# MULTIMODAL CORPORA FOR HUMAN-MACHINE INTERACTION RESEARCH

*Satoshi Nakamura*[1] *Keiko Watanuki*[2] *Toshiyuki Takezawa*[1] *Satoru Hayamizu*[3]

[1]ATR Spoken Language Translation Research Laboratories.
2-2-2, Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0288, Japan

[2]Real World Computing Partnership Multimodal Functions Sharp Laboratories.
in System Technology Development Center, Sharp Corporation, 1-9-2, Nakase, Mihama-ku, Chiba 261-8520, Japan

[3] Electorotechnical Laboratories, 1-1-4 Umezono, Tsukuba, Ibaraki 305-0045, Japan

E-mail: nakamura@slt.atr.co.jp, watanuki@iml.mkhar.sharp.co.jp, takezawa@slt.atr.co.jp, hayamizu@etl.go.jp

## ABSTRACT

In recent years human-machine interaction has increased its importance. One approach to an ideal human-machine interaction is develop a multi-modal system behaves like human-beings. This paper introduces an overview on multimodal corpora which are currently developed in Japan for the purpose. The paper describes database of 1)Multi-modal interaction, 2)Audio-visual speech, 3)Spoken dialogue with multiple speakers, 4)Gesture of sign language and 5)Sound scene data in real acoustic environments.

## 1. INTRODUCTION

In recent years human-machine interaction has increased its importance. One approach to an ideal human-machine interaction is to develop a system with human-like behavior. Speech interaction is the important modality from a viewpoint of carrying an human intension and meanings. However, mouth, eye, face, head movements and gestures also play very important roles in the natural interaction. Although technologies concerning on LVCSR and TTS had accomplished remarkable progress, those technologies are unfortunately insufficient for natural human-machine interaction. Thus it should be the most important for machines to integrate multi-modal information to realize natural human-machine interaction in a real world. However, even on an acoustic modality, conventional research only dealt with monaural clean speech uttered by a single speaker. We should deal with multiple speakers in a real world. Furthermore there are various kinds of noises and reverberation in the real world. There are almost no database on acoustic environments including noises and reverberation. This paper introduces an overview on multimodal corpora which are currently developed in Japan. The paper describes database of 1)Multi-modal interaction, 2)Audio-visual speech, 3)Spoken dialogue with multiple speakers, 4)Gesture of sign language and 5)Sound scene data in real acoustic environments.

## 2. MULTIMODAL INTERACTION DATABASE

### 2.1. Background and Objectives

We, human beings are communicating each other by exchanging verbal and nonverbal information such as facial expressions, gestures, gaze, and intensity and pitch of speech. The analyses of how such various information is integrated in one person, or interacted between a speaker and a listener, provide insights not only for understanding human behavior but also for designing man-machine interfaces. We are interested in developing multimodal man-machine interface in order to make interaction between human and computer more natural and active. To achieve this goal, we have taken an approach of modeling human behaviors in the context of ordinary face-to-face conversations [1, 2]. This section describes a physical configuration of our multimodal interaction database. The database is a collection of two-people conversation data which contain speech, movie, motion data (3D positions for 18 body parts) and speech transcripts. The database will provide a source for analyzing the relation among various multi-modalities observed in one individual, and also the interaction between the two.

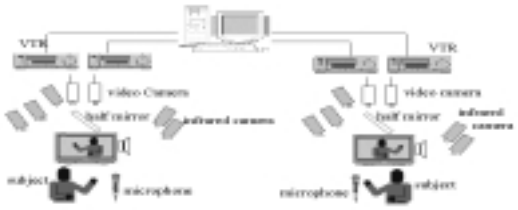


**Figure 1. Overview of Multi-Modal Data Collection System**

### 2.2. Data Collection

In order to construct multimodal interaction database, we have implemented a system which utilizes video and audio recording equipments, and optical motion sensing device to capture verbal and nonverbal information in interpersonal communication[3]. Figure 1 shows an overview of

**Figure 2. Examples of Multi-Modal Image Data**



6th International Conference on Spoken Language Processing (ICSLP 2000)
Beijing, China
October 16-20, 2000

the data collection system. In this system, two persons are seated in different location, and talk each other through 29 inch monitors. The video images displayed on the monitor are reflected by a half mirror, and two video cameras are set behind the mirror: one is for upper body motions, and the other is for zooming out for taking body motions. Each movie is recorded by analog VTRs (betacam). They are controlled by a PC and synchronized. The recorded movie data are then converted to digital movie data. The image size of a movie is 320 x 240, and its sampling rate is 30 frames /sec. The audio sampling rate is 16kHz. Motion sensing device measures 3D positions for the markers attached on the 18 positions on the upper body, including head top, shoulders, elbows and hands. Each marker is taken by 5 infrared cameras for each subject. The marker positions are traced optically at 60 frames /sec. The movie data and the motion data are synchronized with a sound signal at the beginning. For one conversation, two movies and 18 markers' position data are stored in the database as raw data. Based on the raw data, more detailed data can be acquired. For example, from audio data, intensity and pitch of speech are calculated; and speech is transcribed. From motion data, on the assumption that a human body is constructed by 8 segments: right arm, left arm, right upper arm, left upper arm, right hand, left hand, head and body, each segment position and direction can be calculated from the maker positions for the segments.

### 2.3. Multimodal Interaction Database

Such motion, movie and speech data are all synchronized along a common time scale, and stored in a database. Figure 2 shows an example of movies in the database. We have so far collected 14 pairs of spontaneous speech (10min /person x 28 persons). The persons of each pair are well known each other so that their talking style is very natural, friendly and active. Using the database, we have found that there is a significant correlation between the beginning of speaking turn and head movements. We have been continuing analyses for the relation between motions and speech. The database should provide a source for the analyses of human behavior in conversation, and for modeling man-machine interactions to achieve more advanced interfaces between humans and computers.

## 3. AUDIO-VISUAL SPEECH CORPORA

### 3.1. Background and Objectives

Speech recognition performance has been drastically improved recently. However, it is also well-known that the performance will be seriously degraded if the system is exposed in noisy environments. Humans pay attention not only to speaker's speech but also to speaker's mouth in such adverse environments. The lip reading is the extreme case if it is impossible to get any audio signal. This suggests a fact that speech recognition can be improved by incorporating mouth images. This kind of bi-modal integration is available in almost every situation. For instance, the recognition accuracy for voiced consonants /b/,/d/,/g/ can be improved by incorporating lip image information, since lip closure for bilabial phonemes is relatively easy to discriminate using image information. On the other hand, lip movement can play a significant role in human-machine communication. If lip movements are synthesized well enough to do lip-reading, hearing impaired people may be able to recover auditory information from the visualized computer agent.

**Table 1. Audio-visual database**

| | |
|---|---|
| File format | SGI movie |
| Speaker | One female speaker |
| Utterances | ATR 5240 Japanese words |
| Audio | sampling: 16bit, 48 [$kHz$] |
| Visual | sampling rate : 30 [$frames/sec$]<br>size : 160×120<br>color quality : 8 bit RGB |

Bimodal speech processing such as bimodal speech recognition and talking face generation from speech input is a promising approach to natural human interface. Thus database collection is a crucial issue for the research. Two kinds of bimodal speech database are collected[4, 5]. First one is speech and talking face image data for bimodal speech recognition. The second is speech and talking face position data for speech-to-lip generation.

### 3.2. Database for Bimodal Speech Recognition

Table 1 shows the specification of the database. Video recording is performed at a sound-proof room and lighting is set from the front. The head is not fixed but the speaker is requested to attach her back to the seat.

Moreover, the speaker is also requested to close a mouth before and after utterance. We observed the difference of lighting conditions, size of lips, and inclination of a face every utterance words, since the video recording is conducted over two or more days.

### 3.3. Database for Talking Face

Generally speaking, it is necessary to collect more precise data of face positions for the talking face generation. The data also should synchronize to a frame rate of speech processing. Then we collected using the OPTOTRAK system. Speech and 3-D lip position data for male and female speakers of Japanese are recorded at 125Hz using the OPTOTRAK, 3-D position sensoring system. These 3-D positions are transformed into the visual parameters height(X),width(Y) of the outer lip contour and protrusion(Z) based on five parameters of the 3-D lip model. The audio-parameter has 33 dimensions of 16-order Mel-Cepstral coefficients, their delta coefficients and the delta log power.

## 4. MULTI-PARTY MULTILINGUAL CONVERSATION CORPORA

### 4.1. Background and Objectives

ATR Interpreting Telecommunications Research Laboratories collected multiparty multilingual conversational speech data for discussing the issues of communication flow, speaker and hearer identification, the type of speech interaction, and information sharing among translation systems. Two kinds of data collections are conducted. One is to collect Japanese monolingual conversations by four participants. The other is to collect multilingual conversations by four participants of different languages (English, German, Korean and Japanese) through six interpreters as shown in Figure 3.

### 4.2. Domain/tasks in conversations

The domain involves travel conversations at a hotel or a travel agency; this task is selected because of its familiarity to people, and its expected use in future speech translation systems. The task of multiparty conversation involves negotiation on hotel room arrangements among an agent at the headquarters of the travel agency and three agents at each county branch.

In this data collection experiments, the agent at the headquarters of the travel agency is assumed to be the chairperson of the multiparty conversation. However, multi-lateral communication flow is adopted, i.e., every participant could utter to any other participants.

### 4.3. Data collection experiments

As multiparty monolingual conversation data, 16 conversations are collected in a quiet office room. As multiparty multilingual conversation data, 16 conversations are collected in a quiet office room. All of them are recorded on digital audio tapes and Hi8 video tapes. From multiparty monolingual conversation data, Japanese part in three conversations is transcribed. From multiparty multilingual conversation data, English and Japanese parts in three conversations are transcribed.

### 4.4. Characteristics of multiparty conversations

Multi-lateral communication flow is naturally realized in the data collection experiments. In multiparty conversations, speakers tended to express the name of the speaker own and/or the expected hearer explicitly for clarify the message. The signals for interruptions, confirmations and agreements tended to be increased rather than those of dialogues by two participants.

### 4.5. Discussions on the issues of translation

As shown in Figure 3, the message of participants are translated through independent speech translation systems, i.e., interpreters. Let us consider the following case. For example, a Japanese participant uttered something to an English participant. Next, the English participant uttered something to a German participant using a pronoun. An English-to-German interpreter did not always understand the referent of the pronoun. This caused incorrect translation. In this experiment, this kinds of phenomena could be found.

## 5. RWC GESTURE CORPORA

### 5.1. Background and Objectives

It is important for a computer to understand human gestures and to support more natural and smooth communication. A common gesture database as a shared resource is necessary to develop a gesture recognition system. Under the RWC (Real World Computing) Program, a multimodal database which consists of image data of human gestures and corresponding speech data is developed for the research and development of multimodal interaction system[8]. The gestures and speech collected are the ones which convey agreement or disagreement, and indicate direction and relative size, etc. There are 48 subjects included in the database. Then RWC gesture database have been collected. The gestures are words and sentences of Japanese sign language. They have been collected in order to develop and evaluate gesture recognition systems.

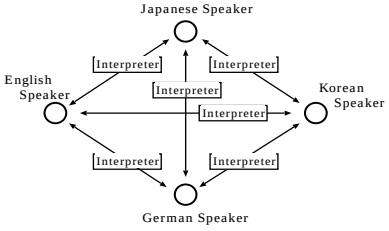


**Figure 3. Overview of Multi-Party Data**

### 5.2. Contents

The gesture database are the image data of motion movies of an upper human doby form the waist up. The gestures are selected from the words and sentences of Japanese sign language.

There are 3 sets of gesture database. The first set consists of the gestures by one video camera. There are two subjects who knows the Japanese sign language. There are the gestures of 300 words and 301 sentences with two or three words in one sentence. The second set consists of the gestures by two video cameras and the image data are stereo motion movies. The distance between two cameras is 0.33 m. There are four subjects who are the experts of Japanese sign language. The gestures are 300 words and they are repeated twice. The third set consists of the gestures by two video cameras as the second set. There are four subjects who are the experts of Japanese sign language. The gestures are 64 sentences. Each sentence consists of three to five words. They are repeated twice.

### 5.3. Recording setup

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

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

### 5.4. Form of distribution

The database will be distributed by DVD-RAMs because the size of the database is large. The images of subjects from the waist up are reduced to half size (vertical 240 X horizontal 320 pixels) and recorded at 30 frames per second. The images are raster-scan images with 24-bit pixels (8 bits per color). Starting point is the upper left. The starting and ending frame of each gesture are manually decided and labeled (tagged) for the word.

Use of this database is limited to research purposes, and is only permitted upon receipt of an application. For

**Table 2. Source sound**

| | | |
|---|---|---|
| Class 1 | Wood | Wooden boards and sticks |
| | Metal | Metal boards and stick |
| | Plastic | Plastic boards and stick |
| Class 2 | Friction Noise | Sound of the saw |
| | Plosive Noise | Break chop sticks |
| | Burst Noise | Claps |
| Class 3 | Metals | Bells |
| | Music instruments | Whistles |
| | Electronic sounds | Telephone Rings |

more detail, visit http://www.rwcp.or.jp/wswg/rwcdb (in Japanese).

## 6. SOUND SCENE DATABASE

### 6.1. Background and Objectives

This project is one of the projects supported by RWCP (Real World Computing Partnership). The objective of the project is to provide common standard databases for research concerning real acoustical environments[6].

It is almost impossible to collect all combinations of the existing sound sources and real acoustical environments. Thus, we started to collect two kinds of sound data. The first data is isolated sounds of environment non-speech sounds and speech sounds. We call the isolated sounds recorded in an anechoic room by the word *dry source* in this paper. The dry source is free from influences of room acoustics. The second data is impulse responses in various acoustical environments. The sound in the environment can be simulated by convolution of the dry sources and the impulse responses. However, there are sounds which is unable to simulate by the convolution such as non point source sounds and moving sound sources. We are planning to collect those sounds using a three dimensional microphone array. The microphone array database enables to extract arbitrary sounds by various beamforming algorithms.

The data is collected in an anechoic room, a variable reverberant room, office environments, where many sound sources exist. Various kinds of sound sources including speech are also collected as target sounds.

### 6.2. Dry Source Database

Dry source is the sound recorded in an anechoic room which is free from room acoustics. The environment sound can be simulated by convolution of the dry source and an impulse response if the transmission channel is linear and stable. We collected three kinds of environment sounds shown in Table 2. The first class is crash sounds of wood, plastic and ceramics. The second class and the third class are composed of sounds occurred when human beings operate on things like spray, saw, claps, coins, books, pipes, telephones, toys, etc. The sounds of the second class are the sounds whose source materials can not be easily associated. Whereas the source materials of the third class sounds can be easily associated uniquely.

We recorded around 100 samples for about 90 kinds of sounds sufficient enough for statistical model training. The recording is conducted in an anechoic room by B&K 4134 microphone and DAT recorder in 48kHz 16bit sampling. SNRs of the data are around 40-50dB.

**Table 3. Recording conditions of impulse Response**

| | |
|---|---|
| A/D, D/A | Pavec MD-8000mk2 64ch 24bit |
| Microphone | 54ch Spherical array<br>14ch Linear array(2.83cm spacing)<br>16ch Circle array |
| Source | Diatone DS-7 loud speaker<br>B&K Type 4128 Head-Torso |
| Source Sounds | Time stretched pulse<br>Phonetically balanced words(216)<br>Phonetically balanced sentences<br>(TIMIT(40), JNUS(50))<br>Real speech |

### 6.3. Impulse Response Database

We collected impulse responses at different locations in different rooms. The sounds are recorded in an anechoic room, a variable reverberant room and offices using 3 kinds of microphone arrays by the Diatone DS-7 loud speaker and B&K Type4128 Head-Torso. Reverberation times of the rooms are from 0.01 to 2 seconds. Table 3 shows recording conditions of impulse responses.

## 7. CONCLUSION

The paper describes database of 1)Multi-modal interaction, 2)Audio-visual speech, 3)Spoken dialogue with multiple speakers, 4)Gesture of sign language and 5)Sound scene data in real acoustic environments. These database is necessary to develop multi-modal systems. We hope these database could play an important role of next technological era of multi-modal information processing.

## 8. REFERENCES


1. Kiyama, J., Watanuki, K. and Togawa F.,: Multimodal Interaction Database and Analysis Environment, Proc. of 1997 RWC Symposium, pp.23-30, 1997.
2. Seki, S., et.al,: Multimodal Agent Interface for Communication, Proc. of 2000 RWC Symposium, pp.135-138, 2000.
3. Watanuki, K., et.al: Collection of Dialogue Data Using Motion Capturing System, Proc. of the 3rd Symposium for the Japanese Association of Sociolinguistic Science, pp.142-143, 1999 (in Japanese).
4. Nakamura, S., Ito, H., Shikano,K., : Stream Weight Optimization of Speech and Lip Image Sequence for Audio-Visual Speech Recognition, Proc. ICSLP2000, pp.965-968 (2000)
5. Kakihara,K.,Nakamura,S.,Shikano,K.,: Speech-to-Face Movement Synthesis Based on HMMs, Proc.ICME2000, (2000)
6. Nakamura, S., Hiyane, K., Asano, F., Nishiura, T., Yamada, T.,: Acoustical Sound Database in Real Environments for Sound Scene Understanding and Hands-Free Speech Recognition, Proc. LREC 2000, (2000).
7. Hayamizu, S., Hasegawa, O., Itou, K., Sakaue, K., Tanaka, K., Nagaya, S., Nakazawa, M., Endoh, T., Togawa, F., Sakamoto, K., Yamamoto, K.,: RWC Multimodal database for interactions by integration of spoken language and visual information, Proceedings of ICSLP, pp.2171-2174 (1996).
8. Hayamizu, S., Nagaya, S., Watanuki, K., Nakazawa, M., Nobe, S., Yoshimura, T.,: A multimodal database of gestures and speech, Proceedings of Eurospeech, pp.2247-2250 (1999).

# MULTIMODAL CORPORA FOR HUMAN-MACHINE INTERACTION RESEARCH

*Satoshi Nakamura*[1], *Keiko Watanuki*[2], *Toshiyuki Takezawa*[1] *and Satoru Hayamizu*[3]

[1]ATR Spoken Language Translation Research Laboratories. 2-2-2, Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0288, Japan

[2]Real World Computing Partnership Multimodal Functions Sharp Laboratories.
in System Technology Development Center, Sharp Corporation, 1-9-2, Nakase, Mihama-ku, Chiba 261-8520, Japan

[3] Electorotechnical Laboratories, 1-1-4 Umezono, Tsukuba, Ibaraki 305-0045, Japan

E-mail:
nakamura@slt.atr.co.jp, watanuki@iml.mkhar.sharp.co.jp, takezawa@slt.atr.co.jp, hayamizu@etl.go.jp

In recent years human-machine interaction has increased its importance. One approach to an ideal human-machine interaction is develop a multi-modal system behaves like human-beings. This paper introduces an overview on multimodal corpora which are currently developed in Japan for the purpose. The paper describes database of 1)Multi-modal interaction, 2)Audio-visual speech, 3)Spoken dialogue with multiple speakers, 4)Gesture of sign language and 5)Sound scene data in real acoustic environments.