A MULTIMODAL DATABASE OF GESTURES AND SPEECH

Satoru Hayamizu¹, Shigeki Nagaya², Keiko Watanuki³, Masayuki Nakazawa¹,
Shuichi Nobe⁴, Takashi Yoshimura¹
1. Electrotechnical Laboratory, 2. Central Research Laboratory, Hitachi,
3. RWC Multimodal Sharp Laboratory, 4. Aoyama Gakuin University
hayamizu@etl.go.jp

ABSTRACT
This paper describes a multimodal database which consists of image data of human gestures and corresponding speech data for the research on multimodal interaction systems. The purpose of this database is to provide an underlying foundation for research and development of multimodal interactive systems. Our primary concern in selecting utterances and gestures for inclusion in the database was to ascertain the kinds of expressions and gestures that artificial systems could produce and recognize. Total 25 kinds of gestures and speech were repeated four times for the recording of each subject. The speech and gestures for a total of 48 subjects were recorded, converted into files and in the first version, the files for 12 subjects were recorded on CD-ROMs.

Keywords: multimodal, database, gesture

1. INTRODUCTION
A key research issue in developing multimodal interactive systems is how to realize interaction between artificial recognition systems and people who spontaneously use diverse gestures and speech in a fully integrative way. But so far little has been done to (1) construct a database supporting research in this area, or to (2) make such a database available as a shared resource to research institutes working in this field [1-4]. We believe the multimodal database is an essential foundation for the research and development of such systems[5-6].

We constructed a multimodal database which consists of image data of human gestures and corresponding speech data for the research on multimodal interaction systems under the Real World Computing Program. Our objective is to build a speech and video database that can be shared among different research groups pursuing similar works that will promote research and development of multimodal interactive systems integrating speech and video data.

It is anticipated that multimodal databases containing the types of interactive objects described here will become increasing important in the future. The objective of this work goes beyond a new methodology or system proposal and prototype. This is because, in order to establish design methods and evaluation criteria for handling man-machine interaction, it is essential to first accumulate instances of interaction that are the subject of research in this area.

Besides the technological obstacles that must be overcome to realize multimodal interaction, there are many issues that must be addressed to build a multimodal database. We must consider the kinds of interaction to be provided by the database, how the interactions are conveyed, how natural human behavior can be enacted for the database, and how the behavior can be recorded as objectively as possible.

And there are many other issues that must be resolved - all involving considerable time and effort - such as how to set up a scene for systematically and efficiently acquiring data for the database, how to set up recording and transmission equipment to record the data while synchronizing the speech and images, how to automate the process of organizing the data in files as much as possible, how to name the files in such a way that the contents and sequence of the files is obvious to users, how to coach the subjects who are reproducing the gestures so they know exactly what is expected of them, and so on.

Following sections describe database contents, collection method, recording setup, file format and conclusion.

2. DATABASE CONTENTS
In analyzing the range of gestures people employ in communicating with other people, there are many types of gestures as follows:

a) Signs (signals).
b) Examples (indicating size, shape, rhythm)
c) Manipulation (continuous control, such as indicating the position of an object).
d) Emotional expression (facial expressions, laughter, etc.)
e) Regulation (receiving permission to speak).
Rather than obtaining qualitative knowledge of human behavior per se, our primary concern in selecting utterances and gestures for inclusion in the database was to ascertain the kinds of expressions and gestures that artificial systems could produce and recognize. In the interests of conducting recognition experiments, we narrowed our selection to the following three categories of gestures and speech:

1) Gestures and utterances indicating agreement and disagreement

These are indicated by nodding the head up and down or shaking the head involving three different variables: number of nods or shakes, amplitude of the gesture, and speed of the gesture. One category of strong disagreement is indicted by gesturing with the hand (i.e., moving the hand palm outward) while shaking the head.

2) Gestures indicating direction

Some directions can be indicated by gesturing with the palm while others are indicated by pointing. The directional gestures included in the database were up, down, left, right, closer, and further back.

3) Gestures indicating relative size and emphasis

Hand gestures and arm gestures should be taken up as separate ones because the region of the hands must first be detected and tracked for an artificial system to recognize them. Our concern at this initial stage is merely to include gestures indicated through arm movements in the database.

A total of 25 kinds of gestures were video recorded including 8 gestures conveying agreement or disagreement, and 17 types of gestures indicating direction and relative size. The 25 kinds of gestures were repeated four times for the recording of each subject. They are listed up as follows:

1) Gestures and utterances conveying agreement / disagreement (1-1 through 1-8 convey the following gestures: 1-1 conveys ordinary agreement; 1-2, 1-3, and 1-4 express strong agreement; 1-5 conveys ordinary disagreement; and 1-6, 1-7, and 1-8 indicate strong disagreement.)

1-1: One medium nod up and down at normal speed while saying hai (yes) one time.
1-2: Two medium nods at normal speed while saying hai (yes) one time.
1-3: One large nod at normal speed while saying hai (yes) one time.
1-4 One medium nod at rapid speed while saying hai (yes) one time.
1-5: One medium shaking of the head at normal speed while saying iie (no) one time.
1-6: One large shaking of the head at normal speed while saying iie (no) one time.
1-7: One medium shaking of the head at rapid speed while saying iie (no) one time.
1-8: One medium shaking of the head and hand at normal speed while saying iie (no) one time.

2) Gestures and utterances conveying direction and size

2-1: Convey upward motion while saying ue, ue (up, up).
2-2: Convey downward motion while saying shita, shita (down, down).
2-3: Convey motion to the right while saying migi, migi (right, right).
2-4: Convey motion to the left while saying hidari, hidari (left, left).
2-5: Convey inward motion while saying temae (closer).
2-6: Convey outward motion while saying muko (over there).
2-7: Indicate the size and shape of a briefcase while saying konna okisa no (one that's this size).
2-8: Convey emphasis while saying kore ga juyo nan desu yo (this is important).
2-9: Indicate length by holding two hands apart while saying kono gurai no okisa no sakana (a fish this big).
2-10: Indicate self by pointing at one's own chest while saying watashi (me).
2-11: Point downward toward the left with the right index finger while saying hidari (left).
2-12: Point downward toward the right with the right index finger while saying migi (right).
2-13: Trace an imaginary circle to the right around a perpendicular axis while saying migi wawari (clockwise).
2-14: Trace an imaginary circle to the left around a perpendicular axis while saying hidari wawari (counter clockwise).
2-15: Indicate stop by holding up one's hand palm outward while saying sutoppu (stop).
2-16: Draw the palms apart indicating enlargement or expansion while saying kakudai (expansion).
2-17: Bring the palms together indicating contraction while saying shukusho (contraction).

3. COLLECTION METHOD

In terms of media and content, we were primarily concerned to record the subjects' speech and the images of the subjects' face and body from the waist up. At the same time we also recorded upper-body images and speech of the partners with whom the subjects were interacting.
Three recording methods were tested in preliminary experiments.

1) Subjects freely improvised gestures and speech as indicated by a written script (large individual variations).

2) Subjects first observe gestures and listen to speech that is prerecorded, then mimic those gestures and speech (little variation).

3) Subjects express themselves through gestures and speech as naturally as possible in interaction with another person.

Here we gave priority to the second approach where subjects mimic prerecorded gestures because the acquisition cost was relatively low with respect to the amount of data collected and because this method produces little variation among the data.

The speech and gestures included in the database were first prerecorded. Subjects then observed the gestures and speech, and were asked to mimic or reenact both the spoken utterances and the observed gestures for inclusion in the database. These were mirror-inverted imitations. Adopting this approach we could efficiently collect fairly consistent data that is well suited for recognition experiments.

4. RECORDING SETUP

4.1 Image Recording Equipment Setup

The video recording equipment was used to record a four-part split screen (showing the subject's face, subject from the waist up, partner from the waist up, and monitor screen) and a solitary image of the subject from the waist up. (Just waist-up images of subjects were recorded on the CD-ROMs.) In positioning the camera and shooting angle, the main concerns were to have the subjects' eyes aligned as much as possible and to have the gestures fit within the confines of the screen. Shots of subjects from the waist up were taken by a video camera and then were recorded using a betacam. Output from the four-part split screen were recorded by separate betacams using a multiviewer. In addition, refresh signals were input to synchronize the various betacams.

A video recording lighting setup was used. Two 1,000-watt lights were used to illuminate for the subjects, and three 500-watt lights were used for background lighting. The subjects were not exposed to direct lighting. Indirect lighting reflected off white reflective surfaces (i.e., reflective paper) on the ceiling and floor was used. White reflective paper was also placed on the desks in front of the subjects to minimize facial shadows.

A uniform blue was used as the background color for both subjects and partners to facilitate image recognition.

In addition, markers were applied to indicate the approximate positions of the arms. For purposes of color separation, red markers were used for the arms, green for the elbows, and white for the shoulders. Yellow-green shirts were worn. These arrangements made it possible to measure position through image processing without the use of elaborate mechanical equipment.

4.2 Sound Recording Equipment Setup

Stereo sound was recorded with the subjects voice over the left channel and the partner's voice over the right channel at the same time as the images were video recorded. DAT recordings were also made for reference purposes.

A head-set type microphone was used for the preliminary trial, but a lapel microphone was adopted for the actual recording to facilitate image processing. Prerecorded speech was used instead of the speech of the partners. For monitoring purposes, a third party could listen in on the dialog through headphones.

5. FILE FORMAT

The recorded files consist of time-synchronized continuous video and audio files. The video files show the subjects from the waist up. The speech files are stereo recording the prerecorded voice on the right channel and the subject's voice on the left channel. In addition, MPEG-1 format video files and AIFF format speech files are included for reference.

1) Image Files

Images of subjects from the waist up were reduced to half size (vertical 240 x horizontal 320 pixels) and recorded at 30 frames per second. There are no headers. Images are raster-scan images with 24-bit pixels (8 bits per color). Starting point is the upper left. Figure 1 shows a sample image file.

2) Audio Files

Stereo sound is converted into two separate files, one for the left channel and the other for the right channel (the speaker's voice is carried on the right channel, and the subject's voice is carried on the left channel). Note that when the subject is gesturing, the speaker is silent, so the right channel is empty at these times. Sampling rate is 16 kHz. There are no headers. The sampling format is 16-bit binary digitizing.
Both image and speech data were labeled (tagged). Start and end position of each utterance were manually attached. Shape, place and orientation of each hand were described also in the tag [7]. Position of face and hands were automatically detected and coded in the tag, too.

The speech and gestures for a total of 48 subjects were recorded, converted into files and in the first version, the files for 12 subjects were recorded on CD-ROMs. Two CD-ROMs are used per subject.

6. CONCLUSION

A multimodal database was presented which consist of image data of human gestures and corresponding speech data for the research on multimodal interaction systems under the Real World Computing Program. The purpose of this database is to provide an underlying foundation for research and development of multimodal interactive systems. Total 25 kinds of gestures and speech were repeated four times for the recording of each subject. The speech and gestures for a total of 48 subjects were recorded, converted into files and in the first version, the files for 12 subjects were recorded on CD-ROMs. Use of this database is limited to research purposes, and is only permitted upon receipt of an application. Further information on the application of the database and its availability can be found at the following URL (http://www.rwcp.or.jp/wswg/rwcdb/mm).

ACKNOWLEDGEMENT

This research was done under the Real World Computing Program. The authors would like to thank Mr. Hiroshi Matumura and Mr. Hiroumi Yamashita for the development of the database and the subjects for their contribution to providing their gestures and speech.

REFERENCES


