Making of a Japanese Viseme Video Database by Multiple High-speed Video Observations

Makoto J. Hirayama
Kanazawa Institute of Technology
7-1 Ohgigaoka, Nonoichi, Ishikawa 921-8501 Japan
E-mail: mako@infor.kanazawa-it.ac.jp

ABSTRACT

A video database was made for extracting Japanese visemes and modeling speech articulators' motions during speech utterances. By using two high-speed video cameras, frontal and lateral views around lips were recorded at up to 300 frames per second. Recorded utterances are short Japanese sentences about 5 seconds duration each at a normal speed. A sentence set used is a phonetically balanced set which includes all Japanese syllables and typical trigram patterns of Japanese phonemes. Experimental configurations and measurement issues are discussed.

1. INTRODUCTION

Speech production makes not only auditory sounds but also visually perceivable images around lips. Lip shapes related to phonemes are so called visemes.

The main engineering applications using visemes are visual speech recognition and visual speech synthesis. Visual speech recognition, i.e. lipreading, used with a normal auditory speech recognition system, is expected to increase its recognition rate of the system, especially in noisy environment such as in train stations (for example, see Hennecke, et. al. [1] for review). Visual speech synthesis, i.e. facial animation, is used for computer graphics movies, video games, user interface systems, and so on.

To make a viseme set, experimental data of lips and other visually observable speech organs are needed. Or, beyond static visemes, temporal data are needed for more realistic modeling of speech organs' movements. Although discussions have been done for making of visual speech database [2], as far as I know, there is no standardized way to make it.

There have been several visual speech recording experiments so far, but they are recorded for individual specific research purposes not for databases and they are usually piecemeal rather than a whole set of variety of phonemes. Also, most of them are at a normal video speed. It is not enough to analyze fast changes of mouth shapes during consonant productions.

Thus, I have been doing high-speed video recording during speech and making a video database for making a modern Japanese viseme set and helping viseme related speech researches. Especially, the experimental data are suitable for analyzing speech organs' movements which have coarticulations. I assume that the data are used for realistic facial animations, although they can be used for other purposes. In this paper, I present configurations of these experiments and the database.

2. EXPERIMENTAL CONFIGURATION

A human subject sits down on a chair and reads short Japanese sentences at normal speed. During the utterances, two high-speed color video camera capture frontal and lateral images of subjects' lips, teeth, a jaw, and a part of a tongue. At the same time, speech acoustics are recorded. Figure 1 shows the equipment setting; (a) A top view drawing of equipment placements; (b) A picture of a front camera, a microphone, and a human subject in the experiment room.

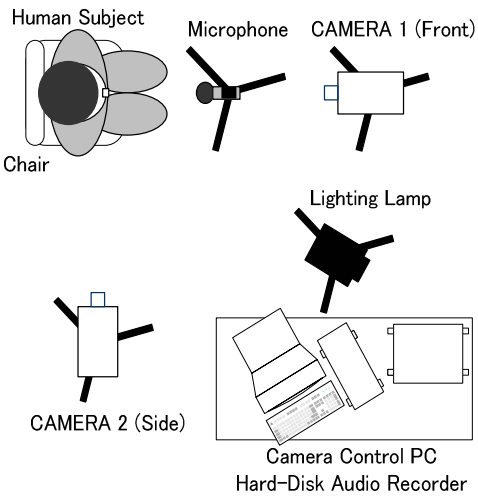
2.1. High-speed Video Camera Settings

The video cameras used for the experiments are high-speed color CCD cameras which support capturing up to 300 frames per second (fps). I used two speeds of 120 and 300 fps. They are 4 times and 10 times faster than a normal video rate of 30 fps in case of a NTSC system. Since Japanese speech at a normal speed contains about 6 Japanese syllables per second or about 12 phonemes of consonants and vowels per second, a 30 fps system captures only 2 or 3 frames per phonemes in average. They are too few to analyze and reproduce speech articulators' trajectories. 120 and 300 fps are much better.

Image resolutions are 256 x 256 pixels at 120 fps and 200 x 200 pixels at 300 fps. These image resolutions are not high compared with a normal video system of 640 x 480 pixels. These limitations are from specifications of the video camera devices. To measure lip positions as finely as possible, cameras are placed closely to a subject face to capture only an area around lips including a chin, not a whole area of his face. Of course, not only a lips area but also a wide area of a bottom half of a whole face are

 ISBN 1-876346-48-5

affected by speech utterances. But, an area around lips is still a central part of visemes which characterize phonemes.



(a) A top view of equipment placements

(b) A front camera, a microphone, and a human subject.

Figure 1: Experimental equipment setting.

Recording durations depend on frame rate and image size settings because captured images from CCD are put into the internal frame memory of the camera device. For 200 x 200 pixels at 300 fps, the maximum duration is 20 seconds. As the speech sentences used for the recording are around 4 or 6 seconds, one sentence is recorded at a time.

Recorded data in the frame memories of the camera devices must be transferred to a personal computer (PC) to watch them. The camera device unit is a PC controlled system and not a stand alone system. Therefore, data transfer from the cameras to a PC is needed every time when a sentence is recorded.

Two cameras are placed in front of and at the right side of a subject face. There is no camera at the left side, that is, I assume symmetry of left and right side views of lip movements although it is not always a truth. The distances of two cameras from the face are almost same so that same scaling can be used to convert pixels to actual lengths. Two cameras are synchronized using a trigger signal input capability of the devices. However, 1 or 2 frames may be shifted by the limitation of device performances. A push switch to make a triggering signal was made to start capturing.

As same as typical high-speed cameras, an extra lighting is needed to obtain lighter images. A video light of 600W halogen lamp is used from a frontal and slightly right direction.

2.2. Audio Recorder Settings

A hard disk recorder is used to record speech acoustics. Using a microphone stand, a condenser microphone is placed in front of, but lower than, a subject's face. Sampling frequency, number of channels, and sampling bits are 44.1 kHz, Mono, and 16 bits.

2.3. Speech Sentences

The sentences are taken from one of widely known speech database made from newspaper articles. The set consists of phonetically balanced short sentences which include all Japanese syllables and their typical trigram patterns. Each sentence consists of 27 Japanese syllables in average and its duration is around 5 seconds. The set includes not only traditional pure Japanese syllables but also foreign ones spoken in current modern Japanese. The number of sentences in one series of a phonetically balanced set is 50. 20 series exist in a whole set of the speech database, therefore there are 1000 sentences in total. However, so far, I use only the 1st 50 sentences to make viseme video database.

2.4. Human Subjects

So far, three native Japanese adult male subjects have been used for the experiments. In future, I would like to extend to variety of subjects.

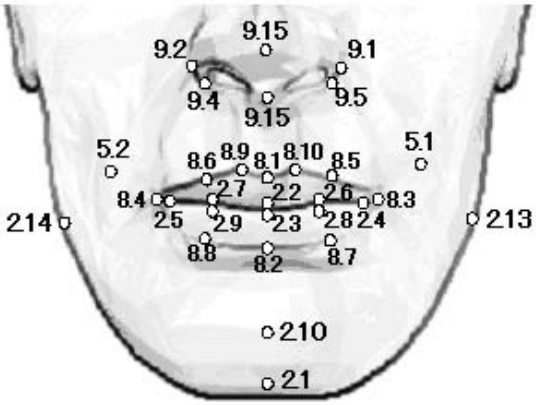
2.5. Marking on a Face

The data in the database are natural raw video images around lips supplied with speech acoustics and sentence texts.

However, to help analyzing movements, an additional configuration setting is added. That is, some points of lips, chin, nose, cheeks, etc. are marked with painting color. Silver dots, whose diameters are about 3 mm, are painted on a face. This technique enables easier tracking of speech articulators' movements. For this purpose, some of feature

ISBN 1-876346-48-5

points (FPs) defined by MPEG-4 [3] are used. Figure 2 (a) shows the 30 marking points used in this experiment. The numbers in the figure are feature point numbers in MPEG-4 standard. MPEG-4 has defined 84 feature points. Figure 2 (b) is a human subject who is marked by silver painting color.



(a) Marking positions and FP numbers

(b) Dot painting on a subject face

Figure 2: MPEG-4 feature points.

3. THE DATABASE SPECIFICATION

The data in the database are natural raw video images around lips supplied with speech acoustics and sentence texts. For one experiment of one subject, data of 50 sentences exist. Video data are stored as sequentially named bitmap files (BMP format). Frontal and lateral data are in separate files. It is possible to convert these files into one of a movie format. So far, neither labeling of temporal frames nor additional analysis have been supplied.

Figure 3 shows an example data from the video database. For this example, the camera configuration is 300 fps at 200 x 200 pixel resolution. The example is 10 sequential frames of utterance /ba/ in a part of a short sentence. This is 10 times more than a normal 30 fps video system. Different from a 30 fps video sytem, transition from /b/ to /a/ is displayed well.

4. DISCUSSION

4.1. Japanese Visemes

There is a well-known Japanese viseme set extracted from stroboscope observations by Fukuda and Hiki [4]. It has defined a viseme set (a list of mouth shaple symbols) of a, e, o, u, i, p, w, t, r, s, sy, y, Vf. The paper also describes temporal tendencies and coarticulation rules.

By watching the video database, it has been confirmed that their analysis are quite accurate. However, the data themselves of their paper have not been available any more so that the results cannot be used as quantitative usages such as facial animations nor further visual speech researches. Thus, this high-speed video data are useful to do such works. As the current paper is mainly focus on the measurement only, further analysis of Japanese visemes will be reported later in subsequent papers.

4.2. A Reason of Video Based Capturing

To record lips movements, methods can be classified into two types, one is marker based tracking and the other is video based image capturing. There are several types of products to do marker tracking depending on marker and sensor types, such as infrared LED tracking, magnetic field sensing, illuminated stickers, and others.

The virtues of the marker based tracking are, first, the tracking are quite accurate, for example, an infrared LED tracking system can measure differences less than 1 mm. And second, data processing is easier than video analysis, because 3-D Cartesian coordinate values of marker positions are automatically output from the system. Before doing these experiments reported in this paper, I had usually used to use marker tracking systems [5].

The virtues of video based capturing is that it does not loose any of visually perceivable phenomena of speech articulators. Thus, I think that video capturing is more suitable for viseme researches because real shapes can be extracted.

4.3. Facial Animation

My primary purpose of this veseme related research is that a realistic and easier facial animation (See Fleming and Dobbs [6] or Parke and Waters [7] for tutorials of facial animations). A talking head will be made and it will be highly compatible with the MPEG-4 standard.

ISBN 1-876346-48-5

Figure 3: /ba/ in a short sentence.

5. CONCLUSION

A video database was made for extracting Japanese visemes and modeling speech articulators' motions during speech utterances. By using two high-speed video cameras, frontal and lateral views around lips were recorded at up to 300 frames per second.

I hope that this video database will be used for viseme researches, the recording configuration will be standardized after discussion, and finally in the future, an accurate international viseme set will be developed such like as the International Phonetic Alphabet.

ACKNOWLEDGMENT

This research is supported by Grants-in-Aid for Scientific Research (KAKENHI) from Ministry of Education, Culture, Sports, Science and Technology (Number 14580431).

REFERENCES

[1] M. E. Hennecke, D. G. Stork, and K. V. Prasad, "Visionary Speech: Looking Ahead to Practical Speechreading Systems, " *Speechreading by Humans and Machines : Models, Systems and Applications*, D. G. Stork and M. E. Hennecke, Eds., pp. 332–349. Berlin Heidelberg: Springer-Verlag, 1996.

[2] M. M. Cohen (moderator), "Databases, Standards and Comparisons," in *Speechreading by Humans and Machines : Models, Systems and Applications*, D. G. Stork and M. E. Hennecke, Eds., pp. 541–548. Berlin Heidelberg: Springer-Verlag, 1996.

[3] ISO/IEC 14496-2 Information technology -- Coding of audio-visual objects -- Part 2: Visual, 2001.

[4] Y. Fukuda and S. Hiki, "Characteristics of the mouth shape in the production of Japanese–Stroboscopic observation," *J. Acoust. Soc. Jpn. (E)* **3**, 2, pp. 75–91, 1982.

[5] E. Vatikiotis-Bateson, K. G. Munhall, M. Hirayama, Y. V. Lee, and D. Terzopoulos, "The Dynamics of Audiovisual Behavior in Speech, " *Speechreading by Humans and Machines : Models, Systems and Applications*, D. G. Stork and M. E. Hennecke, Eds., pp. 541–548. Berlin Heidelberg: Springer-Verlag, 1996.

[6] B. Fleming and D. Dobbs, *Animating Facial Features and Expressions*, Massachusetts: Charles River Media Inc., 1999.

[7] F. I. Parke and K. Waters, *Computer Facial Animation*, Massachusetts: A K Peters, Ltd., 1996.