AUTOMATED LIP-SYNCHRONIZATION

H.M.L.N.K. Herath[1*], B.P.S. Dayananda[1] and J. Wijayakulasooria[2]

[1]Department of Statistics and Computer Science, Faculty of Science, [2]Department of Electronic and Electrical Engineering, Faculty f Engineering, University of Peradeniya

Introduction

'Lip synchronization' is a technical term for matching lip movements with voice. This is the technique that used in many areas such as computer face animation, virtual television program production, multi-modal user interactive systems, talking-heads, 3D animation films, *etc*. The generating of lip movements in a synthetic face, corresponds to a word is the problem which remain unsolved yet.

The basic focus of this research is to develop an automated lip synchronization technique which operates in real time. It is anticipated that the final system should be capable of analyzing a speech signal and produce the coordinates of the critical points of the lip. The research mainly concentrated on capturing mouth / lip movements and corresponding sounds. It also deals with some attempts at capturing facial movements using limited equipment and some techniques in sound processing which can be used for phoneme recognition.

Materials and Methods

The project mainly divides in to three sub projects phases. They are Video data collecting and processing, audio data collecting and processing and finding a relationship between audio and video data. The lip movement and sound recording is carried out using a Canon XL 1 digital video camcorder with 7.2 V DC power supply. The recordings have a frame rate of 24 fps a 320 x 240 resolution. The recordings are in avi format. The video editing is performed using Adobe Premiere. Video data processing and Audio data processing is carried out by the methods available in 'MATLAB Image processing tool box' and 'MATLAB Auditory Tool Box'. 'Elman' Neural network is used for final audio and video mapping.

Video data collecting and processing

Image segmentation is used to extract the lip contour automatically. The recording is carried out by applying black colour paint on the lip. The Original image was converted to a gray scale image (intensity image) and the image intensity values were adjusted. This maps the values in intensity image(I) to new values in image(J) such that 1 % of data is saturated at low and high intensities of image(I).Then the Features in the image were detected by canny edge detection method followed by the reduction line gaps by dilating the image . Next interior gaps were filled and segmented the objects. Finally objects that connected to the border of the image were removed and smoothed the segmented objects by eroding the image.

Audio data collecting and processing

A phoneme is the smallest posited structural unit that distinguishes

meaning, in human language. Forty four phonemes in English language are selected for recordings. Recording phoneme sounds separately is an impossible task. Thus, the words which include phonemes are selected for the recording. For more accuracy two words from each phoneme sound are selected. Lip movement and audio signals are recorded simultaneously. Using Adobe Premier and GoldWave recording was separated as video and audio. Sound processing was carried out using two techniques, generating wave envelop and Generating Mel-Frequency Cepstrum Coefficients (MFCC) (Axelsson and BjÄorhÄall, 2003, Slaney, 1998)

Neural network training

The lip contour is extracted as described in video data collecting and processing .Each lip contour consists of roughly 1600 points. The recording consisted of more than 3000 frames. It is a difficult task to process this huge amount of data using neural network due to large memory requirements. Therefore a single point was selected to carry out the rest of the project. The point was selected considering the mobility of the point. MPEG-4 defines 84 feature points in natural face. (Pandzic, 2002) (Figure 1)

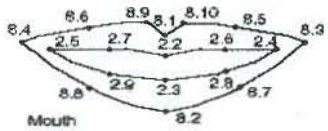

Figure 1. Feature points on mouth

The head move both horizontally and vertically when speaking. Though this is not a large movement when compared to normal speech, when trying to obtain the coordinates of the points it became apparent that the movements were a deterrent. In addition, since there is no steady point, it is difficult to decide on a point to use as the origin (0, 0). The Nose point is used as (0, 0). The nose does not contain muscles and is not mobile when speaking. By considering the feature points on mouth and the coordinate point values, 8.1 point is the most mobile point on mouth. (Figure 2). "Elman" neural network (NN) was selected for audio and video mapping due to its recurrent connection. Normally words are merged , *e.g.* ,the ending phoneme of a word is merged with the starting phoneme of the following word, a phenomenon called co-articulation. The audio data processed by 'Wave envelop' and MFCC and the Y coordinates points were separately fed in to the Elman NN1 and 2 respectively.

Figure 2. Selected point considering the mobility of points

Results and Discussion

The introduced method for lip contour extraction is applied for all the frames in the recording (Figure 3). The proposed method is worked for a constant threshold value and any other constant input argument.

Figure 3. After performing the proposed image segmentation method

According to the tested "Elman" NN1(Figure 4), a huge deviation between the original Y (Y1) and the generated Y coordinated points (Y2) is observed because the lip movement was differed according to the word. Each word has a starting sound and the formation of the lip has happened before the actual sound was generated. The neural network training has been affected by this problem. Hence the tested "Elman" NN2 was obtained (Figure 5), which gives a less deviation, 47.32 % between Y1 and Y2.

Conclusions

The new method has succeeded in extraction of lip movements for certain level. Yet some problems such as shadow detection, detecting gaps between edges were inevitable. To overcome those disturbances usage of monochromatic light, usage of an array of illumination sources can be suggested. But the points extracted from the above method were not used for neural network training. The amount of data was huge and more memory and more speed were needed. It may be appropriate to develop a parallel algorithm to process the data parallel. The speech processing with MFCC is better than wave envelop method for automated lip synchronization systems. The final outcome was improved by using the 'Elman' network ·But the error is high between the original Y coordinated points and the generated Y coordinated points. Therefore, further improvements should be applied to enhance the outcome.

References

Axelsson, A. and BjÄorhÃall, E. (2003).Real Time Speech Driven Face Animation, Masters Thesis at The Image Coding Group, Dept. of Electrical Engineering at LinkÄoping University . 26-30.

Pandzic, I. S. and Forchheimer, R. (2002). MPEG-4 Facial Animation, The Standard, Implementations and Applications, John Wiley and Sons Ltd, The Atrium, Southern Gate, Chichester, West SussexPO19 1UD, England.

Slaney, M. (1998-010). Auditory Toolbox, Version 2 Technical Report Interval Research Corporation.

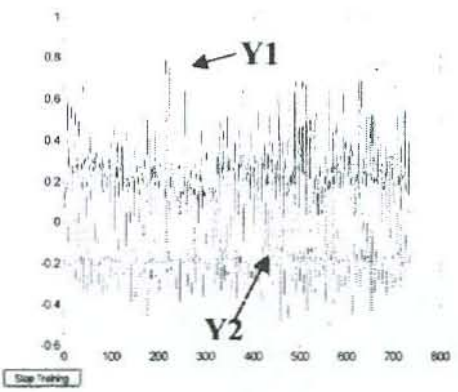

Figure 4 . Tested NN 1 trained by feeding sound envelop with Y1

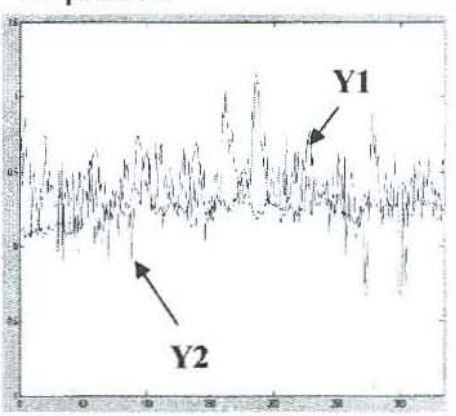

Figure 5 .Tested NN 2 trained by *feeding MFCC with Y1*