LIP FEATURES FOR SPEECH AND SPEAKER RECOGNITION

Roland Auckenthaler† *and Jason Brand* ‡ *and John S. Mason*‡
†Department of Electronics,Technical University Graz,
Inffeldgasse 12,A-8010 GRAZ, AUSTRIA
‡Department of Electrical & Electronic Engineering,
University of Wales Swansea, SA2 8PP,UK
email: {eeaucken, eebrand, J.S.D.Mason}@swansea.ac.uk

ABSTRACT

1 Abstract

This paper implicitly differentiates between the quality of visual representation necessary for speech and speaker recognition and assesses the performance of visual lip features with respect to well established audio features. Blue lip highlighted data is used to show how variations in lip measurements can influence speech and speaker recognition. From these experiments and other researchers results [1] it is postulated that the fine detail of the lips is critical for speaker recognition, but conversely, the same amount of detail does not noticeably improve visual speech recognition. Visual error rates of 26.3% and 70% are achieved for cross-digit speaker and cross-speaker speech recognition respectively.

2 Introduction

In both speech and speaker recognition the audio signal can easily be corrupted by noise, causing recognition degradation. One way to relieve the effects of additive noise is to incorporate visual information from a speakers face, which is often complementary to the acoustic signal [2, 3, 4].

The visual signal, neglected for many years owing to insufficient computing power, has recently been shown to aid speech recognition [5] and speaker recognition [1] in both clean and noisy speech. The influence of the visual modality has also been emphasised in other studies [6].

Acoustic features have been well established for many years and predominantly utilise the properties of the log power spectrum e.g. mel-frequency cepstral coefficients (mfcc) or linear predictive coefficients. Unlike acoustic features, effective visual features are not so well defined, especially when considering both speech and speaker recognition. It is still not known which features are important for speech or speaker recognition or how to extract them.

For visual speech recognition it is accepted that important information is contained in the lip contours [7], but for speaker recognition the feature choice is much less clear. Inspired by the promising results from [1, 8] lip features have been investigated for speech and speaker recognition.

The question posed here relates to the level of person and speech discriminatory information inherent in lip based features and whether this information can be extracted and utilised in an automatic recognition system. This paper presents novel lip features that attempt to capture the fine detail of the exterior lip contour and compares them to some that provide only a gross representation of the lips. The use of chroma-keying also allows the effect of lip variation on speech and speaker recognition to be observed. The importance of consistent lip segmentation and the benefits of visual information for cross digit speaker recognition are clearly demonstrated.

3 Lip Features

In order to achieve their full potential and harness the robust multimodal nature of human speech, it is clear that recognition systems need to overcome the difficult task of extracting visual speech features. Previous speech reading work attempts to broadly classify visual feature extraction methods, as follows: image based, pixel based and non-parametric methods all use pixel level information from an image region containing the lips either directly or after some processing. These methods include the use of blob extraction [2], Fourier transforms and power spectra [9, 5], eigenlips [4], area sieves [10] and optical flow analysis [11]. Here, most of the pixel level detail is retained because it is accepted that the presence of teeth and tongue provides valuable visual speech discrimination. Many features are globally captured and it remains for the system to distinguish between valuable speech and miscellaneous information, such as lighting variations.

Model based techniques, high level methods and parametric features are based on techniques to capture lip contour information. Deformable templates [12], active shape models [10, 13], snakes [14] and parametric curves [1, 15] have all been used in the past to track lip contours and further more extract typical geometric features such

as areas, widths and heights. Included in these methods is the work of Benoit [16] who extensively researched visemes, the visual equivalent of phonemes. However, the discussion as to what constitutes a good visual feature still continues. Clearly, it is not the static information that is of use in speech recognition, but the temporal changes i.e. dynamic features [17, 18], whereas in speaker recognition the situation is less clear.

Artificial blue lip highlighting has been used in the past to enhance the contrast of the lips and the face, to aid lip segmentation [19, 7]. The use of lip highlighting can only be justified if it is lip features that are being investigated and not the extraction process itself. Here we perform speech and speaker recognition experiments to implicitly show the effect of feature variation via the application of lip highlighting.

The motivation for our use of the lips in *speaker* recognition is based on the fact that they might well be used for the task of *speech* recognition. This implies solutions will be forth coming to problems such as lip-tracking and feature extraction. Furthermore it is unclear at this point what the best facial feature might be. In other words it is important to assess just how good lip-features are for person recognition. Previous work [8, 1] suggests that the lips possess approximately the same level of discrimination as the acoustic signal. For example in the case of identical twins Chibelushi [1] reports equal recognition rates from both audio and visual sources. However, it must be noted that Chibelushi's segmentation was not automated, but was performed by hand, the contribution of which is difficult to predict.

4 Geometric Lip Features

The use of geometric features includes colour transformations of the image texture to extract inner mouth contours [2, 20], spatio-temporal gradients to filter out lip contours [21] and the use of optical flow to analyze the dynamics of lip motion [11].

Here we compare the discriminatory properties of two types of geometric feature. The first provides a gross representation of lips using only three measurements, area, width and height. The second is an alternative lip signature, based on lip contours, measured in terms of normalised angle projections. Essentially both can be broadly classified under model based techniques.

Blue lip highlighting is used to facilitate the lip extraction process. Importantly, in the context of speaker recognition, recordings are made over 2 sessions. The aim here is to simulate a robust lip segmentation algorithm. Extraction using lip highlighted data serves two purposes:

1. it makes the difficult task of lip segmentation easier,

2. it provides lip contour variation between sessions allowing this parameter to be investigated.

Using a blue lip colour model, generated from additional training data, the original image is processed and reduced with the aim of retaining only the blue lips. Pixels that are not classified as blue are turned to white. Figure 1 shows an example of some extracted lips.

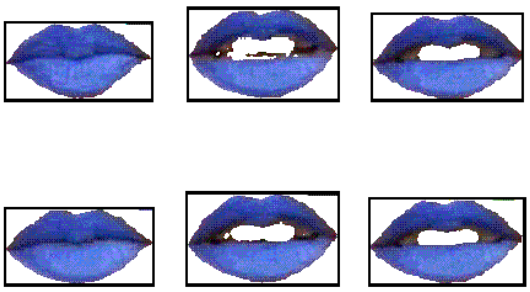

Figure 1: Examples of extracted lips from one speaker

The reduced image is used to locate the centre point of the lips, which are assumed to lie on a line intersecting the lip corners, half way between the image edges. From this point the distance to the outer lip contour and the corresponding angle is traced, resulting in an upper and lower lip signature, figure 2. Examples are shown in figure 2 for upper and lower lips, and it is seen that the two have different characteristics, particularly over the central region.

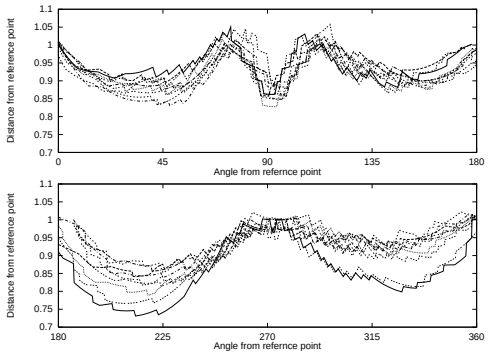

Figure 2: Signatures of upper and lower lips respectively

These signatures do not require the lip corners to be aligned horizontally and therefore are invariant to any head motion in this plane. Transforming the signatures in to the frequency domain via an FFT allows us to select the first **n** bins as features. Two types of visual representation are selected, the first is the FFT magnitude spectra and the second is the FFT real and imaginary components.

The question we seek to answer is whether the signatures are useful in the task of recognition and moreover whether the variation in the lip contours brought about by the application of blue lipstick will have an adverse effect on speech and speaker recognition. This will provide a guideline governing the required accuracy of the lip segmentation algorithm relative to the task in hand

i.e. speech or speaker recognition.

5 Database

Initial experiments are conducted on a small database. 9 persons are chosen from the BT-DAVID audio video database [22] recorded with lip-highlights and four digits are extracted and processed from 2 separate sessions (to avoid discrimination on the grounds of lip-highlighting). The digits are chosen according to reported audio speaker recognition performance. The digits 'four' and 'eight' are reported by Yu [23] to give relatively high error rates in the acoustic domain e.g. 46% in a text dependent task, whereas the digits 'nine' and 'zero' are reported to give excellent speaker recognition results, approaching 0%. The complementary audio and visual recordings are digitised at 8kHz and 25 frames per second, respectively. The facial images are then automatically segmented and lip profiles normalised and extracted. Using this contrasting digit set, the relative merits of the visual features are assessed.

6 Experiments

We postulate that the fine detail of the lip contours is critical for speaker recognition, but not so for speech recognition. Comparisons are made with Chibelushi [1], whose visual feature extraction involve accurate hand segmentation. This method captured the fine detail of the lip contours which we suggest is important for speaker recognition.

Using blue lip data from the same session, a robust lip segmentation algorithm is simulated, allowing us to observe the effects of only small lip variations. Cross-digit, visual speaker recognition experiments are performed using test and training data from the same session. For example, we train using the digit 'four', and we test using the remaining digits 'eight','nine' and 'zero'. The digits are then swapped around and the results are averaged. Within the session, the blue lipstick is only applied once to each speaker and therefore it is assumed that variations in the lip contour signatures are at a minimum. The features can therefore be assessed according to their discriminatory properties and not judged by the performance of the lip segmentation. Complimentary cross-speaker, *speech* recognition experiments are also performed.

For comparison purposes, a less accurate lip segmentation algorithm is simulated by performing recognition on test and training data from separate session. Here, the blue lipstick has been applied in approximately the same manner on two separate occasions. The natural variations in the lip contour are unavoidably introduced in to the features.

As stated previously, lip features are formed via an FFT of the upper and lower lip profile signatures. The visual features include the magnitude spectra, the real and imaginary components and also basic geometric area, width and height measurements. These are compared against one another and with standard mel-frequency cepstral coefficients (mfcc). The classifier is a simple nearest-neighbor system based on euclidean distances. Equivalent audio speech and speaker recognition experiments are also performed for relative comparison purposes.

7 Results and Conclusions

Visual Feature	Same session	Different session
Area,width,height	70.7%	69.5%
FFT magnitude	71.6%	71.9%
FFT Real and Imaginary	72.3%	70.5%
Area/FFT Magnitude	71.6%	68.6%
Area/FFT Real and Imaginary	71.1%	66.4%

Table 1: Cross-speaker visual speech recognition error rates

Visual Feature	Same session	Different session
Area,width,height	32.8%	69.9%
FFT magnitude	44.4%	68.1%
FFT Real and Imaginary	42.6%	69.4%
Area/FFT Magnitude	27.3%	63.4%
Area/FFT Real and Imaginary	26.3%	64.8%

Table 2: Cross-digit visual speaker recognition error rates

Tables 1 and 2 show cross digit and cross speaker visual recognition error rates using data from the same and different sessions. The drop in recognition due to separate sessions is an indication of lip feature variation. This is likely to be caused by a variation in the application of the blue lipstick. From the contrasting speaker recognition results across sessions, 26% to 65%, we can postulate that visual speaker recognition features are very sensitive to slight variations. Conversely, it can be concluded that the same variation does not meaningfully effect the speech recognition results 68% to 70%, which are predictably almost the same. Using a combination of simple geometric measurements (area, width etc.) and parameters from parabolic curves, Chibelushi [1] achieved speaker recognition scores of 4.5% with 1 version training. The same features only managed at best 51% for visual speech recognition with 1 version training. These results show how important the fine detail of the lips is for good speaker recognition performance. However, for speech recognition, there is little point in utilising the fine detail of the lips as gross detail appears to provide much the same information.

Table 2 clearly shows that the best visual speaker feature is a combination of the fine lip contour detail provided by the FFT real and imaginary components and a gross representation provided by the area, width and height, giving a best result of 26.3% error. For speech

Audio Feature	Same session	Different session
14th order mfcc	24.2%	24.5%

Table 3: Cross-speaker audio speech recognition error rates

Audio Feature	Same session	Different session
14th order mfcc	63.4%	76.8%

Table 4: Cross-digit audio speaker recognition error rates

recognition, the gross detail from the area features has a marginal advantage over the lip signatures.

With regards to speaker recognition, tables 2 and 4 show the clear advantage of visual features, 26.3% over audio features, 63.4%, for this cross-digit case. Futhermore it would appear that *cross-session* visual performance, 63.4% is also superior to the complimentary audio case, 76.8%. Conversely, tables 1 and 3 highlight the advantages of audio features, 24.2% over visual features, 66.4% for cross-speaker speech recognition.

References

[1] C. Chibelushi, J. Mason, and F. Deravi. Integration of acoustic and visual speech for speaker recognition. In *Proc. EUROSPEECH*, page 157, 1993.

[2] E. Petajan. Automatic Lip Reading to Enhance Speech Recognition. In *Proc. IEEE Computer Vision and Pattern Recognition*, pages 44–47, 1985.

[3] P.Jourlin, J.Luettin, D.Genoud, and H.Wassner. Acoustic Labial Speaker Verification. *AVBPA-97*, pages 319–334, 1997.

[4] C. Bregler and Y. Konig. Eigenlips for Robust Speech Recognition. In *Proc. ICASSP*, pages 669–672, 1994.

[5] C. Bregler, H. Hild, S. Manke, and A. Waibel. Improved Connected Letter Recognition by Lipreading. In *Proc. ICASSP*, pages 557–560, 1993.

[6] H.McGurk and J.MacDonald. Hearing Lips and Seeing Voices. *Nature*, 264:746, 1976.

[7] Robert Kaucic, Barney Dalton, and Andrew Blake. Real-Time Lip Tracking for Audio-Visual Speech Recognition Applications. In *Proc. ECCV*, 1996.

[8] J.Luettin, N.Thacker, and S.Beer. Speaker Identification by Lipreading. In *Proc. ICSLP*, pages 62–64, 1996.

[9] S. Nakakura, R. Nagi, and K. Shikano. Improved Bimodal Speech Recognition using Tied-Mixture HMM's and 5000 word audio-visual synchronous database. *EUROSPEECH*, page 1623, 1997.

[10] I.Matthews, J.Bangam, and S.Cox. Audiovisual Speech Recognition Using Multiscale Nonlinear Image Decomposition. In *Proc. ICSLP*, pages 38–41, 1996.

[11] K. Mase and A. Pentland. Automatic Lip Reading by Optical FLow Analysis. *Systems and Computers in Japan*, 1991.

[12] D.Chandramohan and P.Silsbee. A Multiple Deformable Template for Visual Speech Recognition. In *Proc. ICSLP*, volume 1, pages 50–53, 1996.

[13] J. Luettin, N. Thacker, and S. Beet. Locating and Tracking Facial Speech Features. *Int. Conf. Pattern Recogntion*, 1996.

[14] D.Terzopoulos and K. Waters. Analysis and Synthesis of Facial Image Sequences Using Physical and Anatomical Models. *IEEE Trans Pattern Analysis and Machine Intelligence*, 15, 1993.

[15] Tarcisio Coianiz, Lorenzo Torresani, and Bruno Caprile. 2D Defromable Models for Visual Speech Analysis. *Proc NATO Advanced Study Institute: Speechreading by Man and Machine*, 1995.

[16] C.Benôit and L.Reveret. A Viseme-based Approach to Labiometrics for Automatic Lipreading. *AVBPA-97*, pages 335–342, 1997.

[17] J. Luettin. Towards Speaker Independent Continuous Speechreading. *EUROSPEECH*, page 1991, 1997.

[18] A. Rogozan and P. Deleglise. Continuous Visual Speech Recognition Using Geometric Lip-Shape Modesl and Neural Networks. *EUROSPEECH*, page 1999, 1997.

[19] L.Reveret. From Raw Images of the lips to Articulatory Parameters: A Viseme Based Prediction. *EUROSPEECH*, page 2011, 1997.

[20] E. Petajan and H. Graf. Robust Face Feature Analysis for Automatic Speechreading and Character Animation. *Speechreading by Man and Machine*, pages 425–436, 1996.

[21] K.V. Prasad, D.G. Stork, and G. Wolff. Processing Video Images for Neural Learning of Lipreading. *Technical Report, Ricoh California Research Center*, 1993.

[22] J.S.D. Mason, F. Deravi, C.C. Chibelushi, and S. Gandon. Project:DAVID (Digital Audio Visual Integrated Database). *Swnasea University, Electrical and Electronic Department*, 1996.

[23] K. Yu, J. Mason, and J. Oglesby. Speaker recognition using hidden Markov models, dynamic time warping and vector quantisation. *IEE proc. vision, image and signal processing*, 142:313–318, 1995.