognition Rate (%)					
Clean	96.3	97.9	98.8	96.7	97.5	98.3
30dB	96.3	97.9	98.8	97.1	97.5	97.9
25dB	96.7	97.9	98.8	97.1	97.5	97.9
20dB	96.3	97.5	98.8	96.7	96.7	97.9
15dB	96.3	96.3	98.3	94.6	96.7	97.9
10dB	95.8	95.4	97.9	93.8	95.8	96.3
5dB	88.3	94.2	94.2	90.0	92.1	92.9
0dB	80.4	83.3	82.1	82.5	81.7	81.7

TABLE VIII
AUDIO-VISUAL SPEECH RECOGNITION FOR SPEAKER-DEPENDENT CASE

N	4	4	4	5	5	5
M	8	16	32	8	16	32
SNR	Average Recognition Rate (%)					
Clean	97.9	99.5	99.5	98.4	99.0	100
30dB	97.9	99.5	99.5	99.0	99.0	100
25dB	97.9	99.5	99.5	99.0	99.0	100
20dB	97.4	99.5	99.5	98.4	99.0	100
15dB	94.4	98.4	99.0	96.3	98.4	99.5
10dB	95.8	97.4	99.0	95.3	96.8	97.4
5dB	87.9	95.3	94.2	90.5	92.6	93.7
0dB	79.5	84.2	80.0	82.6	81.6	82.1

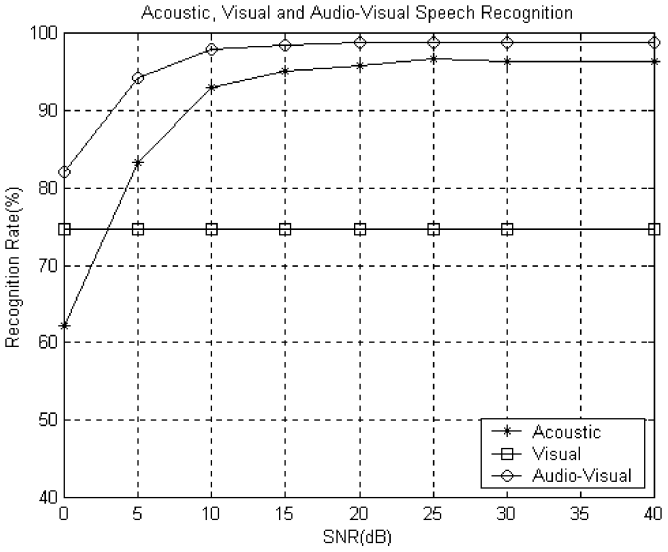

Fig. 6. Recognition rate using different speech information sources under various acoustic SNRs for speaker-independent case.

for speaker-independent case and 19 samples from nine speakers were used for speaker-dependent case. The evaluation results are shown in Tables VII and VIII for speaker independent and dependent cases, respectively. The results for the speaker-dependent case are slightly better than speaker-independent case, but the difference is not significant. From these results, for each digit, the optimum HMM was selected. Among these different digit models, the HMM with four states and 32 Gaussian mixture components produced the best results for speaker-independent case and the HMM with five states and 32 Gaussian mixture components produced the best results for speaker-dependent case.

The best performing HMMs were used for testing the performance of the bimodal speech-recognition system. The average recognition accuracy of acoustic-only, visual-only and audio-visual speech recognition are shown in Figs. 6 and 7 for speaker independent and dependent cases. As noise increases, audio-visual speech recognition produced much better results than the acoustic-only speech recognition for both speaker-dependent and -independent cases. Thus, the higher the noise is, the more the recognition rate is improved compared to the acoustic-only speech recognition.

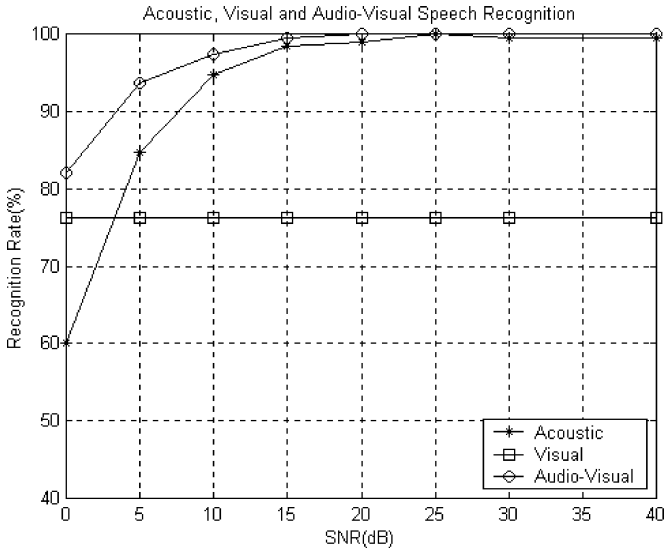

Fig. 7. Recognition rate using different speech information sources under various acoustic SNRs for speaker-dependent case.

From recognition results of audio-visual speech-recognition experiments, the most important conclusion that can be drawn is that, when visual speech information is used together with acoustic speech information, speech recognition becomes more robust to noise and, furthermore, the bimodal speech recognition system out-performs both acoustic-only and visual-only speech-recognition systems by as much as 20% and 8% at $\mathrm{SNR} = 0$ dB, respectively.

During our research, we used HMMs to model the speech signal. In the literature, fuzzy logic and NNs are also used to model the speech signal. However, for the same database, the performances of fuzzy logic and NN-based ASR systems were inferior to HMM-based ASR systems [16]. This inferior performance is mostly due to the fact that HMM can successfully model the sequentially changing behavior of speech by its dynamic state structure. However, for fuzzy logic and NN-based ASR systems, it is not possible to include the time-domain information to the recognition engine, so this makes the modeling of speech signal difficult.

V. CONCLUSION

In this paper, we performed a rigorous analysis and comparison of geometric features that can be used in an audio-visual speech-recognition system. For this purpose, first, single geometric visual features were experimented, and their importance for bimodal speech recognition was determined through experiments conducted on a visual speech recognition system using optimal HMMs. Experimental results showed that geometric visual features along the vertical directions, such as Y and Y', are more important than the ones along the horizontal direction, such as the feature X. We then experimented the combined visual features using optimized HMMs and found that the visual feature vector formed by X, Y, and Θ results in high recognition accuracy.

For bimodal speech-recognition experiments, early integration algorithm is used to form a joint feature vector from acoustic and visual features. Experimental results showed that for limited vocabulary digit recognition, the audio-visual speech-recognition system significantly improved the recognition rate, especially at high noise levels and out-

performed both acoustic-only and visual-only speech recognition systems for both speaker independent and dependent cases. It is important to note that a set of as few as three labial geometric features is sufficient to improve the recognition rate by as much as 20% (from 62%, with acoustic-only information, to 82%, with audio-visual information at $\mathrm{SNR} = 0$ dB).

Finally, we note that, although we used markers to extract the geometric visual features, geometric visual features can be extracted in real time accurately without using any markers [26], [27].

ACKNOWLEDGMENT

The authors would like to thank the reviewers for their valuable comments that helped to improve the presentation of this paper. The authors also appreciate the valuable discussions with Dr. C. E. Siong from Knowles Electronics, Singapore.

REFERENCES

[1] B. P. Yuhas, M. H. Goldstein Jr., T. J. Sejnowski, and R. E. Jenkins, "Neural network models of sensory integration for improved vowel recognition," *Proc. IEEE*, vol. 78, pp. 1658–1668, Oct. 1990.

[2] H. McGurk and J. MacDonald, "Hearing lips and seeing voices," *Nature*, vol. 264, no. 5588, pp. 746–748, 1976.

[3] P. L. Silcbee, "Automatic lipreading," *Biomed. Sci. Instrum.*, vol. 29, pp. 415–22, 1993.

[4] E. Vatikiotis-Bateson, I.-M. Eigsti, S. Yano, and K. Munhall, "Eye movement of perceivers during audiovisual speech perception," *Perception Psychophys.*, vol. 60, no. 6, pp. 926–940, 1998.

[5] H. Yehia, P. Rubin, and E. Vatikiotis-Bateson, "Quantitative association of vocal-tract and facial behavior," *Speech Commun.*, vol. 26, no. 1, pp. 23–43, 1998.

[6] T. Chen and R. R. Rao, "Audio-visual integration in multimodal communication," *Proc. IEEE*, vol. 86, pp. 837–852, May 1998.

[7] J. R. Movellan, "Visual speech recognition with stochastic networks," in *Proc. Neural Inform. Process. Syst. (NIPS): Natural and Synthetic*, vol. 7, 1995, pp. 851–858.

[8] E. D. Petajan, "Automatic lipreading to enhance speech recognition," in *Proc. IEEE Global Telecommun. Conf. (GLOBECOM)*, vol. 1, Atlanta, GA, Nov. 26–29, 1984, pp. 265–272.

[9] E. D. Petajan, N. M. Brooke, B. J. Bischoff, and D. A. Bodoff, "An improved automatic lipreading system to enhance speech recognition," in *Proc. CHI'88*, New York, 1988, pp. 19–25.

[10] D. G. Stork, G. Wolff, and E. Levine, "Neural network lipreading system for improved speech recognition," in *Proc. Int. Joint Conf. Neural Networks (IJCNN)*, vol. 2, Baltimore, MD, 1992, pp. 289–295.

[11] J. Luettin, N. A. Thacker, and S. W. Beet, "Speechreading using shape and intensity information," in *Proc. 4th Int. Conf. Spoken Lang. Process. (ICSLP)*, vol. 1, Philadelphia, PA, 1996, pp. 58–61.

[12] G. Potamianos, H. Cosatto, H. Graf, and D. Roe, "Speaker independent audio-visual database for bimodal ASR," in *Proc. Euro Tutorial Workshop Audio-Visual Speech Process.*, Rhodes, Greece, 1997, pp. 65–68.

[13] S. Basu, C. Neti, N. Rajput, A. Senior, L. Subramaniam, and A. Verma, "Audio-visual large vocabulary continuous speech recognition in the broadcast domain," in *Proc. IEEE 3rd Workshop Multimedia Signal Process.*, Copenhagen, Denmark, Sept. 13–15, 1999, pp. 475–81.

[14] K. Yu, X. Jiang, and H. Bunke, "Lipreading using signal analysis over time," *Signal Process.*, vol. 77, no. 2, pp. 195–208, 1999.

[15] S. Dupont and J. Luettin, "Audio-visual speech modeling for continuous speech recognition," *IEEE Trans. Multimedia*, vol. 2, pp. 141–151, Sept. 2000.

[16] D. G. Stork, "Visionary speech: Looking ahead to practical speechreading systems," in *Speechreading by Humans and Machines*, M. E. Hennecke, Ed. Berlin, Germany: Springer, 1996, pp. 331–351.

[17] J. S. D. Mason, J. Brand, R. Auckenthaler, F. Deravi, and C. Chibelushi, "Lip signatures for automatic person recognition," in *Proc. 2nd Int. Conf. Audio- and Video-Based Biometric Person Authentification*, Washignton, DC, Mar. 22 – 23, 1999, pp. 142–147.

[18] M. N. Kaynak, Q. Zhi, A. D. Cheok, C. C. Ko, S. Cavalier, and K. K. Ng, "Database generation for bimodal speech recognition research," in *Proc. 2nd JSPS-NUS Seminar Integrated Eng.*, Singapore, 2000, pp. 220–227.

[19] P. Cosi, J.-P. Hosom, J. Shalkwyk, S. Sutton, and R. A. Cole, "Connected digit recognition experiments with the OGI toolkit's neural network and HMM-based recognizers," in *Proc. 4th IEEE Workshop Interactive Voice Technol.Telecommun. Applicat. (IVTTA)*, Turin, Italy, 1998, pp. 135–140.

[20] J. Schalkwyk, D. Colton, and M. Fanty, "The CSLU Toolkit for Automatic Speech Recognition," Tech. Rep., Oregon Graduate Inst. Sci. Technol., Ctr. for Spoken Lang. Understanding, Beaverton, 1995.

[21] M. C. Lincoln and A. E. Clark, "Toward pose-independent face recognition," *IEE Colloq. Visual Biometrics*, pp. 5/1–5, 2000.

[22] L. R. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition*. Englewood Cliffs, NJ: Prentice-Hall, 1993, Prentice-Hall Signal Processing Series.

[23] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc. IEEE*, vol. 77, pp. 257–286, Feb. 1989.

[24] C. Becchetti and L. P. Ricotti, *Speech Recognition, Theory and C++ Implementation*. New York: Wiley, 1999.

[25] A. Rogozan and P. Deleglise, "Adaptive fusion of acoustic and visual sources for automatic speech recognition," *Speech Commun.*, vol. 26, no. 1–2, pp. 149–161, 1998.

[26] J. Zhang, M. N. Kaynak, A. D. Cheok, and C. C. Ko, "Real-time lip tracking for virtual lip implementation in virtual environments and computer games," in *Proc. IEEE Int. Conf. Fuzzy Syst.*, vol. 3, Melbourne, Australia, 2001, pp. 1359–1362.

[27] M. T. Chan, Y. Zhang, and T. S. Huang, "Real-time lip tracking and bimodal continuous speech recognition," in *Proc. IEEE 2nd Workshop Multimedia Signal Process.*, Los Angeles, CA, Dec. 1998, pp. 65–70.

A Note on the Robust Stability of Uncertain Stochastic Fuzzy Systems With Time-Delays

Zidong Wang, Daniel W. C. Ho, and Xiaohui Liu

***Abstract*—Takagi–Sugeno (T-S) fuzzy models are now often used to describe complex nonlinear systems in terms of fuzzy sets and fuzzy reasoning applied to a set of linear submodels. In this note, the T-S fuzzy model approach is exploited to establish stability criteria for a class of nonlinear stochastic systems with time delay. Sufficient conditions are derived in the format of linear matrix inequalities (LMIs), such that for all admissible parameter uncertainties, the overall fuzzy system is stochastically exponentially stable in the mean square, independent of the time delay. Therefore, with the numerically attractive Matlab LMI toolbox, the robust stability of the uncertain stochastic fuzzy systems with time delays can be easily checked.**

***Index Terms*—Fuzzy systems, linear matrix inequality (LMI), nonlinear systems, robust stability, stochastic systems, time-delay systems.**

I. INTRODUCTION

Stability analysis of stochastic systems has been well investigated in the past three decades, since stochastic modeling has come to play

Manuscript received September 18, 2003; revised December 18, 2003. This work was supported in part by the EPSRC under Grant GR/S27658/01 and Grant GR/R35018/01, in part by the Nuffield Foundation under Grant NAL/00630/G, in part by the RGC of Hong Kong SAR under Grant CityU/101103, and in part by the Alexander von Humboldt Foundation of Germany. This paper was recommended by Associate Editor T. Sudkamp.

Z. Wang and X. Liu are with the Department of Information Systems and Computing, Brunel University, Uxbridge, Middlesex UB8 3PH, U.K. (e-mail: Zidong.Wang@brunel.ac.uk).

D. W. C. Ho is with Department of Mathematics, City University of Hong Kong, Kowloon, Hong Kong.

Digital Object Identifier 10.1109/TSMCA.2004.826296