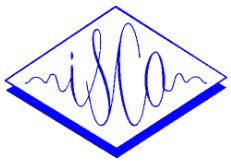

ESCA Workshop on
Audio-Visual Speech Processing
Rhodes, Greece
September 26-27, 2997

SPEAKER INDEPENDENT AUDIO-VISUAL DATABASE FOR BIMODAL ASR

Gerasimos Potamianos, Eric Cosatto, Hans Peter Graf*,* and *David B. Roe*

AT&T Labs–Research, 180 Park Ave, Florham Park, NJ 07932-0971, U.S.A.
*AT&T Labs–Research, 100 Schulz Drive, Red Bank, NJ 07701-7033, U.S.A.
email: {makis,eric,hpg,roe}@research.att.com

ABSTRACT

This paper describes the audio-visual database collected at AT&T Labs–Research for the study of bimodal speech recognition. To date, this database consists of two multiple speaker parts, namely isolated confusable words and connected letters, thus allowing the study of some popular and relatively simple speaker independent audio-visual recognition tasks. In addition, a single speaker connected digits database is collected to facilitate speedy development and testing of various algorithms. Intentionally, no lip markings are used on the subjects during data collection. Development of robust and speaker independent algorithms for mouth location and lip contour extraction is thus necessary in order to obtain informative features about visual speech (visual front end). We describe our approach to this problem, and we report our automatic speech-reading and audio-visual speech recognition results on the single speaker connected digits task.

1. INTRODUCTION

Recently, there has been an increasing interest in methods for enhancing *automatic speech recognition* (ASR) by using *visual* information derived from the speaker's *lips* or *oral cavity*, i.e., *lip-reading*, or *speech-reading* (Petajan, 1984; Stork and Hennecke, 1996). In many cases, and especially when the audio channel is characterized by low *signal to noise ratio* (SNR), ASR systems incorporating both visual and acoustic information have been reported to perform significantly better than their audio only counterparts (Bregler et al., 1993; Adjoudani and Benoit, 1996; Su and Silsbee, 1996). However, in most cases, such results are limited to a single speaker and an isolated or connected word, small vocabulary recognition task. In addition, and in order to ensure a robust *visual front end*, many researchers use intrusive techniques (Petajan, 1984; Adjoudani and Benoit, 1996), such as lip marking.

We believe that successful research in visually assisted speech recognition will follow the *large vocabulary continuous speech recognition* (LVCSR) model, namely *database* driven statistical pattern matching (Rabiner and Juang, 1993). Specifically, we advocate: (i) A *hidden Markov model* approach to statistical pattern matching; (ii) a carefully chosen, robust set of visual and acoustic *features*; and (iii) a large, diverse database that can be used both for training statistical models and for evaluating the accuracy of the lip-reading systems.

Recognizing the fact that part of the limitations of current *audio-visual* (AV) systems is due to the lack of appropriate databases, we have concentrated our initial efforts in collecting an AV database, that satisfies most of the items in the AV research community wish list (Chibelushi et al., 1996; Stork and Hennecke, 1996). In this paper we describe this database, and we report our experimental recognition results on a "test" part of it, namely single speaker *connected digits*. In Section 2 we give details of our workstation based AV database collection system, whereas in Section 3 of our AV database. In Section 4 we describe our visual front end, which, given a video sequence of the speaker's face, extracts appropriate visual features that are subsequently used in the HMM based automatic speech-reading and AV ASR systems, discussed in Section 5. In Section 6 we report our recognition results on the single speaker connected digits task. Our conclusions are drawn in Section 7.

2. THE AUDIO-VISUAL DATABASE COLLECTION SYSTEM

Our AV collection system is based on an SGI Indigo2 workstation with suitable peripherals. It captures YUV 4:2:2 video by digitizing incoming RGB video from a high quality 3CCD camera. Up to 30 secs of uncompressed video are captured at 30 *interleaved frames* per second and a 560×480 pixel resolution. Information is thus available at 60 *fields* per second. In addition, four channels of 16 bit 16 kHz audio are captured, using four desktop microphones of varying quality (9, 14, 18 and 24 dB SNR), which feed their amplified signals directly into the four–channel SGI audio board. A single video file, four audio files, and two files containing precise audio and video time synchronization information (in nsecs), corresponding to the speech segment of every utterance (augmented by 0.75 secs of "silence" at each side) are saved.

The entire collection procedure is software driven and subject friendly: The subject sits facing the microphones, camera, and the workstation monitor, and reads text when this appears on the monitor (see Fig. 1). No lip markings are imposed on the subject. All that is required, is that the camera captures the frontal view of the subject's face.

3. THE AT&T AUDIO-VISUAL DATABASE

We are interested in addressing successively more challenging *speaker independent* bimodal recognition tasks, so our database reflects this rationale. It will eventually consist of four parts, with the first two parts having already been collected (see Table 1): Part 1 consists of a small vocabulary of highly confusable, mostly monosyllabic, *isolated, consonant–vowel*–consonant (C–V–C) words. Part 2 consists of *connected letters*. Part 3 will consist of phonetically balanced North American Business (NAB) sentences (a traditional LVCSR task), whereas Part 4 of spontaneously spoken sentences (story-telling type). Similar tasks to Parts 1 and 2 have been addressed in the literature (Bregler et al., 1993; Su and Silsbee, 1996; Adjoudani and Benoit, 1996), with the limitations discussed

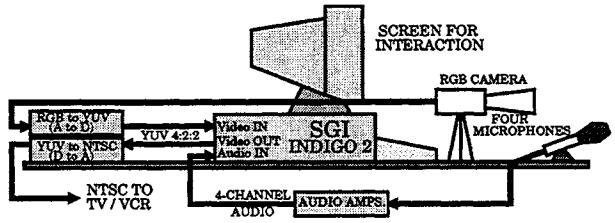

Figure 1: Diagram of our AV database collection system.

Part	Subj	Task	Voc	Words
T	1	connected digits	11	300 × 5
1	50	isolated words	123	1250 × 1
2	50	connected letters	26	1250 × 4
3	N/A	continuous speech	N/A	N/A
4	N/A	spontaneous speech	N/A	N/A

Table 1: Current and future tasks (marked as N/A) of our AV database. Number of subjects (Subj), task, vocabulary size (Voc), and collected words (utterances × number of words per utterance) are depicted.

in our introduction though.

To facilitate speedy development and testing of basic visual feature extraction and integration algorithms for AV speech recognition, we decided to initially address the simple "test" task of single speaker connected digits recognition. Part T ("test") of our database thus consists of 300 digit five-tuples spoken by a single subject. Only a 192 × 160 pixel window around the subject's mouth is captured (see also Fig. 2(a)).

Part 1 of the database consists of 1250 isolated word utterances, spoken by 50 subjects. Each subject records one of five sets consisting of 25 distinct words (see Table 2). These sets are almost disjoint (the total vocabulary size is 123), and consist of mostly C–V–C words, designed to contain as many minimally distinct words as possible. Particular attention was paid to consonants with different places of *articulation*, such as *labial*, *alveolar*, and *glottal*. *Phoneme substitutions* were covered by using pairs of minimally distinct words such as "pin"–"kin", or "bathe"–"beige". At every set, the following sets of minimally distinct phoneme pairs are covered at least once in both final and (wherever possible) initial consonant position: (i) *Unvoiced plosives*: All pairs including /p/, /t/, /k/. (ii) *Voiced plosives*: All pairs including /b/, /d/, /g/. (iii) *Unvoiced fricatives*: All pairs including /f/, /th/, /s/, /sh/. (iv) *Voiced fricatives*: All pairs including /v/, /dh/, /z/, /zh/. (v) *Nasals*: At least two pairs containing /m/, /n/, and /ng/ (/zh/ and /ng/ do not occur in the initial position). In addition, consonant *deletions* were covered in both initial and final positions for all of the above five classes with words such as "near"–"ear". A few other confusable consonants were included in the database, especially the *semivowels* /r/, /l/, and /w/, and the *affricates* /jh/ and /ch/. Vowel *confusions* were covered by designing C–V–C words with only the vowel varying between words: (i) *Front vowels*: V = /iy/, /ih/, /ae/, /eh/. (ii) *Mid vowels*: V = /aa/, /er/, /ah/, /ao/. (iii) *Back vowels*: V = /uw/, /uh/, /ow/. (iv) *Dipthongs*: V = /aw/, /ay/, /oy/, /ey/.

Part 2 of our database consists of 1250 connected letter four-tuple utterances, spoken by the same 50 subjects as in Part 1 (25 utterances per subject). The letters are randomly generated to satisfy a uniform *bi-gram* distribution (equal inter–letter context).

SET 1	SET 2	SET 3	SET 4	SET 5
gnat	pod	den	deer	rogue
sign	bawl	Ruth	ear	bees
side	thin	phone	deal	we
veal	sought	hang	near	roe
haw	sit	puck	cooed	bays
beg	tin	known	cued	lithe
deb	Gaul	putt	cod	leaf
thighs	in	tuck	dun	burrs
leap	soot	cut	van	then
fie	burr	sown	code	live
hawk	Tim	hay	beer	rose
rape	pin	rue	could	math
sigh	cod	woof	bays	boys
bed	suit	pup	cawed	zen
nap	cog	shown	beige	leash
map	bull	dive	dung	mash
thigh	kin	dan	bay	mass
shy	king	down	cud	lee
reap	burp	moan	gear	roam
knack	odd	own	mere	lease
rep	Bert	dies	babe	rove
rip	Kim	done	Ann	leach
rap	cob	don	than	alive
zeal	thing	zone	bathe	liege
dead	seat	roof	dumb	buys

Table 2: The five sets of Part 1 of our AV database.

Our subject population is facially and racially diverse. Among the 50 subjects, there are 10 females, whereas, within the male population, 12 subjects have moustaches and 9 beards. Complexions vary significantly, whereas 21 subjects wear glasses (see also Fig. 3). Thirteen subjects are native American English speakers with the rest having light accents. We plan to increment our subject population with more native speakers and female subjects.

In Parts 1 and 2 of the database the full frontal face of the subject is captured in the 560 × 480 pixel window, whereas the background is green. In the future we plan to make the video information more realistic, by considering a variety of camera placements, partially occluded mouth cases, etc. The current database size is approximately 75 GB. We are currently investigating suitable video compression schemes for it, such as MPEG2.

4. THE VISUAL FRONT END

Since no lip markings are used on the subjects during data collection, a robust and speaker independent visual front end needs to be designed. Our visual front end system consists of two stages (see Fig. 2): First the subject's mouth is located and the *outer lip contour* is extracted. Then, either appropriate geometric features of the lip contour (*geometric feature* approach), or appropriate mouth image transform coefficients (*image transform* approach) are used as visual features.

4.1. Mouth location and lip contour extraction

We use a combination of *shape* and *texture analysis*, as well as *color segmentation* for first finding the location of the mouth and then its precise shape (Graf et al., 1996; Graf et al., 1997). The shape and texture analysis consists of bandpass filtering, morphological operations, and adaptive thresholding for identifying regions of interest (ROI), i.e., regions where facial features may be present. The color analysis divides the color space into a number of color clusters and identifies colors dominating in certain areas. After segmenting the image into the different colors, the resulting segments are filtered for their shapes and grouped into larger areas with a clustering algorithm. This produces a number of ROI, where facial features, or parts of the mouth, may be located. The candidate ROI

(a) (b) (c) (d)

Figure 2: The visual front end. (a): Original image (from Part T of our database). (b): Outer lip contour C and center of mass $(\overline{x}, \overline{y})$ of $\underline{C}$. (c): Two geometric features (mouth height h and width w). (d): Bounding box of the 32×16 pixel image around $(\overline{x}, \overline{y})$.

obtained from the two channels of processing are combined and evaluated with an "n–gram" search (Graf et al., 1996). To deal with the diversity of subjects our system includes several different models of the face and the mouth area. New faces are compared to these models and the most representative one is chosen for the analysis (Graf et al., 1997). A training step is required for the adjustment of several system parameters such as bandpass filter parameters, adaptive thresholds, etc.

We tested our algorithms on the database. In Part T we have been able to reliably locate the mouth in all 26413 frames. In approximately 1% of these frames the lip contour erroneously consists of the upper lip contour only. The system was trained on 100 frames of the same subject. We have then considered sample sequences of all our 50 database subjects (see Fig. 3). When trained on a particular subject, the mouth location is correctly found in more than 98% of the frames of the same subject. When trained on a random set of 10 subjects, the system correctly handles 87% of the 40 remaining subjects (Graf et al., 1997). These results were obtained with the algorithms operating on individual frames and no inter–frame correlation information being used.

4.2. Visual feature extraction

In the geometric feature approach we first appropriately normalize and rotate the outer lip contours, in order to compensate for subject and camera–subject relative location variations. Geometric features are then extracted from the transformed contours $C = \{(x, y)\}$. Four features, namely mouth *height* (h), *width* (w), and the two first order *absolute central moments* ($\mathbf{m}_{10}$, $\mathbf{m}_{01}$) of the contour interior $\underline{C}$, as well as their first and second *derivatives*, were deemed the most informative for automatic speech-reading on Part T of our database. Let

$$f(x,y) = \begin{cases} 1, & \text{if } (x,y) \in C \cup \underline{C}\ ; \\ 0, & \text{otherwise}\ . \end{cases}$$

Then the four features of interest are defined as

$$\mathbf{h} = \max_x \Sigma_y f(x,y)\ , \quad \mathbf{w} = \max_y \Sigma_x f(x,y)\ ,$$

and $\mathbf{m}_{10}$, $\mathbf{m}_{01}$, where

$$\mathbf{m}_{pq} = \Sigma_x \Sigma_y |x - \overline{x}|^p |y - \overline{y}|^q f(x,y)\ . \qquad (1)$$

In (1) $(\overline{x}, \overline{y})$ denotes the *center of mass* of $\underline{C}$,

$$\overline{x} = \frac{\mu_{10}}{\mu_{00}}\ , \ \overline{y} = \frac{\mu_{01}}{\mu_{00}}\ , \text{ where } \mu_{pq} = \Sigma_x \Sigma_y x^p y^q f(x,y)\ . \quad (2)$$

The geometric feature approach, as discussed above, fails to consider important oral cavity information. In addition, the outer lip contours are sometimes inaccurate. To circumvent these shortcomings, we have employed the following image transform approach: Based on

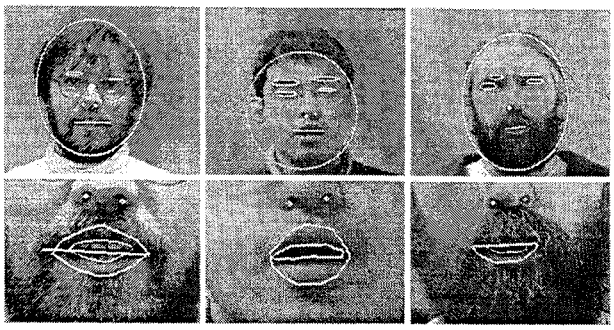

Figure 3: Examples of identified facial features (upper row) and outer and inner lip contours (lower row) for three subjects (from Parts 1 and 2 of our database).

the outer lip contour information, a 32×16 pixel histogram equalized monochrome image is obtained, centered around $(\overline{x}, \overline{y})$ (see (1) and Fig. 2(d)). This image contains most of the oral cavity information. It is then further downsampled to 16×16 pixels, and transformed by means of an appropriate discrete image transform. A number of such transforms have been considered. Among them, the *discrete wavelet transform* (DWT), implemented as in Press et al., 1993, was shown to work the best for automatic speech-reading on part T of our database. Fifteen wavelet coefficients located on the upper left triangular sub-lattice of the image, as well as their first and second derivatives over time, are used as visual features.

5. HMM BASED AUTOMATIC SPEECH-READING AND AUDIO-VISUAL SPEECH RECOGNITION

We employ a simple, continuous HMM based approach for all three, audio only, visual only, and audio-visual (AV), speech recognition tasks (see Fig. 4). The corresponding HMM models of interest have the same architecture, but their observed features differ: In the audio only case we use the traditional 39–dimensional (39–D) *mel–frequency cepstral* coefficient based audio front end (Rabiner and Juang, 1993); in the visual only case we use, either the 12–D geometric, or the 45–D DWT based visual feature vector, whereas in the AV case we use the concatenation of the audio and visual feature vectors. All observation probabilities $Pr[\mathbf{O}|s]$ are modeled as multi–dimensional *Gaussian mixtures*, i.e.,

$$Pr[\mathbf{O}|s] = \sum_{m=1}^{M} c_{sm}\, \mathcal{N}(\mathbf{O}; \mu_{sm}, \Sigma_{sm})\ ,$$

where $\mathbf{O}$ is the observation feature vector (e.g., $\mathbf{O_A}$, $\mathbf{O_V}$, $\mathbf{O_{AV}} = [\mathbf{O_A}, \mathbf{O_V}]$), s is the HMM state, μ_{sm}, Σ_{sm} are the Gaussian mean vectors and diagonal covariance matrices, respectively, and c_{sm} are the mixture weights (Rabiner and Juang, 1993). Notice that our simple AV HMM models follow the *early integration* paradigm in AV fusion (Su and Silsbee, 1996; Adjoudani and Benoit, 1996), with no weighting of the two information channels though.

This approach allows easy training of all models by *bootstrapping* audio only HMM models. However, it requires alignment of the visual features to the audio features, since the latter are available at 100 Hz, whereas the former at 30, or 60 Hz. We employ *linear interpolation* (or *extrapolation*) over time to acquire the "audio–aligned" visual features, by using the audio and visual timing information (see Fig. 4).

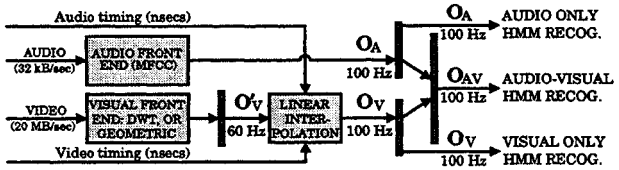

Figure 4: HMM based audio only (A), visual only (V), and audio-visual (AV) automatic speech recognition.

Training	Audio	Visual (a)	Visual (b)
Viterbi	99.8(99.2)	62.3(10.8)	83.5(38.3)
GPD	100.(100.)	67.7(20.8)	93.2(70.8)

Table 3: Test set word (string, in parentheses) accuracies (%) on Part T of our database, achieved by the audio only and two visual only recognizers, (a): based on geometric features; (b): based on the DWT.

The bootstrapping training procedure is essentially the *Viterbi* training algorithm, where the initial model parameters are estimated based not on the training data uniform segmentation, but rather on the segmentation obtained by a well trained original model based on its corresponding front end. In our single speaker connected digits recognition experiments we used bootstrapping to obtain single speaker audio only HMM models, based on well trained speaker independent such models. By bootstrapping the resulting audio only HMM models, we obtained visual only and AV HMM models. All HMM models can be further improved by using *discriminative* training, implemented by means of the *gradient probabilistic descent* (GPD) algorithm (Rabiner and Juang, 1993).

6. RECOGNITION RESULTS

We now report our recognition experiments on Part T of the database (single speaker connected digits). The 300 database digit five-tuples (strings) are partitioned into a training set of 180 digit strings and a test set of 120 digit strings (unseen during HMM training). We use 11 whole word, 8–, or 10–state, left–to–right, continuous HMM models with two mixtures per state, and a single state 16–mixture silence model, for all our audio, visual, and AV HMM models. The test set recognition results (with no grammar) of the audio only and two video only HMM models, based on our two visual front end approaches, are depicted in Table 3. Clearly, the DWT based visual front end is superior to the geometric feature based approach in this case. In addition, GPD trained models perform better than Viterbi trained ones. Notice that our best visual only recognition result of 93.2% compares favorably to connected digits recognition results reported elsewhere (Brooke, 1996).

In Fig. 5 we depict AV recognition results using the AV HMM model discussed in Section 5. The audio signal is corrupted by additive noise of various SNR intensities that consists of spoken digits in the background (by the same speaker). The bimodal recognizer robustness to noise is impressive.

7. CONCLUSIONS

We presented our audio-visual database, aimed towards robust and speaker independent bimodal speech recognition. Currently it consists of 50 subjects uttering isolated confusable words and connected letters. We have additionally collected single speaker connected digits, as a means of testing our audio-visual collection system, visual front end, and recognition algorithms. Our discrete wavelet transform based visual only recognizer achieves a 93.2% recognition accuracy on the digits task, whereas the audio-visual recognizer exhibits significant robustness to noisy audio.

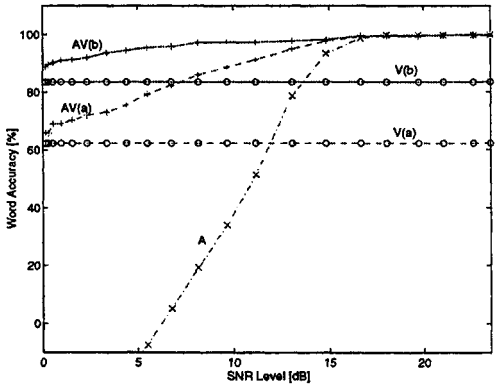

Figure 5: Test set audio only (A), visual only (V), and audio-visual (AV) speech recognition results on Part T of our database, under noisy audio conditions (background digits). Two different visual front ends are used ((a): geometric features; (b): DWT features). All HMM models are Viterbi trained on matching noisy conditions.

8. REFERENCES

A. Adjoudani and C. Benoit (1996), "On the integration of auditory and visual parameters in an HMM-based ASR", in *Stork and Hennecke, 1996*, pp. 461-471.

C. Bregler, H. Hild, S. Manke, and A. Waibel (1993), "Improving connected letter recognition by lipreading", *Proc. Int. Conf. Acoust. Speech Signal Process.*, Minneapolis, Vol. I, pp. 557-560.

N.M. Brooke (1996), "Talking heads and speech recognizers that can see: The computer processing of visual speech signals", in *Stork and Hennecke, 1996*, pp. 351-371.

C.C. Chibelushi, F. Deravi, and J.S.D. Mason (1996), "Survey of audio visual speech databases", *Technical Report*, Electrical and Electronic Engr., Univ. of Wales, Swansea, UK.

H.P. Graf, E. Cosatto, D. Gibbon, M. Kocheisen, and E. Petajan (1996), "Multi-modal system for locating heads and faces", *Proc. Int. Conf. Automatic Face and Gesture Recog.*, Los Alamitos, pp. 88 - 93.

H.P. Graf, E. Cosatto, and G. Potamianos (1997), "Robust recognition of faces and facial features with a multi-modal system", *Proc. Int. Conf. Systems Man Cybern.*, Orlando, in press.

E.D. Petajan (1984), "Automatic lipreading to enhance speech recognition", *Proc. Global Telecom. Conf.*, Atlanta, pp. 265-272.

W.H. Press, B.P. Flannery, S.A. Teukolsky, and W.T. Vetterling (1988), *Numerical Recipes in C. The Art of Scientific Computing* (Cambridge University Press, Cambridge).

L. Rabiner and B.-H. Juang (1993), *Fundamentals of Speech Recognition* (Prentice Hall, Englewood Cliffs).

D.G. Stork and M.E. Hennecke eds. (1996), *Speechreading by Humans and Machines* (Springer, Berlin).

Q. Su and P.L. Silsbee (1996), "Robust audiovisual integration using semicontinuous hidden Markov models", *Proc. Int. Conf. Speech Lang. Process.*, Philadelphia, Vol. 1, pp. 42-45.