AUDIO-VISUAL SPEECH TRANSLATION WITH AUTOMATIC LIP SYNCHRONIZATION AND FACE TRACKING BASED ON 3-D HEAD MODEL

Shigeo Morishima[‡,†], *Shin Ogata*[†,‡], *Kazumasa Murai*[†], *Satoshi Nakamura*[†]

[†] ATR Spoken Language Translation Research Laboratories.
2-2-2 Hikaridai Seika-cho Soraku-gun Kyoto, 619-0288 Japan
[‡] Faculty of Engineering, Seikei University,
3-3-1, Kichijoji-Kitamachi, Musashino city, Tokyo, 180-8633, Japan

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 by connected digit discrimination using data with and without audio-visual lip-synchronicity. The results confirm the sufficient quality of the proposed audio-visual translation system.

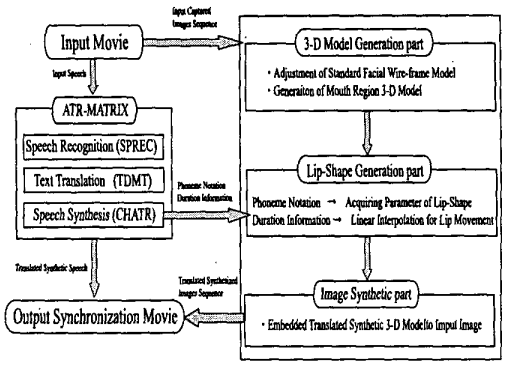


Fig. 1. System Overview

1. INTRODUCTION

There have been demands to realize automatic speech-to-speech translation between different languages. The studies 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 non-verbal information is indispensable for natural human communication. Facial expression plays an important role in transmitting both verbal and non-verbal 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 non-verbal 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 non-verbal information. If we could develop a technology that is able to translate the facial speaking motion synchronized to the translated speech, a natural multi-lingual 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.

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.

2. SYSTEM OVERVIEW

Figure 1 shows an overview of the system in this research. The system is divided 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.

0-7803-7402-9/02/$17.00 ©2002 IEEE

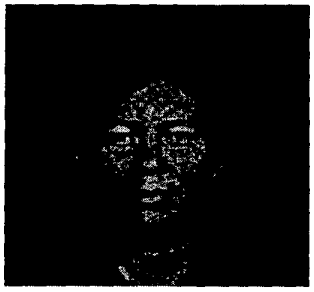


Fig.2(a) Original Image

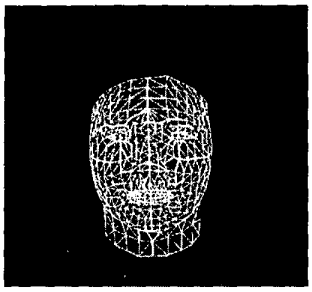


Fig.2(b) Fitting

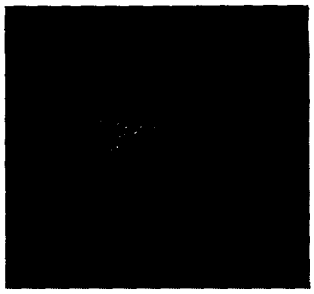


Fig.2(c) Mouth Model

ATR-MATRIX is composed of ATR-SPREC to execute speech recognition, TDMT to handle text-to-text dialog translation, and CHATR[3] to generate synthesized speech. In the process of audio-visual speech translation, 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. 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 used for linear interpolation for smooth lip movement. Here, the lip-shape parameters are defined by a vector derived from the natural face at lattice points on a wire-frame for each phoneme. In the final step of the Image-Translation Part, the translated synthetic mouth region 3-D model is embedded into input images. In this step, the 3-D model colour 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. GENERATION OF A 3D PERSONAL FACE MODEL

Face fitting tool developed by IPA[7] is a tool to generate a 3D face model using one 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. 2 shows a personal face model synthesis process. Fig.2(a) is a original face image, Fig.2(b) shows fitting result of generic face model[7] and Fig.2(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 face model, we used Cyberware[8], the head & face 3D color scanner, and generate a 3D model with a real personal shape using a standard face & head wire-frame model (Fig.2(b)). At first, to fit the standard face model to the Cyberware data, both the standard model and Cyberware data are mapped to 2D plane. Then, we manually fit a standard model's face parts to corresponding Cyberware face parts (Fig. 2(b)). Finally, we replace the coordinates values of the standard model to Cyberware coordinates values, and obtain a real 3D personal face model.

4. GENERATION OF LIP SHAPE IN UTTERANCE

When a talker utters, 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.

4.1. Standard Lip Shape Database

For control of the mouth region's 3-D model, Ito et al. [5] defined we use seven control points on the model[5]. 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 vectors of lattice points on the wire-frame model. Then, we stored these 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.

4.2. Classification of Phonological Rules 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

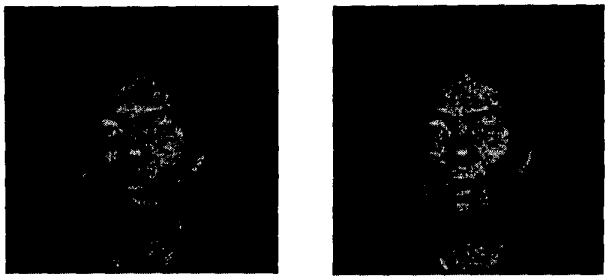
22 kinds of parts based on visemes. In addition to English, we classified Japanese vowel phonemes into 5 kinds of parts. We also prepared a silence interval viseme. The system has as many standard lip-shapes in its database as the number of visemes. Japanese consonant lip-shape data come from 60% of the English consonant standard lip-shape data. In English phonemes, there are kinds of visemes that are composed of multiple visemes. For example, these include [au], [ei] and [ou] of the phonetic alphabet. As stated previously, those visemes are decomposed into standard lip-shapes. We called them multiplicate visemes. Each parameter of phonemic notations from CHATR has duration information. However, the decomposed visemes need to be apportioned by duration information. We experientially apportioned 30% of the duration information to the front part of multiplicate visemes and the residual duration information to the back part of them.

4.3. Linear Interpolation for Lip Movement

The lip-shape database of this system is defined by only the vector of lattice points on a wire-frame. However, there are no data among the standard lip-shapes. We apply linear interpolation for lip movement by using duration information from CHATR. The system must have 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. Thereafter, we assign a 100% weight of the vector to the starting time and a 0% weight to the ending time and interpolate between these times. For the next phoneme, the weight of the vector is transformed from 0% to 100% as well as the current phoneme. By a value of the vector sum of these two weights, the system configures a lip-shape that has a vector unlike any in the database. Although this method is not directly connected with kinesiology, we believe that it provides a realistic lip-shape image.

5. 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 processing using template matching can be divided into the 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, the frontal face image is chosen out of video frame images for the texture mapping. The reason is that the video images used in this experiment are mainly frontal in movement and rotation. Next, we make a 2D template images for every movement and rotation using a 3D model. 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. Finally, we carry out template matching between the template images and an input video frame image, and estimate movement and rotation values so that a matching error becomes minimum.



(a) Original Image (b) Translated Synthetic Image

Fig. 3. Translated Speaking Face Image

Here, (RV,GV,BV) is a R,G,B value of and input video frame image, and (RT,GT,BT) is a R,G,B value of the template image. 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. Therefore, we set initial values of the position and angle as those in a past frame, and search for desired movement and rotation from 3n-1 hypothesis near the starting point[6].

6. EVALUATION EXPERIMENTS

We carried out subjective experiments to evaluate effectiveness of the proposed algorithm. Fig.3 shows an example of the translated speaking face image. 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 digits from 4 to 7 digits. We tested using speech and speaking face movies in speech:(original or synthetic)x image:(original or synthetic) for (matched or mismatched) conditions by 10 subjects. The original speech and synthetic speech by CHATR are used with condition of audio SNR=-6,-12,-18dB using white Gaussian noise. Fig.4 shows the results. According to the low audio SNR, the subjective discrimination rates degrades. However, the proposed system enhances the perception rates compared to the case "synthetic speech" and "original speech and original face with mismatched utterance" in SNR=-12dB condition. It is also found that mismatched combination such as original speech and synthetic face, and original speech and mismatched original face sometimes hurts human perception. In order to assure effectiveness of the audi-visual translation, futher improvements will be necessary.

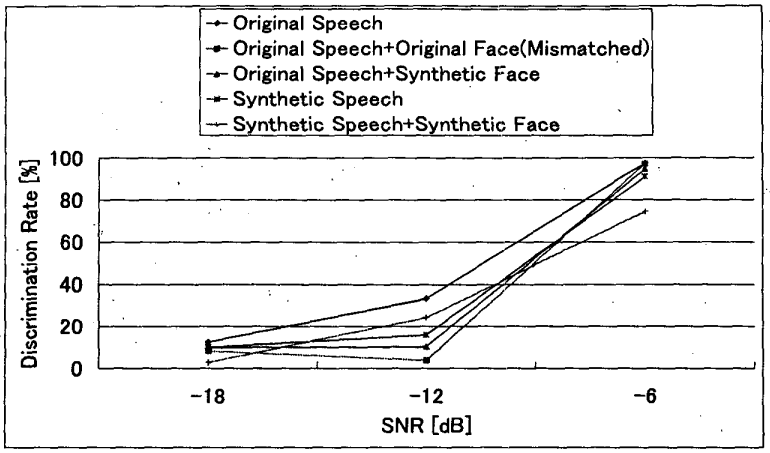


Fig. 4. Subjective Digit Discrimination Rate

7. 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. For further improvement of this system, we need to model a tongue by using in this 3-D model. The tongue model also needs to be controlled by parameters, and it will help the entire 3-D model to express images more precisely. At same time, we must retrace the method of the linear interpolation of the lip-shape. 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.

8. REFERENCES

[1] T. Takezawa, T. Morimoto, Y. Sagisaka, N. Campbell,H. Iida, F. Sugaya, A. Yokoo, and S. Yamamoto, "Japanese-to-English speech translation system:ATR-MATRIX" Proc. International Conference of Spoken Language Processing, ICSLP, pp. 957-960, 1998.

[2] H.P. Graf, E.Cosatto, T.Ezzat, "Face Analysis for the Synthesis of Photo-Realistic Talking Heads", Proc. 4th International Conference on Automatic Face and Gesture Recognition, pp. 189-194, 2000.

[3] N.Campbell, A.W.Black, "Chatr:A multi-lingual speech re-sequencing synthesis system", IEICE Technical Report, sp96-7, pp. 45, 1995.

[4] T. Kuratate, H. Yehia, E. Vatikiotis-Bateson, " KINEMATICS-BASED SYNTHESIS OF REALISTIC TRACKING FACE", Proc. International Conference on Auditory-Visual Speech, Processing, AVSP",8, pp. 185-190, 1998

[5] K. Ito, T. Misawa, J. Muto, S. Morishima, "3-D Lip Expression Generation by using New Lip Parameters", IEICE Technical Report, A-16-24, pp. 328, 2000.

[6] T. Misawa, S. Nakamura, S. Morishima, "Automatic Face Tracking And Model Match Move In Video Sequence Using 3-D Face Model", Proc. International Conference of Multi-Media, Expo, ICME 2001, 2001

[7] Facial Image Processing System for Human-like "Kansei" Agent web site : http://www.tokyo.image-lab.or.jp/aa/ipa/

[8] Cyberware Head & Face Color 3D Scanner Web Site : http://www.cyberware.com/products/index.html