A SIMPLE SYSTEM FOR MEASURING AUDIOVISUAL SPEECH

Jordi Robert-Ribes and Bruce Millar

Computer Sciences Laboratory
Research School of Information Sciences and Engineering
Australian National University

ABSTRACT

A system which allows the collection of audio-visual speech data is described. The system is described at the level of the hardware and software involved, and at the level of the precision of the data collected. The key issues of audio-video synchronisation and the standards required for the study of speech, the formal description of both audio and video signals, and the extraction and manipulation of audio and video features are examined in some detail. An example of the performance of the system is presented as the output of a novel software analysis package which has been developed for use with this system. Further developments of this software package are outlined.

INTRODUCTION

The recognition that the production of speech is indeed a multimedia event and that the perception of speech involves the use of multimodal cues is not new, but the facility to treat speech in both its audio and video components and to model both auditory and visual mechanisms for its interpretation has brought this full-orbed approach to the study of speech more and more to the fore in recent times. Specifically, the study of the visual component of speech, represented by lip movements, is a new and growing domain. Measuring the shape of the lip aperture and the movements of the lips represents the foundational data for such studies. Such measurements can be made very accurately with sophisticated equipment, both hardware and software, that is available to only few research teams (Lallouache, 1990; Le Goff et al., 1994; Cosi and Caldognetto, 1996). The aim of the current project is to investigate the potential for generating useful data using the relatively unsophisticated hardware that is becoming increasingly available on multimedia workstations (a PC running under Windows). Such data will be useful only if it has acceptable accuracy and can be acquired in an automated fashion.

In this paper we examine two groups of issues. The first group comprises those issues relating to the quality of audiovisual speech data including the necessary synchronisation of the audio and video streams as well as adequate description of the audio and video environments. The second group comprises issues relating to the extraction and presentation of speech-related features in the two data streams including the identification of the lips and the reduction of the lip-image to parametric form.

DEFINING AUDIO-VISUAL QUALITY

It is clearly easier to achieve high quality data using sophisticated audio and video interfaces which are characterised by both high speed and high resolution and whose synchronisation is aided by special purpose electronics. Our concern has been to explore the use of audio-visual speech processing in a common human-computer interface environment using a minimum of special-purpose hardware and software. Hence we examine a methodology for recording useful audio-visual speech signals with a PC equipped with a video board and an audio board. Another constraint of this environment is the potential lack of control over the lighting conditions and the position of the speaker from both microphone and camera. We have therefore paid some attention to the issue of self-calibrating techniques which enable the description of the data without resorting to rigid control of the human-computer interface. Accurate specification of recording conditions is necessary in order to generate useful audiovisual databases. This paper focusses only on the specification of the visual component as the auditory component has been addressed in Millar (1992).

Synchronisation - The main issue of concern at the front-end of the recording procedure is the synchronisation between the audio and video data streams. Therefore we must ensure that the audio and video signals are synchronous in the file without any delay from one stream to the other. The source of such delays relate to differences in the initialisation of the audio capture and the video capture hardware and the capacity of the system to maintain the data flows for both data capture subsystems.

While optimum selection of data capture hardware can significantly reduce this problem, we decided that an extrinsic check on synchronisation was justified. The traditional "clapper board" of the film industry was rejected in favour of the metronome. The metronome creates an audio output in synchronisation with the arm of the metronome being vertical. The audio "clicks" are clearly detected in the audio stream and the vertical position of the arm of the metronome is clearly seen in the video stream. Relative to the clapper board technique the metronome provides the extra benefit of continuous video motion through the audio event as well as consistent velocity of motion. These features make it possible to not only to determine within which video image the audio impulse appears but also can give approximate temporal position within that image time span. The necessary measures to allow this synchrony check to be made were achieved by recording a metronome in the background of the speaker. If required we can then resynchronise the audio and video streams of the file by aligning the metronome generated signals.

Luminence and Size - The formal description of the recording conditions is very important in every data collection for further analysis. The two major conditions that we need to consider for video speech are the ambient lighting conditions, and the camera perspective on the face. The ambient light can be described by the luminence, chrominance, and saturation of the light impinging on the face from all directions, and the camera perspective by the azimuth, elevation, orientation, and distance with respect to the lips and its lens characteristics.

In order to provide such descriptors we have devised a novel method to capture the lighting conditions at the effective position of the speaker's face and to determine that effective position with respect to the camera with one simple video exposure. We have developed an artificial face which generates all the necessary information for the normalisation of real facial video data. This information comprises measures of the quantity and quality of the light and its direction, and measures that enable calculation of the physical size represented by one pixel in both the vertical and horizontal dimensions. The artificial face comprises five square panels each of dimension 100mm. Four of the panels are each connected to one side of a central panel and each subtend an angle of 45 degrees to that panel. Figure 1 shows the front and lateral views of this object. It is covered with white paper (CIE Whiteness = 112; Opacity = 91%; Moisture content = 4.8%).

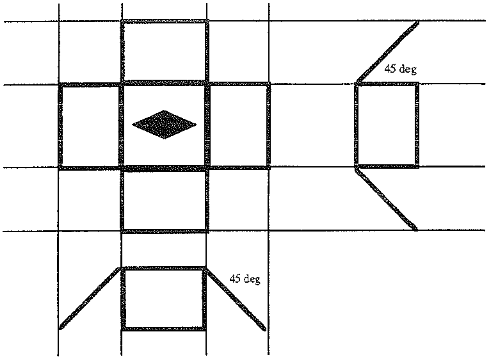

Figure 1

The front panel has a centrally located blue diamond (horizontal size = 90mm; vertical size = 40mm) attached to it. This diamond is used to calculate the size of each pixel by comparing its size in pixels to its known physical size. Once the diamond is located the positions of the centres of the other four panels can be simply calculated. The RGB values of the pixels at each of these positions as well as on the white portion of the front panel describe the quality, quantity and direction of the light. A simple 15 byte value representing either RGB values or the equivalent chrominance, luminance and saturation can thus be logged for each recording session to allow normalisation of lighting differences between sessions.

EXTRACTING AUDIO-VISUAL FEATURES

The extraction of audio-visual features can proceed once an authentic representation of audio-visual data has been achieved. The extraction process must allow calculation of acoustic and facial-image parameters from the basic audio signal and video image. It is also important that these calculated parameters are integrated with the formal description of the recording conditions into a linked data structure so that all subsequent analysis of the data can maintain connectivity with the source of the data.

While there is a large literature on the derivation of well-accepted and robust audio speech parameters from the acoustic signal, the same cannot be claimed for the derivation of video speech parameters from the visual signal. While the literature certainly suggests that the lips are the key visual features carrying speech information, it seems likely that other visual features carry more modest amounts of speech information, or serve to enhance that carried by the lips. Such features as facial muscular distinction, strongly defined nostrils, and the presence of spectacles could all contribute in the latter sense. However, the most crucial feature is surely the degree of "lip definition" depending on lip size and lip colour relative to the colours of both surrounding skin and the oral orifice. Image processing techniques which can cope robustly with such intrinsic facial features are required before lip shape parameters can be derived directly from a natural video image.

At the present time it is necessary to apply artificial lip definition in order to explore our aim of linking together a range of speech processing and image processing algorithms which can be applied to an audio-video data structure. We describe the initial version of a software system to achieve this and indicate its state of development and current capability.

MULTI-MEDIA SPEECH MEASUREMENT SYSTEM

A Multimedia Speech Measurement (MSM) software system has been developed in order to manage, process and display data collected on a standard "multimedia" personal computer system. MSM runs on a PC equipped with the standard Video for Windows software supplied with the Windows operating system. The hardware used to record audiovisual signals comprises a Videolabs FlexiCam camera with stereo microphone and a wide angle (5mm) lens. The video channel is recorded through a FastElectronics FPS60 video board and the audio channel through an IBM M-audio capture and playback board. The standard software for the FPS60 board allows the recording of AVI (Audio Video Interleaved) files comprising one video and two audio streams. AVI files are a standard type of audiovisual files used by Microsoft Windows video applications.

The aim of the MSM software is to facilitate the creation of accurate and well-described audiovisual data, to allow calculation of acoustic and lip-shape parameters from basic signal and image measures, to integrate these measurements and the formal description of the recording conditions into a combined data structure, and to display all these data in a flexible manner.

In version 1 of MSM certain functions are performed with external software, such as AdobePremiere, however, the design of MSM allows an easy implementation of such features as MSM modules in the near future. The remainder of this paper describes its current mode of operation and points to future upgrades.

Recording procedure

Audio-visual data is recorded using the Adobe Premiere software. Our current hardware setting does not allow video and audio capture simultaneously. Therefore we digitise the video images on-line directly to an AVI file. We use an audio tape for recording the audio which is then digitised off-line and integrated into the AVI file. Synchronisation was performed by manual deletion of leading audio samples to bring the audio click of the metronome to the appropriate position within the time-span of one image as judged by the visual position of the metronome arm.

Synchronisation accuracy - The synchronisation estimate between the audio stream and the video stream of the file obtained at the recording stage is achieved by using the signals produced by the metronome. Any misalignment of the audio click and the vertical metronome arm can be corrected by adding or deleting some of the initial audio samples. The accuracy of this process depends on the clarity and resolution of the metronome arm and its fiducial marker and the speed of motion of the arm. It has been found that with metronome mounted as close as possible to the shoulder of the speaker and operating on a high speed (with a period of about 350 ms), a resolution of approximately 2 ms is possible. Such accuracy is deemed adequate for most speech behaviour.

Size and luminence normalisation data - At the start of each recording session the artificial face is placed in the position adopted by the speakers face with the central panel perpendicular to the axis to the camera lens. MSM queries the physical size of the diamond (in metres), detects the left, top, right and bottom points and calculates the width and height expressed in pixels, and then produces the vertical and horizontal scaling factors required. This computation can be performed on a single video image or averaged over several to reduce variance due to the manual positioning of the artificial face.

Chroma-keying - Once the data has been captured, the chroma-keying technique is used to set the artificially coloured lip image to black. This is currently performed using the facilities of Adobe Premiere. The user choses one particular colour in the image (in this case blue) and turns into black all the pixels similar to that colour. This operation could be done in several spaces, such as RGB (Red-Green-Blue) or the chrominance space. Mégard (1995) showed the advantage of using the chrominance space for transforming blue lips into black. When the modified image is transformed back into RGB form (as in the AVI format) it provides significant computational advantages for the tracking of lip position and shape.

Lip measures - The lip-measure algorithm is applied to each image of the video speech data file. It extracts from the AVI file one image and the acoustic samples corresponding to the time span of that image. The video data is processed and the acoustic data is passed to the audio processes. This software can easily process files with different image spatial or temporal resolution and different acoustic formats as all such facets are defined in the AVI file header and are accessed by the MSM system.

The lip measures are the inner and outer lip width (Aint,Aext) and height (Bint,Bext). In order to calculate these measures, the program performs the following steps: (1) it detects and tracks the inner and outer contours of the black lips, (2) it finds the left, top, right and bottom points of both contours, and (3) it calculates the distances between left and right, and between top and bottom for both contours.

These distances are the widths and heights of each contour. They can then be transformed from pixels into metres by using the normalisation factors calculated at a previous stage. The accuracy of these measures depends on (1) the precision of the artificial lip-colouring, (2) the video spatial resolution, (3) the distance between the speaker and the camera, and (4) the algorithm.

Acoustic measures - The acoustic measures of signal energy and the fundamental frequency of voiced excitation are computed using standard algorithms on a frame by frame basis. These values are stored as one value per frame per parameter for display and further analysis purposes.

Display and storage of the output measures - The MSM software allows the output of the measures in two ways: as a display on the screen and exported into a text file.

When displayed on the screen the user can choose to display the video images, the audio waveform, the lip measures and/or the audio measures in the same window with variable temporal definition. Figure 2 shows an example of such a display. The images were 384x288 pixels and each meter corresponded to 1340 pixels vertically and 1500 pixels horizontally.

When exported into a text file, the program generates a text file that contains one line per image and as many columns as the number of measures calculated. In addition it puts in a first column the number of the image and in the first line the name of the measures that have been calculated. Figure 3 shows an example of such a file.

FUTURE DEVELOPMENTS

In order to achieve the outlined aims of MSM, the future developments will be concentrated in two main directions. Firstly, the chroma-keying procedures will be introduced as a module to MSM. Secondly, the measures and details on recording conditions will be stored in the original AVI file by means of creating text streams to that file.

REFERENCES

Cosi P. and Caldognetto E.M. (1996). "Lips and jaw movements for vowels and consonants: Spatio-temporal characteristics and bimodal recognition applications". In D. Stork and M. Hennecke (Eds), Speechreading by Humans and Machines, NATO ASI Series, Springer Verlag, Berlin.

Lallouache M.T. (1990). "In poste 'visage-parole'. Acquisition et traitement des contours labiaux". XVIII Journées d'Études sur la Parole, Montreal, 282-286.

Le Goff B., Guiard-Marigny T., Cohen M., and Benoît, C. (1994). "Real-time analysis-synthesis and intelligibility of talking faces". Proceedings of the Second ESCA/IEEE Workshop on Speech Synthesis, New Paltz, New York, U.S.A., Sept. 1994

Mégard C. (1995). "Elaboration et évaluation d'un système de traitement numérique de la couleur pour l'extraction des lèvres en temps réel". DEA thesis, Institut de la Communication Parlée, Grenoble (France).

Millar J.B. (1992) "The description of spoken language", Proceedings of 4th Australian International Conference on Speech Science and Technology, Brisbane, pp.80-85.

Acknowledgments

We acknowledge the provision of software for contour tracking by the ICP (Grenoble, France), which has been modified for use in this study. The first author acknowledges the provision of a funded Visiting Fellowship by the Research School of Information Sciences and Engineering, Australian National University.

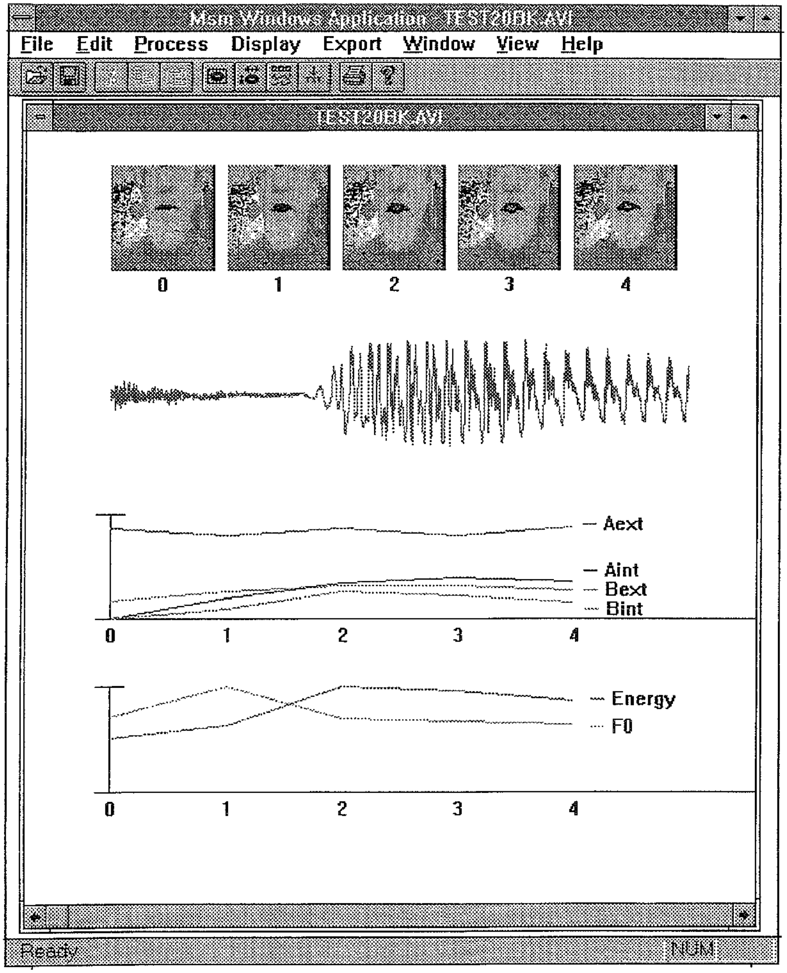
Msm Windows Application - TEST20BK.AVI
File Edit Process Display Export Window View Help
TEST20BK.AVI
0
1
2
3
4
Aext
Aint
Bext
Bint
0
1
2
3
4
Energy
F0
0
1
2
3
4
Ready
NUM

Figure 2

Image	Energy	F0	Aext	Bext	Aint	Bint
0	18.396	153	0.133	0.0282	0	0
1	22.677	216	0.127	0.0415	0.0324	0.0175
2	35.923	151	0.134	0.0498	0.0526	0.0295
3	34.768	145	0.124	0.0498	0.0620	0.0320
4	31.341	141	0.135	0.0435	0.0557	0.0254

Figure 3