Multi-modal Translation System and its Evaluation

Shigeo MORISHIMA\(^{(1,2)}\), Satoshi NAKAMURA\(^{(2)}\)

\(^{(1)}\)Faculty of Engineering, Seikei University.
shigeo@ee.seikei.ac.jp

\(^{(2)}\)ATR, Spoken Language Translation Research Laboratories.

Abstract

Speech-to-speech translation has been studied to realize natural human communication beyond language barriers. Toward further multi-modal natural communication, visual information such as face and lip movements will be necessary. In this paper, we introduce a multi-modal English-to-Japanese and Japanese-to-English translation system that also translates the speaker's speech motion while synchronizing it to the translated speech. To retain the speaker's facial expression, we substitute only the speech organ's image with the synthesized one, which is made by a three-dimensional wire-frame model that is adaptable to any speaker. Our approach enables image synthesis and translation with an extremely small database. We conduct subjective evaluation tests using the connected digit discrimination test using data with and without audio-visual lip-synchronization. The results confirm the significant quality of the proposed audio-visual translation system and the importance of lip-synchronization.

1. Introduction

There have been demands to realize automatic speech-to-speech translation between different languages. In Japan, ATR had engaged to study especially on speech-to-speech translation technologies since 1986. The researches have been carried out to achieve translation of spoken dialogue among different languages. The research has been expanding to the technology for more extensive multiple domains, multiple languages, distant-talking speech, and daily conversation.

A speech translation system has been studied mainly for verbal information. However, both verbal and nonverbal information is indispensable for natural human communication. Facial expression plays an important role in transmitting both verbal and nonverbal information in face-to-face communication. Lip movements transmit speech information along audio speech. For example, stand-in speech in movies has the problem that it does not match the lip movements of the facial image. Face movements are also necessary to transmit nonverbal information of the speaker. In the case of making the entire facial image by computer graphics, it seems to be difficult to send messages of nonverbal information. If we could develop a technology that is able to translate the facial speaking motion synchronized to the translated speech, a natural multilingual multi-modal speech translation could be realized. There has been some researches\(^{[2]}\) on facial image generation to transform lip-shape based on concatenating variable units from huge database. However, since images generally contain much larger information than those of sounds, it is difficult to prepare large image databases. Thus conventional systems need to limit speakers.

Therefore, In this paper, we propose a method that uses both artificially generated images based on a 3-D wire-frame head model in the speaker's mouth region and captured images from the video camera for the other regions for natural speaking face generation.

We also propose a method to generate 3D personal face model with real personal face shape, and to track the face motion like movement and rotation automatically for audio-visual speech translation. The method enables to detect movement and rotation of the head given the three dimensional shape of the face, by template matching using a 3D personal face wire-frame model. We describe the method to generate a 3D personal face model, an automatic face tracking algorithm, and evaluation experiments of tracking accuracy. Finally we will demonstrate generated mouth
motions that the speaker has never spoken and show subjective evaluation result by connected digit discrimination using data with and without audio-visual lip-synchronization.

2. System overview

Fig. 1 shows an overview of the system in this research. The system is divided broadly into two parts: one is the Speech-Translation Part and the Image-Translation Part. The Speech-Translation Part is composed of ATR-MATRIX[1], which was developed in ATR-ITL. ATR-MATRIX is composed of ATR-SPREC to execute speech recognition, TDMT to handle text-to-text translation, and CHATR[3] to generate synthesized speech. The two parameters of phoneme notation and duration information, which are outputs from CHATR, are applied to facial image translation.

The first step of the Image-Translation Part is to make a 3-D model of the mouth region for each speaker by fitting a standard facial wire-frame model to an input image. Because of the differences in facial bone structures, it is necessary to prepare a personal model for each speaker, but this process is required only once for each speaker.

The second step of the Image-Translation Part is to generate lip movements for the corresponding utterance. The 3-D model is transformed by controlling the acquired lip-shape parameters so that they correspond to the phoneme notations from the database used at the speech synthesis stage. Duration information is also applied and interpolated by linear interpolation for smooth lip movement. Here, the lip-shape parameters are defined by a momentum vector derived from the natural face at lattice points on a wire-frame for each phoneme. Therefore, this database does not need speaker adaptation.

In the final step of the Image-Translation Part, the translated synthetic mouth region’s 3-D model is embedded into input images. In this step, the 3-D model’s color and scale are adjusted to the input images. Even if an input movie (image sequence) is moving during an utterance, we can acquire natural synthetic images because the 3-D model has geometry information. Consequently, the system outputs a lip-synchronized face movie to the translated synthetic speech and image sequence at 30 frames/sec.

3. Personal face model

It is necessary to make an accurate 3-D model that has the target person’s features for the face recreation by computer graphics. In addition, there is demand for a 3-D model that doesn’t need heavy calculation.

In our research, we use a 3-D head model[5][6] shown in
fig. 2, and tried to make a 3-D model of the mouth region. This 3-D head model is composed of about 1,500 triangular patches and has about 800 lattice points.

Face fitting tool developed by IPA[9] is a tool to generate a 3D face model using one’s photograph. But the manual fitting algorithm of this tool is very difficult and requires a lot of time for users to generate a 3D model with real personal face, although it is able to generate a model with nearly real personal shape with many photographs. Fig. 3 shows a personal face model. Fig.3(a) is an original face image, fig.3(b) shows fitting result of generic face model and fig.3(c) is a mouth part model constructed by a personal model used for mouth synthesis for lip synchronization.

In order to raise accuracy of face tracking using the 3D personal face model, we used a range scanner like Cyberware[10]. This is the head & face 3D color scanner which can capture both range data and texture shown in fig. 4.

We can generate a 3D model with a real personal shape using a standard face wire-frame model. First, to fit the standard face model to the Cyberware data, both the generic face model and Cyberware data are mapped to 2D cylindrical plane. Then, we manually fit a standard model’s face parts to corresponding Cyberware face parts by using texture data. This process is shown in fig. 5. Finally, we replace the coordinates values of the standard model to Cyberware range data coordinates values, and obtain an accurate 3D personal face model shown in fig.6.

### 4. Automatic face tracking

Many tracking algorithms have been studied by many researchers for a long time, and a lot of algorithms are applied to track a mouth edge, an eyes edge, and so on. However, because of blurring feature points between frames, or occlusion of the feature points by rotation of a head etc., these algorithms were not able to do accurate tracking.

In this chapter, we describe about an automatic face tracking algorithm using a 3D face model. Tracking process using template matching can be divided into three steps.

At first, texture mapping of one of the video frame images is carried out to the 3D individual face shape model created in Section 3. Here, a frontal face image is chosen out of video frames for the texture mapping. Because the target is video phone application, so the rotation angle in video sequence is very small.

Next, we make a 2D template images for every translation and rotation using a 3D model shown in fig.7. Here, in order to reduce a matching error, a mouth region is excluded in a
template image. Thereby, even while the person in a video image is speaking something, tracking can be carried out more stably. Approximation is that so much expression change does not happen in all video sequences.

Finally, we carry out template matching between the template images and an input video frame image and estimate translation and rotation values so that a matching error becomes minimum. This process is illustrated in fig. 8.

We show a flow chart for the search process of a face position and a rotation angle in one frame in Fig. 9. The template matching for the tracking is carried out using euclid error function in RGB value in all pixel normalized by pixel number within template.

Since template matching is performed only in the face region except the blue back of template images and thus the number of pixels are different for each template image, we apply normalization in the error function by the number of pixels.

5. Evaluation of tracking

We carried out tracking experiments to evaluate effectiveness of the proposed algorithm.

5.1 Measurement by OPTOTRAK[7]

To evaluate the accuracy of our tracking algorithm, we measure the face movement in video sequence using OPTOTRAK, the motion measurement system. We measured following head movement.

(1) Rotation of X axis. (2) Rotation of Y axis.
(3) Rotation of Z axis.
(4) Movement of X direction.

Henceforth, we treat the data obtained by OPTOTRAK as correct answer value for tracking.

5.2 Evaluation of the tracking

As an example of a tracking result, the comparison graph of rotation angle to Y axis is shown in a Fig. 10. An average of the error of the angle between the angle obtained by our algorithm and that by OPTOTRAK is about 0.477[degree].

This graph shows that the error increases as a rotation angle becomes large. This is because the front image is mapped on the 3D model.

Example of model match move into video frame are shown in Fig. 11. Top raw is original video frame chosen from sequence randomly. 2nd raw is synthetic face according to estimated position and rotation angle information by our algorithm. 3rd raw is generated image by replacing original face with synthetic one. By subjective test, the quality of synthesized image sequence looks so natural that no body can detect the replacement with synthetic face.
6. Lip shape in utterance

When a person says something, the lips and jaw move simultaneously. In particular, the movements of the lips are closely related to the phonological process, so the 3-D model must be controlled accurately.

As with our research, Kuratate et al.[4] tried to measure the kinematical data by using markers on the test subject’s face. This approach has the advantage of accurate measurements and flexible control. However, it depends on the speaker and requires heavy computation. Here, we propose a method by unit concatenation based on the 3-D model, since the lip-shape database is adaptable to any speaker.

6.1 Standard lip shape

For accurate control of the mouth region’s 3-D model, Ito et al. [6] defined seven control points on the model. Those points could be controlled by geometric movement rules based on the bone and muscle structure.

In this research, we prepared reference lip-shape images from the front and side. Then, we transformed the wire-frame model to approximate the reference images. In this process, we acquired momentum vectors of lattice points on the wire-frame model. Then, we stored these momentum vectors in the lip-shape database. This database is normalized by the mouth region’s size, so we do not need speaker adaptation. Thus, this system has realized talking face generation with a small database.

6.2 Lip shape classification by VISEME

Viseme is a word created from “phoneme”, which is the smallest linguistic sound unit. Visemes are generally also defined for lip movement information like [au] and [ei] of the phonetic alphabet, but in this research we decomposed those visemes further into shorter and more static units.

We classified English phonemes into 22 kinds of parts based on visemes. In addition to English, we classified Japanese vowel phonemes into 5 kinds of parts. We also prepared a silence interval viseme.

The system has as many standard lip-shapes in its database as the number of visemes. Japanese consonant lip-shape data come from 60% of the English consonant standard lip-shape data.

6.3 Utterance animation

The lip-shape database of this system is defined by only the momentum vector of lattice points on a wire-frame. However, there are no transient data among the standard lip-shapes. The system must have momentum vectors of the lattice point data on the wire-frame model while phonemes are being uttered. Therefore, we defined that the 3-D model configures a standard lip-shapes when a phoneme is uttered at any point in time. This point is normally the starting point of phoneme utterance and we defined keyframe at the starting point of each phoneme segment and interpolated by linear or sinusoidal curve between keyframes. Although this method is not directly connected with kinesiology, we believe that it provides a realistic lip-shape image.

7. Evaluation experiment

We carried out subjective experiments to evaluate effectiveness of the proposed image synthesis algorithm. In order to clarify the effectiveness of the proposed system, we...
carry out subjective digit discrimination perception tests. The test audio-visual samples are composed of connected 6 digits.

We tested using original speech and speaking face movies in speech. The original speech are used with condition of audio SNR=6,-12,-18dB using white Gaussian noise. Fig.12 shows the results. Subjects are composed of 10 students and they are asked to memorize each number sequence like phone number. Total number of sequence is 15. In each time, test sequence is selected randomly and heard by head-set. This test is performed in web page.

In every case, according to the low audio SNR, the subjective discrimination rates degrades. “VOICE ONLY” is only playback of speech without video. Even in case of SNR -6dB, discrimination rate is not 100%. However, by adding matched face movie, the rate becomes to 100% in all cases. “ORIGINAL” is combination of original voice and video-captured natural face image. In this case, even in -18dB high discrimination rate can be achieved. “LINEAR” is the case of linear interpolation of keyframes located a basic mouth shape and “SINUSOIDAL” is the case of non-linear interpolation using sinusoidal curve between keyframe. “MISMATCH” is using voice and mismatched video-captured face. Discrimination rate is drastically degrade in case of mismatch between voice and image even in -6dB.

Anyway, as a result nonlinear interpolation considering coarticulation is good scoring and proposed system significantly enhances the perception rates. This method gives good standard for evaluation of lip-synchronization. Better interpolation method for lip-synchronization will be researched and it will give a standard method for lip-sync problem.

8. Conclusion

As a result of this research, the proposed system can create any lip-shape with an extremely small database, and it is also speaker-independent. It also retains the speaker’s original facial expression by using input images other than the mouth region. Furthermore, this facial-image translation system, which is capable of multi-modal English-to-Japanese and Japanese-to-English translation, has been realized by applying the parameters from CHATR.

Furthermore, this system only works off-line in this research. To operate the system on-line, greater speed is necessary. In addition, because of the different durations between original speech and translated speech, a method that controls duration information from the image synthesis part to the speech synthesis part needs to be developed.

Evaluation of lip-synchronization method is proposed and it will give a standard method for lip-sync problem.

References


![Fig.12: Subjective Digit Discrimination Rate Plot](image-url)