Combined Gesture-Speech Analysis and Synthesis

Principle Investigator and Candidates:

Principle Investigators:

- Research Asst. Mehmet Emre Sargin
- Asst. Prof. Engin Erzin
- Asst. Prof. Yucel Yemez
- Prof. A. Murat Tekalp

Date: 1/3/2005

Abstract:

Multimodal speech and speaker modeling and recognition are widely accepted as vital aspects of state of the art human-machine interaction systems. While correlations between speech and lip motion as well as speech and facial expressions are widely studied, relatively little work has been done to investigate the correlations between speech and gesture.

Detection and modeling of head, hand and arm gestures of a speaker have been studied extensively in [1--4] and these gestures were shown to carry linguistic information [5, 6]. A typical example is the head gesture while saying “yes”. In this project, it is desired to find the statistical correlation between gestures and speech. Speech features are selected as (Mel Frequency Cepstrum Coefficients) MFCC. Gesture features are composed of positions of hand, elbow and head with respect to a predefined point. In this sense, prior to the detection of gestures, discrete symbol sets for gesture will be determined using Hidden Markov Models (HMM). Using these discrete symbol sets, sequence of gesture features will be clustered and probable gestures will be detected. The correlation between gestures and speech will be modeled by applying input-output HMM’s to the temporal information embedded in sequence of gesture elements and speech elements (phonemes, words, phrases). This correlation can be used to fuse gesture and speech modalities for more natural synthesis of talking avatars from text. Synthesis of gesture that is correlated with speech has many applications in edutainment (i.e. video games, 3-D animations).

1. Project Objective

In face to face conversations humans not only communicate with their voice but also use gestures to express and emphasize their feelings. Thus, in edutainment applications, humans expect interactive conversations that animated person’s speech is aided and complemented by other sensory modalities, including expression, gaze, gesture, grasp, signing, emotion, and factors beyond the textual equivalent of speech. In other words, the more animations include correlated gestures, the more they seem natural to humans. In this sense, ultimate goal of this project is synthesis of natural gestures that are correlated with speech.

2. Background Information

This project can be divided into several tasks. These tasks are; preparation of the database, gesture extraction, gesture-speech correlation analysis and synthesis of natural gesture. Basic block diagram of project is given in Figure 1.

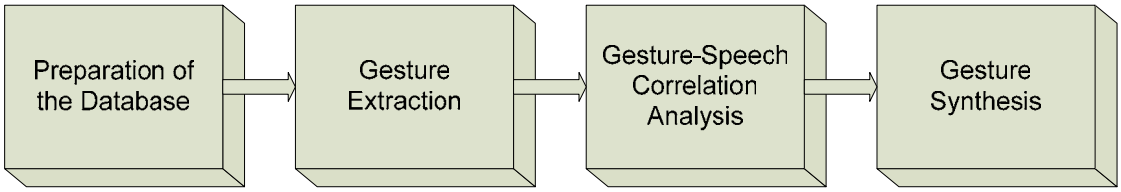

Figure 1

- *Preparation of the Database:*
    - *Objectives:* Participants will prepare a database for gesture analysis. The database can be prepared by capturing a sequence from TV, video or camera.
    - *Requirements:* According to the capturing device, participants are expected to be familiar with using that capture device (i.e. PCI Capture Card, DVCAM, etc.). Participants should also be familiar with using video editing interfaces (i.e. VirtualDub etc.).
- *Gesture Extraction:*
    - *Objectives:* Participants will extract gesture features of the sequences in the database.
    - *Requirements:* Familiarity with OpenCV and/or Matlab image-video processing toolboxes will be helpful for this task. Participants are expected to have some background in pattern recognition, object detection and tracking.
- *Gesture-Speech Correlation Analysis:*
    - *Objectives:* Participants will firstly extract speech features of the sequences in the database. Then, combined gesture-speech features will be used to model the correlation.
    - *Requirements:* As MFCC coefficients are going to be used to represent speech features, it will be helpful to have some experience with speech processing toolboxes (HTK, Matlab, etc.) [7]. Modeling of the gesture-speech correlation will be obtained by using HMM's. Therefore, it is expected that participants have at least some intuition about HMM [8].
- *Gesture Synthesis:*
    - *Objectives:* Given a speech sequence, participants will synthesize the speaker's gesture.
    - *Requirements:* Determination of proper gesture will be done by using the correlation model and visualization of the gestures will be achieved by animating the stick model of the body [9]. Animation of stick model can be done by using OpenGL or any other graphics API.

3. Detailed technical description

3.1. Technical Description

The work packages involved in this project are preparation of a gesture-speech database; detection of statistically significant, repetitive, smallest gesture elements (called gesture units); determination of the correlation between gesture units and speech features; generic modeling of gesture units for a specific speaker and gesture synthesis of that specific speaker given his speech sequence.

3.1.1. Preparation of a Gesture-Speech Database

In order to detect gestures that are correlated with speech, the vital work task is the selection of gesture-speech database. In this study, the gestures of a specific person will be investigated. The video database related with that specific person will include the gestures that she frequently uses. The visual data should preferably be such that locations of head, arm, elbows, etc. should easily be detectable and traceable. A convenient possibility is to build up a database by using the sequences captured from a standup show video.

3.1.2. Detection of Gesture Elements

In order to be able to establish a representative collection of discrete symbol sets for gesture, the underlying physics for formation of gestures accompanying speech, in particular hand and arm gestures, should be taken into account. A gesture can be expressed in terms of well-ordered gesture primitives based on muscle activity; in that way it becomes possible to examine correlations between ordered groups of gesture elements and ordered groups of speech-sound elements and to verify the premise that there is a direct and physiological relation between speech and gesture.
Analysis of gesture in terms of muscle-activation patterns can be achieved by using the fact that each gesture is composed of perfectly coordinated movements. These coordinated movements can be classified into the following categories:

- Abduction: The arm is moved away from the side of the body and moves through a semi-circle to arrive finally at a position pointing vertically upwards above the shoulder.
- Adduction: The arm is moved in the contrary direction across the line of the body and upwards as far as the skeletal structure permits i.e. the arm is swung across the body and up.
- Extension: The arm is straightened, unbending at the elbow to stretch forward;
- Flexion: The reverse of extension, the arm bending at the elbow, at the limit to almost complete approximation of the forearm and upper arm.
- External Rotation: The turning of the upper arm on the shoulder or the forearm on the elbow outward i.e. clockwise.
- Internal Rotation: The reverse of external rotation, with the upper arm turning inward on the shoulder or the forearm turning inward at the elbow;
- Supination: The complex of movements of the arm and the hand which result in the hand being turned palm upwards;
- Pronation: The inverse complex of movements of the arm and the hand resulting in the hand being turned palm downwards.

More simply, the movements of the arm can be presented in terms of the body's co-ordinates, front-back, side-side, and up-down as:

- The arm is moved forward and upward until it ends pointing upwards above the head.
- The arm is moved out to the side and upwards, again ending pointing upwards above the shoulder.
- The arm is moved inwards and across the body, moving up at the same time as far as the skeletal system permits, with the forearm at the end pointing up in front of the face and head.
- The arm is rotated inwards or outwards on its axis.
- The arm is bent or straightened at the elbow.

In this project mainly head and arm gestures will be investigated. Therefore features should contain the motion information of arms and head. The positions of hand, elbow and shoulders are first detected. Then, the coordinates of these points with respect to a predefined origin can be used to form the feature vectors. The feature points will be driven from a stick model, shown in Figure 2, that will be fitted to torso, arms and head.

Figure 2

3.1.3. Gesture Speech Correlation Analysis

As mentioned in the previous section; head, hand and arm gestures of humans are produced by concatenation of small gesture elements. Humans also produce speech in a similar fashion which is composed of small speech elements called phonemes. Actually, coordination of voluntary motor movements like speaking and producing gestures, is in the responsibility of the same part of the central nervous system called cerebellum. In this sense, the tools that are used for speech modeling can also be used in gesture modeling. Therefore, hidden Markov model (HMM) architecture can be used for clustering and modeling of gestures. After clustering gesture features to gesture elements, speech is clustered similarly. As a result, a sequence of gesture-speech elements is obtained. The correlation within and between speech and gestures are obtained by using temporal information embedded in the gesture-speech sequence. Temporal modeling is necessary also to appropriately handle probable delays that may occur between gestures and corresponding speech utterances. Input-output HMM’s may serve well for this kind of temporal modeling.

3.1.4. Synthesis of Gestures

Using the statistical model obtained in Section 3.1.3, given a speech sequence, corresponding gesture sequence can be predicted. Predicted gesture elements can be used

to animate a body stick model. Proposed gesture animation scheme of a waving stick model can be seen in Figure 3.

Figure 3

3.2. Resources

Participants are expected to use several available image/video processing software tools and write computer programs when needed. Therefore, a computer is to be appropriated to each participant. Internet connection to each computer would also be useful.
The existing software that can be used in this project is open source and free. However software development applications such that Matlab and Microsoft Visual Studio require license. Possible open source and licensed software can be listed below:

- Open source applications:
    - VirtualDub (Video Editing)
    - HTK (HMM Software)
    - OpenCV (Image Processing Library)
- Licensed Applications:
    - Microsoft Visual Studio .NET
    - Matlab 7.0

3.3. Project Management

The project manager selected from the principle investigators is responsible for the project management. The project leader will split the participating researchers into groups according to their areas of interests. Each group will be responsible for one task of the project explained in Section 2.
The first week of the project will focus on providing the necessary theoretical background and skills. This will be accomplished in the form of lectures for theoretical aspects and small programming assignments for practice and related computing skills. Lectures and assignments will be organized as morning and afternoon sessions. Lecture schedule is given in Section 4.
In the following weeks, the project manager and participators will meet at least two times a week to exchange information about the progress. These meetings will also provide inter-group communication which will be useful for determining the input/output relations between each task. If necessary, the project manager will give extra lectures during these meetings.

4. Work Plan

Timeline of the project can be seen in Figure 4.

Figure 4

Schedule of the lectures can be seen in Figure 5.

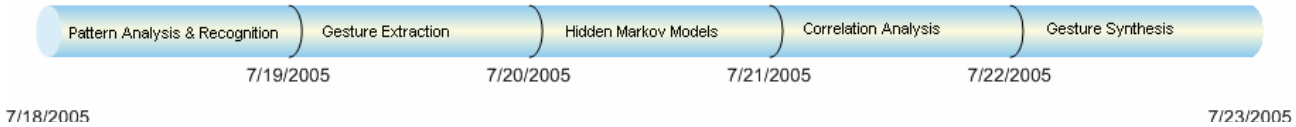

Figure 5

5. Benefits of the research

At the end of this project, participants are expected to improve their knowledge and skills on:

- Hidden Markov Models
- Using Hidden Markov Model Toolkit
- Using OpenCV Library
- Using Matlab Image Processing Toolkit
- Statistical modeling
- Speech analysis through Hidden Markov Models
- Multimodal signal processing
- Team work and collaborative study

6. Profile of team

6.1. Leaders

Research Asst. Mehmet Emre Sargin
Asst. Prof. Engin Erzin/ Asst. Prof. Yucel Yemez

6.2. Staff proposed by the leader

Prof. A. Murat Tekalp

6.3. Other researchers needed

We look for researchers who have some background in image, video processing and pattern recognition. Experience on background modeling, foreground segmentation and object tracking and using OpenCV and IPL (image processing libraries) would also be useful.

7. References

[1] Jie Yao and Jeremy R. Cooperstock, "Arm Gesture Detection in a Classroom Environment," Proc. WACV'02 pp. 153-157, 2002.

[2] Y. Azoz, L. Devi. R. Sharma, "Tracking Hand Dynamics in Unconstrained Environments," Proc. Int. Conference on Automatic Face and Gesture Recognition'98 pp. 274-279, 1998.

[3] S. Malassiotis, N. Aifanti, M.G. Strintzis, "A Gesture Recognition System Using 3D Data," Proc. Int. Symposium on 3D Data Processing Visualization and Transmission'02 pp. 190-193,2002.

[4] J-M. Chung, N. Ohnishi, "Cue Circles: Image Feature for Measuring 3-D Motion of Articulated Objects Using Sequential Image Pair," Proc. Int. Conference on Automatic Face and Gesture Recognition'98 pp. 474-479, 1998.

[5] S. Kettebekov, M. Yeasin, R. Sharma, "Prosody based co-analysis for continuous recognition of coverbal gestures,"Proc. ICMI'02 pp.161-166, 2002.

[6] F. Quek, D. McNeill, R. Ansari, X-F. Ma, R. Bryll, S. Duncan, K.E. McCullough "Gesture cues for conversational interaction in monocular video," Proc. Int. Workshop on Recognition, Analysis, and Tracking of Faces and Gestures in Real-Time Systems'99 pp. 119-126, 1999.

[7] For detailed information visit: http://htk.eng.cam.ac.uk

[8] Rabiner, L.; Juang, B., "An introduction to hidden Markov models" ASSP Magazine, IEEE, Vol.3, Iss.1, pp. 4- 16, Jan 1986

[9] Jae-Moon Chung; Ohnishi, N., "Cue circles: image feature for measuring 3-D motion of articulated objects using sequential image pair" Automatic Face and Gesture Recognition, 1998. Proceedings. Third IEEE International Conference on, Vol., Iss., pp. 474-479, 14-16 Apr 1998