VALIDATION OF AN AUTOMATIC LIP-TRACKING ALGORITHM AND DESIGN OF A DATABASE FOR AUDIO-VIDEO SPEECH PROCESSING

Roland Goecke[1], Quynh Nhu Tran[2], J Bruce Millar[1], Alexander Zelinsky[3] and Jordi Robert-Ribes[4]

[1]Computer Sciences Laboratory and [3]Robotic Systems Laboratory, Research School of Information Sciences and Engineering, Australian National University, Canberra ACT 0200
[2]Information Systems Group, University of Newcastle, Callaghan NSW 2308
[4]Cable & Wireless Optus, 101 Miller St, North Sydney NSW 2060
E-Mail: Roland.Goecke@anu.edu.au

ABSTRACT: We have recently proposed a new algorithm for the automatic extraction of lip feature points. Based on their positions, parameters describing the shape of the mouth are derived. Since the algorithm is based on a stereo vision face tracking system, all measurements are in real-world distances. In this paper, we evaluate the accuracy of the automatic feature extraction algorithm by comparing its results with a manual feature extraction process. The results show an average error of about 1-2mm for the internal mouth width and height. In the second part of the paper, we present the design of an AV speech database for Australian English for future experiments on the correlation of audio and video speech signals.

INTRODUCTION

Automatic speech recognition (ASR) systems typically use statistical models of spoken language and allow continuous speech recognition in reasonably good acoustic conditions. In noisy conditions, however, these systems can fail unpredictably. 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. An automatic speech-reading, or lip-reading, system as part of an audio-video speech processing (AVSP) system is expected to improve ASR in noisy conditions and can thus lead one step closer to more natural human-computer interactions.

While there exist well-established ways of analysing the acoustic speech signal, it still is an open research issue which parameters describe the visual speech information best. Two main approaches to extracting visible speech features can be identified. First, *implicit feature extraction methods* use the raw image data as input to a recognition engine (Hidden Markov Model, artificial neural network) which learns the pixel patterns associated with certain lip movements (Meier et al., 1996, Movellan & Chadderdon, 1996). Optical flow techniques also fall into this category (Mase & Pentland, 1991). A principal component analysis (PCA) is often used to reduce the dimensionality of the input vector and to define the main directions of variation. In the second approach, *explicit feature extraction methods* use image processing techniques to find the location of certain feature points (e.g. lip corners) in the image. Methods range from simple image-based approaches, such as thresholding (Petajan, 1984) or integral projection (Yang et al., 1998), to complex model-based approaches, such as active contour models (Bregler & Omohundro, 1994), active shape models (Luettin et al., 1996), or 3D lip models (Revéret & Benoît, 1998, Basu et al., 1998).

We have recently (Goecke et al., 2000) proposed a new algorithm for the explicit extraction of the positions of the lip corners as well as the mid-points of upper and lower lip on the inner lip contour (Figure 1). When the mouth is fully closed, the inner lip contour is not visible. In that case, we define the shadow line between the lips as the inner lip contour. The algorithm uses a combination of colour image data and knowledge about the structure of the mouth area. It classifies the openness of the mouth into three discrete states. In addition, the presence of teeth is also detected. We want to use this explicit feature extraction algorithm to analyse the correlation between the audio and video part of the speech data. We are particularly interested in determining which visual speech parameters carry redundant information and can thus be used to improve recognition rates of ASR systems in noisy conditions.

AUTOMATIC LIP FEATURE EXTRACTION

Our lip-tracking algorithm builds on a completely non-intrusive stereo vision face tracking system on a Pentium II PC (Newman et al., 2000) which estimates the head pose in 3D.

Figure 1. Examples of lip features being tracked on the inner lip contour

Stereo vision has the advantage of obtaining measurements in real-world coordinates and has not been used for AVSP before to our knowledge. The speaker is well-illuminated by a single light source directly below the cameras to avoid inaccuracies caused by shadows. The mouth area in both camera images is automatically determined and further processing limited to that area (Figure 2). A combination of intensity and saturation values is used to find the feature points in the stereo images in three steps. First, classify the mouth according to its openness into one of the three states: *closed*, *partially open*, or *wide open*. This determines the techniques used in the next steps. Secondly, find the lip corners in each image of the stereo image pair. Then calculate their respective 3D positions using the known camera parameters. Thirdly, find the exact positions of the lip mid-points in 3D.

Figure 2. Automatically identified mouth regions in a stereo image pair

Our parameter set consists of *mouth width*, *mouth height*, *protrusion of upper lip*, *protrusion of lower lip*, and the *ratio of mouth width to mouth height*. The protrusion parameters are determined as the shortest distance from the mid-lip points to the straight line from one lip corner to the other. The ratio parameter is included as a potentially more useful way of describing the protrusion because preliminary experiments showed that the other two protrusion parameters can be unreliable. Furthermore, each frame is labelled on the visibility of the upper and/or lower teeth. A confidence measure based on the difference between the corresponding 2D mouth width and mouth height distances in the stereo images is calculated. If the 2D distances in the two images do not agree, then that is an indication that the algorithm has failed and the results are marked as unreliable.

VALIDATION OF THE AUTOMATIC FEATURE EXTRACTION ALGORITHM

Visual inspection of the extracted feature positions shows a high degree of accuracy (Figure 1). The algorithm fails only in a few frames which are well detected by the confidence measure. To quantify the error, a ground-truth would be required but cannot be obtained for practical reasons. However, a software tool was developed (Tran, 2000) which compares the automatically extracted feature positions with the results from a manual extraction by the user. Although the manual selection of features potentially introduces a new source of error, the user, it gives a clear indication on the goodness of the automatic feature extraction.

So far, the algorithm has been tested on three speakers. Each subject was asked to speak three sequences exhibiting calibratory mouth movements:

1. ba ba ba ...
2. e o e o e o ...

3. Joe took father's green shoe bench out.

The first sequence maximises vertical lip movement, while the second sequence emphasises horizontal lip movement (rounding and stretching). The third sequence was taken from the design of the XM2VTSDB database (Messer et al., 1999). It covers all viseme and phoneme categories in the English language. Each sequence is about 4s long.

The user clicks the mouse on the position of each of the four lip feature points in each frame of the sequences. While this is a tedious and error prone process for longer sequences, it is feasible for these short sequences. The parameter set is computed in the same way in the automatic extraction algorithm. Then, for each feature point and for each parameter, the average absolute error and the standard deviation is calculated. The results are shown for each of the three sequences in Tables 1-3.

	Mouth Width	***Mouth Height***	***Protrusion Upper Lip***	***Protrusion Lower Lip***	***Width / Height***
Subject MH	1.7 ± 2.1mm	1.9 ± 2.4mm	7.8 ± 9.5mm	6.2 ± 7.8mm	1.3 ± 2.1
Subject TG	2.8 ± 10.2mm	1.9 ± 2.6mm	7.1 ± 8.4mm	6.0 ± 7.6mm	5.1 ± 8.6
Subject RG	0.8 ± 1.0mm	0.8 ± 1.0mm	6.7 ± 8.6mm	4.8 ± 6.2mm	4.1 ± 11.1

Table 1. Average absolute error and standard deviation for sequence 1: "ba ba ba ..."

	Mouth Width	***Mouth Height***	***Protrusion Upper Lip***	***Protrusion Lower Lip***	***Width / Height***
Subject MH	2.6 ± 4.0mm	1.4 ± 1.8mm	3.5 ± 4.5mm	5.9 ± 7.3mm	2.0 ± 3.9
Subject TG	1.6 ± 2.9mm	1.1 ± 1.4mm	6.0 ± 7.6mm	5.3 ± 7.4mm	0.7 ± 1.2
Subject RG	1.7 ± 2.5mm	1.9 ± 2.6mm	4.2 ± 6.0mm	4.8 ± 7.0mm	0.5 ± 0.7

Table 2. Average absolute error and standard deviation for sequence 2: "e o e o e o ..."

	Mouth Width	***Mouth Height***	***Protrusion Upper Lip***	***Protrusion Lower Lip***	***Width / Height***
Subject MH	2.3 ± 3.0mm	1.7 ± 2.3mm	6.6 ± 8.5mm	5.8 ± 7.5mm	1.5 ± 2.6
Subject TG	1.2 ± 1.6mm	2.2 ± 2.7mm	4.9 ± 6.5mm	5.0 ± 6.6mm	1.9 ± 2.8
Subject RG	1.9 ± 2.8mm	1.2 ± 1.6mm	4.4 ± 5.7mm	3.4 ± 4.5mm	0.9 ± 2.1

Table 3. Average absolute error and standard deviation for sequence 3: "Joe took ..."

The comparison shows that the manual and automatic feature extractions yield similar results. They only differ at about 1-2mm for the mouth width and mouth height. This a very accurate result given that we use a totally non-intrusive lip tracking algorithm with no additional markers or made-up lips. The standard deviation is 1-3mm except for subject TG in sequence 1 and subject MH in sequence 2. Their larger values result from outliers where the algorithm failed to find a feature position close to the true position. However, these failures were identified by the confidence measure. If we exclude these outliers from the analysis, than the results from these two sequences are similar to the others.

Looking at each feature point separately reveals that the difference in the vertical position is less than 1 pixel. However, the difference in the horizontal position is 2-3 pixels on average. Using the stereo disparity to recover depth information requires the same point being identified in both images. Incorrectly placed feature points lead to less accurate depth information which affect the accuracy of our parameters. The algorithm was less accurate than the human observer in the horizontal placement of the lip corners for a partially open mouth because the inner lip contour could not be clearly distinguished from the lip flesh in the lip corner area. On the other hand, the algorithm was more accurate than the human observer in the horizontal placement of the mid-points of the lips. This is because it calculates the horizontal position as being the one of the middle point between the lip corners and thus ensures a corresponding horizontal position in both stereo images, whereas the user

is allowed to move the mouse freely. A more restrictive approach in which the human observer is guided in a similar way as the automatic approach would yield even more accurate results. The error for the protrusion parameters is considerably larger. The reasons need to be investigated in more detail but problems in finding corresponding points in both images correctly will affect the protrusion parameters more strongly. The width-to-height ratio is more accurate but also has its limitations. When the mouth is closed, small changes in the measured mouth height trigger large changes in the ratio. Furthermore, the difference between a small round mouth and a wide open mouth cannot be told from the ratio because it is in both cases close to 1.

DESIGN OF THE AUDIO-VIDEO SPEECH DATABASE

We want to use our lip-tracking algorithm to investigate the correlation between audio and video representations of spoken language. It has been shown in the past that the incorporation of information about visible speech articulation in an ASR system can improve the recognition rate (cp. Stork & Hennecke, 1996). This suggests that there is redundant information in the two modalities that can be used if one of the two channels is affected by noise. All practical systems face noise, be it acoustic noise, such as background noise, or visual noise, such as changes in illumination. However, little work has been done on which audio and video parameters correlate most. The ultimate aim of the project is to develop an adaptive algorithm which will flexibly rely more strongly on one or the other modality depending on their respective noise levels. This requires a-priori knowledge of the varying correlation of audio and video speech signals for a variety of speakers. We have therefore built an AV speech database for the analysis of the underlying correlation. The data corpus was systematically designed to cover the range of phonemes and visemes in Australian English in different situations while keeping the size reasonably small for practical reasons. The database contains the following sequences.

I. Definition of the face model

A sequence is needed for establishing the face model for the stereo vision face tracking system. Facial landmarks, like eye corners, are selected manually from frontal as well as left and right 45° views, before the face tracking system automatically computes a 3D face model from these landmarks. The face model is needed for estimating the general head pose.

II. Calibration sequences

The sequences "ba ba ba ..." and "i o i o i o ...", each repeated for about 10 seconds, are used for calibrating the speakers. The first of the two sequences gives insight into the amount of vertical lip movement, while the second sequence emphasises horizontal lip movement. These sequences can also be used for determining the visual expressiveness of the speaker. For each sequence, the average maximum mouth height and width, respectively, are determined for each speaker. The mean value and the standard deviation of each of those two parameters over all speakers can then be computed. Speakers with values in the margin of the overall distribution can be excluded from the subsequent analysis if desired.

III. Words

This is the core part of the database for the correlation analysis. There are 44 phonemes (24 consonant and 20 vowel ones) and 11 visemes (7 consonant and 4 vowel ones) in Australian English (Plant and Macrae, 1977). According to the ANDOSL design, the phonemes can be categorised into oral stops (/p/, /b/, /t/, /d/, /k/, /g/), fricatives (/f/, /v/, /T/, /D/, /s/, /z/, /S/, /Z/, /h/), affricatives (/tS/, /dZ/), nasals (/m/, /n/, /N/), liquids and glides (/l/, /r/, /w/, /j/), short vowels (/I/, /U/, /E/, /@/, /O/, /V/, /A/), long vowels (/i:/, /u:/, /@:/, /o:/, /a:/), and diphthongs (/ei/, /e@/, /@u/, /oi/, /ai/, /au/, /i@/, /u@/). Following Plant and Macrae (1977) and Plant (1980), visemes of Australian English can be classified into bi-labials (/p/, /b/, /m/), labio-dentals (/f/, /v/), inter-dentals (/T/, /D/), labio-velar glides (/w/, /9/), palatals (/S/, /tS/, /dZ/), alveolar non-fricatives and plosives and velar plosives (/l/, /n/, /j/, /g/, /k/), alveolar fricatives and plosives (/s/, /d/, /t/), front non-open vowels and front-close-onset diphthongs (/i/, /i:/, /I/, /E/, /i@/), open vowels and open-onset diphthongs (/A/, /ai/, /e:/, /V/, /@/, /ei/), back/central non-open vowels and diphthongs containing these vocalic positions (/@:/, /oi/, /u:/, /iu/, /U/), and back/central open vowels and diphthongs containing these vocalic positions (/O/, /@u/, /au/).

The phonemes and visemes are studied in a VCV- or CVC-context to be free of any phonological or lexical restrictions. The vowel context is the wide open "ar". The voiced bi-labial /b/ is used as the consonant context. Having a bi-labial context simplifies the visual analysis. Using /b/ instead of /p/

lengthens each word and thus gives more data to analyse, in particular for short vowels. On a negative 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 having a bi-labial opening and closing before and after the word for visual segmentation offsets the disadvantages from the coarticulation.

In order to overcome the typical articulation patterns associated with reading words from a list, each word is put in a carrier phrase which reads "You grab /word/ beer." This is particularly important for the VCV-words which have no bi-labial context otherwise. The list, ordered into phoneme categories, then consists of:

- Short vowels: Bib, Bub (as in “should”), Beb, Bob, Bub (as in “cup”), Bab
- Long vowels: Beeb, Boob, Berb, Borb, Barb
- Diphthongs: Babe, Bareb, Boyb, Bibe, Bowb, Beerb, Bobe
- Oral stops: Arpar, Arbar, Artar, Ardar, Arkar, Argar
- Fricatives: Arfar, Arvar, Arthar (as in "thin"), Arthar (as in "that"), Arsar, Arzar, Arshar
- Affricatives: Archar, Arjar
- Nasals: Armar, Arnar, Arngar
- Liquids and glides: Arlar, Ara, Arwar, Aryar.

Close examination of this list will reveal that a few phonemes are missing. These are the voiced fricative /Z/ as in "azure" and the diphthong /u@/ as in "tour". These phonemes have a low occurence in Australian English. It was therefore quite likely that speakers would not pronounce the prompts for these two phonemes correctly which will negatively affect the analysis. They were thus omitted in the word list. Besides, these two phonemes were rather difficult to achieve in the suggested CVC- and VCV-contexts. Furthermore, the neutral vowel /@/ and the neutral consonant /h/ were left out because they add little to the correlation analysis. Thus we have a total number of 40 words.

IV. Digits
This sequence and the next one can be used as examples of applying the knowledge gained from the analysis of the data in the previous word sequences to short sequences that are more application-driven. The database includes a sequence of digits spoken in order from 0 to 9. Again, each digit is put in the same carrier phrase "You grab /digit/ beer." to ensure lip closure before and after the digit.

V. Continuous speech
Finally, the database includes an example of continuous speech for each speaker. This is in the form of the sentences: "Joe took father's green shoe bench out. Yesterday morning on my tour, I heard wolves here. Thin hair of azure colour is pointless." Analysing these sentences in more depth reveals that all phonemes and visemes of Australian English as listed above occur in them.

This data corpus comprises 10 speakers and equal numbers of female and male speakers ensure a fair gender balance. All speakers are native Australian English speakers and were 23-40 years old at the time of the recordings.

CONCLUSIONS AND FUTURE WORK

We have presented a way of validating the results from an automatic lip feature extraction algorithm. The results from a comparison with a manual feature extraction indicate strongly that the algorithm is able to automatically, reliably and accurately find the features in which we are interested. Internal mouth width and height were measured with an accuracy of about 1-2mm. The protrusion parameters for upper and lower lip in their current way are of limited use. This is due to the fact that when the lips are protruded, all parts move more or less together forward. Thus, there is little or no change of the protrusion of the mid-points of the lips compared to the lip corners. The protrusion should instead be measured against some other facial part which is more stationary, such as the nose tip, for example.

We have built an AV speech database for Australian English. This data corpus will be used to investigate the correlation between audio and video in spoken language. The analysis will start with the parameter set described but we also want to examine how these parameters change over time, i.e. what are their velocity and acceleration patterns. As others have suggested, motion might be a more important cue than shape in AVSP (e.g. Vatikiotis-Bateson et al., 1996). We are particularly interested

in determining where the redundant information in AV data lies. Once the correlation between audio and video parameters is known, this knowledge can be applied to ASR systems to improve their recognition rate in noisy environments where today's systems often fail.

REFERENCES

Basu, S., Oliver, N. & Pentland, A. (1998) "3D lip shapes from video: A combined physical-statistical model", Speech Communication 26(1-2), 131-148.

Bregler, C. & Omohundro, S.M. (1994) "Surface learning with applications to lipreading", In Cowan, J.D., Tesauro, G. & Alspector, J. (editors), Advances in Neural Information Processing Systems 6, 43-50.

Goecke, R., Millar, J.B., Zelinsky, A. & Robert-Ribes, J. (2000) "Automatic Extraction of Lip Feature Points", Proceedings of ACRA 2000, Melbourne, in press.

Luettin, J., Thacker, N.A. & Beet, S.W. (1996) "Active Shape Models for Visual Speech Feature Extraction", n Stork, D.G., & Hennecke, M.E. (editors), Speechreading by Humans and Machines, NATO ASI Series 150, 383-390.

Mase, K. & Pentland, A. (1991) "Automatic lipreading by optical-flow analysis", Systems and Computer in Japan 22(6), 67-76.

Meier, U., Huerst, W. & Duchnowski, P. (1996) "Adaptive Bimodal Sensor Fusion for Automatic Speechreading", Proceedings of ICASSP'96, http://werner.ira.uka.de/ISL. publications.html

Messer, K., Matas, J., Kittler, J., Luettin, J. & Maitre, G. (1999) "XM2VTSDB: The Extended M2VTS Database", Proceedings of AVBPA'99, Washington D.C.

Movellan, J.R. & Chadderdon, G. (1996) "Channel separability in the audio-visual integration of speech: A Bayesian approach", In Stork, D.G., & Hennecke, M.E. (editors), Speechreading by Humans and Machines, NATO ASI Series 150, 473-487.

Newman, R., Matsumoto, Y., Rougeaux, S. & Zelinsky, A. (2000) "Real-Time Stereo Tracking for Head Pose and Gaze Estimation", Proceedings of Int. Conf. on Automatic Face and Gesture Recognition FG2000, Grenoble, France, 122-128.

Petajan, E.D. (1984) "Automatic Lipreading to Enhance Speech Recognition", PhD thesis, University of Illinois at Urbana-Champaign, 1984.

Plant, G.L. (1980) “Visual Identification of Australian Vowels and Diphthongs”, Australian Journal of Audiology 2(2), 83-91.

Plant, G.L. & Macrae, J.J. (1977) "Visual Perception of Australian Consonants, Vowels and Diphthongs", Australian Teacher of the Deaf 18, 46-50.

Revéret, L. & Benoît, C. (1998) "A new 3D Lip Model for Analysis and Synthesis of Lip Motion", In Burnham, D., Robert-Ribes, J. & Vatikiotis-Bateson, E. (editors), Proceedings of AVSP'98, 207-212.

Stork, D.G. & Hennecke, M.E. (1996) "Speechreading by Humans and Machines", NATO ASI Series 150.

Tran, Q.N. (2000) "Show Me Your Lips", Tech. report, Computer Sciences Laboratory, RSISE, Australian National University.

Vatikiotis-Bateson, E., Munhall, K.G., Hirayama, M., Lee, Y.V. & Terzopoulos, D. (1996) "The Dynamics of Audiovisual Behaviour in Speech", In Stork, D.G., & Hennecke, M.E. (editors), Speechreading by Humans and Machines, NATO ASI Series 150, 221-232.

Yang, J., Stiefelhagen, R., Meier, U. & Waibel, A. (1998) "Real-time Face and Facial Feature Tracking and Applications", Proceedings of AVSP'98, Terrigal, Australia, 79-84.