



COVER SHEET

Dean, David and Lucey, Patrick and Sridharan, Sridha (2005) Audio-visual speaker identification using the CUAVE database. In Vatikiotis-Bateson, Eric and Burnham, Denis and Fels, Sidney, Eds. Proceedings Auditory-Visual Speech Processing 2005, British Columbia, Canada.

Accessed from <http://eprints.qut.edu.au>

Copyright 2005 the authors

AUDIO-VISUAL SPEAKER IDENTIFICATION USING THE CUAVE DATABASE

David Dean, Patrick Lucey and Sridha Sridharan

Speech, Audio, Image and Video Research Laboratory
Queensland University of Technology
GPO Box 2434, Brisbane 4001, Australia
ddean@ieee.org, {p.lucey, s.sridharan}@qut.edu.au

ABSTRACT

The freely available nature of the CUAVE database allows it to provide a valuable platform to form benchmarks and compare research. This paper shows that the CUAVE database can successfully be used to test speaker identification systems, with performance comparable to existing systems implemented on other databases. Additionally, this research shows that the optimal configuration for decision-fusion of an audio-visual speaker identification system relies heavily on the video modality in all but clean speech conditions.

1. INTRODUCTION

Clemson University’s CUAVE [1] database is a relatively new, and small entrant to the field of audio-visual databases, alongside existing databases such as XM2VTS [2], and M2VTS [3]. However, one great advantage of CUAVE is that it is freely available, so its use in forming benchmarks will become very valuable to the research community in a number of fields. CUAVE was originally designed with audio-visual speech recognition in mind, and it has been used for both regular [4] and simultaneous speaker speech recognition [5]. However other fields of research such as audio-visual mutual information [6] and eye-tracking [7] have also used this database to good effect.

In this paper, we implement an audio-visual speaker identification (AVSPI) system using the CUAVE database. The CUAVE database has not previously been used for speaker recognition experiments, with most existing research [8, 9] using the XM2VTS database for this task. One of the main advantages provided by XM2VTS over CUAVE is the large number of subjects (295 compared to CUAVE’s 36) means it is better suited to evaluating the speaker identification task. We believe, however, that the freely available nature of CUAVE will provide a valuable platform for comparison of AVSPI research, even if the individual systems cannot be evaluated nearly as completely on 36 speakers as it could on 295.

The system implemented in this paper is a text-dependent, decision-fusion, audio-visual speaker recogni-

tion system. A block diagram of this system is shown in Figure 1. Decision fusion was chosen because it can be weighted in regards to the reliability of each mode, which is not possible with feature fusion.

2. EXPERIMENTAL SETUP

The stationary-speech sections of the CUAVE database were used for this experiment, however the first stationary continuous speech section was omitted because there was often still significant movement while the speaker moved from the profile view to the front-one view during the early speech-events. This choice of sequences from CUAVE resulted in 7 ten-digit sequences for each of the 36 speakers of which 2 were continuous and 5 were spoken with the words isolated.

For each speaker, 1 continuous and 4 isolated segments were chosen as the training set, with 1 continuous and 1 isolated segment used for testing. The testing segments were corrupted with speech-babble noise at $\{-6, -3, 0, 3, 6, 9, 12\}$ dB signal-to-noise ratio (SNR) to investigate the response of the system to noisy train/test mismatch.

Testing was performed in a text-dependent manner, with the text used being each of the digits in the test set. For this reason, the test set was further segmented into words using the word segmentation data supplied with the CUAVE database.

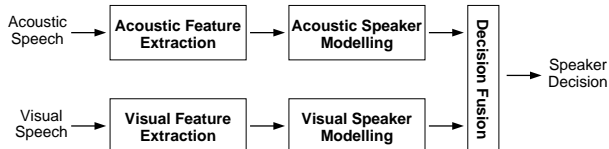


Figure 1: Block Diagram of decision-fusion AVSPI system

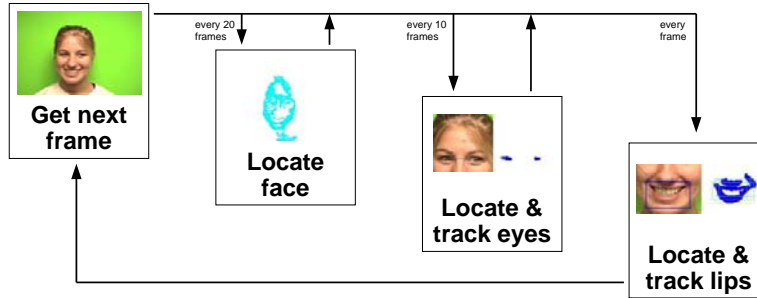


Figure 2: Overview of lip tracking system.

3. ACOUSTIC FEATURES

Mel frequency cepstral coefficients (MFCCs) were used to represent the acoustic features in these experiments because of their general application to both speech and speaker recognition. Each feature vector consisted of first 15 MFCCs, normalised energy coefficient, and the first and second time derivatives of those 16 features to result in a 48 dimensional feature vector. These features were calculated every 10 milliseconds using 25 millisecond Hamming-windowed speech signals.

4. VISUAL FEATURES

The visual speech features extracted for this research consisted of PCA, or *eigenlip*, based feature vectors extracted from the lip region of interest (ROI). These features were chosen because they have been shown to work well for visual speaker recognition on other databases [10].

4.1. Lip location and tracking

To extract features from the lip ROI, this region first needs to be located in each frame of the video. For this research, this was performed in three main stages, face location, eye location and lip location. As shown in Figure 2, each stage was used to help form a search region for the next stage.

4.1.1. Face Location

Before face location was performed on the videos, 10 manually selected skin points for each speaker are used to form thresholds for the Red, Green and Blue (r, g, b) values in colour-space for skin segmentation. The thresholds for each colour-space were calculated from the skin points as

$$\mu_c - \sigma_c \leq p_c \leq \mu_c + \sigma_c, \quad (1)$$

Where $c \in \{r, g, b\}$, μ_c and σ_c are the mean and standard deviation of the 10 points in colour-space c , and p_c is the value of the pixel being thresholded in colour-space c .

Once the thresholds were calculated, they were used for skin segmentation of the video to generate a bounding box of the face region within the frames every 20 frames, and this face location was remembered in the intermediate frames. While there were some false positives from shirt and hair for some speakers, they were not serious enough to harm the eye location and tracking.

4.1.2. Eye Location and Tracking

When transformed into $YCbCr$ space, the eye region of face images exhibit a high concentration of blue-chrominance, and a low concentration of red-chrominance. Therefore eye detection can be done in the $Cr - Cb$ space with reasonable results. However, eyebrows often appear as false positives and can degrade results. To remove the influence of eyebrows the $Cr - Cb$ image can be shifted vertically and subtracted from the original $Cr - Cb$ image. This will cancel the eyebrow minima by subtracting the eye minima, whereas the eye minima will be subtracted by the high values in the skin region and receive a large negative value suitable for thresholding [11].

To locate the eyes from the face region from the previous stage, the top half of the face region was designated as the eye search-area, which was then searched using the shifted $Cr - Cb$ algorithm for the eye locations. The possible eye candidates were searched for two points that were not too large, too close horizontally, and not too distant vertically. Finally the two candidates which had the largest horizontal distance were chosen to be the eye locations. This process was performed every 10 frames, and the locations were remembered in the intermediate frames.

To ensure that the eye locations were correct, the located position was compared to the previous eye location. If the locations had varied more than 30 pixels, they were assumed to be in error and ignored, thereby keeping the previous eye locations. The previous eye locations were also kept if no eyes were found.

While this algorithm worked well for the majority of the sequences we used from the CUAVE database, a significant

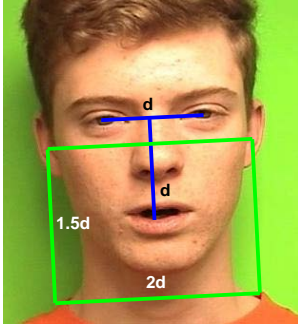


Figure 3: Calculating lip search region from eye locations.

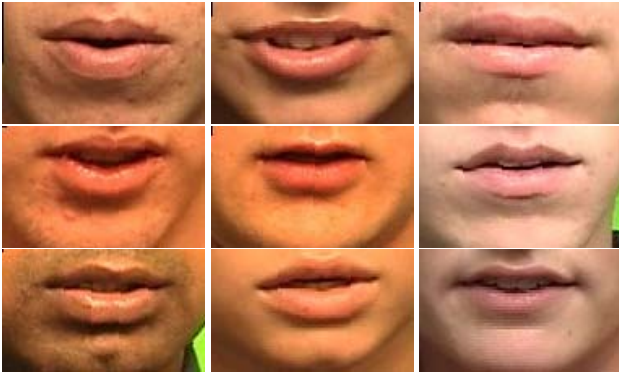


Figure 4: Sample lip ROIs from the CUAVE database.

portion had problems with not locating eyes, or incorrect eye locations. These sequences were labelled manually with the eye locations every 50 frames.

4.1.3. Lip Location and Tracking

Once the eye locations have been found, they are used to calculate a lip search region, as shown in Figure 3. The lip search region is then rotation-normalised, converted to R/G colour-space, and thresholded between two multiples, determined empirically, of the average value within the region. The lip candidates from the thresholding are examined to remove unlikely lip locations (eg. too small, wrong shape). A search-window of 125×75 pixels is then scanned over the lip candidate image to find the windows with the highest concentration of lip candidate regions. The final lip ROI is chosen as the lowest, most central of these windows.

To handle situations where incorrect lip location occurs, the new location is compared to the old location, and rejected if it strays too far. Also, to smooth out the movement of the lip ROI, the final lip ROI is calculated by performing a moving average on the last 10 lip ROIs. Some examples of the captured lip ROIs are shown in Figure 4.

4.2. Visual feature extraction

Fifty representative eigenlips were trained based on 1,000 lip frames randomly chosen from both the training and testing set. These eigenlips were then used to project every lip frame in both the training and testing sets into 50-dimensional PCA-space. The PCA-features were therefore extracted at the same rate as the video frames, 29.97 fps, or approximately 1 frame every 33.4 ms.

5. MODELLING AND FUSION

In its simplest form, speaker identification is the process of choosing the correct speaker from a set of possible speakers trained previously. This is referred to as closed-set identification, and this is the task undertaken by these experiments. A more complicated form of speaker identification is open-set identification, where the possibility that the tested speaker is not within the set of trained speakers is considered, but this will not be considered for these experiments. For these experiments, text-dependent speaker modelling will be performed, meaning that the speakers say the same utterance for both training and testing. 10 experiments are therefore run, one for each digit in the CUAVE database.

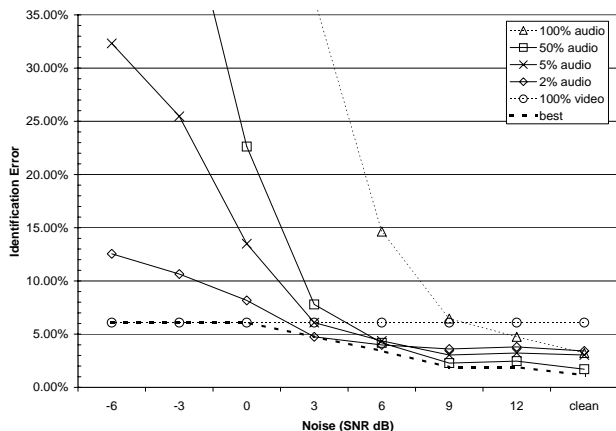
Using phoneme transcriptions obtained using earlier research on the CUAVE database [12], speaker independent HMMs were trained for each phoneme separately in for both the audio and visual modalities. The structure of these HMMs were identical for both modalities, with 3 hidden states for the phone models, and 1 hidden state for the short pause model. These speaker independent models were then adapted using MLLR adaption into speaker-dependent models. The HMM Toolkit, HTK [13] was used to train and test HMMs for these experiments.

Once the speaker-dependent (SD) phone models were created, the test audio and video streams were segmented into the individual digits using the digit transcriptions provided with the CUAVE database. For each specific digit (i.e., ‘one’, ‘two’, ...) the SD phone models corresponding to the specified word for each speaker, in both audio and video, were examined to determine the top 10 most likely speakers, along with a score for each speaker. This process was also repeated in the audio modality over test audio that had been corrupted with speech-babble noise at a range of signal-to-noise ratios.

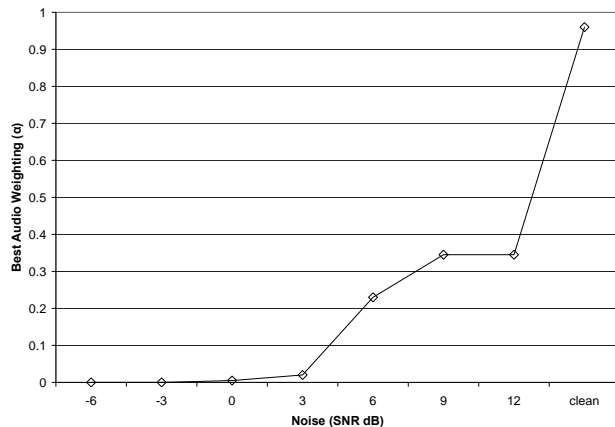
Once the top 10 speakers for each sequence were determined in both modalities, the results were recalculated using a simple weighted sum late fusion,

$$\hat{s}_F = \alpha \times \hat{s}_A + (1 - \alpha) \times \hat{s}_V \quad (2)$$

where \hat{s}_i is the score in mode i , normalised to the range $0 \rightarrow 1$.



(a) Speaker identification error



(b) α with lowest identification error

Figure 5: Response of AVSPI system to speech-babble noise.

The fusion experiments were carried out over the entire test set, including noisy audio. Optimum α values were obtained for each signal-to-noise ratio by testing every value of α from 0 to 1 in steps of 0.01.

6. RESULTS

The response of this fused system to speech-babble audio corruption over a range of signal-to-noise ratios is shown in Figure 5(a). The dashed line indicates the lowest error rates of any α value for a particular noise level. This line indicates the performance that could be obtained with perfect adaptive fusion, where the fusion parameter α could be determined by estimating the noisiness of a signal. Adaptive fusion would allow the fused system to perform as well as the visual system in noisy speech, and better than both systems in clean speech.

It can be seen that the best non-adaptive, or static, fusion performance of this system can be obtained with relatively low value for the fusion parameter α . As an example, the performance corresponding to $\alpha = 0.02$, or 2% audio, is nearly as good as the visual (at least when compared to the audio) in very noisy speech, and only slightly worse than the audio in clean speech. This indication of little dependency on the acoustic domain can be backed up by looking at the values of α that produce the lowest error for each of the noise levels in Figure 5(b). Other than clean speech, it can clearly be seen that better performance can be obtained at all noise levels by placing a high dependence upon the visual modality.

These results have shown that the visual domain is very good at speaker recognition, and this high level of performance can help the acoustic modality at all noise levels. Because the PCA-based visual features are extracted from

a region around the lips, as shown in Figure 4, a lot of static speaker-specific information is also captured with the more speech-related, and dynamic lip-configuration information. This speaker-specific information includes things like the colour of the skin and lips, the presence of facial hair and any other distinguishing marks within that region of the face. It is this speaker-related static information, more so than the speech-related dynamic information, that accounts for the good performance of the visual speaker recognition task, and its corresponding positive impact on the fusion results.

7. CONCLUSION

This paper has shown that the CUAVE database is a suitable platform for implementing an audio-visual speaker identification system. The performance of this system appears to be in the same magnitude as existing AVSPI research [8, 9] on the XM2VTS database, and the performance doesn't appear to be excessively inflated due to the low number of speakers compared to XM2VTS. However, to evaluate the relative performance completely this system should be implemented on the XM2VTS database as well. Additionally, research into the data requirements to adequately test AVSPI systems would be valuable to determine if CUAVE can adequately test an AVSPI system.

Additionally this research has shown that decision-fusion, audio-visual speaker recognition systems can achieve best performance by relying heavily on the video in all but the cleanest of speech. This high performance in the visual domain is caused by primarily by static, speaker-dependent information in the lip ROI. This information allows fusion to provide significant improvement on acoustic speaker recognition in both noisy and clean speech.

8. ACKNOWLEDGEMENTS

We would like to thank Clemson University for freely supplying us their CUAVE audio-visual database [1] for our research.

9. REFERENCES

- [1] E. Patterson, S. Gurbuz, Z. Tufekci, and J. Gowdy, "Cuave: a new audio-visual database for multimodal human-computer interface research," in *Acoustics, Speech, and Signal Processing, 2002. Proceedings. (ICASSP '02). IEEE International Conference on*, vol. 2, 2002, pp. 2017–2020.
- [2] K. Messer, J. Matas, J. Kittler, J. Luetttin, and G. Maitre, "Xm2vtsdb: The extended m2vts database," in *Audio and Video-based Biometric Person Authentication (AVBPA '99), Second International Conference on*, Washington D.C., 1999, pp. 72–77.
- [3] S. Pigeon. (1998, 5/4/2004) M2vts multimodal face database. [Online]. Available: <http://www.tele.ucl.ac.be/PROJECTS/M2VTS/m2fdb.html>
- [4] J. Gowdy, A. Subramanya, C. Bartels, and J. Bilmes, "Dbn based multi-stream models for audio-visual speech recognition," in *Acoustics, Speech, and Signal Processing, 2004. Proceedings. (ICASSP '04). IEEE International Conference on*, vol. 1, 2004, pp. I-993–6 vol.1.
- [5] E. Patterson and J. Gowdy, "An audio-visual approach to simultaneous-speaker speech recognition," in *Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03). 2003 IEEE International Conference on*, vol. 5, 2003, pp. V-780–3 vol.5.
- [6] H. Nock, G. Iyengar, and C. Neti, "Speaker localisation using audio-visual synchrony: An empirical study," in *CIVR 2003*, 2003.
- [7] S. Amarnag, R. Kumaran, and J. Gowdy, "Real time eye tracking for human computer interfaces," in *Multimedia and Expo, 2003. ICME '03. Proceedings. 2003 International Conference on*, vol. 3, 2003, pp. III-557–60 vol.3.
- [8] N. Fox and R. B. Reilly, "Audio-visual speaker identification based on the use of dynamic audio and visual features," in *Audio-and Video-Based Biometric Person Authentication (AVBPA 2003), 4th International Conference on*, ser. Lecture Notes in Computer Science, vol. 2688. Guildford, UK: Springer-Verlag Heidelberg, 2003, pp. 743–751.
- [9] A. V. Nefian, L. H. Liang, T. Fu, and X. X. Liu, "A bayesian approach to audio-visual speaker identification," in *Audio-and Video-Based Biometric Person Authentication (AVBPA 2003), 4th International Conference on*, ser. Lecture Notes in Computer Science, vol. 2688. Guildford, UK: Springer-Verlag Heidelberg, 2003, pp. 761–769.
- [10] A. Kanak, E. Erzin, Y. Yemez, and A. Tekalp, "Joint audio-video processing for biometric speaker identification," in *Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03). 2003 IEEE International Conference on*, vol. 2, 2003, pp. II-377–80 vol.2.
- [11] D. Butler, C. McCool, M. McKay, S. Lowther, V. Chandran, and S. Sridharan, "Robust face localisation using motion, colour and fusion," in *Proceedings of the Seventh International Conference on Digital Image Computing: Techniques and Applications, DICTA 2003*, C. Sun, H. Talbot, S. Ourselin, and T. Adriaansen, Eds. Macquarie University, Sydney, Australia: CSIRO Publishing, 2003.
- [12] P. Lucey, T. Martin, and S. Sridharan, "Confusability of phonemes grouped according to their viseme classes in noisy environments," in *SST 2004*, Sydney, Australia, 2004.
- [13] S. Young, G. Evermann, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, *The HTK Book*, 3rd ed. Cambridge, UK: Cambridge University Engineering Department., 2002.