User Dependant Speech Based Lip Synchronization

S.M.I. Wickrama[1] and J.V. Wijayakulasuooriya[2*]

[1]*Computing Centre,* [2]*Department of Electrical and Electronic Engineering, Faculty of Engineering, University of Peradeniya*

Introduction

User dependant speech based lip synchronization is an area that is being researched thoroughly. Combination of the auditory and visual speech recognition is more accurate than only auditory or only visual. Consequently, there has been a large amount of research on incorporating bimodality of a speech into the human-computer interaction interfaces. A speech-driven face animation is one of the research topics in this area, since using natural voice for the animation of synthetic faces remains a challenging area of research in computer animation.

Since the FAPs (Facial Animation Parameters) are required to animate faces of different sizes and proportions, the FAP values are defined in face animation parameter units (FAPU). The mouth is one of the most difficult face features to analyze and track. It has a very versatile shape and almost every muscle of the lower face drives its motion, unlike some areas of the face such as hair line or eyebrows which become mobile only when the user uses certain tones and expressions.

The aim of this research is to develop a speech based lip synchronization technique which operates in real time. The final system should be capable of analyzing a speech signal and reproducing the coordinates of the critical points of the lip. The first part of the research was dedicated to investigating models for the automatic lip synchronization by speech signal analysis and to finding a method which is suitable for recording speech and facial movements with the minimum of equipment.

The research dealt with a method by which the key points (critical points) of the lip could be read as coordinates and stored in a suitable format and also a method by which to process speech signals, so it can be fed into a neural network for quick results and easy training.

It is hoped that by further developing this such a method could be used to in the animation industry and for web communication, such as webcasting and user friendly avatars.

Methodology

The first step in this research was to find a suitable recording method for the lip movements and corresponding sound capture. For this, several methods were tried out to see which was the most appropriate for the required project. The recording was done using a Canon XL 1 digital video camcorder with 7.2 V DC power supply. The recordings have a frame rate of 24 frames per second (fps) and a 320x240 resolution. The first recording was in Moving Picture Experts Group (MPEG) format but it was later discovered that Audio Video Interleave (avi) was the better format for picture frame / sound editing.

Most of the avi manipulation and specific frame extraction was done in Adobe Premiere. The coordinates of the key points marked on the lip were obtained using computer graphic methods of a Matlab program. For finer co-ordinate manipulations and storing of data in a more accessible form for those who do not have Matlab (Ver 7.0), Microsoft Excel spread sheets were used. The wave files, in wave and avi format were also processed using Matlab, Adobe Premiere and Goldwave. The neural networks were constructed using Matlab and the output files were also implemented in Matlab graphics (which uses OpenGL which is C based.).

Data collection and pre-processing

The work initially started aiming to develop a system that was capable of simulating a 3D face movement with real time user independent speech. However, the work was limited to 2D face movement capture and lip synchronization for user dependant speech signals due to difficulties that arose during the project. These difficulties included extracting key coordinates from the recording, different levels of frame clarity, the lack of proper lighting equipment for the recordings, etc.

The research mainly concentrated on capturing mouth / lip movements and corresponding sounds. It also dealt with some attempts at

capturing facial movements using limited equipment and some techniques in sound processing which can be used for phoneme recognition.

Several methods were tried out to find a suitable arrangement for the recording to be made. These methods included using the two mirror approach (Lin, I-Chen, Jeng-Sheng, Ouhyoung, Ming 2003 and Whit, Amelia and Lees, 1999) for 3D motion capture, full frontal face capture and finally recording only the mouth and nose area. The marker coordinates on the lip contours were extracted from the recordings and stored in Ms Excel worksheets.

Then, speech analysis was done using cross-correlation of the phonemes used to see if this process could be used for Phoneme recognition. Also, a technique based on Artificial Neural Networks (ANN) was tried to create a user dependant system. In this technique, the speech analysis consisted of programming neural networks by feeding direct speech made up of 44 phonemes and using the critical points of the lip movements as the desired out come.

When the result proved unsatisfactory several other speech processing methods were used to extract the features from the speech signal. These methods included reducing the waves into its wave envelop pattern and extracting the energy levels of the wave using Discrete Fourier Transform.

Experimental results

Cross-correlation

Speech analysis was done using cross-correlation of the phonemes used to see if this process could be used for Phoneme recognition.

From this method
:accurately recognizing a sound is 56.18%.
:25.00% phonemes gave ambiguous results.
:18.18 % of the phonemes were not recognizable at all.

It is believed that the recognition accuracy of 56.18% is due to the pronunciation distinctness between the phonemes such as /p/, /t/ and /th/, as well as to the small class size. Certain of phonemes such as /k/ consisting of the unvoiced stop, are not that easily recognized. Sounds with the soft 'sss' sounds are extremely difficult to accurately recognize. It is believed that this is caused by the similarity in pronunciation between these phonemes.

ANN

A technique based on Artificial Neural Networks (ANN) was tried to create a user dependant system. In this technique, the speech analysis consisted of programming neural networks by feeding direct speech made up of 44 phonemes and using the critical points of the lip movements as the desired out come. Feeding direct sound to the ANN produced large errors and was therefore deemed unsuitable.

When the result from the first method proved insufficient several other speech processing methods were used to extract the features from the speech signal. The second attempt was to train an ANN using the wave pattern envelop and the corresponding key-point coordinates. This method gave better results than the earlier method but was still not satisfactory.

The final method was to extract the energy levels of the wave using Discrete Fourier Transform and training an ANN using this data with again, the key point co-ordinates as the desired outcome. This method gave for more encouraging results as the error when training ANN was very small.

Discussion

The cross- correlation method can be used to categorize phonemes into three different groups and to identify them as being in one of these groups but it is not possible to use that method for actual identification of a specific phoneme from an audio speech signal.

The two original methods (feeding direct sound to the ANN and training the ANN using sound envelops) do not produce satisfactory results but using the Fourier transform to process sound waves showed promise. When the ANN where trained using this method gave more satisfactory results than any of the previously tried methods.