EURASIP Journal on Applied Signal Processing 2002:11, 1189–1201

Moving-Talker, Speaker-Independent Feature Study, and Baseline Results Using the CUAVE Multimodal Speech Corpus

Eric K. Patterson
Department of Electrical and Computer Engineering, Clemson University, Clemson, SC 29634, USA
Email: epatter@eng.clemson.edu

Sabri Gurbuz
Department of Electrical and Computer Engineering, Clemson University, Clemson, SC 29634, USA
Email: sabrig@eng.clemson.edu

Zekeriya Tufekci
Department of Electrical and Computer Engineering, Clemson University, Clemson, SC 29634, USA
Email: ztufekci@eng.clemson.edu

John N. Gowdy
Department of Electrical and Computer Engineering, Clemson University, Clemson, SC 29634, USA
Email: jgowdy@eng.clemson.edu

Received 30 November 2001 and in revised form 10 May 2002

Strides in computer technology and the search for deeper, more powerful techniques in signal processing have brought multimodal research to the forefront in recent years. Audio-visual speech processing has become an important part of this research because it holds great potential for overcoming certain problems of traditional audio-only methods. Difficulties, due to background noise and multiple speakers in an application environment, are significantly reduced by the additional information provided by visual features. This paper presents information on a new audio-visual database, a feature study on moving speakers, and on baseline results for the whole speaker group. Although a few databases have been collected in this area, none has emerged as a standard for comparison. Also, efforts to date have often been limited, focusing on cropped video or stationary speakers. This paper seeks to introduce a challenging audio-visual database that is flexible and fairly comprehensive, yet easily available to researchers on one DVD. The Clemson University Audio-Visual Experiments (CUAVE) database is a speaker-independent corpus of both connected and continuous digit strings totaling over 7000 utterances. It contains a wide variety of speakers and is designed to meet several goals discussed in this paper. One of these goals is to allow testing of adverse conditions such as moving talkers and speaker pairs. A feature study of connected digit strings is also discussed. It compares stationary and moving talkers in a speaker-independent grouping. An image-processing-based contour technique, an image transform method, and a deformable template scheme are used in this comparison to obtain visual features. This paper also presents methods and results in an attempt to make these techniques more robust to speaker movement. Finally, initial baseline speaker-independent results are included using all speakers, and conclusions as well as suggested areas of research are given.

Keywords and phrases: audio-visual speech recognition, speechreading, multimodal database.

1. INTRODUCTION

Over the past decade, multimodal signal processing has been an increasing area of interest for researchers. Over recent years, the potential of multimodal signal processing has grown as computing power has increased. Faster processing allows the consideration of methods which use separate audio and video modalities for improved results in many applications. Audio-visual speech processing has shown great potential, particularly in speech recognition. The addition of information from lipreading or other features helps making up for information lost due to corrupting influences in the audio. Due to 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, 4, 5, 6, 7].

Because of the young age of this area of research as well as difficulties associated with the high volumes of data necessary for simultaneous video and audio, the creation and distribution of audio-visual databases have been somewhat limited to date. Most researchers have been forced to record their own data. This has been limited in many ways to either cropped video, stationary speakers, or video with aids for feature segmentation. In order to allow for better comparison and for researchers to enter the area of study more easily, available databases that meet certain criteria are necessary. The Clemson University Audio-Visual Experiments (CUAVE) corpus is a new audio-visual database that has been designed to meet some of these criteria.[1] It is a speaker-independent database of connected (or isolated, if desired) and continuous digit strings of high quality video and audio of a representative group of speakers and is easily available on one DVD. Section 2 of this paper presents a discussion of the database and associated goals. Section 2.1 discusses a brief survey of audio-visual databases and presents motivation for a new database. Next, Section 2.2 includes design goals of the CUAVE database and specifics of the data collection and corpus format.

One of the goals of the CUAVE database is to include realistic conditions for testing robust audio-visual schemes. One of these considerations is the movement of speakers. Methods that do not require a fixed talker are necessary. Recordings from the database are grouped into tasks where speakers are mostly stationary and tasks where speakers intentionally move while speaking. Movements include nodding the head in different directions, moving back and forth and side-to-side in the field of view, and in some cases rotation of the head. This results in more difficult testing conditions. A feature study, discussion, and results that compare stationary and moving talkers in a speaker-independent grouping are included. Features included are an image-processing-based contour method that is naturally affine invariant, an image transform method that is also modified to improve rotation robustness, and a new variation of the deformable template that is also inherently affine invariant. Section 3 includes a discussion of the different feature extraction methods, the overall recognition system, training and testing from the database, and results of the different techniques.

Finally, one of the main purposes of all speech corpora is to allow the comparison of methods and results in order to stimulate research and fuel advances in speech processing. This is a main consideration of the CUAVE database, easily distributable on one DVD. It is a fully-labeled, medium-sized task that facilitates more rapid development and testing of novel audio-visual techniques. To this end, baseline results are included using straightforward methods for a speaker-independent grouping of all the speakers in a connnected-digit string task. Section 4 presents these results. Finally, Section 5 closes with some observations and suggestions about possible areas of research in audio-visual speech processing.

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

2. THE CUAVE SPEECH CORPUS

This section discusses the motivation and design of the CUAVE corpus. It includes a brief review of known audio-visual databases. Information on the design goals of this database are included along with the corpus format. Finally, a single speaker case that was performed in preparation before collecting this database is discussed. This case presents isolated word recognition in various noisy conditions, comparing an SNR-based fusion with a noise-and-SNR-based fusion method. Basing fusion on the noise type is shown to improve results.

2.1. Audio-visual databases and research criteria

There have been a few efforts in creating databases for the audio-visual research community. Tulips1 is a twelve subject database of the first four English digits recorded in 8-bit grayscale at 100×75 resolution [8]. AVLetters includes the English alphabet recorded three times by ten talkers in 25 fps grayscale [9]. DAVID is a larger database including various recordings of thirty-one speakers over five sessions including digits, alphabets, vowel-consonant-vowel syllable utterances, and some video conference commands distributed on multiple SVHS tapes [10]. It is recorded in color and has some lip highlighting. The M2VTS database and the expanded XM2VTSDB are geared more toward person authentication and include 37 and 295 speakers, respectively, including head rotation shots, two sequences of digits, and one sentence for each speaker [11]. The M2VTS database is available on HI-8 tapes, and XM2VTSDB is available on 20 1-hour MiniDV cassettes. There have also been some proprietary efforts by AT&T and by IBM [6, 12]. The AT&T database was to include connected digits and isolated consonant-vowel-consonant words. There were also plans for large-vocabulary continuous sentences. However, the database was never completed nor released. The IBM database is the most comprehensive task to date, and fortunately, front-end features from the IBM database are becoming available but there are currently no plans to release the large number of videos [12]. Large-vocabulary continuous speech recognition is an ultimate goal of research. Because audio-visual speech processing is a fairly young discipline, though, there are many important techniques open to research that can be performed more easily and rapidly on a medium-sized task. To 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. It is flexible, representative, and easily available. Section 2.2 discusses the design goals, collection methods, and format of the corpus.

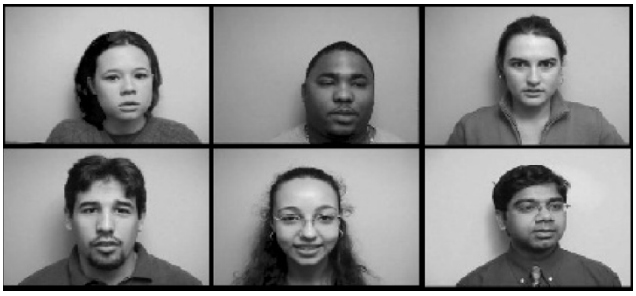

Figure 1: Sample speakers from the database.

2.2. Design goals and corpus format

The major design criteria were to create a flexible, realistic, 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 that could be made available through contact information listed on our website. Aside from being a speaker-independent collection of isolated and connected digits (zero through nine), 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 and manually labeled to improve training and testing possibilities. This section discusses the collection and formatting of the database performed to meet the aforementioned goals.

The database consists of two major sections: one of individual speakers and one of speaker pairs. The first major part, with individual speakers, consists of 36 speakers. The selection of individuals was not tightly controlled, but was chosen so that there is an even representation of male and female speakers, unfortunately a rarity in research databases, and also so that different skin tones and accents are present as well as other visual features such as glasses, facial hair, and hats. Speakers also have a wide variety of skin and lip tones as well as face and lip shapes. See Figure 1 for a sample of images from some of the CUAVE speakers.

Individual speakers were framed above the shoulders and recorded speaking connected and continuous digit strings. Initially, 50 connected digits are spoken while standing still. As this was not forced, there is some small natural movement among these speakers. Also, at a few points, some speakers actually lapse into continuous digits. Aside from these lapses, the connected-digit tasks may be treated as isolated digits, if desired, since label data is included for all digits. This section is for general training and testing without more difficult conditions. Video is full color with no aids given for face/lip segmentation. Secondly, each individual speaker was asked to intentionally move around while talking. This includes side-to-side, back and forth, and/or tilting movement of the head while speaking 30 connected digits. There is occasionally slight rotation in some of the speakers as well. This group of moving talkers is an important part of the database to allow for testing of affine invariant visual features. In addition to this connected-digit section, there is also a continuous-digit section with movement as well. So far, much research has been limited to low resolution and pre-segmented video of only the lip region. Including the speaker's upper body and segments with movement will allow for more realistic testing. The third and fourth parts spoken by each individual include profile views and continuous-digit strings facing the camera. Both profiles are recorded while speaking 10 connected digits for each side. The final continuous digits were listed as telephone numbers on the prompting to elicit a more natural response among readers. There are 6 strings. The first 3 are stationary and the final 3 are moving for a total of 60 continuous digits for each speaker. See Table 1 for a summary of these tasks.

Table 1: Summary of CUAVE tasks: individuals and pairs, stationary and moving, connected and continuous digits.

Part	Task	Movement	Number of digits	Mode
(1) Individual	1	Still	50×36 speakers	Connected
	2	Moving	30×36 speakers	Connected
	3	Profile	20×36 speakers	Connected
	4	Still	30×36 speakers	Continuous
	5	Moving	30×36 speakers	Continuous
(2) Pairs	6	Still	$(30 \times 2) \times 20$ pairs	Continuous

The second major section of the database includes 20 pairs of speakers. Again, see Table 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. In addition, techniques that will help with the difficult problem of speech babble can be tested here. (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

Figure 2: Sample speaker pair from the database.

versa. For the third sequence, both speaker A and B overlap each other while speaking each person's separate digit sequence. See Figure 2 for an image from one of the speaker-pair videos.

The database was recorded in an isolated sound booth at a resolution of 720×480 with the NTSC standard of 29.97 fps, using a 1-megapixel-CCD, MiniDV camera. (Frames are interpolated from separate fields.) Several microphones were tested. An on-camera microphone produced the best results: audio that was clear from clicking or popping due to speaker movement and video where the microphone did not block the view of the speaker. The video is full color with no visual aids for lip or facial feature segmentation. 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 mall, or a moving car, etc. See Figure 3. In this manner, not only can audio noise be added for testing but video *noise* such as other speakers' faces or moving objects may be added as well that will allow for testing of robust feature segmentation and tracking algorithms. We plan to include standard video backgrounds (such as those recorded in a shopping mall, crowded room, or moving automobile) in an upcoming release.

Nearly three hours of data from about 50 speakers were recorded onto MiniDV tapes and transferred into the computer using the IEEE 1394 interface (FireWire). The recordings were edited into the selection of 36 representative speakers (17 female and 19 male) and 20 pairs. Disruptive mistakes were removed but occasional vocalized pauses and mistakes in speech were kept for realistic test purposes. The data was then compressed into individual MPEG-2 files for each speaker and group. It has been shown that this does not significantly affect the collection of visual features for lipreading [13]. The MPEG-2 files are encoded at a data-rate of 5,000 kbps with multiplexed 16-bit, stereo audio at a 44 kHz sampling rate. An accompanying set of downsampled *wav* format audio files, checked for synchronization, are included at a 16-bit, mono rate of 16 kHz. The data is fully-labeled manually at the millisecond level. HTK-compatible label files are included with the database [14]. 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 label data (and possibly some database 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 audio-visual databases. The next section discusses a single-speaker test case conducted before collecting the database.

2.3. Speaker-independent test case using noise-based late integration

Before recording the database, we conducted a test study in the area of visual features and data fusion related to audio-visual speech recognition using a single speaker for a 10-word isolated speech task [15, 16]. The audio-visual recognizer used was a late-integration system with separate audio and visual hidden Markov model (HMM) based speech recognizers. The audio recognizer used sixteen Mel frequency discrete wavelet coefficients extracted from speech sampled at 16 kHz [17]. The video recognizer used red-green plane division and thresholding before edge detection of the lip boundaries [18]. Geometric features based on the oral cavity and affine invariant Fourier descriptors were passed to the visual recognizer [15]. All the noises from the NOISEX database [19] were added at various SNRs, and the system was trained on clean speech and tested under noisy conditions. Initial results for the single speaker were good and led us to choose some of the database design goals mentioned in Section 2.2, such as the inclusion of recordings with moving speakers and continuous digit strings. See Table 2 for a summary of these results. (Here, the field of view was cropped to lips only to ease preliminary work.) The purpose of this study was to investigate the effects of noise on decision fusion. Late integration was used and recognition rates were found over all noises and multiple SNRs. See Table 3. For each noise type and level, optimal values of data fusion using a *fuzzy-and* rule were determined [1, 16]. Results obtained using the fusion values based on noise type and SNR are shown to give slightly superior results to those obtained with fusion based solely on SNR.

3. FEATURE STUDY USING 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 and 2 of the CUAVE database (see Table 1). The task includes individuals speaking connected-digit strings while either stationary or moving. The visual features used are affine-invariant Fourier descriptors (AIFDs) [15], the discrete cosine transform (DCT), an improved rotation-corrected application of the DCT, and a naturally affine-invariant B-Spline template (BST) scheme. (See Table 1 for a summary of the tasks.) Details of the overall system are presented in Section 3.1. The techniques for segmenting and tracking the face and lips are discussed in Section 3.2. Visual-feature extraction is discussed

Figure 3: Examples of varied video backgrounds.

Table 2: Recognition rates (percentage correct) averaged over noises for single-speaker case.

	Audio only	Video only	SNR-based AV	Noise & SNR-based AV
Clean	100%	87.0%	100%	100%
18 dB	96.5%	87.0%	99.1%	99.3 %
12 dB	80.9%	87.0%	96.4%	96.9%
6 dB	52.6%	87.0%	91.5%	92.9%
0 dB	29.1%	87.0%	87.6%	89.9%
−6 dB	21.0%	87.0%	87.0%	88.4%

Table 3: Performance comparison of (SNR-based — noise-based) late integration using noises from NOISEX-92.

Noise	Clean	18 dB	12 dB	6 dB	0 dB	−6 dB
Speech	100% — 100%	99% — 99%	94% — 95%	91% — 91%	85% — 87%	87% — 87%
Lynx	100% — 100%	99% — 99%	94% — 95%	90% — 91%	85% — 87%	87% — 87%
Op Room	100% — 100%	100% — 100%	97% — 97%	92% — 94%	90% — 90%	87% — 88%
Machine Gun	100% — 100%	100% — 100%	99% — 99 %	96% — 97%	92% — 94%	87% — 91%
STITEL	100% — 100%	98% — 98%	94% — 95%	85% — 89%	81% — 87%	87% — 87%
F16	100% — 100%	98% — 99%	97% — 97%	91% — 92%	88% — 89%	87% — 89%
Factory	100% — 100%	99% — 99%	97% — 98%	90% — 92%	88% — 91%	87% — 88%
Car	100% — 100%	100% — 100%	99% — 99%	97% — 97%	92% — 93%	87% — 90 %

in Sections 3.3, 3.4, and 3.5. Lastly, Section 3.6 presents a comparison of moving and stationary talker results using the various feature methods.

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 who yielded more robust lip contour extraction was chosen. 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 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.

Before each of the following visual features was tested, a brief initial test was performed to demonstrate the importance of dynamic features as supported by other researchers [13, 18]. Table 4 contains results from a difficult two-speaker test case. Recognition models were trained on DCT features from one female speaker and tested on one male speaker with facial hair. Static and difference features as well as the combination were compared.

Table 4: Word accuracy for preliminary difficult test case demonstrates usefulness of difference features.

DCT features	Accuracy
35 Static	6%
35 Difference	18%
15 Difference	10%
15 Stat./20 Diff.	10%

As shown in previous work and in this case, different coefficients seem to provide more useful information, particularly when dealing with a speaker-independent case. Shape information along with information on lip movement is more applicable in a speaker-dependent case. Methods for

speaker adaptation, image warping, or codebook schemes might improve the use of shape information. For dynamic coefficients, some work also demonstrates that better results can be obtained by averaging over several adjacent frames for longer temporal windows [20].

For this study, coefficients are only differenced over one frame, as some previous work has also shown no apparent improvement with longer temporal windows [13]. 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 [20] to yield better results than a smaller region. All features are passed as different coefficients at a rate of 29.97 Hz to match the NTSC standard. As these are speechreading results only, no interpolation to an audio frame rate is performed. Techniques for locating the face and lip regions are described next.

3.2. Detecting the face and lips

Dynamic thresholding of red-green division used to initially segment the lips in our earlier work did not prove to be good for feature segmentation over the database of many speakers. For this reason, we manually segmented the face and lip regions for all speakers, as shown in Figure 4, to obtain the mean and variance of the nonface, face, and lip color distributions. Since R, G, and 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. The estimated Gaussian distributions for the nonface, face, and lip classes are shown in Figure 5. These are used to construct a classifier based on the Bayes classification rule [21],

$$p(\mathbf{x}|\omega_1)P(\omega_1) \overset{?}{<>} (\mathbf{x}|\omega_2)P(\omega_2), \tag{1}$$

where $\mathbf{x}$ is the feature vector (RGB levels), ω is each class (nonface versus face, or face versus lips), $p(\mathbf{x}|\omega_i)$ is estimated by the respective estimated Gaussian density, and $P(\omega_i)$ is estimated by the ratio of class pixels present in training data. 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 except for a few speakers whose facial tones are ruddy and very similar to their lip tones.

For speed in image-processing routines, the image is broken into 10×10 pixel blocks. Each block is averaged and classified as nonface or face, then as face or lips. Determining the location of the speaker's lips in the video frame is central to determining all visual features presented here, although some are more sensitive to exact location. In order to identify the lip region, the face is first located in each frame. This is based on a straightforward template search of the classified blocks, locating the largest region classified as *face*. The segmentation from this step is shown in Figure 6. The current search

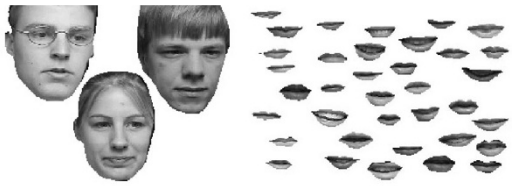

Figure 4: Manually segmented face and lips for distribution training.

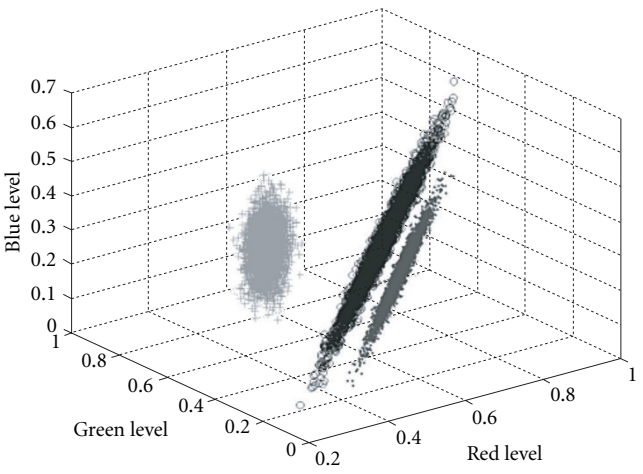

Figure 5: Gaussian estimation of nonface, face, and lip color distributions.

template is rectangular to improve the speed over using an oval template that might have to be rotated and scaled to match speaker movement. This performs well for the current setup that does not have *environment* backgrounds substituted for the green screen via chroma keying. Figure 7 displays a *located* face. There are some extreme cases of speaker movement such as large tilting of the head that can cause poor face location. This would likely be solved by the slower scaled/rotated oval template.

After locating the probable face region, this area is searched for the best template match on lip classified pixels. Pixels are used in locating the lip region for more accuracy. A *center of mass* check is also employed to help centering the tracker on the lips. Again, a rectangular template is used for speed rather than an oval template. Heuristics, such as checking block classifications near the current template, help preventing misclassifying red shirt collars as lips. See Figure 8. A final overlaid lip segmentation is shown in Figure 9. After the lip region is located, it is passed in two forms to the chosen visual feature extraction routine. A smaller box around the location of the lips is passed to be searched for lip corners and angle, a larger box is passed as well. It has been shown that performing the image transform on a slightly larger region including the jaw as well as the lips yields improved results in speechreading [20]. Figure 10 demonstrates the regions passed on to the feature routines. The routines have been designed so that they may be expanded to track multiple

Figure 6: Segmentation of face from nonface and lips within face region.

Figure 8: Sensitivity of classification scheme to red clothing, improved by searching within the face region and using Heuristics.

Figure 7: Locating the face region.

Figure 9: Final lip segmentation.

speakers, as in the CUAVE speaker-pair cases. The following subsections discuss each of the feature methods compared in this work.

3.3. 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 AIFD is given in [15, 22]. A binary image of the ROI passed from the liptracker is created using the classification rule in (1). This is similar to the overlaid white pixels shown in Figure 9. The Robert's cross operator is used to estimate lip edges as shown in Figure 11. The outer lip edge is traversed to assemble pixel coordinates to which AIF technique will be applied.

A difficulty of lip pixel segmentation is that in certain cases, lip pixels are occasionally lost or gained by improper classification. In this case, the lip contour may include extra pixels or not enough ones such as in Figure 12 where there is separation between the upper and lower lips. This separation would cause a final mouth contour to be generated which may only include the lower lip such as in Figure 13 rather than the good contour detected and shown in Figure 14. This affects all the feature-extraction techniques to some degree, since the lip tracker is based on similar principles. However, contour estimation methods are the most affected. (The DCT method presented in the next subsection is the least affected by this, though, helping account somewhat for its better performance.) These effects could possibly be minimized by adding more mixtures to the color models, using an adaptive C-means-type scheme for color segmentation within regions, or including additional information such as texture. Additional mixtures would be more important for extracting the inner lip contour around teeth and shadow pixels. For the comparison in this paper, the single-mixture model suffices, since all feature schemes are based on the same liptracker routine. In fact, using individually trained mixtures for each person rather than a single mixture overall did not yield improved recognition results in our tests even though segmentation appeared somewhat more accurate.

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. Remaining 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, rotate, and point-order. (Although we employed a parametric-contour estimation of lip contours before AIFD calculation in our earlier work for improved results, this is not yet implemented in this system [15]. Results here are for directly estimated lip contours with no curve estimation or smoothing.)

3.4. Image transform method

The larger ROI passed from the liptracker is downsampled into a 16×16 grayscale intensity image matrix, as shown in Figure 15. The DCT was chosen as the image transform instead of other transform methods because of its information

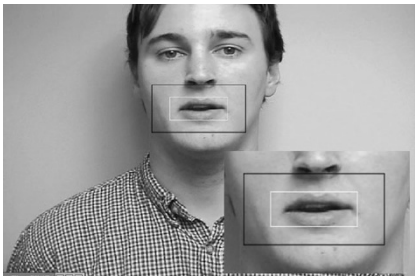

Figure 10: Locating the smaller and larger lip region.

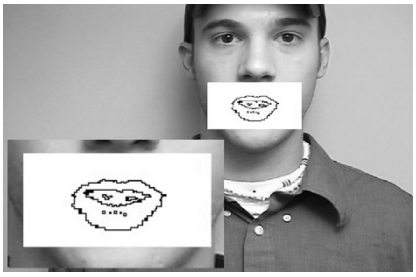

Figure 11: Mouth contour after edge detection.

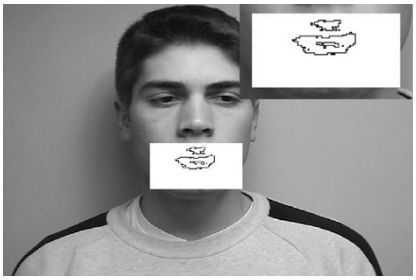

Figure 12: Poor mouth contour after edge detection.

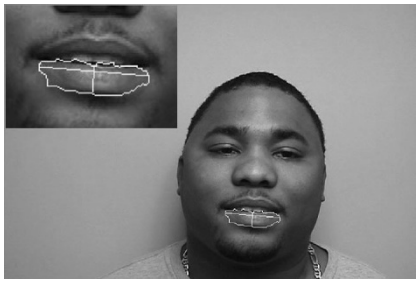

Figure 13: Poor final mouth contour for AIFD calculation.

packing properties and strong results presented by other research [13, 20]. The separable 2D DCT was used in this work

$$Y = C^T X C, \tag{2}$$

where

$$\begin{aligned} C(n,k) &= \frac{1}{\sqrt{N}}, \quad k = 0,\ 0 \le n \le (N-1), \\ C(n,k) &= \sqrt{\frac{2}{N}} \cos \frac{\pi(2n+1)k}{2N}, \\ &\quad 1 \le k \le (N-1),\ 0 \le n \le (N-1). \end{aligned} \tag{3}$$

The matrix X is the downsampled image matrix and Y is the resulting transformed matrix. The upper left block (6×6) of Y 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.

As shown in Section 3.6, the DCT was not robust to speaker movement. To try 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 head with an elliptical template search. All image pixels that would then form a box parallel to the angled lips are chosen for the X matrix to which the DCT is applied. Figure 16 demonstrates estimation of the rotation-correction factor, and Figure 17 shows a downsampled image matrix based on the estimated angle. The head is slightly tilted but the lips are *normalized* back to horizontal lips. This correction is an attempt to remove one possible affine transformation is rotation. Translation variance has already been minimized, hopefully, by the liptracker and *center of mass* algorithm that should center the downsampled image matrix around the lips. Scale is not currently taken into account although a good possibility would be to scale the downsampled image block by the width of the speaker's face. The approximate measurement can be estimated slightly above a speaker's mouth where the width of the face should not change due to jaw movement. This should roughly scale each person's lips and improve robustness to back and forth movement and also improve speaker-independent performance.

Although the rotation correction improves the DCT's robustness to speaker movement, stationary performance actually drops due to the dependence on estimating the lip corners. Sensitivity to lip segmentation has been introduced as in the contour methods, and performance drops to about their level. In order to improve this, a smoothed rc-DCT that *smooths* the angle estimate between frames to minimize improper estimates or erratic changes was tested. The DCT, rc-DCT, and the smoothed rc-DCT results are compared against the image processing and template-based contour methods in Section 3.6.

3.5. Deformable template method

Asomewhat different approach is taken in this deformable template method than in other approaches based on parametric curves (B-Spline, Bézier, elliptical), snakes, or active contour models. The main goal 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

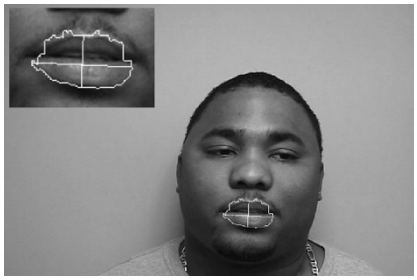

Figure 14: Final mouth contour for AIFD calculation.

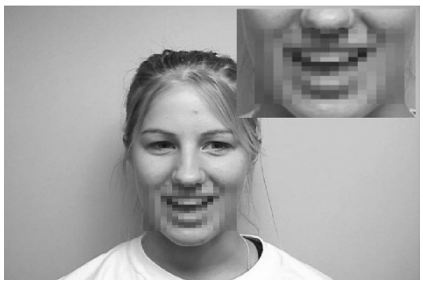

Figure 15: Downsampling for DCT application.

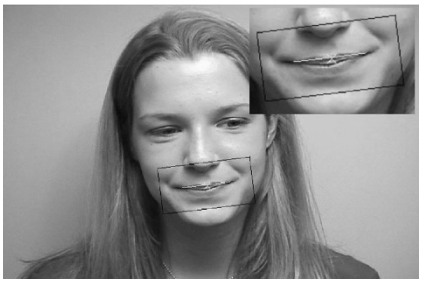

Figure 16: Rotation correction before applying DCT.

Figure 17: Rotation-corrected, downsampled image region for DCT application.

sensitive to improper lip segmentation. The deformable template principle is the same; lip contours are guided by minimizing a cost function $C(R)$ over the area enclosed by the curves, forming region R [23]

$$C(\mathbf{R}) = \sum_{(x,y)\in R} \log \frac{P(\mathbf{o}(x,y)|\omega_{\text{face}})}{P(\mathbf{o}(x,y)|\omega_{\text{lips}})}. \tag{4}$$

The method presented here may be implemented with B-Spline curves. For the tables of results, this technique will be referred to as B-Spline template (BST). Currently, we use splines with a uniform knot vector, and the Bernstein basis also known as Bézier curves [24]. Eight control vertices (CVs) are used, four for the top lip and four for the bottom one. (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 18. A Bézier curve generated by CVs is shown in Figure 19. CVs are also known as B_i points, where $B_i = (x_i, y_i)$, that control the parametric curves by the following formulas:

$$\begin{aligned} P(t) &= \sum_{i=0}^{n} B_i J_{n,i}(t), \quad 0 \le t \le 1, \\ J_{n,i} &= \binom{n}{i} t^i (1-t)^{n-i}. \end{aligned} \tag{5}$$

There are two sets of 4 B_i points: B^t_{0-3} for the top curve and B^b_{0-3} 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. The corners and angle of the lips are estimated as discussed in the previous section. The control points B_0 and B_3 for the top and bottom are 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 based on changing the width, height, location, etc. of the template are generated. 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. First, the location is searched by moving all the CVs throughout the larger box to confirm that the lips are accurately centered within the template. Next, the CVs are changed to minimize $C(R)$ in respect to the width of the template. After the width, the height of the top lip and bottom lip are compared against the cost function. Finally, the shape can be changed by searching for the spacing of B_1 and B_2 based on $C(R)$, for both the top and bottom independently. The horizontal spacing from the center of the lips between B_1 and B_2 was assumed to be symmetric, as no useful information is seemingly gained from asymmetric lip positions. Once B^t_{0-3} and B^b_{0-3} that minimize $C(R)$ are determined, 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. Currently, scaling is not implemented, but could be done so in a similar manner as discussed in the previous subsection. The advantages of this technique are simplicity, slightly less sensitivity to lip segmentation,

Figure 18: Control vertices for estimated lip curves.

Figure 19: Estimated lip contour using control vertices for Bézier curves.

Figure 20: Well estimated lip contour from Bézier curves.

Figure 21: Poorly estimated lip contour from Bézier curves.

and inherent affine invariance. Another possibility is that CVs are automatically generated that could be passed to control curves for a facial animation or lip synchronization task based on a recorded speaker. Example of curves that minimized $C(R)$ are shown in Figures 20 and 21. The former represents a good match of the cost minimization to the actual lip curve. The latter is slightly off although moving information may still convey useful information. During experimentation, more exact contour estimation did not always seem to translate directly into improved recognition performance.

3.6. Stationary and moving results

Each of the visual feature methods were used on the test setup described in Section 3.1. The grouping is completely speaker-independent set, training on one group and testing on a completely separate group. Task 1 (see Table 1) is used for stationary testing and training. Task 2 is used for moving-speaker testing and training. The results are presented in Table 5 over a stationary set, a moving set with models trained on the stationary set, and a moving set with models trained on a moving set. It is interesting to note that training on the stationary set produces better performance for the moving cases. It is likely that features are corrupted more during the speaker movement, thus producing poor training data. Although results may not strictly be compared to other results, the range of these is on the order of results from other medium-sized, connected/continuous, speaker-independent speechreading tasks [13, 20]. 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 [13] 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 Section 3.4 both improved results more on the moving case and nearly regained stationary performance. The BST features performed on par with the AIFDs. Another interesting note, though, is that they actually earned the best performance on the moving group, 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 speechreading. Unfortunately, they are still quite sensitive to the lip segmentation scheme through the cost function, as shown in Section 4. One possibility that should significantly improve this dependence is to increase the number of mixtures in the Gaussian density estimation for the lips to include tongue, teeth, and shadows. This should reduce *patchiness* in lip pixel classification, thus improving the accuracy of minimizing the cost function.

Table 5: Comparison of performance (word accuracy) on stationary and moving speakers, digit strings.

Features	Stationary	Moving (trn. stat.)	Moving (trn. mov.)
AIFD	22.40%	21.89%	14.79%
DCT	27.71%	19.62%	17.70%
rc-DCT	22.57%	21.53%	21.05%
Smoothed rc-DCT	27.43%	23.92%	21.53%
BST	22.86%	24.46%	21.20%

4. SPEAKER-INDEPENDENT RESULTS OVER ALL SPEAKERS

In this section, we include baseline results 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, Mel-Frequency Cepstral Coefficients (MFCC)-based HMM recognizer implemented in HTK using 8-state models and only one mixture per state, we obtained 92.25% correct recognition with a word accuracy of 87.25% (slightly lower due to some insertions between actual digits). With some tuning and the addition of more mixtures, recognition near 100% should be attainable. Here, we also include visual speechreading results over 36 speakers using the same training and testing groups as for the audio. Results are included for DCT, rc-DCT, smoothed rc-DCT, and BST features as in the previous test. AIFDs are not currently 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 on the moving task. Again the rc-DCT performs better on moving, but loses stationary performance. The smoothed rc-DCT performs the best here on moving speakers but does not 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.

Table 6: Baseline Speechreading results (word accuracy) over all speakers.

Features	Stationary	Moving
DCT	29.00%	21.12%
rc-DCT	25.95%	22.39%
Smoothed rc-DCT	26.47%	24.73%
BST	23.85%	21.48%

5. 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 faciliate 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. An example application could be in an automobile where a variety of noises and levels are encountered. A camera could be easily mounted to keep the driver in the field of view for an audio-visual speech recognizer. Dynamic estimations classifying the background noise could be used to improve recognition results. A feature study of image-processing-based contours, image transform, and deformable template methods has also been detailed. It included results on stationary and moving talkers and also attempts to lessen the effect of speaker movement on the various visual feature sets. Contour methods are more severely limited by models used for feature segmentation and tracking. However, simple changes may yield improved results in the template method. Future work may show if these can compare with image transform results. Scale correction will also be included for the various feature sets to determine its effect. Finally, baseline results have been included for two of the database tasks to encourage comparison of techniques among researchers.

There are several areas that are wide open for audio-visual speech recognition research that may be tested on this database. A very important area is visual feature extraction. A variety of speakers and speaker movement has been included for this end. Methods are needed that either significantly strengthen feature tracking under practical conditions or that create new features for speechreading. 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 speechreading 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, multistream 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. Hopefully, the CUAVE database will facilitate more widespread research in these areas.

REFERENCES

[1] D. W. Massaro, *Perceiving Talking Faces: From Speech Perception to a Behavioral Principle*, MIT Press, Cambridge, Mass, USA, 1997.

[2] Q. Summerfield, "Lipreading and audio-visual speech perception," *Philos. Trans. Roy. Soc. London Ser. B*, vol. 335, no. 1273, pp. 71–78, 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, Washington, DC, USA, October 1988.

[4] P. L. Silsbee and A. C. Bovik, "Computer lipreading for improved accuracy in automatic speech recognition," *IEEE Trans. Speech, and Audio Processing*, vol. 4, no. 5, pp. 337–351, 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 Trans. Speech, and Audio Processing*, vol. 7, no. 6, pp. 629–642, 1999.

[6] C. Neti, G. Potamianos, J. Luettin, et al., "Audio-visual speech recognition final workshop 2000 report," Tech. Rep., Center for Language and Speech Processing, Johns Hopkins University, Baltimore, Md, USA, October 2000.

[7] G. Potamianos, C. Neti, G. Iyengar, and E. Helmuth, "Large-vocabulary audio-visual speech recognition by machines and humans," in *Eurospeech*, Aalborg, Denmark, September 2001.

[8] J. R. Movellan, "Visual speech recognition with stochastic networks," in *Advances in Neutral Information Processing Systems*, G. Tesauro, D. Toruetzky, and T. Leen, Eds., vol. 7, M. I. T. Press, Cambridge, Mass, USA, 1995.

[9] I. Matthews, *Features for audio-visual speech recognition*, Ph.D. thesis, School of Information Systems, University of East Anglia, Norwich, UK, 1998.

[10] C. C. Chibelushi, S. Gandon, J. S. D. Mason, F. Deravi, and R. D. Johnston, "Design issues for a digital audio-visual integrated database," in *IEE Colloquium on Integrated Audio-Visual Processing for Recognition, Synthesis and Communication*, pp. 7/1–7/7, Savoy Place, London, November 1996.

[11] K. Messer, J. Matas, J. Kittler, J. Luettin, and G. Maitre, "XM2VTSDB: the extended M2VTS database," in *2nd International Conference on Audio and Video-Based Biometric Person Authentication*, pp. 72–77, Washington, DC, USA, March 1999.

[12] G. Potamianos, E. Cosatto, H. P. Graf, and D. B. Roe, "Speaker independent audio-visual database for bimodal ASR," in *Proceedings of European Tutorial and Research Workshop on Audio-Visual Speech Processing*, pp. 65–68, Rhodes, Greece, 1997.

[13] G. Potamianos, H. P. Graf, and E. Cosatto, "An image transform approach for HMM based automatic lipreading," in *Proc. IEEE International Conference on Image Processing*, vol. 111, pp. 173–177, Chicago, Ill, USA, 1998.

[14] S. Young, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, *The HTK Book (Version 2.1)*, Entropic Cambridge Research Laboratory, Cambridge, UK, 1997.

[15] 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 *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, Salt Lake City, Utah, USA, May 2001.

[16] E. K. Patterson, S. Gurbuz, Z. Tufekci, and J. N. Gowdy, "Noise-based audio-visual fusion for robust speech recognition," in *International Conference on Auditory-Visual Speech Processing*, Scheelsminde, Denmark, September 2001.

[17] Z. Tufekci and J. N. Gowdy, "Mel-scaled discrete wavelet coefficients for speech recognition," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, Istanbul, Turkey, 2000.

[18] G. I. Chiou and J. Hwang, "Lipreading from color video," *IEEE Trans. Image Processing*, vol. 6, no. 8, pp. 1192–1195, 1997.

[19] A. P. Varga, H. J. M. Steenekan, M. Tomlinson, and D. Jones, "The NOISEX-92 study on the effect of additive noise on automatic speech recognition," Tech. Rep., DRA Speech Research Unit, Malvern, UK, 1992.

[20] G. Potamianos and C. Neti, "Improved ROI and within frame discriminant features for lipreading," in *Proc. IEEE International Conference on Image Processing*, Thessaloniki, Greece, 2001.

[21] S. Theodoridis and K. Koutroumbas, *Pattern Recognition*, Academic Press, New York, NY, USA, 1999.

[22] K. Arbter, W. E. Snyder, H. Burkhardt, and G. Hirzinger, "Application of affine-invariant Fourier descriptors to recognition of 3-D objects," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 12, no. 7, pp. 640–647, 1990.

[23] T. Chen, "Audiovisual speech processing: Lip reading and lip synchronization," *IEEE Signal Processing Magazine*, vol. 18, no. 1, 2001.

[24] D. Rogers, *An Introduction to NURBS*, Morgan Kaufmann Publishers, San Francisco, Calif, USA, 2001.

Eric K. Patterson was born in Beaufort, South Carolina, USA in 1972. He received the B.S. degree in computer engineering from Clemson University in 1995. He also received the Ph.D. degree in computer engineering under the supervision of Dr. John Gowdy from Clemson University in 2002. During his graduate study, he conducted research in audio-visual speech recognition, chaos theory, audio background noise, and educational tools for signal processing distance learning. He is currently an Assistant Professor of computer science at The University of North Carolina at Wilmington. His interests include multimodal speech recognition, interactive multimedia systems, computer animation, and other digital art.

Sabri Gurbuz was born in Andirin, K.Maras, Turkey, on May 5, 1969. He received the B.S. degree in electronic and telecommunication engineering from Istanbul Technical University, Turkey in 1989. He also received his M.S. degree in electrical and computer engineering from Clemson University, Clemson, South Carolina in 1997 with the Image Analysis and Acquisition Laboratory under the supervision of Professor Robert J. Schalkoff on image segmentation using trainable fuzzy classifiers. In June 2002, he received his Ph.D. degree entitled *Robust and Efficient Techniques for Audio-Visual Speech Recognition*, in electrical and computer engineering in the same university with the Digital Speech and Audio Processing Laboratory, under the supervision of Professor John N. Gowdy. In the Fall of 2001, he has been an intern with the Digital Television Systems Research Group in Philips Research, Briarcliff, New York, USA and filed a patent for estimating objective video sharpness quality. His research interests span the areas of stereo computer vision, image and signal processing, multimedia signal processing, automatic speech recognition, audio-visual speech, and speaker recognition. He has published over 15 papers in these areas and has one patent filed and one in the process. He is a member of IEEE.

Zekeriya Tufekci received his B.S. degree in electrical engineering from Hacettepe University, Ankara, Turkey, in 1987 and the M.S. and Ph.D. degrees in electrical engineering from Clemson University, Clemson, South Carolina, USA in 1996 and 2001, respectively. He joined the Department of Electrical and Electronics Engineering, Izmir Institute of Technology, where he is currently Assistant Professor. His research interests include robust speech recognition, audio-visual speech recognition, speaker recognition, wavelet, neural networks, and fuzzy logic.

John N. Gowdy is Professor and Chair of Electrical and Computer Engineering at Clemson University. He has performed research in the speech area for over 30 years. He has worked in the areas of speech recognition, speaker recognition, speech coding, and speech synthesis. He teaches courses in digital signal processing, speech signal processing, and microcomputer interfacing at Clemson University.