CUAVE: A NEW AUDIO-VISUAL DATABASE FOR MULTIMODAL HUMAN-COMPUTER INTERFACE RESEARCH

E.K. Patterson, S. Gurbuz, Z. Tufekci, and J.N. Gowdy

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

ABSTRACT

Multimodal signal processing has become an important topic of research for overcoming certain problems of audio-only speech processing. Audio-visual speech recognition is one area with great potential. Difficulties due to background noise and multiple speakers are significantly reduced by the additional information provided by extra visual features. Despite a few efforts to create databases in this area, none has emerged as a standard for comparison for several possible reasons. This paper seeks to introduce a new audio-visual database that is flexible and fairly comprehensive, yet easily available to researchers on one DVD.

The CUAVE database is a speaker-independent corpus of over 7,000 utterances of both connected and isolated digits. It is designed to meet several goals that are discussed in this paper. The most notable are availability of the database, flexibility for use of the audio-visual data, and realistic considerations in the recordings (such as speaker movement). Another important focus of the database is the inclusion of pairs of simultaneous speakers, the first documented database of this kind. The overall goal of this project is to facilitate more widespread audio-visual research through an easily available database. For information on obtaining CUAVE, please visit our webpage (http://ece.clemson.edu/speech).

1. INTRODUCTION

Over the past decade, multimodal signal processing has been an increasing area of interest for researchers. The power of computing has increased to the level where separate modalities such as audio and video can be used in a complementary method to improve desired results. Audio-visual speech processing has shown great potential, particularly in areas such as speech recognition and speaker authentication. For speech recognition the addition of information from lipreading or other features helps make up for information lost due to corrupting influences. Because of this, audio-visual speech recognition can outperform audio-only recognizers in environments where there are background noises or other speakers. Researchers have demonstrated the relationship between lipreading and human understanding [1, 2] as well as produced performance increases with multimodal systems [3, 4, 5, 6].

Because of the relative novelty of the research area as well as difficulties associated with the high volumes of data necessary for simultaneous video and audio, the creation of audio-visual databases has been limited. Because of this, most researchers in the area have had to record their own data. In order to allow for better comparison and for researchers to enter the area of study more easily, available databases are necessary that meet certain criteria. This paper presents the Clemson University Audio Visual Experiments (CUAVE) database that has been designed to help meet some of these criteria. It is a speaker independent database of isolated and connected digits of high quality video and audio of a representative group of speakers and is easily available on one DVD. Section 2 of this paper discusses a brief survey of audio-visual databases and presents motivation for a new database. Next, Section 3 presents design goals of the CUAVE database and specifics of the data collection and corpus format. Section 4 presents some of our work in audio-visual speech recognition as well as some preliminary results and discussion about the database. Finally, Section 5 closes with some observations and suggestions about possible areas of research in audio-visual speech processing.

2. AUDIO-VISUAL DATABASES AND RESEARCH CRITERIA

There has been some effort 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 100x75 resolution [7]. AVLetters includes the English alphabet recorded three times by ten talkers in 25 fps grayscale [8]. DAVID is a larger database including various recordings of thirty-one speakers over five sessions including digts, alphabets, vowel-consonant-vowel syllable utterances, and some video conference commands distributed on multiple SVHS tapes [9]. 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 [10]. The M2VTS database is available on HI-8 tapes, and XM2VTSDB is available on twenty 1-hour MiniDV cassettes. There have also been some proprietary efforts by AT&T and by IBM [11, 6]. The AT&T database was to include connected digits and isolated consonant-vowel-consonant words. There were also plans for large-vocabulary continuous sentences. The database was never finished nor released, however. Front-end features from the IBM database are becoming available, but there are currently no plans to release the videos [11].

Large-vocabulary, continuous speech recognition is an ultimate goal of research. Because audio-visual speech processing is

Fig. 1. Sample speakers from the database.

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 and testing, CUAVE was produced as a speaker independent database consisting of isolated and connected digits in different situations. It is flexible, representative, and easily available. Section 3 discusses the design goals, collection methods, and format of the corpus.

3. 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 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, manually labeled to improve training and testing methods. 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 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, a rarity in 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. See Figure 1 for a sample of images from some of the speakers.

Fig. 2. Sample speaker pair from the database.

Each speaker was framed including the shoulders and head and recorded speaking digits in several different styles. Initially, 50 isolated digits are spoken while standing naturally still (not forced). This is for general training and testing without very difficult conditions. Video is full color with no aids given for face/lip segmentation. Secondly, each individual speaker was asked to move side-to-side, back-and-forth, or tilt the head while speaking 30 isolated digits. This is an important part of the database to allow for testing of affine-invariant visual features. In addition to this isolated section, there is also a connected-digit section with movement as well. So far, much research has been limited to low resolution, 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 20 isolated digits with both profile views, and then 60 connected digits facing the camera again. The final 60 connected digits include telephone-number-like sequences, 30 while standing still and 30 while moving.

The second major section of the database includes 20 pairs of speakers. Its 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, particulary with audio information only. Video features correlated with speech features should allow solutions to this problem. In addition, techniques can be tested here that will help with the difficult problem of speech babble. (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. Person A speaks a connected-digit sequence, followed by speaker B and vice-versa a second time. For the third sequence, both speaker A and B overlap each other while speaking each persons 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 720x480 with the NTSC standard of 29.97 fps, using a 1-megapixel-CCD, MiniDV camera. 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 back-

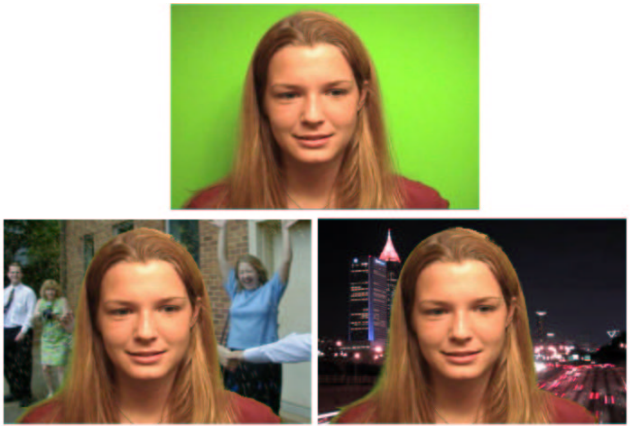

Fig. 3. Example of varied video backgrounds for speakers.

ground 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 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 brought into the computer using the IEEE 1394 interface. 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 [12]. 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 [13]. 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 some work in preparation of collecting the database and preliminary work on the actual corpus.

4. PRELIMINARY WORK RELATED TO CUAVE

Before recording the database, we conducted research 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 [14] [15]. The audio-visual recognizer used was a late-integration system with separate audio and visual HMM-based speech recognizers. The audio recognizer used sixteen Mel-frequency discrete wavelet coefficients extracted from speech sampled at 16 kHz [16]. The video recognizer used red-green plane division and thresholding before edge-detection of the lip boundaries [17]. Geometric features based on the oral cavity and affine-invariant Fourier descriptors were passed to the visual recognizer [14]. (We believe that the new database will provide a good test bed for further research on affine-invariant methods.) Speech noise was added from the NOISEX database [18] 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 the database design goals mentioned in Section 3. See Table 4 for a summary of these results. (Here, the field-of-view was cropped to lips-only to ease preliminary work).

	audio %	video %	av %
clean	100	87	100
18 db	96.5	87	99.3
12 db	80.9	87	96.9
6 db	52.6	87	92.9
0 db	29.1	87	89.8
-6 db	21.0	87	88.4

Table 1. Recognition Rates Averaged Over Noises for a Single Speaker

We have begun initial work using the completed CUAVE database and include some baseline results here. Our first experiments have been on the audio of the 36-speaker isolated digit task. The 36 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, we obtained 92% correct recognition with an accuracy of 87% (slightly lower due mostly to insertions between actual spoken digits). With some tuning and the addition of more mixtures, recognition near 100% should be attainable. The next task areas for baseline results consist of visual-feature extraction, lipreading, and data fusion. We plan to begin testing with the affine-invariant Fourier descriptor visual features used in our earlier work [14] and possibly expand for comparison to other techniques such as principal components analysis, snakes, or B-splines. An arbitrary thresholding of red/green division used to initially segment the lips in our earlier work is not sufficient for a database of many speakers. For this reason, we have manually segmented the face and lip regions for all 36 speakers, such as demonstrated in Figure 4, to obtain Gaussian approximations of the face and lip color distributions. These have shown promising initial results for face/lip segmentation over the whole database. We are performing MPEG-2 decoding under Linux along with an implementation of our visual feature extraction routines. Based on developmental results, we believe that our final algorithms will execute in near realtime, including the decoding. In upcoming releases of the database, we plan to make the color distribution information available, as well as source code for MPEG-2-based tools to facilitate use of the included videos.

5. CONCLUSIONS AND RESEARCH SUGGESTIONS

This paper presents a flexible, speaker-indpendent 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

Fig. 4. Manually segmented face and lips for distribution training.

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. Hopefully, the database may be used to progress audio-visual speech recognition methods. As it is a representative, medium-sized digits task, it may also be used for testing in other areas as well, such as speaker recognition, lip synchronization, etc.

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. Also, 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. 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.

6. 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, 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, 1999.

[6] C. Neti, G. Potamianos, J. Luettin, 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] J.R. Movellan, "Visual speech recognition with stochastic networks," in *Advances in Neutral Information Processing Systems*, Toruetzky D. Tesauro, G. and T. Leen, Eds., vol. 7. MIT Press, Cambridge, 1995.

[8] I. Matthews, *Features for Audio-Visual Speech Recognition*, Ph.D. thesis, School of Information Systems, University of East Anglia, UK, 1998.

[9] 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*, Savoy Place, London, Nov. 1996, number 1996/213, pp. 7/1–7/7.

[10] K. Messer, J. Matas, J. Kittler, J. Luettin, and G. Maitre, "Xm2vtsdb: The extended m2vts database," in *Second International Conference on Audio and Video-Based Biometric Person Authentication, Washington, D.C.*, 1999.

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

[12] G. Potamianos, H.P. Graf, and E. Cosatto, "An image transform approach for hmm based automatic lipreading," in *Proceedings of the ICIP, Chicago*, 1998.

[13] S. Young, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, *HTK Book*, Cambridge University Press, 1997.

[14] 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.

[15] 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, Denmark*, 2001.

[16] Z. Tufekci and J.N. Gowdy, "Mel-scaled discrete wavelet coefficients for speech recognition," in *Proceedings of ICASSP*, 2000.

[17] G. I. Chiou and J. Hwang, "Lipreading from color video," *IEEE Transactions on Image Processing*, vol. 6, no. 8, 1997.

[18] 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, 1992.