ANALYSIS OF AUDIO-VIDEO CORRELATION IN VOWELS IN AUSTRALIAN ENGLISH

Roland Goecke[1]*, J Bruce Millar*[1]*, Alexander Zelinsky*[2]*, and Jordi Robert-Ribes*[3]

[1]Computer Sciences Laboratory and [2]Robotic Systems Laboratory, Research School of Information Sciences and Engineering, Australian National University, Canberra ACT 0200, Australia
[3]Cable & Wireless Optus, 101 Miller St, North Sydney NSW 2060, Australia

E-Mail: Roland.Goecke@anu.edu.au URL: http://cslab.anu.edu.au/~rgoecke

ABSTRACT

This paper investigates the statistical relationship between acoustic and visual speech features for vowels. We extract such features from our stereo vision AV speech data corpus of Australian English. A principal component analysis is performed to determine which data points of the parameter curve for each feature are the most important ones to represent the shape of each curve. This is followed by a canonical correlation analysis to determine which principal components, and hence which data points of which features, correlate most across the two modalities. Several strong correlations are reported between acoustic and visual features. In particular, F1 and F2 and mouth height were strongly correlated. Knowledge about the correlation of acoustic and visual features can be used to predict the presence of acoustic features from visual features in order to improve the recognition rate of automatic speech recognition systems in environments with acoustic noise.

1. INTRODUCTION

Although automatic speech recognition (ASR) systems have become common tools in human-computer interaction (HCI), they still have some limitations with respect to the environment in which they can be used. Current commercially available ASR systems employ statistical models of spoken language and enable continuous speech recognition in reasonably good acoustic conditions. However, they can fail unpredictably in noisy conditions. One way of overcoming some of the limitations of audio-only ASR systems is to use the additional visual information of the act of speaking [1-5].

The aim in audio-video (AV) ASR in adverse conditions is to be able to replace less reliable acoustic measurements with more reliable visual measurements. This requires a-priori knowledge of the correlation between acoustic and visual speech features for all phonemes. We explore in this paper how this knowledge can be established.

Yehia et al [6] found a strong correlation (80-91%) between the shape of the vocal-tract and the position of facial feature points around the mouth and lower face. They also found that a large part (72-85%) of the variance observed in acoustic parameters can be determined from vocal-tract and facial data together. And even the facial data alone performed well in accounting for the acoustic parameter variance. The drawbacks of their study were the small number of speakers looked at (only 2) and the use of intrusive measurement techniques. They used small transducers placed on tongue, lips and teeth for electromagnetic tracking of the vocal-tract motion and infrared LEDs on the lower face half for optical tracking of the facial motion. While Yehia et al focussed on the relation between acoustic and facial parameters for speech production and animation, we look at the issue from an ASR perspective.

We have recently presented a novel algorithm for the explicit extraction of lip feature points based on a stereo vision head-tracking system [7]. Both head-tracking and lip-tracking are completely non-intrusive and do not use any facial markers. We have also recorded an AV speech data corpus (with 10 speakers) containing all phonemes and visemes in Australian English for the analysis of the correlation between acoustic and visual features [8]. Section 2 briefly describes the experimental design for the recordings and the techniques used for extracting the features used in the correlation analysis. Section 3 details the methods used in the statistical analysis of the relation between acoustic and visual features. The results are presented in Section 4 and discussed in Section 5. Section 6 concludes with a discussion of the future direction of this work.

2. EXPERIMENTAL DESIGN

2.1. Stereo Vision System

The lip-tracking system is integrated with the stereo vision head tracking system (Figure 1). A stereo vision system has the advantage that depth information can be recovered and that measurements are therefore in 3D, giving real-world distances rather than simply the 2D image coordinates of a mono-vision system.

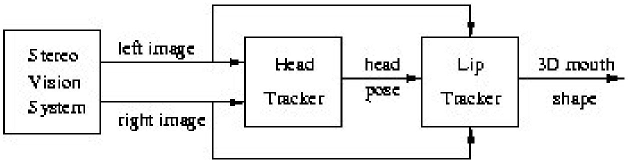

Figure 1: Overview of the stereo vision system

The head tracking system is based on template matching using normalised cross-correlation and is able to track the person's movements at a frame rate of 15-30Hz. The system consists of two calibrated standard, colour analog NTSC video cameras. The camera outputs are multiplexed at half the vertical resolution into a single 512x480 image (Figure 2) before being acquired by a Hitachi IP5005 video card on a Pentium II (300MHz CPU) every 33ms. Details on the system can be found in [9].

The lip-tracking algorithm is applied to the mouth windows which are automatically determined during the head tracking based on the head pose (Figure 2). We combine colour information from the images with knowledge about the structure of the mouth area for different degrees of mouth openness in the algorithm. For example, when the mouth is open, we often expect to see teeth, so we can specifically look for them which improves the robustness of the lip-tracking.

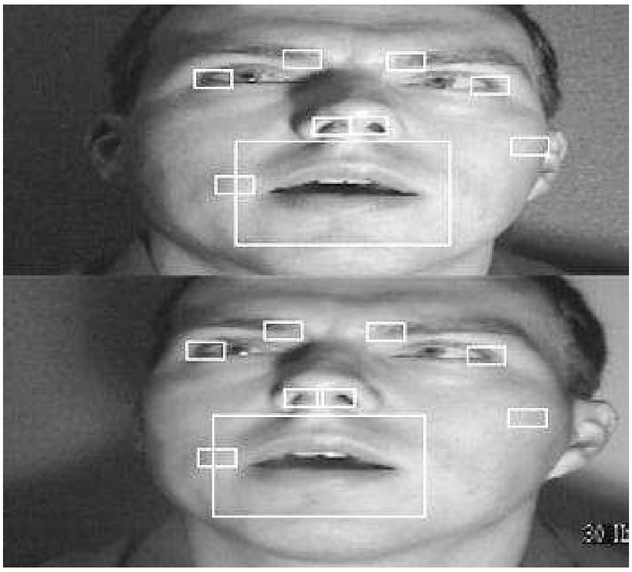

Figure 2: Stereo image with head-tracking templates and mouth windows for lip-tracking.

The algorithm extracts the 3D positions of the ***two lip corners*** and the ***mid-point of upper and lower lips***. Since every person has differently shaped lips, we use the inner lip contour so that the personal characteristic shape of the lips has minimal effect on the measurements. From these four lip points, we derive a feature set which gives a first-order description of the shape of the mouth: ***mouth width***, ***mouth height***, and ***protrusion of upper and lower lip***. Furthermore, the algorithm labels each frame on the appearance of upper and lower teeth. Figure 3 shows two examples of lip-tracking results. Video clips of the lip-tracking process can be found at our homepage (http://cslab.anu.edu.au/~rgoecke). A detailed description of the lip-tracking algorithm can be found in [7].

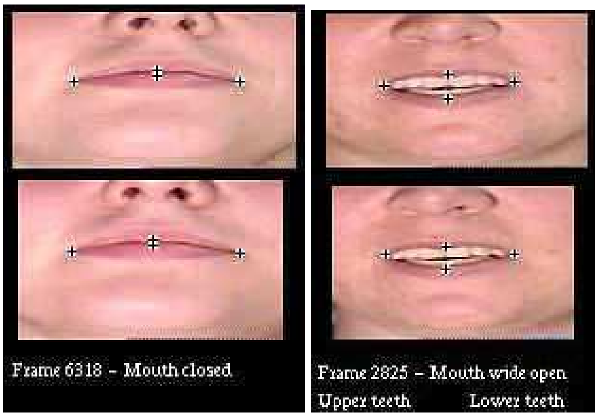

Figure 3: Lip-tracking results.

2.2. AV Speech Data Corpus

We have recorded an AV speech data corpus for Australian English (AuE) using the stereo vision system just described. The speakers sat in front of a stereo camera pair with an omnidirectional microphone attached 20-25cm below their mouth (Fig. 4). The face was well illuminated by a light source just below the cameras so that no shadows appear on a speaker's face. Recordings were made to digital video (DV) tape because of its ability to playback the recordings many times without a loss of quality. The DV standard also comprises a digital audio component. In our case, the recordings were made at 30Hz video frame rate and 16bit 48kHz mono audio rate in a controlled acoustic environment (almost no external noise, some air conditioning and computer noise in the background).

The data corpus comprises 10 native speakers (5 female and 5 male speakers). It was designed to cover all phonemes and visemes in AuE except for the neutral vowel /@/ because of its great audio variability and the neutral consonant /h/ which adds little to the correlation analysis. In addition, the voiced fricative /Z/ and the diphthong /u@/ were also omitted because they have a low occurence in AuE. The core part of the corpus consists of 40 sequences per speaker containing consonant-vowel-consonant- (CVC-) or vowel-consonant-vowel- (VCV-) words with the phoneme of interest in the central position. These words were put in a carrier phrase ("You grab ***word*** beer.") to overcome articulation patterns associated with reading words from a list. The carrier phrase also facilitates the visual segmentation through the use of bi-labial closings before and after the CVC- or VCV-word. Full details on the design of the AV speech data corpus can be found in [8].

Figure 4: Setup for AV Recordings.

2.3. Acoustic Feature Extraction

First, the audio component was grabbed from the DV tape and stored as a .wav file without changing the bit rate and sampling rate settings. Then the ESPS signal processing toolkit was used for the extraction of the acoustic features. The acoustic features extracted were the voice source excitation frequency f0, the formant frequencies F1-F3 and RMS energy. The ESPS function *get_f0* was used to extract the f0 and RMS energy values. It uses a 30ms Hanning window with 5ms overlap. The ESPS function *formant* was used to extract F1, F2, and F3 from the audio data by linear prediction analysis. The ESPS standard parameters of 49ms window length and $\cos^4$ window type were applied in the analysis.

3. METHODS FOR ANALYSIS

In this paper, we focus on the analysis of the AV correlation in vowels in AuE. While our AV speech data corpus also contains sequences of the consonants in Australian English, their AV correlation analysis will be reported on in a different paper. This categorisation into vowels and consonants is based on the phonemes since we collected data from the phonemes of AuE rather than the visemes. Of course, visemes and phonemes are related but the sets have a different structure. There is no 1-1 map between phonemes and visemes. Leaving diphthongs aside, there are sequences for 6 short and 5 long vowels of AuE in the data corpus.

First, a linear correlation was performed on the visual feature set to see if any of the features were correlated. Not surprisingly, it turned out that the protrusion of upper and lower lip were highly correlated ($r > 0.97$) because in normal speech both lips are moved backward and forward simultaneously. Therefore, we will look only at one protrusion parameter in the analysis, the upper lip protrusion, for example. Table 1 shows the feature set used in the statistical analysis.

Acoustic	Visual
Voice source excitation f0	Mouth height
Formant frequency F1	Mouth width
Formant frequency F2	Lip protrusion
Formant frequency F3	
RMS energy	

Table 1: Overview of the acoustic and visual features

For the vowels, the CVC-word was in the form of **/bXb/** where **X** is substituted for the particular vowel. The parameters of all features were extracted for the period of time from the beginning of the lip opening after the initial bi-labial **/b/** until the end of the lip opening just before the second bi-labial **/b/**. These time periods were about 200ms (or 6 video frames) for short vowels and about 300ms (or 9 video frames) for long vowels. However, durations also vary across speakers for each vowel. In order to have the same number of points on the time scale for all utterances of a vowel from all speakers, a piecewise linear interpolation was performed to give each parameter curve a base of 50 points for acoustic features and 10 points for visual features. We used the R statistical system [10] for the interpolation and all subsequent statistical analyses. Note that our use of the PCA to represent the shape of a curve is immune to the fact that acoustic and visual features have a different number of data points. Obviously, the analysis frame rate is different for acoustic features (every 10ms) and visual features (every 33ms).

Having a bi-labial context simplifies the visual analysis. Using /b/ instead of /p/ lengthens each word and thus results in more data to analyse, which particularly important in the case of short vowels. On the other side, a bi-labial context causes strong coarticulation effects in the formants. However, these are quite predictable for /b/ and we believe the advantages of a bi-labial context for visual segmentation offset the disadvantages. We limited coarticulation deliberately to constrain the size and complexity of our data set.

For each parameter data point and for each phoneme, a principal component analysis (PCA) was performed across the 10 speakers. It is important to understand that this use of the PCA technique is a way to represent the shape of a parameter curve. For example, the first principal component (PC) will then tell us the linear combination of which data points accounts for the largest amount of the variance in the parameter curves of all speakers for that one particular phoneme. This use of the PCA technique is different from applying it to all, let's say, acoustic features in order to reduce the number of features by selecting the ones that account for most of the variance.

Then the top two PCs of each feature representing about 70% of the variance were taken and two features (= 2x2 PCs) at a time combined as acoustic and visual PC vectors. Finally, a canonical correlation on these PC vectors was done to quantify which PCs, and subsequently which data points of which features, were correlated across the two modalities. This also reveals temporal information about the relationship between acoustic and visual features. We tested several combinations of PC vectors but we have not yet tested all possible combinations. Figure 5 shows a schematic example of the mouth height parameter curve. Black dots on the parameter curve represent data points with correlation values above a certain threshold, for example 0.9.

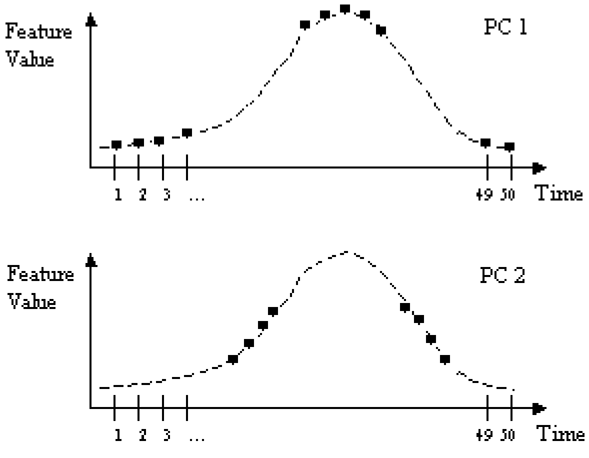

Figure 5: Schematic example of relationship between data points and two principal components.

Canonical correlation is a form of correlation relating two sets of variables. Similar to factor analysis, there is more than one canonical correlation, each representing orthogonally separate patterns of relationships between the two sets. The first canonical correlation is always the one which explains most of the relationship. We only look at the first correlation in our analyses.

4. RESULTS

4.1. Results of PCA

The first PC of f0 and the formant features typically accounted for most of the variance in the first and in the central data points. To a lesser extent, the first PC was also related to the data points at the end of the vowel. The second PC of these acoustic features accounted for the variance in the data points surrounding the central ones.

The picture is different for the RMS energy feature. Here, the first PC was much more dominant than for the formant features (about 80% compared to about 40-50% of the variance). It used a linear combination of practically all data points to achieve this.

The first two PCs of the mouth height and lip protrusion features exhibit a behaviour similar to the one of the formant features. However, for the mouth width feature, the first PC was rather related to the central data points and the second PC to the first and the last data points.

4.2. Results of Canonical Correlation

Table 2 shows, as an example, the r-values of the first canonical correlation with the first two PCs of F1 and F2 as audio variables and the first two PCs of mouth height and mouth width as video variables. Generally, we found that combinations of the first two PCs of F1, F2, and F3 correlated strongly with the PCs of mouth height. We also found a strong correlation between the first PC of f0 and the PCs of the mouth height, but not for the second PC of f0. The PCs of F2 and F3 also correlated well with the first PC of the mouth width. Finally, there was a strong correlation between RMS energy and F2 on one side and mouth height on the other side. No indication of a linear relationship between the data points of the protrusion parameter and the data points of any of the acoustic features was found. Short and long vowels showed similar results.

Short Vowels		**Long Vowels**	
/A/	0.99	/a:/	0.96
/E/	0.96	/@:/	0.98
/I/	0.95	/i:/	0.96
/O/	0.98	/o:/	0.94
/U/	0.96	/u:/	0.97
/V/	0.94		

Table 2: r-values of first canonical correlation with the first two PCs of F1 and F2 as audio variables and the first two PCs of mouth height and mouth width.

5. DISCUSSION

It must be said that this is an exploratory investigation. We are aware of the fact that a data corpus with 10 speakers is still of a fairly small size for such an analysis. In particular, one would like to incorporate more than the first two PCs to account for a higher percentage of variance in the utterances for each of the vowels. We have immediate plans to extend our data corpus by adding more speakers so that our analysis can include more PCs of each feature to give a more accurate picture.

One can also argue that a piecewise linear interpolation in the process of giving all parameter curves the same number of data points is less effective than a spline interpolation which would smooth the curve and thus reduce the effects of measurement errors. This needs to be investigated. However, for the current analysis we chose a piecewise linear interpolation for simplicity.

We looked at individual phoneme correlations because of the possible many-to-one articulatory-to-acoustic transforms. It has long been known that different articulatory movements can produce acoustically very similar results.

A strong correlation was found between some acoustic and visual features. In particular, the data points of F1 and F2 and the data points of the mouth height were strongly related. Articulatory-to-acoustic speech production theories model are known to model the acoustical consequences of the degree of mouth opening well. Opening or closing the lips has almost immediate acoustical consequences and hence it is no surprise to find a high correlation between the formant frequencies and the mouth height feature. However, since it may be possible that a speaker changes the behaviour of the vocal tract behind the lips to compensate for the acoustical consequences of the lip movements, it is useful to really find these strong correlations.

We would have expected the lip protrusion feature to be of more significance. However, fairly diverse parameter curves were found when looking at each vowel separately across the speakers. This diversity could either be due to speakers using different lip positions to produce the same sound or it could point to inaccuracies in the extraction of the lip protrusion feature from the video signal. It is known that different articulatory movements can still produce the same acoustic result. It is possible that this plays a role here as well. These results need further investigation.

6. CONCLUSION AND FUTURE WORK

To summarise, we performed an exploratory analysis of the statistical relationship between acoustic and visual speech features. A PCA was used to represent the shape of the parameter curves and then a canonical correlation analysis was done using the first two PCs of each feature with two features combined at a time to form the set of variables for this analysis. Strong correlations between the data points of F1 and F2 and the data points of mouth height were found. F3, RMS energy, and mouth width also showed some high correlation values. The lip protrusion feature appeared to be of little significance. The reason for the diversity of shapes experienced in the lip protrusion parameter curves must be further investigated. As mentioned before, not all possible combinations of features have yet been looked at but we will do so in the future.

Strong correlations between some acoustic and visual features means that this knowledge could be used in an AV ASR system to predict the presence of these acoustic features from the visual features. This is of particular importance for the use of ASR systems in environments with acoustic noise.

The data corpus needs to be extended by adding more speakers so that more PCs can be included in the canonical correlation analysis, thus incorporating more of the variance in the features. It can also be argued that repetitions from the same speakers would be of value to look into the intra-subject variability. This is common in speech corpora for speaker recognition but only to a lesser extent in corpora designed for ASR. We will also explore the use of a spline interpolation technique for producing the same number of data points for each utterance before the PCA. Smoother parameter curves may lead to more accurate results.

ACKNOWLEDGEMENT

The authors would like to thank the reviewer of this paper whose comments have helped to make this a better paper.

7. REFERENCES

1. Adjoudani, A., and Benoît, C. “On the Integration of Auditory and Visual Parameters in an HMM-based ASR” In Stork, D.G. and Hennecke, M.E. (editors): *Speechreading by Humans and Machines*, Vol. 150, NATO ASI Series, Springer-Verlag, 1996, p461-471.

2. Bregler, C., and Konig, Y. “’Eigenlips’ for Robust Speech Recognition” *Proc. of ICASSP’94*, Vol. II, Adelaide, Australia, 1994, p669-672.

3. Matthews, I., Cootes, T., Cox, S., Harvey, R., and Bangham, J.A. “Lipreading using shape, shading and scale” *Proc. of AVSP’98*, Terrigal, Australia, 1998, p73-78.

4. Petajan, E.D. “Automatic Lipreading to Enhance Speech Recognition” PhD thesis, University of Illinois at Urbana-Champaign, 1984.

5. Stork, D.G., and Hennecke, M.E. (editors) “Speechreading by Humans and Machines”, Vol. 150, NATO ASI Series, Springer-Verlag, 1996.

6. Yehia, H., Rubin, P., and Vatikiotis-Bateson, E. “Quantitative association of vocal-tract and facial behavior” *Speech Communniation*, 26 (1-2): 23-43.

7. Goecke, R., Millar, J.B., Zelinsky, A., and Robert-Ribes, J. “Automatic Extraction of Lip Feature Points” *Proc. of the Australian Conference on Robotics and Automation ACRA2000*, Melbourne, Australia, 2000, 31-36.

8. Goecke, R., Tran, Q.N., Millar, J.B., Zelinsky, A., and Robert-Ribes, J. “Validation of an Automatic Lip-Tracking Algorithm and Design of a Database for Audio-Video Speech Processing” *Proc. of the 8th Australian International Conference on Speech Science and Technology SST-2000*, Canberra, Australia, 2000, 92-97.

9. Newman, R., Matsumoto, Y., Rougeaux, S., and Zelinsky, A. “Real-time stereo tracking for head pose and gaze estimation” *Proc. of Automatic Face and Gesture Recognition FG2000*, Grenoble, France, 2000.

10. Gentleman, R., and Ihaka, R. “The R Project for Statistical Computing” http://www.r-project.org, 2000.