EURASIP Journal on Applied Signal Processing 2004:11, 1637–1647

# Multimodal Translation System Using Texture-Mapped Lip-Sync Images for Video Mail and Automatic Dubbing Applications

**Shigeo Morishima**

*School of Science and Engineering, Waseda University, Tokyo 169-8555, Japan*
*Email: shigeo@waseda.jp*

*ATR Spoken Language Translation Research Laboratories, Kyoto 619-0288, Japan*

**Satoshi Nakamura**

*ATR Spoken Language Translation Research Laboratories, Kyoto 619-0288, Japan*
*Email: satoshi.nakamura@atr.jp*

*Received 25 November 2002; Revised 16 January 2004*

We introduce a multimodal English-to-Japanese and Japanese-to-English translation system that also translates the speaker's speech motion by synchronizing it to the translated speech. This system also introduces both a face synthesis technique that can generate any viseme lip shape and a face tracking technique that can estimate the original position and rotation of a speaker's face in an image sequence. To retain the speaker's facial expression, we substitute only the speech organ's image with the synthesized one, which is made by a 3D wire-frame model that is adaptable to any speaker. Our approach provides translated image synthesis with an extremely small database. The tracking motion of the face from a video image is performed by template matching. In this system, the translation and rotation of the face are detected by using a 3D personal face model whose texture is captured from a video frame. We also propose a method to customize the personal face model by using our GUI tool. By combining these techniques and the translated voice synthesis technique, an automatic multimodal translation can be achieved that is suitable for video mail or automatic dubbing systems into other languages.

**Keywords and phrases:** audio-visual speech translation, lip-sync talking head, face tracking with 3D template, video mail and automatic dubbing, texture-mapped facial animation, personal face model.

## 1. INTRODUCTION

The facial expression is thought to send most of the nonverbal information in ordinary conversation. From this viewpoint, many researches have been carried on face-to-face communication using a 3D personal face model, sometimes called an "Avatar" in cyberspace [1].

For spoken language translation, ATR-MATRIX (ATR's multiligual automatic translation system for information exchange) [2] has been developed for the limited domain of hotel reservations between Japanese and English. A speech translation system has been developed for verbal information, although it does not take into account articulation and intonation. Verbal information is the central element in human communications, but the facial expression also plays an important role in transmitting information in face-to-face communication. For example, dubbed speech in movies has the problem that it does not match the lip movements of the facial image. In the case of making the entire facial image by computer graphics, it is difficult to send messages of original nonverbal information. If we could develop a technology that is able to translate facial speaking motion synchronized to translated speech where facial expressions and impressions are stored as effectively as the original, a natural multi-lingual tool could be realized.

There has been some research [3] on facial image generation to transform lip shapes based on concatenating variable units from a huge database. However, since images generally contain much larger information than that of sounds, it is difficult to prepare large image databases. Thus conventional systems need to limit speakers.

Therefore, we propose a method that uses a 3D wireframe model to approximate a speaker's mouth region and captured images from the other regions of the face. This approach permits image synthesis and translation while storing the speaker's facial expressions in a small database.

If we replace only the mouth part of an original image sequence, the translation and rotation of head have to

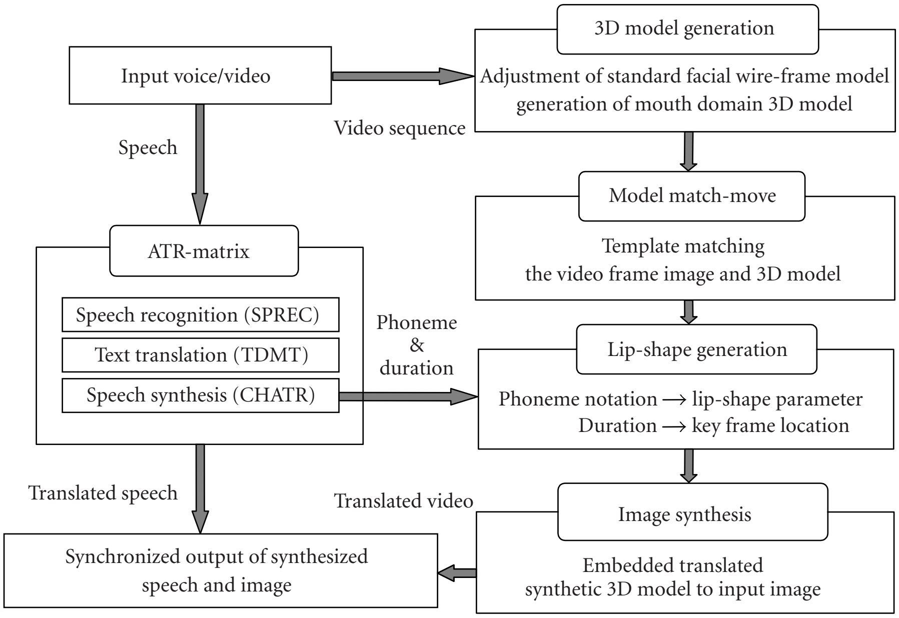

3D model generation
Input voice/video
Adjustment of standard facial wire-frame model generation of mouth domain 3D model
Video sequence
Speech
Model match-move
Template matching the video frame image and 3D model
ATR-matrix
Speech recognition (SPREC)
Text translation (TDMT)
Speech synthesis (CHATR)
Phoneme & duration
Lip-shape generation
Phoneme notation ⟶ lip-shape parameter
Duration ⟶ key frame location
Translated speech
Translated video
Image synthesis
Synchronized output of synthesized speech and image
Embedded translated synthetic 3D model to input image

Figure 1: Overview of the system.

be estimated accurately while keeping smooth motion between frames. We propose a method to generate a 3D face model with a real personal face shape and to track face motion such as translation and rotation automatically for audio-visual speech translation. The method enables the proposed system to detect movement and rotation of the head from the 3D shape of the face by template matching using a 3D personal face wire-frame model.

We describe a speech translation system, the method to generate a 3D personal face model, an automatic face-tracking algorithm, and experiments to evaluate tracking accuracy. Finally, we show generated mouth motions that were never spoken by a speaker and introduce a method to evaluate lip synchronization.

## 2. OVERVIEW OF MULTIMODAL TRANSLATION

Figure 1 shows an overview of the system developed in this research. The system is divided broadly into two parts: the *speech translation part* and the *image translation part*.

The speech translation part is composed of ATR-MATRIX [2], which was developed in ATR-ITL. ATR-MATRIX is composed of ATR-SPREC to execute speech recognition, transfer-driven machine translation (TDMT) to handle text-to-text translation, and CHATR [4] 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 3D 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 3D 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 linearly 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 3D model is embedded into input images. In this step, the 3D 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 3D model has geometry information.

Consequently, the system outputs a lip-synchronized face movie for the translated synthetic speech and image sequence at 30 frames/second.

## 3. SPEECH TRANSLATION SYSTEM

The system is based on the speech-to-speech translation system developed at ATR [2]. This system is called ATR-MATRIX. The system consists of speech recognition, language translation, and speech synthesis modules. The speech recognition module is able to recognize naturally spoken utterances in the source language. The language translation module is able to translate the recognized utterances to sentences in the target language. Finally, the translated sentences are synthesized by the text-to-speech synthesis module. In the following section, each of the modules is described.

### 3.1. Speech recognition system

The long research history and continuous efforts of data collection at ATR have made a statistical model-based speech recognition module possible. The module is speaker-independent and able to recognize naturally spoken utterances. In particular, the system drives multiple acoustic models in parallel in order to handle differences in gender and speaking styles.

Speech recognition is achieved by the maximum a posterior (MAP) decoder, which maximizes the following probability:

$$\begin{aligned}\hat{W} &= \arg\max P(W|O)\\ &= \arg\max P(O|W)P(W).\end{aligned} \tag{1}$$

Here, $P(O|W)$ and $P(W)$ are called "acoustic model probability" and "language model probability, respectively."

Parameters of the acoustic model and the language model are estimated by speech data and text data. For the acoustic model, a hidden Markov model (HMM) is widely used. However, the conventional HMM has problems in generating optimal state structures. We devised a method called HMnet (hidden Markov network), which is a data-driven automatic state network generation algorithm. This algorithm iteratively increases the state network by splitting one state into two states by considering the phonetic contexts so as to increase likelihood [5]. Speech data is sampled at 16 kHz and 16 bits. Short-time Fourier analysis with a 20-millisecond-long window every 10 milliseconds is adopted. Then, after a Mel-frequency bandpass filter and log compression, the twelfth-order Mel-frequency cepstrum coefficients and their first- and second-order time derivatives are extracted. For acoustic model training, we used 167 male and 240 female speech samples of travel dialogue and phonetically balanced sentences. Total length of the training data is 28 hours. Using this data, the structure and parameters of the HMnet acoustic model are determined. The estimated model is composed of 1400 states with 5 Gaussian mixtures for speech and 3 states with 10 Gaussian mixtures for the silence model.

For the language model, the statistical approach called the "$N$-gram language model" is also widely used. This model characterizes a probability of the word occurrence by the conditional probability-based previous word history. A trigram language model defined by the previous two words is widely used. The length of "$N$" should be determined by considering the trade-off between the number of parameters and the amount of training data. Once we get a text corpus, the $N$-gram language model can be easily estimated. For the word triplets that occur infrequently, probability smoothing is applied. In our system, a word-class-based $N$-gram model is used. This method reduces the training data problems and the unseen triplets problem by using a word class as part-of-speech. For language model training, we used 7,000 transcribed texts from real natural dialogues in travel domain. The total number of words is 27,000. Using this text corpus, the class-based $N$-gram language model is estimated for 700 classes.

Finally, the speech recognition system searches the optimal word sequence using the acoustic models and the language models. The search is a time-synchronous two-pass search after converting the word vocabulary into a tree lexicon. The multiple acoustic models can be used in the search but get pruned by considering likelihoods scores.

The performances of speaker-independent recognition in the travel arrangement domain were evaluated. The word error rates for face-to-face dialogue speech, bilingual speech, and the machine-friendly speech are 13.4%, 10.1%, and 5.2%, respectively.

### 3.2. Speech synthesis system

The speech synthesis system generates natural speech from the translated texts. The speech synthesis system developed at ATR is called CHATR [6]. The CHATR synthesis relies on the fact that a speech segment can be uniquely described by the joint specification of its phonemic and prosodic environmental characteristics. The synthesizer performs a retrieval function, first predicting the information that is needed to complete a specification from an arbitrary level of input and then indicating the database segments that best match the predicted target specifications. The basic requirement for input is a sequence of phone labels, with associated fundamental frequency, amplitudes, and durations for each. If only words are specified in the input, then their component phones will be generated from a lexicon or by rule; if no prosodic specification is given, then a default intonation will be predicted from the information available.

The CHATR preprocessing of a new source database has two stages. First, an analysis stage takes as its input an arbitrary speech corpus with an orthographical transcription and then produces a feature vector describing the prosodic and acoustic attributes of each phone in that corpus. Second, a weight-training stage takes as its input the feature vector and a waveform representation and then produces a set of weight vectors that describe the contribution of each feature toward predicting the best match to a given target specification.

At synthesis time, the selection stage takes as its input the feature vectors, the weight vectors, and a specification of the target utterance to produce an index into the speech corpus for random-access replay to produce the target utterance.

### 3.3. Language translation system

The translation subsystem uses an example-based approach to handle spoken language [7]. Spoken language translation faces problems different from those of written language translation. The main requirements are (1) techniques for handling ungrammatical expressions, (2) a means for processing contextual expressions, (3) robust methods for speech recognition errors, and (4) real-time speed for smooth communication.

The backbone of ATR's approach is the translation model called TDMT [8], which was developed within an example-based paradigm. TDMT's constituent boundary parsing [9] provides efficiency and robustness. We have also explored the processing of contextual phenomena and a method for

Table 1: Quality and time.

| Language conversion | Japanese-to-English | Japanese-to-German | Japanese-to-Korean | English-to-Japanese |
|---|---|---|---|---|
| A (%) | 43.4 | 45.8 | 71.0 | 52.1 |
| A + B (%) | 74.0 | 65.9 | 92.7 | 88.1 |
| A + B + C (%) | 85.0 | 86.4 | 98.0 | 95.3 |
| Time (seconds) | 0.09 | 0.13 | 0.05 | 0.05 |

dealing with recognition errors and have made much progress in these explorations.

In TDMT, translation is mainly performed by a transfer process that applies pieces of transfer knowledge of the language pair to an input utterance. The transfer process is the same for each language pair, that is, Japanese-English, Japanese-Korean, Japanese-German, and Japanese-Chinese, whereas morphological analysis and generation processes are provided for each language, that is, Japanese, English, Korean, German, and Chinese.

The transfer process involves the derivation of possible source structures by a constituent boundary parser (CB-parser) [9] and a mapping to target structures. When a structural ambiguity occurs, the best structure is determined according to the total semantic distances of all possible structures. Currently, the TDMT system addresses dialogues in the *travel domain*, such as travel scheduling, hotel reservations, and trouble-shooting. We have applied TDMT to four language pairs: Japanese-English, Japanese-Korean [10], Japanese-German [11], and Japanese-Chinese [12]. Training and test utterances were randomly selected for each dialogue from our speech and language data collection, which includes about 40 000 utterances in the travel domain. The coverage of our training data differs among the language pairs and varies between about 3.5% and about 9%. A system dealing with spoken dialogues is required to realize a quick and informative response that supports smooth communication. Even if the response is somewhat broken, there is no chance for manual pre-/postediting of input/output utterances. In other words, both speed and informativity are vital to a spoken-language translation system. Thus, we evaluated TDMT's translation results for both *time* and *quality*.

Three native speakers of each target language manually graded translations for 23 dialogues (330 Japanese utterances and 344 English utterances, each about 10 words). During the evaluation, the native speakers were given information not only about the utterance itself but also about the previous context. The use of context in an evaluation, which is different from typical translation evaluations, is adopted because the users of the spoken-dialogue system consider a situation naturally in real conversation.

Each utterance was assigned one of four ranks for translation quality:

(A) perfect: no problem in either information or grammar;
(B) fair: easy to understand with some unimportant information missing or flawed grammar;
(C) acceptable: broken but understandable with effort;
(D) nonsense: important information has been translated incorrectly.

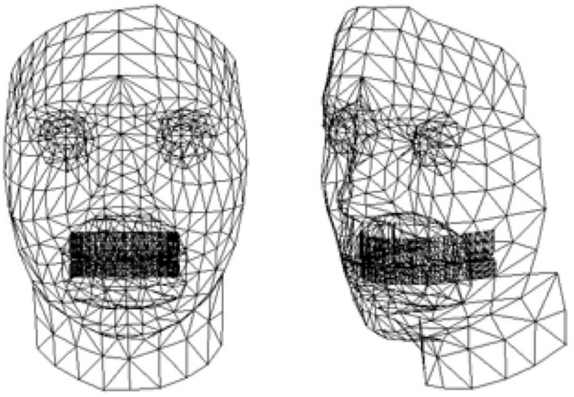


Figure 2: 3D head model.

Table 1 shows the latest evaluation results for TDMT, where the "acceptability ratio" is the sum of the (A), (B), and (C) ranks. The JE and JG translations achieved about 85% acceptability, and the JK and EJ translations achieved about 95% acceptability. JK's superiority is due to the linguistic similarity between the two languages; EJ's superiority is due to the relatively loose grammatical restrictions of Japanese.

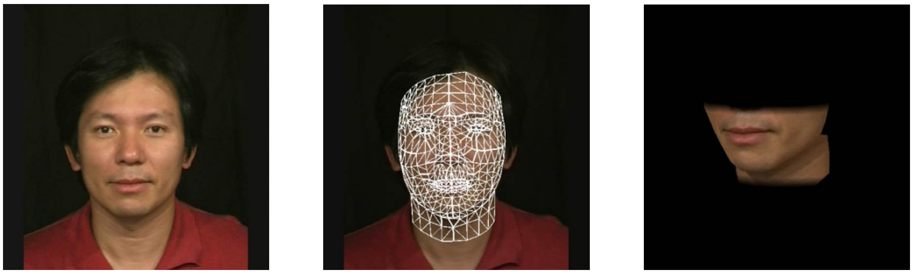
The translation speed was measured on a PC/AT Pentium II/450 MHz computer with 1 GB of memory. The translation time did not include the time needed for a morphological analysis, which is much faster than a translation. Although the speed depends on the amount of knowledge and the utterance length, the average translation times were around 0.1 seconds. Thus, TDMT can be considered efficient.

## 4. GENERATING PERSONAL FACE MODEL

It is necessary to make an accurate 3D model that has the target person's features for the face recreation by computer graphics. In addition, there is demand for a 3D model that does not need heavy calculation load for synthesis because this model is used for both generating a face image and tracking face location, size, and angle.

In our research, we used the 3D head model [13, 14] shown in Figure 2 and tried to make a 3D model of the mouth region. This 3D head model is composed of about 1,500 triangular patches and has about 800 lattice points.

The face fitting tool developed by IPA (Facial image processing system for Human-like "kansei" Agent, http://www.tokyo.image-lab.or.jb/aa/ipa/) is often used to generate a 3D face model using one photograph. However, the manual fitting algorithm of this tool is very difficult and requires a lot of time for users to generate a 3D model with a real personal



(a) (b) (c)

Figure 3: 3D model generation process. (a) Input image. (b) Fitting result. (c) Mouth model.

Figure 4: Head and face 3D color range scanner.

face, although it is able to generate a model with a nearly real personal shape along with many photographs. Figure 3 shows a personal face model. Figure 3a is an original face image, Figure 3b shows the fitting result of a generic face model, and Figure 3c is the mouth model constructed by a personal model used for the mouth synthesis in lip synchronization.

In order to raise the accuracy of face tracking by using the 3D personal face model, we used a range scanner like *Cyberware* [14], shown in Figure 4. This is a head-and-face 3D color scanner that can capture both range data and texture as shown in Figure 5.

We can generate a 3D model with a real personal shape by using a standard face wire-frame model. First, to fit the standard face model to the *Cyberware* data, both the generic face model and the *Cyberware* data are mapped to a 2D cylindrical plane. Then, we manually fit a standard model's face parts to the corresponding *Cyberware* face parts by using texture data. This process is shown in Figure 6. Finally, we replace the coordinate values of the standard model to *Cyberware* range data coordinates values and obtain an accurate 3D personal face model shown in Figure 7.

The face fitting tool provides a GUI that helps the user to fit a generic face wire-frame model onto texture face data accurately and consistently with coarse-to-fine feature points selection.

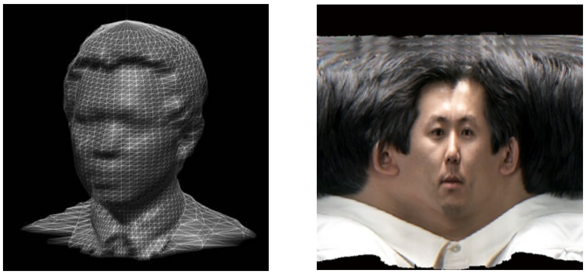
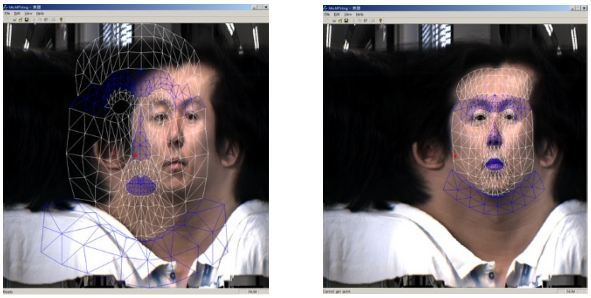


Figure 5: Acquired shape and texture.



(a) (b)

Figure 6: Face parts fitting on 2D plane. (a) Before fitting. (b) After fitting.

## 5. AUTOMATIC FACE TRACKING

Many tracking algorithms have been studied for a long time, and many of them have been applied to tracking a mouth edge, an eye edge, and so on. However, because of such problems as blurring of the feature points between frames or occlusion of the feature points by rotation of a head, these algorithms have not been able to provide accurate tracking.

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

First, texture mapping of one of the video frame images is carried out using the 3D individual face shape model created in Section 3. Here, a frontal face image is chosen from the video frames for the texture mapping.

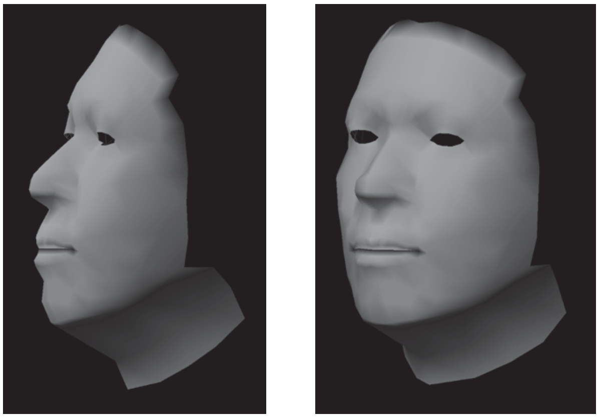


Figure 7: Generated 3D personal model.

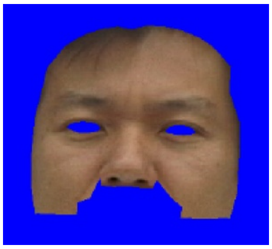


Figure 8: Template face image.

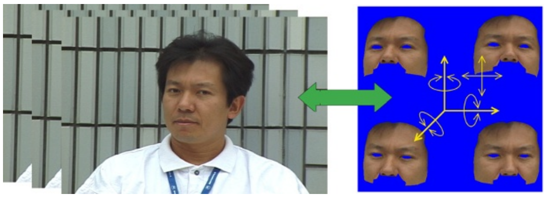


Figure 9: Template matching mechanism.

Next, we make 2D template images for every translation and rotation by using the 3D model shown in Figure 8. Here, in order to reduce matching errors, the mouth region is excluded from a template image. Consequently, even while the person in a video image is speaking something, tracking can be carried out more stably.

Expression change also can be handled by modifying a 3D template with the face synthesis module. The face synthesis module in the IPA face tool (http://www.tokyo.image-lab.or.jb/aa/ipa) can generate a stereotype face expression and also introduce personal character.

Currently, the test video image sequence includes slight and ordinary expression change but does not include an emotional expression. Therefore, modifying the face template is not currently considered, and the face template is treated as a rigid body.

Finally, we carry out template matching between the template images and an input video frame image and estimate translation and rotation values so that matching errors become minimum. This process is illustrated in Figure 9.

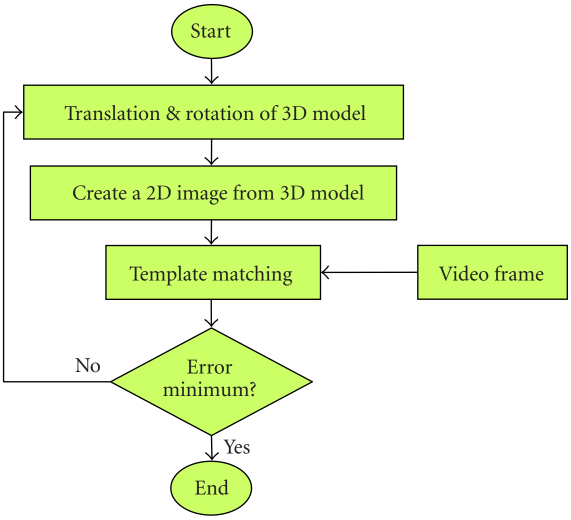

Start
Translation & rotation of 3D model
Create a 2D image from 3D model
Template matching
Video frame
Error minimum?
No
Yes
End

Figure 10: Flow of face tracking.

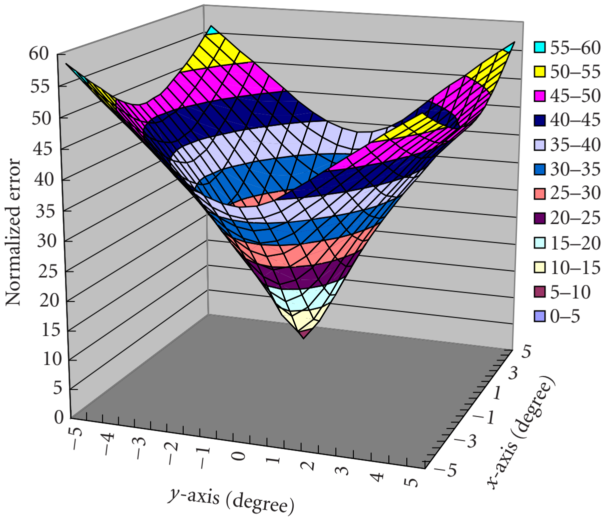


Figure 11: Error graph for rotation.

We show a flow chart for the search process of a face position and a rotation angle in one frame in Figure 10. The template matching for the tracking is carried out by using a euclid error function in the RGB value of all pixels normalized by the pixel number within a template.

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

By searching for a certain area, we obtain an error graph as shown in Figure 11. An approximation shows that there is only one global minimum. Therefore, we set initial values of the position and angle to those in the previous frame and search for desired movement and rotation from a $3^n - 1$ hypothesis near the starting point. We show a conceptual figure of the minimum error search in Figure 12.

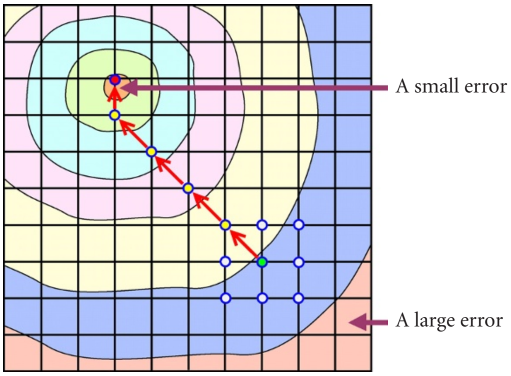


FIGURE 12: $3^n - 1$ gradient error search.

## 6. EVALUATION OF FACE TRACKING

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

### 6.1. Measurement by OPTOTRAK

To evaluate the accuracy of our tracking algorithm, we measured the face movement in a video sequence using *OPTOTRAK* (see [15]), the motion measurement system. We measured the following head movements:

(1) rotation of $x$-axis,
(2) rotation of $y$-axis,
(3) rotation of $z$-axis,
(4) movement of $x$ direction.

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

### 6.2. Evaluation of the tracking

As an example of a tracking result, a graph computing the rotation angle to $y$-axis is shown in Figure 13. The average of the angle error between the angle obtained by our algorithm and that by OPTOTRAK is about 0.477 (degree).

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

An example of a model matching movement in a video frame is shown in Figure 14. The top row is the original video frame chosen from the sequence randomly. The second row is a synthetic face according to the position and rotation angle estimated by our algorithm. The third row is the image generated by replacing the original face with a synthetic one. From a subjective test, the quality of the synthesized image sequence looks so natural that it is impossible to distinguish the replacement face from the original one.

### 6.3. Processing speed

The system configuration is as follows:

(i) CPU: Xeon 2 GHz;
(ii) Memory: 1 GB;
(iii) OS: Microsoft Windows 2000 Professional;
(iv) VGA: 3D labs Wild Cat 5110.

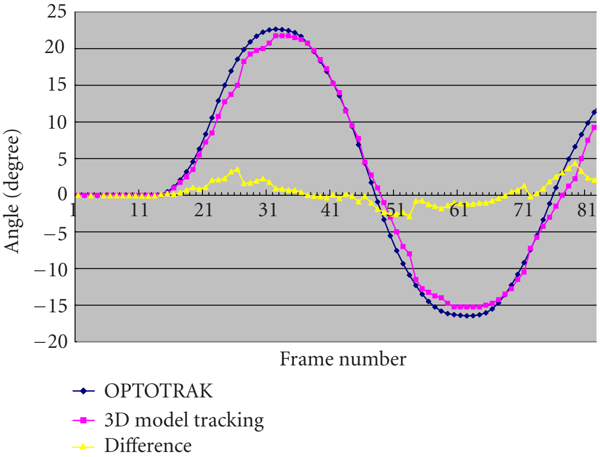


FIGURE 13: Evaluation of rotation angle with $y$-axis.

In the first frame of the video sequence, it takes about 30 seconds because a full screen search is needed. In the succeeding frames with little head motion, the searching region is limited locally to the previous position so this becomes 3 seconds. When the head motion becomes bigger, the searching path becomes deeper and convergence takes a longer time of up to 10 seconds.

Currently, this is too slow to realize a real-time application, but the delay time is only one video frame theoretically, so a higher-speed CPU and video card can overcome this problem in the future.

## 7. 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 3D model must be controlled accurately.

As with our research, Kuratate et al. [16] tried to measure the kinematical data by using markers on the test subject's face. This approach has the advantage of accurate measurement and flexible control. However, it depends on the speaker and requires heavy computation. Here, we propose a method by unit concatenation based on the 3D model, since the lip-shape database is adaptable to any speaker.

### 7.1. Standard lip shape

For accurate control of the mouth region's 3D model, Ito et al. [14] defined seven control points on the model. These are shown in Figure 15. 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 achieved talking face generation with a small database.

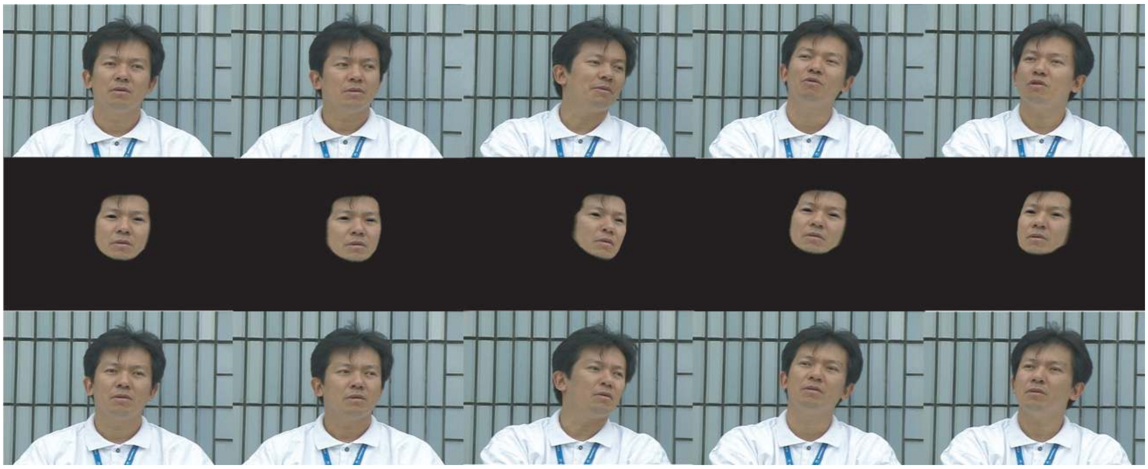


Figure 14: Examples of model match-move.

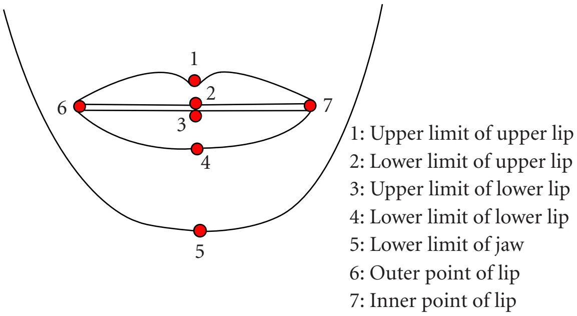

1
2
6
7
3
4
5
1: Upper limit of upper lip
2: Lower limit of upper lip
3: Upper limit of lower lip
4: Lower limit of lower lip
5: Lower limit of jaw
6: Outer point of lip
7: Inner point of lip

Figure 15: Positions of control points.

### 7.2. 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 silent interval viseme. Table 2 shows the correspondences of these 28 visemes to the phonemic notation outputs from CHATR.

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 standard English consonant lip-shape data.

In English phonemes, some kinds of visemes 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 call these multiplicate visemes.

Each parameter of phonemic notations from CHATR has duration information. Furthermore, the decomposed visemes need to be apportioned by duration information. We experimentally apportioned 30% of the duration information to the front part of multiplicate visemes and the residual duration information to the back part of them.

### 7.3. 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. In this section, we describe an interpolation method for lip movement by using duration information from CHATR.

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 3D model configures a standard lip shape when a phoneme is uttered at any point in time. This point is normally the starting point of a phoneme utterance, and we defined the keyframe at the starting point of each phoneme segment.

Thereafter, we assign a 100% weight of the momentum vector to the starting time and a 0% weight to the ending time and interpolate these times by a sinusoidal curve between them.

For the next phoneme, the weight of the momentum 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. The sinusoidal interpolation is expressed as follows. When a keyframe lip-shape vector is defined as $V_n$ located at $t = t_n$, and a previous keyframe vector is defined as $V_{n-1}$ at $t = t_{n-1}$, an interpolation between these keyframes is realized by the weight $W_n$, and the lip-shape vector is described as $M(t)$:

$$M(t) = W_n(V_{n-1} - V_n) + 0.5(V_{n-1} + V_n), \quad t = [t_{n-1}, t_n],$$
$$\text{where } W_n = 0.5\cos\left(\frac{t - t_{n-1}}{t_n - t_{n-1}}\right)\pi. \tag{2}$$

Table 2: Classification of visemes.

| Viseme number | Phoneme notation from CHATR | |
|---|---|---|
| 1 | /ae/ | English |
| 2 | /ah/, /ax/ | |
| 3 | /A/ | |
| 4 | /aa/ | |
| 5 | /er/, /ah r/ | |
| 6 | /iy/, /ih/ | |
| 7 | /uh/ | |
| 8 | /uw/ | |
| 9 | /eh/ | |
| 10 | /oh/, /ao/ | |
| 11 | /ax r/ | |
| 12 | /l/ | |
| 13 | /r/ | |
| 14 | /b/, /p/, /m/ | |
| 15 | /t/ | |
| 16 | /d/, /n/ | |
| 17 | /k/, /g/, /hh/, /ng/ | |
| 18 | /f/, /v/ | |
| 19 | /s/, /z/, /sh/, /zh/, /ts/, /dz/, /ch/, /jh/ | |
| 20 | /th/, /dh/ | |
| 21 | /y/ | |
| 22 | /w/ | |
| 23 | /a/, /A/ | Japanese |
| 24 | /i/, /I/ | |
| 25 | /u/, /U/ | |
| 26 | /e/, /E/ | |
| 27 | /o/, /O/ | |
| 28 | /#/ | Silence interval |

## 8. EVALUATION EXPERIMENTS

We carried out subjective experiments to evaluate effectiveness of the proposed image synthesis algorithm. Figures 16 and 17 show examples of the translated speaking face image. In order to clarify the effectiveness of the proposed system, we carried out subjective digit discrimination perception tests. The test audio-visual samples are composed of connected 4 to 7 digits in Japanese.

We tested using original speech and speaking face movies in speech. The original speech is used under the audio conditions of SNR = $-6, -12, -18$ dB using white Gaussian noise. Figure 18 shows the results. Subjects are 12 Japanese students (10 males and 2 females) in the same laboratory. Discrimination rate means the rate users can recognize each digit accurately by listening with headphones.

In every case, according to the low audio SNR, the subjective discrimination rates degrade. "Voice only" is only playback of speech without video. Even in the case of SNR = $-6$ dB, the discrimination rate is not 100%. However, by adding a matched face movie, the rate becomes 100% in all cases. "Original" is a combination of the original voice and the video-captured natural face image. In this case, even at $-18$ dB, a high discrimination rate can be achieved. "Linear" indicates linear interpolation of keyframe parameters of the basic mouth shape. Lip-shape vector $M(t)$ is expressed as

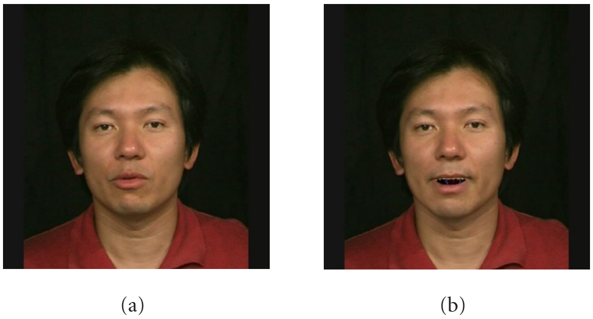

(a) (b)

Figure 16: Translated synthetic image from Japanese to English. (a) Original image. (b) Synthetic image.

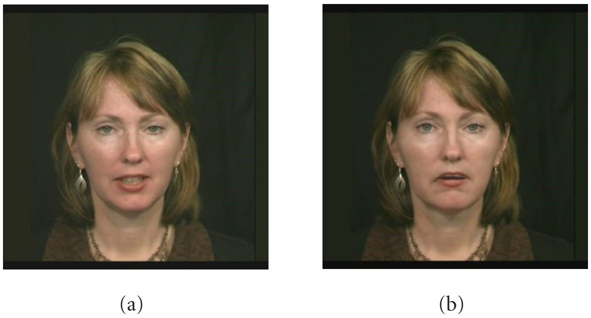

(a) (b)

Figure 17: Translated synthetic image from English to Japanese. (a) Original image. (b) Synthetic image.

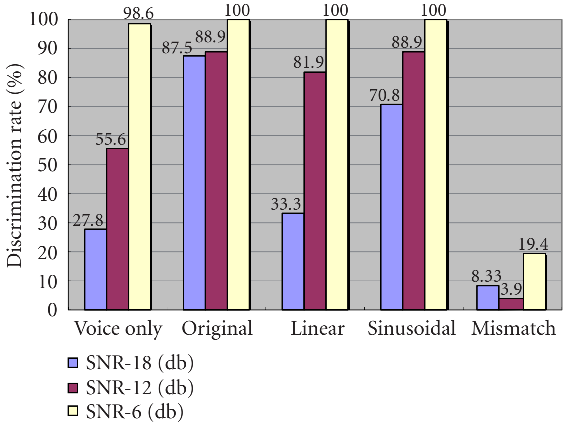


Figure 18: Subjective digit discrimination rate: evaluation test result.

follows:

$$M(t) = V_{n-1} + \alpha t \quad t = [t_{n-1}, t_n], \tag{3}$$

where $\alpha = (V_n - V_{n-1})/(t_n - t_{n-1})$. "Sinusoidal" is nonlinear interpolation using a sinusoidal curve between keyframes as described in Subsection 7.3.

"Mismatch" is using a digit voice and an unsynchronized video-captured face saying another digit number. The discrimination rate drastically degrades in the case of "Mismatch" between voice and image, even at −6 dB.

As a result nonlinear interpolation using a sinusoidal curve while considering coarticulation is able to reach a high score, and the proposed system significantly enhances perception rates. This method provides a good standard for evaluation of lip synchronization. A better interpolation method for lip synchronization will be pursued in order to more closely match the original image sequence.

## 9. CONCLUSIONS

As a result of this research, we propose a multimodal translation system that is effective for video-mail or applications for automatic dubbing into other languages. For a video phone application, a few seconds delay is inevitable depending on speech recognition and translation algorithm, and this will never be overcome theoretically, therefore real-time telecommunication cannot be realized. Video tracking requires high cost now, but this will be overcome by increases of CPU power.

Currently, speech recognition and machine translation strongly depend on context. However, the size of context will grow bigger and bigger, and context-independent systems will be realized in the future by changing databases.

Our proposed system can create any lip shape with an extremely small database, and it is also speaker-independent. It retains the speaker's original facial expression by using input images besides those of the mouth region. Furthermore, this facial-image translation system, which is capable of multimodal English-to-Japanese and Japanese-to-English translation, has been realized by applying the parameters from CHATR. This is a hybrid structure of the image-based and CG-based approaches to replace only the part related to verbal information.

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.

A method to evaluate lip synchronization was proposed and it will provide a standard method for lip-sync performance evaluation.

## REFERENCES

[1] T. Yotsukura, E. Fujii, T. Kobayashi, and S. Morishima, "Generation of a life-like agent in cyberspace using media conversion," IEICE Tech. Rep. MVE97-103, Institute of Electronics, Information and Communication Engineers, Tokyo, Japan, pp. 75–82, 1998.

[2] T. Takezawa, T. Morimoto, Y. Sagisaka, et al., "A Japanese-to-English speech translation system: ATR-MATRIX," in *Proc. 5th International Conference on Spoken Language Processing (ICSLP '98)*, pp. 2779–2782, Sydney, Australia, November–December 1998.

[3] H. P. Graf, E. Cosatto, and T. Ezzat, "Face analysis for the synthesis of photo-realistic talking heads," in *Proc. 4th IEEE International Conference on Automatic Face and Gesture Recognition (FG '00)*, pp. 189–194, Grenoble, France, March 2000.

[4] N. Campbell and A. W. Black, "CHATR: a multi-lingual speech re-sequencing synthesis system," IEICE Tech. Rep. SP96-7, Institute of Electronics, Information and Communication Engineers, Tokyo, Japan, pp. 45–52, 1995.

[5] M. Ostendorf and H. Singer, "HMM topology design using maximum likelihood successive state splitting," *Computer Speech and Language*, vol. 11, no. 1, pp. 17–41, 1997.

[6] W. N. Campbell, "CHATR: A high definition speech re-sequencing system," in *Proc. 3rd ASA/ASJ Joint Meeting*, pp. 1223–1228, Honolulu, Hawaii, USA, December 1996.

[7] E. Sumita, S. Yamada, K. Yamamoto, et al., "Solutions to problems inherent in spoken-language translation: The ATR-MATRIX approach," in *Proc. 7th Machine Translation Summit VII*, pp. 229–235, Singapore, Singapore, September 1999.

[8] O. Furuse, J. Kawai, H. Iida, S. Akamine, and D. Kim, "Multi-lingual spoken-language translation utilizing translation examples," in *Proc. 3rd Natural Language Processing Pacific Rim Symposium (NLPRS '95)*, pp. 544–549, Seoul, Korea, December 1995.

[9] O. Furuse and H. Iida, "Incremental translation utilizing constituent boundary patterns," in *Proc. 16th International Conference on Computational Linguistics (COLING '96)*, vol. 1, pp. 412–417, Copenhagen, Denmark, August 1996.

[10] E. Sumita and H. Iida, "Experiments and prospects of example-based machine translation," in *Proc. 29th Annual Meeting of the Association for Computational Linguistics (ACL '91)*, pp. 185–192, Berkeley, Calif, USA, June 1991.

[11] M. Paul, E. Sumita, and H. Iida, "Field structure and generation in transfer-driven machine translation," in *Proc. 4th Annual Meeting of the NLP*, pp. 504–507, Fukuoka, Japan, 1998.

[12] K. Yamamoto and E. Sumita, "Feasibility study for ellipsis resolution in dialogues by machine learning techniques," in *Proc. 17th International Conference on Computational Linguistics-Annual Meeting of the Association for Computational Linguistics (COLING-ACL '98)*, pp. 1428–1435, Montreal, Quebec, Canada, August 1998.

[13] K. Ito, T. Misawa, J. Muto, and S. Morishima, "3D head model generation using multi-angle images and facial expression generation," IEICE Tech. Rep. 582, Institute of Electronics, Information and Communication Engineers, Tokyo, Japan, pp. 7–12, 2000.

[14] K. Ito, T. Misawa, J. Muto, and S. Morishima, "3D lip expression generation by using new lip parameters," IEICE Tech. Rep. A-16-24, Institute of Electronics, Information and Communication Engineers, Tokyo, Japan, pp. 328, 2000.

[15] T. Misawa, K. Murai, S. Nakamura, and S. Morishima, "Automatic face tracking and model match-move automatic face tracking and model match-move in video sequence using 3D face model in video sequence using 3D face model," in *Proc. IEEE International Conference on Multimedia and Expo (ICME '01)*, pp. 234–236, Tokyo, Japan, August 2001.

[16] T. Kuratate, H. Yehia, and E. Vatikiotis-Bateson, "Kinematics-based synthesis of realistic talking faces," in *Proc. International Conference On Auditory-Visual Speech Processing (AVSP '98)*, pp. 185–190, Terrigal, New South Wales, Australia, December 1998.

**Shigeo Morishima** was born in Japan on August 20, 1959. He received the B.S., M.S., and Ph.D. degrees, all in electrical engineering, from the University of Tokyo, Japan, in 1982, 1984, and 1987, respectively. From 1987 to 2001, he was an Associate Professor and from 2001 to 2004, a Professor at Seikei University, Tokyo. Currently, he is a Professor at School of Science and Engineering, Waseda University. His research interests include computer graphics, computer vision, multimodal signal processing, and human computer interaction. Dr. Morishima is a Member of the IEEE, ACM SIGGRAPH, and the Institute of Electronics, Information and Communication Engineers, Japan (IEICE-J). He is a Trustee of Japanese Academy of Facial Studies. He received the IEICE-J Achievement Award in May, 1992 and the Interaction 2001 Best Paper Award from the Information Processing Society of Japan in February 2001. He was having a sabbatical staying at University of Toronto from 1994 to 1995 as a Visiting Professor. He is now a Temporary Lecturer at Meiji University and Seikei University, Japan. Also, he is a Visiting Researcher at ATR Spoken Language Translation Research Laboratories since 2001 and ATR Media Information Science Laboratory since 1999.

**Satoshi Nakamura** was born in Japan on August 4, 1958. He received the Ph.D. degree in information science from Kyoto University in 1992. He worked with ATR Interpreting Telephony Research Laboratories from 1986 to 1989. From 1994 to 2000, he was an Associate Professor at the Graduate School of Information Science, Nara Institute of Science and Technology, Japan. In 1996, he was a Visiting Research Professor of the CAIP Center of Rutgers University of New Jersey, USA. He is currently the Head of Acoustics and Speech Research Department at ATR Spoken Language Translation Laboratories, Japan. He also serves as an Honorary Professor at University Karlsruhe, Germany, since 2004. His current research interests include speech recognition, speech translation, spoken dialogue systems, stochastic modeling of speech, and microphone arrays. He received the Awaya award from the Acoustical Society of Japan in 1992, and the Interaction 2001 Best Paper Award from the Information Processing Society of Japan in 2001. He served as an Editor for the Journal of the IEICE Information from 2000 to 2002. He is currently a Member of the Speech Technical Committee of the IEEE Signal Processing Society.