



COVER SHEET

Dean, David and Lucey, Patrick and Sridharan, Sridha and Wark, Tim (2005)
Comparing Audio and Visual Information for Speech Processing. In *Proceedings The Eighth International Symposium on Signal Processing and Its Applications*, pages pp. 58-61, Sydney, Australia.

Copyright 2005 IEEE

COMPARING AUDIO AND VISUAL INFORMATION FOR SPEECH PROCESSING

David Dean*, Patrick Lucey*, Sridha Sridharan* and Tim Wark†

*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

†Queensland University of Technology &
e-Health Research Centre/CSIRO ICT Centre
Brisbane 4001, Australia
tim.wark@csiro.au

ABSTRACT

This paper examines the utility of audio-visual speech for the two related tasks of speech and speaker recognition. A study of the confusion that exists between speaker and speech elements was performed to show that principal component analysis (PCA) based visual speech is considerably better for the task of speaker recognition than for speech. Decision fusion speech and speaker recognition engines were also tested under various levels of acoustic degradation to find that the optimal fusion configuration for speaker recognition was substantially different than that for speech. These results highlight the problem of employing similar visual features for both speech and speaker recognition.

1. INTRODUCTION

Traditionally, the use of speech to recognise either words or speakers has been performed only in the acoustic modality. Whilst this area of research is fairly mature, there are still major problems with performance in real world environments, particular under high levels of acoustic noise [1]. Audio-visual speech processing (both speech and speaker recognition) attempts to alleviate these problems through the addition of the visual modality to acoustic speech processing [2].

Audio-visual speech processing is a field in its relative infancy, and its application to the tasks of speech and speaker recognition overlap in many areas. In fact, the same configuration can be used for both *speaker-dependent* word recognition, and *text-dependent* speaker recognition. In both configurations, speaker-dependent word (or sub-word) models are trained, and the choice of models for recognition denote the application. Speech recognition is performed by choosing amongst the words for a particular speaker, and speaker recognition is performed by choosing amongst the speakers for a particular word.

Little research has been done into how these two applications (speech and speaker recognition) differ in areas other than the models chosen for recognition. One of the areas where the two applications should differ is the reliance

on each modality. It is generally accepted in acoustic speech processing that cepstral features derived from the acoustic data will work equally well for both speech and speaker recognition [3]. No such consensus has been reached for visual speech processing.

In this paper we investigate the suitability of each domain for the related tasks of speech and speaker recognition. This is important because it gives an indication of how the complete audio-visual speech and speaker recognition systems will perform as the audio is degraded. If good performance can be obtained for the chosen task in the visual domain, the fusion will perform far better at high noise levels than if the visual domain performs poorly. The relative strengths of each domain also provide a starting point for adaptive fusion, providing knowledge of how reliable a domain is likely to be for a particular task.

2. EXPERIMENTAL SETUP

2.1. Training and Testing Datasets

For this experiment, training and evaluation data was extracted from the individual speaker section of Clemson University's CUAVE audio-visual database [4]. This database was chosen because, although relatively new, it is the only freely available audio-visual database for researchers to use. The freely available nature of this data makes it ideal for forming benchmarks and comparing research.

Each individual speaker in the CUAVE database has a single MPEG2 file containing 16 separate sequences. These sequences consist of the digits 'zero' to 'nine' for the isolated-word sections, and 6 different permutations of the same digits for each one of the 6 continuous-word sections.

For these experiments the data was split into the individual sequences, and only the sequences where the face remained stationary throughout were used. Of the 7 sequences (5 isolated and 2 continuous) for each speaker left available from the database, the training set was chosen as the first 4 isolated and the first continuous, with the remaining 2 sequences left for testing. Testing sequences were also arti-

ficially corrupted with speech babble noise at -6, -3, 0, 3, 6, 9 and 12 dB signal-to-noise ratio (SNR) to examine the effects of train/test mismatch on the experiments.

2.2. Feature Extraction

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.

PCA-based features were chosen for the visual domain because they were easily implemented, and should show minimal difference in performance with other popular visual feature extraction techniques such as linear discriminant analysis (LDA) or discrete cosine transform (DCT) for visual speech recognition [5, 6].

To extract the visual features, the lips first had to be located in each frame of the video image. This was performed in three stages: face location, eye location and actual lip location. Simple chromatic-based skin-segmentation was used to determine the approximate location of the face region every 20 frames. The top half of this location was searched to locate the eyes using a shifted $Cr - Cb$ algorithm [7] every 10 frames. Sequences that didn't eye-track successfully (around 40%) were manually eye-tracked every 50 frames. The eye locations were used to rotation-normalise the image, and geometrically form a lip-search-region, which was converted into R/G colour-space, thresholded, and searched for the most likely lip window by maximising the sum of the values within that window.

Once the lips were located for each frame of the training and testing datasets, PCA, or *eigenlips*, was used to reduce the raw pixels down to a manageable number of features. PCA was performed on all the training lip images to create a PCA subspace. Each lip image was then projected into this subspace, and the highest 20 linear modes of mouth variation were kept. Delta and acceleration features were also used to take advantage of the dynamic nature of visual speech, resulting in a 60 dimensional feature vector.

2.3. Speaker Dependent Phoneme Modelling

Both experiments performed here are based on speaker-dependent, hidden Markov model (HMM) based, phoneme models. These models were generated from the training dataset using a CUAVE phoneme transcription generated from earlier research. The phoneme models were first trained in a speaker-independent manner, then adapted to each speaker using maximum likelihood linear regression

(MLLR) adaption. The HMM topology consisted of 3 hidden states for each phone in both modalities, with 1 hidden state for the short-pause model, as suggested in [3].

3. CONFUSION OF PHONEMES AND SPEAKERS

To investigate the suitability of each domain for both tasks, each test sequence was transcribed free-form using the trained speaker dependent phoneme models. At no point in the transcribing process was the set of possible models limited in either speaker or phoneme classes.

Once the complete test dataset had been transcribed in this manner, the confusion between different phones and speakers was examined to determine the suitability for speaker or speech recognition. As an example, to evaluate the use of both domains for speech recognition, the confusion between phonemes can be examined, as shown by the confusion tables in Table 1. By examining the diagonals of these tables, corresponding to the instances of correct identification, it can clearly be seen that the acoustic domain is far better suited to speech recognition than the visual domain.

The confusion was also calculated between speakers (as opposed to phonemes in Table 1), and over the various levels of noise corruption in the test dataset to arrive at the identification rates shown in Table 2.

From these identification rates, we can see that, as expected, the MFCC-based acoustic speech can handle both speech and speaker recognition almost equally well, with speech rating slightly higher. Also, while the identification rates are severely diminished with noise, the two rates stay at similar values relative to each other. However, PCA-based visual speech is shown to be biased towards the speaker recognition task.

These results suggest that while both speech and speaker recognition using the acoustic modality degrade severely in noisy conditions, the speech recognition task cannot rely as heavily on the visual domain as speaker recognition can to improve the performance of the system in degraded audio conditions.

4. FUSION CONFIGURATION FOR SPEECH AND SPEAKER IDENTIFICATION

As discussed earlier, speaker-dependent word identification and text-dependent speaker identification systems have the same configuration, and the choice of models for testing denotes the application. The systems used for this experiment used the speaker-dependent HMM phone-models trained in Section 2 to recognise the words from a known speaker, or the speakers from a known word separately in each modality. The 10 most likely decisions from both modalities are

(a) Clean Acoustic Speech

	(d)	sp	ah	ao	ay	eh	ey	f	ih	iy	k	n	ow	r	s	t	th	uw	v	w	z
(i)		184	5	1	19	3	3	8	8	2	9	65	3	5	23	15	1	1	18	9	2
sp	65	337		1	10	2	1		1	1	55	1		1	6	4	1	3	2	2	
ah	4	1	66		2																1
ao	1		2	66				1				1									
ay	1	2			132						4	1									
eh	7		1		1	67			2					1							
ey		2					65								1						1
f	3	45		1	3	3		45	3	1	8	1		5		2		3	1	3	
ih	1	1							133	1	2	1	2	2	1		1				
iy	2	4								68	2			2	1						
k	2	2				2					61				5						1
n	23	18	1	2	5	2						189	1	1				1	1	9	
ow	5	1				1					2	66	1					4			
r	11	7						2	1			1	69		2	2					
s	10	5					3							154	1						13
t	11	15				1			1	3	4			5	76						1
th	7	26			1	1	3				10	8	2	4	4						1
uw	1							3			4	1	1				72				
v	20	5				4					1	7	2		2	1				84	1
w	3		4	2	1		1		1		5	2									55
z	8				1			1							13						24

(b) Clean Visual Speech

	(d)	sp	ah	ao	ay	eh	ey	f	ih	iy	k	n	ow	r	s	t	th	uw	v	w	z
(i)		16	2	4	11	7	3	11	7	7	3	5	7	7	14	11	2	10	10	3	
sp	128	239	4	4	15	4	14	6	9	5	7	12	4	5	11	9	1	9	4	1	
ah	26	1	29	1	6				1	1			3	3	1	1	1	1	1		1
ao	16	1	1	42	3			1	1	1		1	1	1				1	1		1
ay	28	1	5	1	85	2	3		2	2					4	1	1			2	1
eh	27				6	37		3	2	2	2		1	1	1						1
ey	22				4		35	1	2	1											2
f	21	6			1	2		70	2			4		2	9	3		2	3		
ih	56	6		1	3	4		3	51		1	2	2	1	3	3	1	1	4	1	
iy	13	2	2	2	6		2	2	1	30	1	4	2	2	1	3		1	1	3	
k	35	1			2	5					20	3		1							1
n	112	6	1	3	1	4	5	2	8	9	2	68	5	2	4	7		3	5	4	
ow	22	2		1	4	1					2	1	38	3				1	1	2	1
r	38	3		2	1	1	1	1				4		24	5	3	3	2	4	1	
s	61	3	1	2	5	2	2	2	2	1	2	8	1	1	66	10	3	3	5	3	
t	43	3	2		2	2	2	2	1		1	3	5	7	41						2
th	31	1	2		4		2	3	3		1	4	2		1		11				2
uw	24	1		2	1	1			4	3		3	4		1					35	1
v	44	2			3		2		2	3		5	2	2	4	1			2	53	1
w	23		1					6	1			1	1	4	1	2					31
z	35			1	1							1	1		3			2	1		2

Table 1: Phoneme confusion table in the (a) acoustic domain and (b) visual domain. Column headers indicate actual phonemes, row headers indicate transcribed phonemes, and (i) and (d) refer to phoneme insertions and deletions respectively. The diagonals correspond to correct identification, and are shown in **bold**.

Acoustic			
Noise (SNR)	Phoneme	Speaker	Both
-6 dB	6.0%	2.9%	1.4%
-3 dB	8.9%	6.5%	4.2%
0 dB	15.6%	13.8%	10.8%
3 dB	25.9%	24.7%	21.6%
6 dB	38.3%	34.6%	31.6%
9 dB	52.8%	44.6%	41.7%
12 B	63.8%	52.7%	49.2%
Clean	71.2%	65.3%	61.2%
Visual			
Noise (SNR)	Phoneme	Speaker	Both
Clean	38.7%	58.5%	37.8%

Table 2: Likelihood of phoneme and speaker identification using speaker-dependent phone models.

then combined using weighted-sum decision fusion,

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

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

The choice of α denotes the perceived reliability of each modality, with $\alpha = 0$ being video input only, and $\alpha = 1$ is audio only.

The response of each system to speech-babble noise in the acoustic domain over a selected range of α values is shown in Figure 1. As can be seen from these graphs, the performance of acoustic-only for both task is basically equal, while the visual-only performance is clearly better for speaker recognition than for speech. These graphs also show that speech recognition fusion is far more likely to be *catastrophic*, meaning worse than either audio or video

alone, at all noise levels, while speaker recognition fusion is only catastrophic at high noise levels, such as below 3 dB SNR for $\alpha = 0.5$.

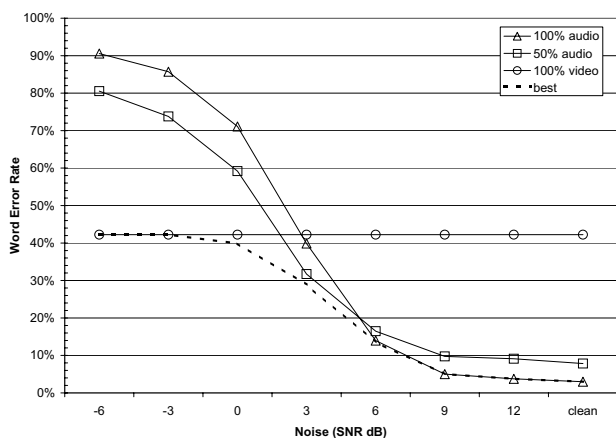
The line in both graphs labelled ‘best’ indicates the lowest error rate obtained for any α value at each level of acoustic noise. This is the performance of a perfect *adaptive* system, one which can determine the noise level and adjust the α -value accordingly. The α -values that correspond to the lowest error for each noise level for both tasks is shown in Figure 2, which clearly shows that speaker recognition has a much higher reliance on visual information at all levels when compared to speech recognition.

5. DISCUSSION AND CONCLUSION

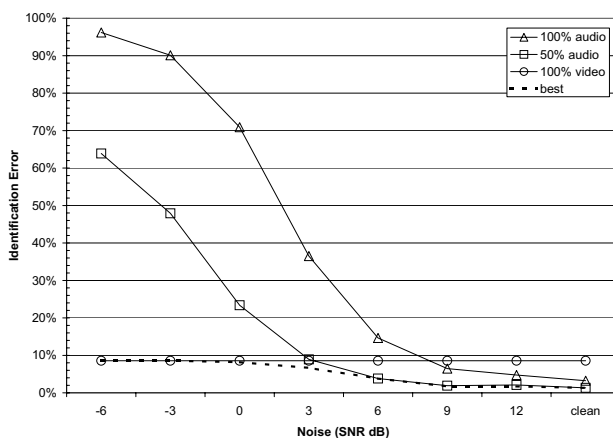
In this paper we have compared speaker dependent word recognition and text dependent speaker recognition using a common audio-visual speech processing configuration. The research has shown that, while MFCC audio features can be used equally well in either speech or speaker recognition, PCA-based visual features are mostly speaker-dependent, and therefore should be used with care in visual speech processing.

As PCA-based visual features are extracted from a region around the lips, a large amount of static speaker-specific information is also captured with the more speech-related, dynamic lip-configuration information. This speaker-specific information includes characteristics such as the colour of the skin and lips, or the presence of facial hair.

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 when compared to the speech recognition task. The



(a) Word identification error rates



(b) Speaker identification error rates

Figure 1: Noise response of word and speaker identification systems.

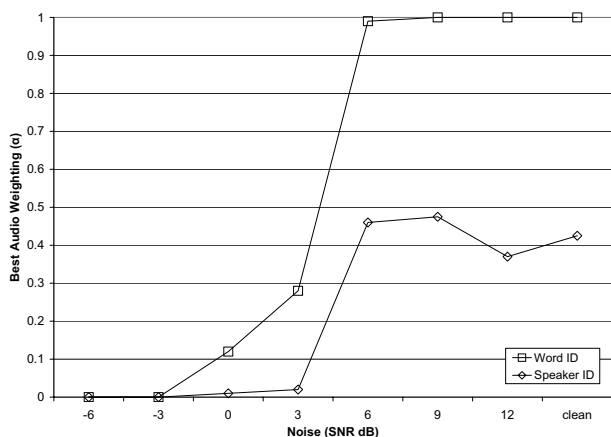


Figure 2: α values with the highest performance for word and speaker identification.

corresponding poor performance of the speech recognition task indicates that the dynamic lip-configuration information needed for visual speech recognition is not captured adequately using a PCA-based representation.

These results show that, even at low levels of acoustic noise, PCA-based visual features can provide similar or better performance than MFCC-based acoustic features for speaker recognition. Accordingly, adaptive fusion for this task should be biased towards the visual domain for best performance. Conversely, speech recognition will have a higher reliance on the audio alone, resulting in much worse fused performance at high noise levels.

Further study needs to be performed in methods of improving the visual modality for speech recognition. By focussing more on the dynamic speech-related information by

using methods such as mean-image removal, optical flow or contour representations it should be possible to obtain better performance for the recognition of audio-visual speech.

6. ACKNOWLEDGEMENTS

The authors wish to thank Clemson University for freely supplying their CUAVE audio-visual database [4] for our research.

7. REFERENCES

- [1] T. Wark and S. Sridharan, "Adaptive fusion of speech and lip information for robust speaker identification," *Digital Signal Processing*, vol. 11, no. 3, pp. 169–186, 2001.
- [2] C. Chibelushi, F. Deravi, and J. Mason, "A review of speech-based bimodal recognition," *Multimedia, IEEE Transactions on*, vol. 4, no. 1, pp. 23–37, 2002.
- [3] 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.
- [4] 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.
- [5] M. S. Gray, J. Movellan, and T. J. Sejnowski, "Dynamic features for visual speechreading: A systematic comparison," in *Neural Information Processing Systems*, 1996, pp. 751–757.
- [6] G. Potamianos, A. Verma, C. Neti, G. Iyengar, and S. Basu, "A cascade image transform for speaker independent automatic speechreading," in *Multimedia and Expo, 2000. ICME 2000. 2000 IEEE International Conference on*, vol. 2, 2000, pp. 1097–1100 vol.2.
- [7] 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.