interspeech 2008 incorporating SST 08

Cascading Appearance-Based Features for Visual Speaker Verification

David Dean, Sridha Sridharan and Patrick Lucey

Speech, Audio, Image and Video Research Laboratory, Queensland University of Technology
Brisbane, Australia

ddean@ieee.org, {s.sridharan, p.lucey}@qut.edu.au

Abstract

The cascading appearance-based (CAB) feature extraction technique has established itself as the state of the art in extracting dynamic visual speech features for speech recognition. In this paper, we will focus on investigating the effectiveness of this technique for the related speaker verification application. By investigating the speaker verification ability of each stage of the cascade we will demonstrate that the same steps taken to reduce static speaker and environmental information for the speech recognition application also provide similar improvements for speaker recognition. These results suggest that visual speaker recognition can improve considerable when conducted solely through a consideration of the dynamic speech information rather than the static appearance of the speaker's mouth region.

Index Terms: visual speaker recognition, visual speech recognition, cascading appearance-based features

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, particularly under high levels of acoustic noise. Audio-visual speech processing (AVSP) (covering both speech and speaker recognition) attempts to alleviate these problems through the addition of the visual modality to acoustic speech processing [1].

One of the most important factors in the final performance of an AVSP system is the choice of speech-based feature extraction techniques for the acoustic and visual modalities. While such feature-extraction research is very mature for the acoustic modality, the comparative novelty of visual speech processing has not yet resulted in a general consensus on the extraction of suitable visual features [2].

Because visual speech is fundamentally represented by the movement of the visual articulators, many feature extraction techniques focus on these movements rather than the stationary appearance within each frame. This approach has been shown to work very well for speech recognition [3], but it is not clear that it would apply for speaker recognition where static features, such as skin colour or facial hair, may be useful for identity purposes [4].

The extraction of visual speech features directly from the mouth region-of-interest (ROI) of a talking face has been shown to outperform geometric or contour-based feature extraction techniques for visual speech recognition [2]. However, a downside of these *appearance-based* feature extraction techniques is that each frame of the ROI contains a large amount of static speaker or environmental specific information that is unrelated to the movements of the visible articulators. Dynamic visual feature extraction techniques are designed to take the static ROI images and emphasize the dynamic nature of the visual speech over the stationary appearance within each frame. A number of techniques have been developed to extract these dynamic features, from simple approaches like difference images, or the use of delta and acceleration coefficients to more complicated techniques such as optical flow [1].

This research was supported by a grant from the Australian Research Council (ARC) Linkage Project LP0562101.

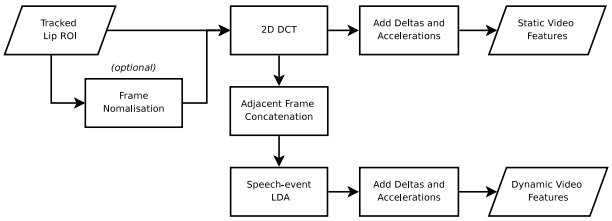


Figure 1: *Overview of the cascading appearance-based feature extraction system used for this paper. This is a simplified version of Potamianos et al.'s original cascade [3].*

The current state-of-the-art in dynamic visual speech feature extraction is a multi-stage cascade of appearance-based (CAB) feature extraction techniques developed by Potamianos et al. [3], which has been shown to work well for speaker independent speech recognition [2]. While CAB features have been demonstrated for speaker recognition by Nefian et al [5], no detailed study of the effects of the each stage cascade on speaker recognition has yet been conducted, and will therefore form the focus of this paper.

2. Cascading appearance-based features

2.1. Overview

An outline of the CAB feature extraction system used for this paper is shown in Figure 1, and can be seen to have three main stages:

1. Frames are (optionally) first normalised to remove irrelevant speaker or environmental information,
2. Static features are extracted for each individual frame, and
3. Dynamic features are generated from the static features over a window of several frames

In order to examine the utility of the CAB features for speaker verification, features will be extracted from both the static and

dynamic stages of the cascade, with and without the frame normalisation. This will allow the usefulness of each stage of the cascade to be evaluated for the speaker verification task.

2.2. Frame normalisation

Before the static features can be extracted from each frame's ROI, an image normalisation step is first performed to remove any irrelevant information, such as illumination or speaker variances. In Potamianos et al.'s original implementation of the cascade [3], this step was performed using feature normalisation after static feature extraction, but image normalisation has been shown to work slightly better due to the ability to handle variations in speaker appearance, illumination and pose as part of a wider pre-processing front-end [6]. As such, image mean normalisation was chosen over feature mean normalisation for these experiments.

This image normalisation step consists of calculating the mean ROI image over an entire utterance which can then be subtracted pixel-by-pixel from each ROI image before the ROI is presented to the static feature extraction stage.

The motivation behind normalising the ROI in this manner comes from the notion that a large amount of speaker appearance-based information is collected in the standard appearance-based feature extraction techniques [7], and that this information would not be useful for modelling speech events. Of course, it is quite possible that this information would be useful for the speaker recognition application, so a version of the cascade will also be tested without this normalisation step to investigate this effect.

2.3. Static feature extraction

Once the ROI has been (optionally) normalised, static visual speech features can then be extracted. The main aim of feature extraction is to provide compression of the raw pixel values in the ROI whilst still maintaining good separation of the differing speech events. Discrete cosine transform (DCT) based feature extraction was chosen for Potamianos et al.'s original cascade [3] as well as this implementation in this paper, as they can more easily be calculated than other other feature extraction techniques such as PCA [2].

For the static feature extraction stage of this cascade, the top D^S coefficients were taken in a zig-zag pattern in the two-dimension DCT of the ROI. For the evaluation of the static features, delta and acceleration components were added to result in a $3 \times D^S$ dimensional feature space, but only the primary D^S features were used as input to the dynamic feature extraction stage.

2.4. Dynamic feature extraction

To extract the dynamic visual features that have been shown to improve human perception of speech, this stage of the cascade extracts linear discriminant analysis (LDA) based features over a range of consecutive ROIs. By operating over a number of consecutive frames centred on each frame under consideration, the LDA stage emphasizes the dynamic over the static features of the visual speech. The input to the LDA algorithm for the concatenated ROI features around $\mathbf{o}_t^S$ is therefore given as

$$\mathbf{o}_t^C = \left[\mathbf{o}_{t-J}^S, \ldots, \mathbf{o}_t^S, \ldots, \mathbf{o}_{t+J}^S\right] \quad (1)$$

Where $\mathbf{o}_t^S$ is the static video features at time t and J is the number of frames being concatenated on each side of the central frame. It can be seen that this results in a feature vector of size $D^C = (2j+1)\, D^S$.

Once the LDA transformation matrix was calculated using training data, it can then be used to transform the static observation vectors from the DCT stage of the cascade to form the dynamic visual speech features used to train and test the models for speaker verification. The final dynamic feature vector dimensionality can be reduced by only choosing the first D^D eigenvectors from the calculated LDA transformation matrix before transforming the concatenated static features.

Session \ Configuration	1	2	3	4	5	6	7	8	9	10	11	12
1	Train		Train		Train		Eval	Test	Eval	Test	Eval	Test
2	Train		Eval	Test	Eval	Test	Train		Train		Test	Eval
3	Eval	Test	Train		Test	Eval	Train		Test	Eval	Train	
4	Test	Eval	Test	Eval	Train		Test	Eval	Train		Train	

Table 1: *XM2VTS dataset configurations used in these experiments*

3. Experimental setup

3.1. Training and testing datasets

For this experiment, training, testing and evaluation data were extracted from the digit-video sections of the XM2VTS database [8]. The training and testing configurations used for these experiments was based on the XM2VTSDB protocol [9], but adapted to allow more tests than provided by the protocol. Each of the 295 speakers in the database has four separate sessions of video where the speaker speaks two sequences of two sentences of ten digits. In each of the configurations, two sessions were used for training, one for evaluation and one for testing, allowing for 12 configurations in total, as shown in Table 1. By comparison, the XM2VTSDB protocol only allows for the first configuration.

These experiments were performed as verification experiments, where the speaker would attempt to enter the system by claiming the identity of a particular client. To perform this task, the speakers were split into two groups: clients, who claimed their own identity; and impostors, who claimed the identity of one of the clients.

As per the XM2VTSDB protocol, 200 speakers were designated clients, and 95 were used as impostors. For each client testing sequence (2 per session), 20 sequences were chosen at random from the impostor set allowing for a total of 400 (200×2) client tests and 8000 ($200 \times 2 \times 20$) impostor tests for each configuration. Over all 12 configurations, 4800 client tests and 96000 impostor tests are performed.

3.2. Feature extraction

In order to provide a baseline for the visual speaker verification experiments, Perceptual linear prediction (PLP) based cepstral features were extracted from the acoustic speech. Each acoustic feature vector consisted of the first 13 PLPs including the zeroth, and the first and second time derivatives of those 13 features resulting in a 39 dimensional feature vector. These features were calculated every 10 milliseconds using 25 millisecond Hamming-windowed speech signals.

Visual features were extracted from a manually tracked lip region-of-interest (ROI) from 25 fps (40 milliseconds / frame) video data. Manual tracking of the locations of the eyes and lips were performed every 50 frames, and the remainder of the

frames were interpolated from the manual tracking. The eye locations were used to normalise the in-plane rotation of the lips. A rectangular region-of-interest, 120 pixels wide and 80 pixels tall, centered around the lips was extracted from each frame in the video. Each ROI was then reduced to 20% of its original size (24×16 pixels) and converted to grayscale.

Following the ROI extraction, the mean ROI over the utterance is optionally removed as the first stage of the CAB feature extraction. The top 20 static DCT-based features are then extracted as the second stage, and deltas and acceleration coefficients were added to resulting in 60 dimensional video feature vectors. Subsequently, to incorporate dynamic speech information, 7 neighboring such features (without the temporal-derivative coefficients) over ± 3 adjacent frames were concatenated, and were projected via LDA to 20 dimensional dynamic visual feature vectors. The delta and acceleration coefficients of this vector were then incorporated, resulting in 60 dimensional visual feature vectors at the final stage of the cascade.

As a result of the application of the CAB feature extraction process, 4 video features were available for testing the speaker verification system. Including the acoustic features results in 5 features tested for speaker verification:

- acoustic PLP features (A-PLP)
- static video without normalisation (V-DCT)
- static video with normalisation (V-MRDCT)
- dynamic video without normalisation (V-LDA-DCT)
- dynamic video with normalisation (V-LDA-MRDCT)

The A-PLP, V-DCT and V-MRDCT features were designed such that they could be extracted from any given utterance without any prior knowledge of the type of data they were working with, allowing their feature vectors to be used for each of the configurations of the XM2VTS database. However, because the LDA-derived features, V-LDA-MRDCT and V-LDA-DCT were trained based on acoustic speech events in the training sessions of the framework, each unique training configuration of the framework had to use a differing set of LDA-derived visual feature vectors. As a result, each sequence being tested had 6 different feature representations for V-LDA-MRDCT and V-LDA-DCT based upon which XM2VTS configuration was being tested.

3.3. Speaker modelling

In order to test the speaker verification system, both text-dependent and text-independent speaker models were trained and tested against all 12 configurations of the XM2VTS database defined earlier.

Both HMM and GMM speaker-dependent models were generated by adapting background models to each individual speaker. The background models were generated using the training sequences for each configuration over both clients and impostors. These models were then adapted to each individual client speaker's training sequences using maximum a posterior (MAP) adaptation [10].

GMM models were trained over all training sequences, whereas HMM models were trained for each word. Empirical experiments were performed on a single configuration to determine the best topology for each datatype, which are shown in Table 2. HMM training was performed using HTK [11], and GMM training with internally developed utilities.

	GMM	HMM	
Model	Mixtures	Mixtures	States
A-PLP	256	11	8
V-DCT	8	9	16
V-MRDCT	128	9	16
V-LDA-DCT	32	9	16
V-LDA-MRDCT	128	9	16

Table 2: *Best performing HMM and GMM topologies chosen for each datatype.*

4. Results and Discussion

The results of the text-dependent and text-independent speaker verification experiments are shown in Figures 2(a) and (b) respectively. Speaker verification scores were calculated by comparing scores obtained with the speaker specific models and the background models and plotting the difference between the two using detection error trade-off (DET) plots to investigate the relative false alarm rate and misses that can be obtained with each datatype under consideration.

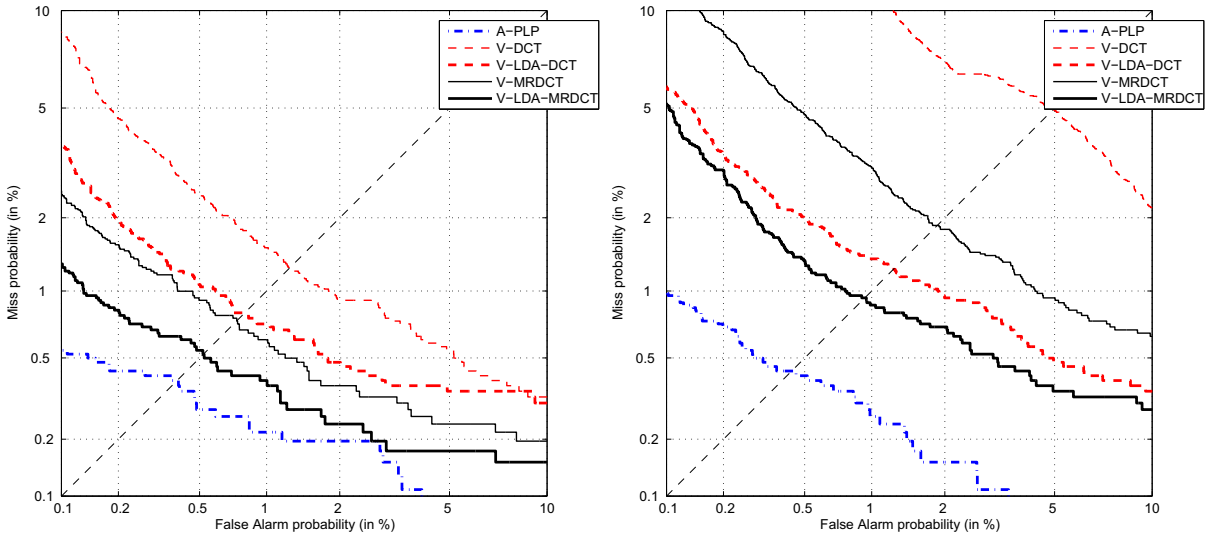
From both the text-dependent and text-independent speaker verification results shown here, it can be seen that both the frame normalisation (DCT vs MRDCT) and application of speech based LDA within the cascade provide a benefit to speaker verification. In both cases, the lowest error rates occur when both the frame normalisation and LDA stages of the cascade are applied, resulting in the V-LDA-MRDCT video features.

Interestingly, these experiments show that the same features that have been shown to perform well for the task of speech recognition by other researchers [2] also perform very well for speaker verification. All of the stages of the cascade that provided benefits by removing speaker- and session-specific information also provided similar benefits for the speaker verification experiments, even though the original intent of the cascade was to provide a form of normalisation across speakers and subsequently improve speech recognition in unknown speakers.

These results suggest that for visual speaker recognition, the behavioural nature of speech could be at least as important, and possible more-so than the physiological characteristics [12]. That is, it may be easier to recognise speakers by how they speak, than by their appearance while they speak. This also has the benefit that as static appearance is less important, environment conditions such as illumination, and within speaker variations such as facial hair or makeup, become less of an issue provided they do not change throughout an utterance, and as long as the extraction of dynamic features can still be performed adequately.

5. Conclusion

This paper investigated the CAB feature extraction process introduced by Potamianos et al. [3] to improve speech recognition for the related task of speaker verification. The experiments conducted within this paper found that this process also provided considerable improvement for speaker verification, even though the aim of the CAB process was to remove or downplay the speaker (and environmental) specific information in order to improve the recognition of speaker-independent speech. That the speaker verification experiments improved in performance as static information was removed suggests that dynamic visual information can play a very important role in visual (and audio-



(a) Text-dependent (HMM)

(b) Text-independent (GMM)

Figure 2: *Detection error trade-off (DET) plots for speaker verification.*

visual) person recognition, particular when the facial movements are speech related.

Of course, face recognition is a very mature area of research that has shown that static recognition of faces can provide good performance, and the possibility certainly exists of using a combination of static face and dynamic features to represent the visual modality with a minimum loss of information. Some promising versions of such systems have been developed [5], but this area is still a relatively new area of research.

6. Acknowledgments

The research on which this paper is based acknowledges the use of the Extended Multimodal Face Database and associated documentation. Further details of this software can be found in [8] or at `http://www.ee.surrey.ac.uk/Research/VSSP/xm2vtsdb/`.

7. References

[1] 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.

[2] G. Potamianos, C. Neti, G. Gravier, A. Garg, and A. Senior, "Recent advances in the automatic recognition of audiovisual speech," *Proceedings of the IEEE*, vol. 91, no. 9, pp. 1306–1326, 2003.

[3] 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.

[4] J. Mason and J. Brand, "The role of dynamics in visual speech biometrics," in *Acoustics, Speech, and Signal Processing, 2002. Proceedings. (ICASSP '02). IEEE International Conference on*, vol. 4, 2002, pp. IV–4076–IV–4079 vol.4.

[5] 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. Guildfork, UK: Springer-Verlag Heidelberg, 2003, pp. 761–769.

[6] P. Lucey, "Lipreading across multiple views," Ph.D. dissertation, Queensland University of Technology, Brisbane, Australia, 2007.

[7] A. G. Chitu, L. J. Rothkrantz, J. C. Wojdel, and P. Wiggers, "Comparison between different feature extraction techniques for audio-visual speech recognition," *Journal on Multimodal User Interfaces*, vol. 1, no. 1, 2007.

[8] K. Messer, J. Matas, J. Kittler, J. Luettin, 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.

[9] J. Luettin and G. Maitre, "Evaluation protocol for the extended M2VTS database (XM2VTSDB)," IDIAP, Tech. Rep., 1998.

[10] C.-H. Lee and J.-L. Gauvain, "Speaker adaptation based on MAP estimation of HMM parameters," in *Acoustics, Speech, and Signal Processing, 1993. ICASSP-93., 1993 IEEE International Conference on*, vol. 2, 1993, pp. 558–561 vol.2.

[11] 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.

[12] J. Brand, J. Mason, and S. Colomb, "Visual speech: A physiological or behavioural biometric?" in *Audio- and Video-based Biometric Person Authentication (AVBPA 2001), 3rd International Conference on*, Halmstad, Sweden, 2001, pp. 157–168.