

EXPERIMENTS USING THE CUAVE AUDIO-VISUAL SPEECH CORPUS IN ROTATION AND SCALE CORRECTION OF VISUAL SPEECH FEATURES

E.K. Patterson and J.N. Gowdy

Department of Electrical and Computer Engineering
Clemson University
Clemson, SC 29634, USA
{epatter, jgowdy}@eng.clemson.edu
<http://ece.clemson.edu/speech>

Abstract

Audio-visual speech recognition has become an important area of research because it holds great potential for overcoming certain problems of traditional audio methods. The effects of corrupting data from background noises and other speakers may be reduced by the additional information provided by visual features. To date, though, much study in this area has been limited to a tightly constrained environment with few established databases for comparison. This paper presents initial experiments on a new audio-visual database designed to promote research in robust audio-visual speech recognition.

The Clemson University Audio Visual Experiments corpus is a challenging audio-visual database that is flexible and fairly comprehensive, yet compact enough to facilitate research progress and be made easily available on one DVD. It is a corpus of digit strings dictated by a wide variety of speakers and speaker pairs and is designed to allow testing of adverse conditions such as moving talkers and simultaneous speech. A feature study of connected digit strings is discussed in this paper that compares three feature methods using techniques for rotation and scale correction over groups of stationary and moving talkers in a speaker-independent grouping.

For information on obtaining CUAVE, please visit our webpage (<http://ece.clemson.edu/speech>).

1. Introduction

Over recent years the potential of multimodal signal processing has grown as computing power has increased. Audio-visual speech processing has shown great potential, particularly in speech recognition. The addition of information from lipreading or other features helps make up for information lost due to corrupting influences in the audio. Because of this, audio-visual speech recognition can outperform audio-only recognizers, particularly in environments where there are background noises

or other speakers. Researchers have demonstrated the relationship between lipreading and human understanding [1, 2] and have produced performance increases with multimodal systems [3–7].

Because of difficulties associated with the high volumes of data necessary for simultaneous video and audio and because of the short time that research has been conducted in this area, the creation and distribution of related databases have been limited to date. Many researchers have been forced to record their own data. This has often been limited to either cropped video, stationary speakers, or video with aids for feature segmentation. The CUAVE corpus is a new audio-visual database that has been designed to help meet some of these criteria. The main purpose is to allow testing of more realistic speaker data that includes a variety of visual features and speaker movement.

One of the goals of the CUAVE database is to include realistic conditions for testing of robust methods. One of these considerations is the movement of speakers. Methods are necessary that do not require a fixed talker. Recordings from the database are grouped into tasks where speakers are mostly stationary and tasks in which speakers intentionally move while speaking. This paper includes results from initial experiments comparing several feature methods on both stationary and moving speakers. Methods are discussed that attempt to improve results by correcting visual features for differences in angle and scale.

2. The CUAVE Audio-Visual Database

To help meet the need for a more widespread testbed for audio-visual development, CUAVE was produced as a speaker-independent database consisting of connected and continuous digits spoken in different situations. Some of the main strengths are the inclusion of speaker movement and simultaneous speaker pairs. The database is presented more thoroughly in [8].

<i>Part</i>	<i>Task</i>	<i>Movement</i>	<i>Number of Digits</i>
<i>1. Individual</i>	1	Still	50 x 36 speakers
	2	Moving	30 x 36 speakers
	3	Profile	20 x 36 speakers
	4	Still	30 x 36 speakers
	5	Moving	30 x 36 speakers
<i>2. Pairs</i>	6	Still	(30 x 2) x 20 pairs

Table 2.1: Summary of CUAVE Tasks

2.1. Design Goals and Corpus Format

The major design criteria were to create a challenging yet easily distributable database that allows for representative and fairly comprehensive testing. Because DVD readers for computers have become very economical recently, the choice was made to design CUAVE to fit on one DVD-data disc. CUAVE is designed to enhance research in two important areas: audio-visual speech recognition that is robust to speaker movement and also recognition that is capable of distinguishing multiple simultaneous speakers. The database is also fully, manually labeled to improve training and testing possibilities.

The database includes two major sections, one of 36 individual speakers and one of 20 speaker pairs. The selection of individuals was not tightly controlled but chosen so that there is a roughly even representation of male and female speakers and also so that different skin tones and accents are present. There are also other features such as glasses, facial hair, and hats. A wide variety of skin and lip tones as well as face and lip shapes is present.

Table 2.1 includes a summary of the different recorded tasks in the database. The first part including individuals has various recordings of digit strings. Speakers were either asked to remain stationary in the frame of view or move around depending on the task. The frame includes the shoulders and head, and during moving tasks, speakers move side-to-side, front-to-back, and tilt their head. There is also an occasional turn of the head.

The second major section includes 20 pairs who simultaneously speak continuous digit strings similar to telephone numbers. Again, see Table 2.1. The goal is to allow for testing of multispeaker solutions. These include distinguishing a single speaker from others as well as the ability to simultaneously recognize speech from two talkers. This is obviously a difficult task, particularly with audio information only. Video features correlated with speech features should facilitate solutions to this problem. (One such application could be a shopping-mall kiosk that distinguishes a user from other shoppers nearby while giving voice-guided information.) The two speakers in the group section are labeled persons A and B. There are three sequences per person. Person A speaks a continuous-digit sequence, followed by speaker B and

vice-versa. For the third sequence, both speaker A and B overlap each other while speaking each person's separate digit sequence.

The recording environment was controlled to produce high-quality, color video and sound:

- Recorded in an isolated sound booth.
- NTSC standard 29.97 fps at 720x480 resolution.
- MiniDV converted to MPEG-2 at 5,000 kbps.
- 16-bit sound, 44 kHz stereo and 16 kHz mono.
- Distributed on one DVD with label files.

Lighting was controlled, and a green background was used to allow chroma-keying of different backgrounds. This serves two purposes. If desired, the green background can be used as an aid in segmenting the face region, but more importantly, it can be used to add video backgrounds from different scenes, such as a crowd or a moving car to allow for testing of robust feature segmentation and tracking algorithms. We plan to include standard video backgrounds (such as recorded in a shopping mall, crowded room, or moving automobile) in an upcoming release. The data-rate and final selection of speakers and groups was chosen so that a medium-sized database of high-quality, digital video and audio, as well as HTK-compatible [9] label data and some tools could be released on one 4.7 GB DVD. This helps with the difficulty of distributing and working with the unruly volume of data associated with high-quality video.

3. Initial Study on Angle and Scale with Stationary and Moving Speakers

This section discusses a moving-talker, speaker-independent feature study over several features. Image-processing, image-transform, and template-matching methods are employed over Part 1, Tasks 1-2 of the CUAVE database. The task includes individuals speaking connected-digit strings while either stationary or moving. The visual features used are affine-invariant Fourier descriptors (AIFDs) [10], the 2-D separable discrete cosine transform (DCT), an improved rotation-corrected application of the DCT, and a Bézier search template (BST) scheme. Methods for adjusting the scale of visual features are also discussed with experimental results.

3.1. System Details

This subsection describes the setup for the stationary versus moving talker comparison over various visual features. The image-processing method for AIFDs detailed below is the most sensitive to the weakness of a single-mixture, uniform color model. Because of this, an attempt was made to generate a more fair comparison of

results. A group of 14 speakers was chosen that yielded more robust lip-contour extraction. The group was arranged arbitrarily into 7 training speakers and 7 testing speakers for a completely speaker-independent study. Part 1 of the database, tasks 1 and 2, were used for this study. (See Table 2.1.) For each set of visual features, HMMs were trained using HTK with the same speakers. The models were simple single-mixture models with eight states per model.

For this study, coefficients are only differenced over one frame, as some previous work has also shown no apparent improvement with longer temporal windows [11]. All visual feature schemes here begin with the same face and lip tracking algorithms. Smaller and larger regions-of-interest (ROIs) are passed to each feature routine. The smaller region tracks tightly around the lips and can be searched to locate lip features, such as the lip corners, angle, or width. The larger includes the jaw, specifically to give a larger region for application of the DCT, as shown in [12] to yield better results than a smaller region. All features are passed as difference coefficients at a rate of 29.97 Hz, to match the NTSC standard. As these are speech-reading results only, no interpolation to an audio frame rate is performed.

A Bayesian classifier based on R,G,B Gaussian mixtures is used to segment the face and lip regions for lip localization. Since R,G,B values are used, a measure of intensity is also included. This is relatively consistent in this test case, although this could be difficult in practical applications, as the intensity may vary. This classification is performed on image blocks or pixels, and the class with the larger value is chosen. Currently, only a single Gaussian mixture is used. This results in good class separation over all but a few speakers whose facial tones are ruddy and very similar to their lip tones. After locating the probable face region, this area is searched for the best template match on lip-classified pixels. A “center-of-mass” check is also employed to help center the tracker on the lips. The final lip region and a larger region of roughly the jaw area are both made available to any of the feature-extraction algorithms.

3.2. Image-Processing Contour Method

The feature-extraction method discussed here involves the application of affine-invariant techniques to Fourier coefficients of lip contour points estimated by image-processing techniques. More information about the development and use of affine-invariant Fourier descriptors (AIFD) is given in [10, 13]. The outer-lip edge within a binary image of the ROI passed from the liptracker is traversed to assemble pixel coordinates such as shown in Figure 3.1. Once the lip contour coordinates are determined, the Discrete Fourier Transform (DFT) is applied. The redundant coefficients are removed. The zeroth coefficient is then discarded to leave shift-invariant information. Re-

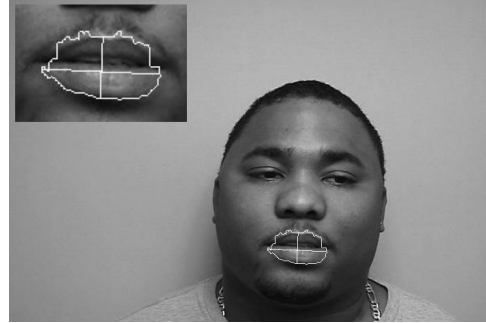


Figure 3.1: Final Mouth Contour for AIFD Calculation.



Figure 3.2: Downsampling for DCT Application.

maining coefficients are divided by an additional arbitrary coefficient to eliminate possible rotation or scaling effects. Finally, the absolute value of the coefficients is taken to remove phase information and, thus, eliminate differences in the starting point of the contour-coordinate listing. This leaves a set of coefficients (AIFDs) that are invariant to shift, scale, rotation, and point-ordering.

3.3. Image Transform Method

The larger ROI passed from the liptracker is downsampled into a 16x16 grayscale intensity image matrix, as shown in Figure 3.2. The DCT was chosen as the image transform instead of other transform methods because of its information packing properties and strong results presented by other research [11, 12]. The separable 2-D DCT was used in this work. The upper left block (6x6) of the transform matrix is used for feature coefficients, with the exception of the zeroth element that is dropped to perform feature mean subtraction for better speaker-independent results. The DCT by itself is not robust to speaker movement. To attempt to improve this, a rotation-corrected version of the image block was passed to the DCT (rc-DCT). The angle for rotation correction is determined by searching for the lip corners and estimating the vertical angle of the lips/head from the two lip corners. This was chosen for speed, versus estimating the tilt of the whole

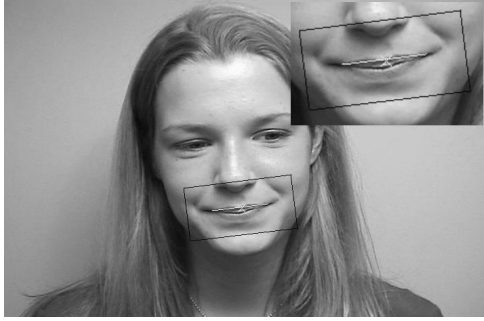


Figure 3.3: Rotation Correction Before Applying DCT.



Figure 3.4: Rotation-Corrected, Downsampled Image Region for DCT Application.

head with an elliptical template search. All image pixels that would then form a box parallel to the angled lips are chosen for the matrix to which the DCT is applied. Figure 3.3 demonstrates estimation of the rotation-correction factor, and Figure 3.4 shows a downsampled image matrix based on the estimated angle.

An improved version of the rc-DCT algorithm was also tested that “smooths” the angle estimate between frames to minimize improper estimates or erratic changes. The DCT, rc-DCT, and smoothed rc-DCT results are compared against the AIFD and BST results in Section 3.5.



Figure 3.5: Control Vertices for Estimated Lip Curves.

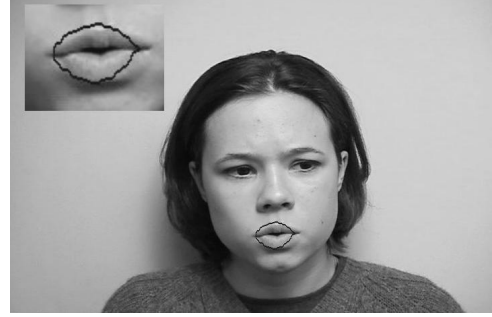


Figure 3.6: Estimated Lip Contour Using Control Vertices for Bézier Curves.

3.4. Deformable-Template Method

The main goal of the deformable-template approach used is to capture information from lip movement in a reference-coordinate system, thus directly producing affine-invariant results. The method of determining the lip movement is simple, direct, and slightly less sensitive to improper lip segmentation than more traditional image-processing techniques. The deformable template here is based on Bézier curves [14]. Lip contours are guided by minimizing a cost function $C(R)$ over the area enclosed by the curves, region R [15]:

$$C(\mathbf{R}) = \sum_{(x,y) \in R} \log \frac{P(\mathbf{o}(x,y)|\omega_{face})}{P(\mathbf{o}(x,y)|\omega_{lips})}. \quad (1)$$

In the tables of results, this technique will be referred to as Bézier Search Template (BST). Eight control vertices (CVs) are used, four for the top lip, and four for the bottom. (Because the top and bottom points for the lip corners are the same, this actually reduces to six distinct points.) CVs are demonstrated in Figure 3.5. Bézier curves generated by CVs are shown in Figure 3.6. CVs are also known as B_i points, where $B_i = (x_i, y_i)$, that control the parametric curves by the following formulas:

$$P(t) = \sum_{i=0}^n B_i J_{n,i}(t), 0 \leq t \leq 1 \quad (2)$$

$$J_{n,i} = \binom{n}{i} t^i (1-t)^{n-i} \quad (3)$$

There are two sets of 4 B_i points: B_{0-3}^t for the top curve, and B_{0-3}^b for the bottom curve. The range of values for t is usually normalized to the 0 to 1 range in parametric curve formulas. Here, the width of the lips is used to normalize this. B_0 and B_3 for the top and bottom are initially set to the corners of the lips in the image coordinate system. Based on the angle estimation, these points are rotated into a relative, reference coordinate system, with B_0 as the origin. (Affine transforms may be applied to CVs without affecting the underlying curves). New B_i values for searches are generated based on changing

<i>Features</i>	<i>Stationary</i>	<i>Moving</i>
AIFD	22.40 %	21.89 %
DCT	27.71 %	19.62 %
rc-DCT	22.57 %	21.53 %
smoothed rc-DCT	27.43 %	23.92 %
BST	22.86 %	24.46 %

Table 3.2: Comparison of Performance (Word Accuracy) on Stationary and Moving Speakers, Digit Strings.

the width, height, location, etc. of the template. These are then transformed back to the image-coordinate system to generate resulting curves. Several searches are performed, while minimizing $C(R)$ within these curves in the image-coordinate system. Once B_{0-3}^l and B_{0-3}^b are determined that minimize $C(R)$, the actual reference coordinate system CV (x, y) values are differenced from the previous frame and passed on as visual features that capture the moving shape information of the lips. In this respect, angle and translation variance are eliminated.

3.5. Stationary and Moving Results for Subgroup and Rotation Correction

Each of the visual feature methods were used on the test setup described in 3.1. The grouping is completely speaker independent, training on one group and testing on a completely separate group. Task 1 (see Table 2.1) is used for stationary testing and training. Task 2 is used for moving-speaker testing and training. The results are presented in Table 3.2 over a stationary set and a moving set with models trained on the stationary set. Although results may not strictly be compared to other results, the range of these is on the order of results from other medium-sized, continuous, speaker independent speech-reading tasks [11, 12]. Results presented are obtained using single-mixture, eight-state HMMs. The AIFD features were shown to perform nearly equally well on stationary and moving speakers as hoped. Confirming the conclusion in [11] that an image transform approach yields better performance than lip-contour methods, the DCT features outperform the AIFD-contour features in this test system. A large part of this is likely due to sensitivity to the lip segmentation and tracking algorithms. The DCT is much less sensitive, as the larger block can be more easily located than precise lip contours. DCT performance drops substantially, though, under moving speaker conditions. Implementing the rotation-correction did improve the performance of the rc-DCT on the moving-speaker case, however stationary performance dropped significantly. This is due to the dependence on the lip segmentation introduced by the lip-angle estimation. Implementing the smoothing factor as discussed in Subsection 3.3 both improved results more on the moving case, and nearly regained sta-

tionary performance. The BST features performed on-par with the AIFDs. Another interesting note, though is that they actually earned the best performance on moving speakers in this smaller testset, surpassing the smoothed rc-DCT and even their own stationary performance. They seem to be fairly robust to speaker movement, still capturing the lip-motions important for speech-reading. They are still, however, sensitive to exact lip segmentation and construction of the cost function.

3.6. Speaker-Independent Results Using All Speakers

In this section results are over the whole database for Tasks 1 and 2, stationary and moving tasks, respectively. The 36 individual speakers were divided arbitrarily into a set of 18 training speakers and 18 different test talkers for a completely speaker-independent grouping. With a simple, MFCC-based HMM recognizer implemented in HTK using 8-state models and only one mixture per state, we obtained an audio-only word accuracy of 87.25%. The visual results in all studies here are also obtained with fairly simple HTK models that could be improved with multiple-mixtures and boundary re-estimation to increase recognition accuracies in further study. Results are included for DCT, rc-DCT, smoothed rc-DCT, and BST features as in the previous test. AIFDs are not included because the current implementation is very sensitive to differences that a single-mixture color model does not represent well. In fact, the BST features which are somewhat less-sensitive also show a performance drop over the whole group. This is particularly true for the moving results, where the BST features performed well in the prior test. The DCT gained the highest score on the stationary task with 29% word accuracy. Performance drops to the level of the BST features, though, on the moving task. Again the rc-DCT performs better on moving speakers, but loses performance on stationary speakers. The smoothed rc-DCT performs the best here on moving speakers but doesn't quite restore the full performance of the DCT on stationary speakers. This is also probably affected by the additional speakers who do not fit the color model as well as the prior test group. Estimating the lip angle to correct the DCT suffers when lip segmentation is poor. Overall, results seem to indicate that contour methods might perform as well as transform methods if robust enough. The difficulty is creating speaker independent models that perform accurate lip segmentation under the many varying conditions. This is perhaps one of the most difficult problems to overcome before practical systems may be developed.

3.7. Attempting Scale Correction

The earlier subsections detail recognition results using various visual features and attempts to improve affine-invariance of these features. The affine-invariant Fourier

<i>Features</i>	<i>Stationary</i>	<i>Moving</i>
DCT	29.00 %	21.12 %
rc-DCT	25.95 %	22.39 %
smoothed rc-DCT	26.47 %	24.73 %
BST	23.85 %	21.48 %

Table 3.3: Baseline Speechreading Results (Word Accuracy) over All Speakers.

descriptors (AIFDs) by their development are invariant to scale as well as rotation and shift; the discrete cosine transform (DCT) and Bézier Search Template (BST) features, however, are not naturally scale-invariant. This subsection presents attempts to improve results based on scaling the image block used for the DCT and the control vertices used to describe the Bézier lip contours.

There are two possible sources for improvements by making features invariant to scale. The first possibility of change in scale comes from a speaker’s head movement back-and-forth from the camera. This movement should cause a degradation in single-speaker or speaker dependent performance of visual speech recognition. The second possibility of differences in scale arises from the various sizes of heads and lips among speakers. Improving invariance to scale in this regard could possibly improve speaker independence of visual features.

In order to scale the DCT and BST features, a reference measurement to scale against needs to be made. In general, facial features remain relatively to scale along with the size of head among various humans [16]. Based on this assumption, the method chosen estimates the width of a speaker’s head within the face-and-lip tracker routine. Once the speaker’s lips are located, the width of the face is estimated just above them by searching the breadth of pixels classified as “face.” (Another assumption is made that face width should not routinely change at or above this point due to jaw movement while speaking.) Once this measurement is estimated, a scale ratio is based upon an arbitrary width, roughly that of one of the smallest faces in the database. Based on this ratio, the DCT or BST features are scaled. For the DCT features, the size of the image block that is converted to grayscale and downsampled is expanded or contracted using the ratio. The 16x16 2-D DCT is still used, but based on the new image block. For the BST features, the location of the control vertices in the reference coordinate system are scaled using the ratio. Difference features are still passed as before.

The first trial used a single width estimate, just above the lip region. Recognition tests were performed using the stationary and moving speaker tasks as in the previous subsections. Recognition accuracy results are given in Table 3.4 using all the speakers from the database for the DCT. The first results based on the single width estimate

<i>Features</i>	<i>Stationary</i>	<i>Moving</i>
DCT, Measured Scale	23.96 %	22.13 %
DCT, Single Width	22.51 %	21.02 %
DCT, Averaged Width	23.74 %	22.13 %
srcDCT, Averaged Width	22.80 %	22.60 %
DCT, No Scale	29.00 %	22.12 %
srcDCT, No Scale	26.47 %	24.73 %
BST, Averaged Width	23.08 %	21.70 %
BST, No Scale	23.85 %	21.48 %

Table 3.4: Word Accuracy Based on Various Scale Estimates, Over All Speakers.

are labeled “DCT, Single Width.” Compared with the original DCT results (“DCT, No Scale”), the scaled performance is lower on the stationary speakers (22.51% v. 29.00%) but nearly the same on moving speakers (21.02% v. 22.12%). This decrease on stationary speakers while almost holding performance on moving speakers suggests that no significant “speaker-independence” is gained by the scaling, since results actually drop when speakers are not moving. No significant change in performance on moving speakers suggests that there may be little forward-to-backward movement among speakers, which is true based on the database recordings. Confined to a small recording booth, backward movement was the most limited motion during the “moving-speaker” task recordings. More side-to-side and angular movement of the head is present in most speakers.

Because of these possibilities, a second scale measurement was attempted. The width of each person’s face was measured manually over initial frames of each speaker’s recordings. The features were then scaled accordingly in the same manner as before. Results are also included in Table 3.4. The stationary DCT performance here is improved (23.96%) but still falls short of the original DCT performance (29.00%). The moving performance, however, now is equal to (insignificantly greater than) the standard DCT performance. This seems to support the conclusions that there is little forward-to-backward movement and that no speaker-independent improvements are made. Assuming little-to-no forward-to-backward movement, a manually measured scale should be accurate (not suffering from estimate errors) and thus improve speaker-independence of the feature set. One possibility, though, is that the HMM models used for recognition do not “generalize” as well during training when the features are scaled for each speaker.

The final method attempted for scaling was an improvement of the width estimate used in the first method. The scaling of the face is performed using the average width of a speaker’s face based on four dynamic width estimates across the middle of the face, as shown in Figure 3.1. These again are estimated after locating the lips



Figure 3.1: Estimating Scale on Averaged Width of Face

in the video frame. An improvement is seen in the recognition performance, as shown in Table 3.4. The performance on both stationary and moving speakers is almost identical to that of the manually measured technique (23.74% and 22.13%, respectively). Again, no improvements are seen overall, though, suggesting that the scale does not appear to add much “speaker-independence” or movement-invariance, at least in the case of this speech corpus. A scaling technique would likely be much more important if the speaker’s distance from the camera were allowed to vary more. A sitting speaker, such as in a car or in front of a computer, will likely have little movement that affects scale. It becomes much more useful information, though, for a speaker that may be walking about a room, for instance.

The averaged-width method was also tested over all the database using both stationary and moving speakers. These results are included last in Table 3.4. The performance is nearly equal to that of the BST features with no scale adjustment. Interestingly, there is a slight drop in stationary performance and slight increase in moving performance.

Finally, all the feature sets were tested over the smaller test group that allows a better comparison with the AIFD features. These results are shown in Table 3.5. In these results, the scaled DCT demonstrates some improvement over the plain DCT (22.97% v. 19.62%) on moving speakers, although results drop slightly on stationary speakers. The performance improvement is not as high as that gained by adding the rotation correction, though (22.97% v. 23.92%). (Rotation also seems to be more prevalent than back-and-forth motion). The rotation-and-scale corrected DCT results for moving speakers remain the same as the rotation-corrected results, but stationary results again drop. A similar pattern is shown by the BST performance.

Overall, these results suggest that adjustment based on scale does not appear to significantly improve moving-speaker performance when forward-backward motion is relatively limited. Head rotation seems to be more typical among speakers and thus more important to correct.

<i>Features</i>	<i>Stationary</i>	<i>Moving</i>
AIFD (already scaled)	22.40 %	21.89 %
DCT scaled	23.71 %	22.97 %
DCT, No Scale	27.71 %	19.62 %
srcDCT scaled	24.29 %	23.92 %
srcDCT, No Scale	27.43 %	23.92 %
BST, scaled	20.29 %	23.37 %
BST, No Scale	22.86 %	24.46 %

Table 3.5: Word Accuracy Based on Averaged Width Scale Estimate Using Test Group.

Also, no improvement in speaker-independent performance is demonstrated by these results. In fact, the performance on stationary speakers tended to drop when scaling was introduced, possibly due to misestimation of the scale or lack of generality in trained models. The measured-scale-DCT performance on the stationary speakers supports the possibility of loss of model generality, as measured scaling should be accurate for non-moving speakers.

4. Conclusions and Research Directions

This paper presents a flexible, speaker-independent audio-visual speech corpus that is easily available on one DVD-data disc. The goal of the CUAVE database is to facilitate multimodal research and to provide a basis for comparison, as well as to provide test data for affine-invariant feature methods and multiple-speaker testing. Our website, listed in the title section, contains sample data and includes current contact information for obtaining the database. As it is a representative, medium-sized digits task, the database may also be used for testing in other areas apart from speech recognition. These may include speaker recognition, lip synchronization, visual-speaker synthesis, etc. Also, results are given that suggest that data fusion can benefit by considering noise-type as well as level. Experiments on the database using image-processing-based contours, image transform, and deformable-template methods as visual features have also been detailed. Results on stationary and moving talkers and also attempts to lessen the effect of speaker movement by normalizing visual features for angle and scale have been presented. Angle-based corrections improved results on moving speakers, but care needs to be taken to prevent loss of performance on stationary speakers. Scale-correction did not appear particularly significant as a method for improving speaker-independent results.

There are several areas that are wide-open for audio-visual speech recognition research. A very important area is visual-feature extraction. A variety of speakers and speaker movement has been included in our database for this end. Methods are needed that either significantly strengthen feature-tracking under practical conditions or

that create new features for speech-reading. Better features may include some other measures of the mouth, teeth, tongue, jaw, psychoacoustic considerations, or eye-direction to name a few possibilities. Techniques for speaker-independent speech-reading are also needed. These may include some form of speaker adaptation, model adaptation, image warping, use of feature codebooks, or some of the many other methods employed for audio speaker adaptation. Finally, data fusion is an important area of research. Improved techniques for multimodal, multi-stream HMMs could also provide important strides, particularly in continuous audio-visual speech recognition. Other methods, such as hybrid fusion systems may be considered. Dynamic considerations will be important for improving data fusion for practical environments. Finally, the ability to distinguish and separate speakers is important for powerful interfaces that may be desired where multiple speakers are present, such as in public areas or automobiles with passengers.

5. References

- [1] D. W. Massaro, *Perceiving Talking Faces: From Speech Perception to a Behavioral Principle*, Cambridge, MA: MIT Press, 1997.
- [2] Q. Summerfield, "Lipreading and audio-visual speech perception," *Phil. Trans. R. Soc.*, vol. 335, 1992.
- [3] E. Petajan, B. Bischoff, D. Bodoff, and N. Brooke, "An improved automatic lipreading system to enhance speech recognition," in *ACM SIGGHI*, pp. 19-25, 1988.
- [4] P. L. Silsbee and A. C. Bovik, "Computer lipreading for improved accuracy in automatic speech recognition," *IEEE Transactions on Speech, and Audio Processing*, vol. 4, no. 5, September 1996.
- [5] P. Teissier, J. Robert-Ribes, J. Schwartz, and A. Guérin-Dugué, "Comparing models for audiovisual fusion in a noisy-vowel recognition task," *IEEE Transactions on Speech, and Audio Processing*, vol. 7, no. 6, November 1999.
- [6] C. Neti, G. Potamianos, J. Luetin, I. Matthews, H. Glotin, D. Vergyri, J. Sison, A. Mashari, and J. Zhou, "Audio-visual speech recognition final workshop 2000 report," Center for Language and Speech Processing, Johns Hopkins University, Baltimore, 2000.
- [7] G. Potamianos, C. Neti, G. Iyengar, and E. Helmuth, "Large-vocabulary audio-visual speech recognition by machines and humans," in *Eurospeech, Denmark*, 2001.
- [8] E. Patterson, S. Gurbuz, Z. Tufekci, and J. Gowdy, "Moving-talker, speaker-independent feature study and baseline results using the cuave multimodal speech corpus," *EURASIP Journal on Applied Signal Processing*, , no. 11, 2002.
- [9] S. Young, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, *The HTK Book*, Entropic Cambridge Research Laboratory Ltd., version 2.1, 1997.
- [10] S. Gurbuz, Z. Tufekci, E. K. Patterson, and J. N. Gowdy, "Application of affine-invariant fourier descriptors to lipreading for audio-visual speech recognition," in *Proceedings of ICASSP*, 2001.
- [11] G. Potamianos, H.P. Graf, and E. Cosatto, "An image transform approach for hmm based automatic lipreading," in *Proceedings of the ICIP, Chicago*, 1998.
- [12] G. Potamianos and C. Neti, "Improved roi and within frame discriminant features for lipreading," in *ICIP, Thessaloniki, Greece*, 2001.
- [13] K. Arbter, W. E. Snyder, H. Burkhardt, and G. Hirzinger, "Application of affine-invariant fourier descriptors to recognition of 3-d objects," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 12, no. 7, 1990.
- [14] D. Rogers, *An Introduction to NURBS*, Morgan Kaufmann Publishers, 2001.
- [15] T. Chen, "Audiovisual speech processing: Lip reading and lip synchronization," *IEEE Signal Processing Magazine*, vol. 18, no. 1, January 2001.
- [16] S. Simblet, *Anatomy for the Artist*, Doring Kenderley, 2001.