Statistical Multimodal Integration for Audio–Visual Speech Processing

Satoshi Nakamura

Invited Paper

***Abstract*—Sensory information is indispensable for living things. It is also important for living things to integrate multiple types of senses to understand their surroundings. In human communications, human beings must further integrate the multimodal senses of audition and vision to understand intention. In this paper, we describe speech related modalities since speech is the most important media to transmit human intention. To date, there have been a lot of studies concerning technologies in speech communications, but performance levels still have room for improvement. For instance, although speech recognition has achieved remarkable progress, the speech recognition performance still seriously degrades in acoustically adverse environments. On the other hand, perceptual research has proved the existence of the complementary integration of audio speech and visual face movements in human perception mechanisms. Such research has stimulated attempts to apply visual face information to speech recognition and synthesis. This paper introduces works on audio–visual speech recognition, speech to lip movement mapping for audio–visual speech synthesis, and audio–visual speech translation.**

***Index Terms*—Audio–visual, bimodal, hidden Markov model (HMM), multimodal, speech recognition, speech synthesis, speech translation.**

I. Introduction

HUMAN beings utilize multiple types of sensory information to communicate with each other. The usage of multiple types of sensory information is essential for more accurate, robust, natural, and friendly interaction. These types of information are also required for computers to realize natural and friendly interfaces, which are currently unreliable and unfriendly, with human beings. In this paper, the term "multimodality" is used to represent multiple types of sensory information for recognition and presentation by human beings and machines.

The multimodal processing occurring in human–machine interaction is normally defined as a type of processing that integrates multiple input and output means of audition and vision. The intention of a human being is usually carried by his/her audio speech. However, due to the insufficient speech recognition performance of today's machines, various kinds of input and output media have had to be developed such as the computer keyboard, mouse, and text and graphics for machines. The multimodalities, which are sensory information for these media, are used to increase the degree of user friendliness and improve the accuracy through complementary use for human-machine interaction. The term "multimodality" in this paper also includes these modalities.

A more detailed look at multimodalities shows that multimodality relationships can be classified into two kinds, synchronous and asynchronous. Synchronous multimodalities should be dealt with synchronously like audio speech and visual mouth movements. These modalities are generated from the same information source simultaneously. In contrast, asynchronous multimodalities can be dealt with asynchronously like mouse pointing and keyboard input. These modalities do not require synchronicity and carry different kinds of information. In this paper, we focus on synchronous multimodalities, specifically an audio information modality of speech and an image information modality of a face for audio–visual speech recognition and synthesis. There are strong demands to improve the speech recognition performance and the speech synthesis ineligibility since speech recognition and speech synthesis are becoming core technologies for human-machine interaction. We use the term "audio speech" as audio signals of speech, and "visual speech" as facial image signals.

Human audio speech and visual speech both originate from movements of the speech organs triggered by motor commands from the brain. Accordingly, such speech signals represent the information of an utterance in different ways. Therefore, these audio and visual speech modalities have strong correlations and complementary relationships. This paper describes research on the integrated processing and trans-modal processing of synchronous multimodalities for audio–visual speech processing. First, we describe the integrated processing of audio–visual speech recognition, and second, we describe the trans-modal processing from audio speech to visual speech in audio–visual speech synthesis.

As mentioned above, audio speech signals are generated by movements of the speech organs. Accordingly, there are strong correlations between the audio speech and visual speech. These modalities sometimes compensate for the insufficiencies of each other. For instance, phonemes /b//d/ are very difficult to discriminate by audio speech information alone, while such discrimi-

Manuscript received September 21, 2001; revised November 14, 2001. This work was supported in part by the Telecommunications Advancement Organization of Japan.

The author is with ATR Spoken Language Translation Research Laboratories, Kyoto 619-0288, Japan (e-mail: satoshi.nakamura@atr.co.jp).

Publisher Item Identifier S 1045-9227(02)04412-0.

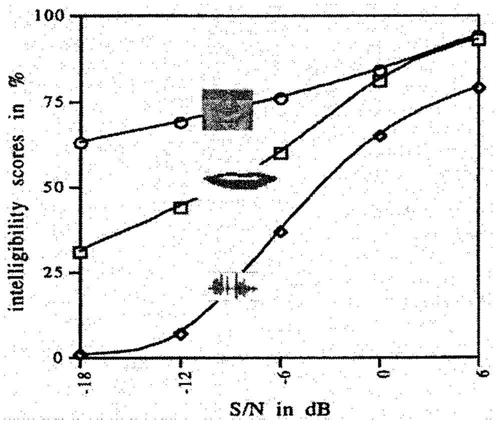


Fig. 1. Audio-visual scores observed with natural speech under different background noise conditions. Bottom curve: Audio alone condition. Middle curve: Audio+natural lips. Top curve: Audio+natural face. Benoit *et al.* [4].

nation is quite easy by visual speech information. There have been studies on audio–visual integration in human perception. McGurk showed that human beings can perceive /da/ with audio /ba/ and visual /ga/ stimuli [1]. This phenomenon implies that there is an effective audio–visual integration mechanism in the human perception of speech sounds. In face-to-face situations, human beings communicate with each other in very robust ways by integrating audio–visual information. Another experiment by Bateson [2] showed that the speaker's viewpoint moves from the partner's eyes to his/her mouth according to a decrease in the acoustic signal to noise ratio (SNR). There was also research [3], [4] showing that the human perceptual intelligibility can be improved by showing the speaker's face in an acoustically noisy environment. Fig. 1 is a figure from the experiments of Benoit *et al.* [4]. This figure shows that the presentation of face and mouth information can help human perception under very low acoustic SNR conditions. More details can be found in the many papers in perceptual research [5]–[8]. These reports imply that automatic audio speech recognition and audio speech synthesis can be improved by utilizing visual speech modalities.

This paper describes the integration of audio–visual modalities for robust speech recognition in Section II. The section introduces a model that can deal with the loose synchronicity of modalities and a method that can optimize modality weights. Section III describes audio–visual speech synthesis based on hidden Markov models (HMMs). The section also introduces a method based on the audio–visual joint probability. Finally, Section IV introduces audio–visual speech translation.

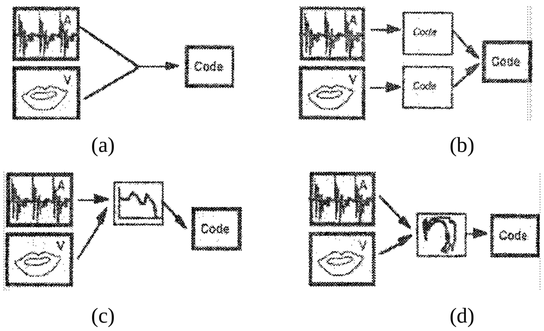
II. Integration of Audio–Visual Modalities for Robust Speech Recognition

A. Perceptual Model for Audio–Visual Integration

Generally speaking, the audio speech recognition performance has been drastically improved recently. However, it is also well known that the performance of a system will seriously degrade if the system is exposed to a noisy environment. Humans pay attention not only to the speaker's audio speech but also to the speaker's mouth in such an adverse environment. Lip reading can be considered the extreme case if it is impossible to get any audio signal. This suggests that audio speech recognition can be improved by incorporating visual mouth images. This type of multimodal integration is available for almost every face-to-face situation. For synchronous multimodal processing like audio–visual speech recognition, the following elemental methodologies should be developed:

1) image segmentation of the region of interest (ROI) in a facial image sequence and end-point detection of the audio speech utterance;
2) feature extraction of audio and visual speech information;
3) integrated modeling of synchronous and asynchronous multimodalities;
4) adaptation on information reliability for new environments.

In this paper, we focus on synchronous multimodal integration modeling. For other items, readers are recommended to refer to textbooks and other survey papers [5], [9], [7], [8], [10].



(a) (b)

(c) (d)

Fig. 2. The four basic models of audio–visual integration: a-DI, b-SI, c-DR and d-MR. [64].

In Fig. 2, we introduce a figure of Schwartz *et al.* [11] to explain human perception models for audio–visual speech perception. The first model (a) is called the direct identification model (DI) or early integration model. Early integration is the most general approach. The input signals are transmitted directly to a bimodal coder (classifier). The second model (b) is called the separate identification model (SI) or late integration model. This model assumes two parallel recognition processes. Then, the phonemes or word likelihood values from each modality are fused. The third model (c) is the dominant recording model (DR). This model considers the auditory modality as being the dominant modality for speech perception. The visual input is recoded into a representation of the dominant modality. The last model (d) is called the motor recoding model (MR). This model assumes that both inputs are projected into a articulatory model space and fused in that space.

B. Automatic Audio–Visual Speech Recognition

In multimodal speech recognition, research had been done mostly on the first and second integration models [12]–[33], [31], [7], [34]–[49]. We use the terms of early integration model for the DI model and late integration model for the SI model hereafter.

The late integration model was exemplified by Petajan's early work [50]. This model fuses classification results from separate classifiers for audio and visual signals. The works by [50], [18], [19], [25], [51], [52] can be classified into this category.

The early integration model is the opposite model of the late integration model. The acoustic data and visual data are input as a single composite pattern vector and then classified. Typical early integration architectures assume a strict time synchronicity of auditory and visual events. The works by [53], [18], [54], [22], [55], [25] can be classified into this category. There have also been studies on intermediate architectures by [24], [56], [57], [33], [32], [31], [37], [7], [58]–[62], [48], [49].

On the other hand, almost no work has yet been reported on the MR architecture. For the DR model, an early recognizer has used artificial neural networks (ANNs) to recode facial images into acoustic spectral envelopes [9]. Here, we focus on an early integration scheme and a late integration scheme.

Early Integration: Early integration is a kind of direct integration or feature fusion. This scheme extracts feature vectors from both audio speech and visual speech and concatenates them into one feature vector sequence. The one set of HMMs of concatenated feature vectors is trained and used. For adaptation, the stream weights are adjusted to optimize the likelihood or minimize the classification errors.

The weighting is carried out as follows:

$$b_{ij}(o_t) = b_{ij}^{\text{audio}}(o_t)^{\lambda_{\text{audio}}} \times b_{ij}^{\text{visual}}(o_t)^{\lambda_{\text{visual}}}. \quad (1)$$

Here, $b_{ij}(o_t)$, $b_{ij}^{\text{audio}}(o_t)$, and $b_{ij}^{\text{visual}}(o_t)$ are the output probabilities of the transition from state i to state j for the composite vector, audio vector, and visual vector, respectively. λ_{audio} and λ_{visual} are the weighting coefficients for the audio vector and visual vector, respectively, which satisfy

$$\lambda_{\text{audio}} + \lambda_{\text{visual}} = 1. \quad (2)$$

Fig. 3 depicts this early integration process.

Late Integration: Late integration is also called decision fusion or separate identification. This scheme extracts feature vector sets separately and uses two sets of HMMs. The results of the HMMs for audio and visual speech are combined with reliability weights

$$P(X|M_i) = P(X_{\text{audio}}|M_i^{\text{audio}})^{\lambda_{\text{audio}}} \times P(X_{\text{visual}}|M_i^{\text{visual}})^{\lambda_{\text{visual}}}. \quad (3)$$

Here, $P(X|M_i)$, $P(X_{\text{audio}}|M_i^{\text{audio}})$, and $P(X_{\text{visual}}|M_i^{\text{visual}})$ are the probabilities of composite vector sequence X from late integration, audio vector sequence X_{audio} from audio HMMs M_i^{audio}, and visual vector sequence X_{visual} from visual HMMs M_i^{visual}, respectively. The i is the word number. The weighting coefficients also satisfy

$$\lambda_{\text{audio}} + \lambda_{\text{visual}} = 1. \quad (4)$$

Fig. 4 depicts this late integration process. Figs. 5 and 6 show 54 French nonsense word recognition results by Adjoudani *et al.* [29]. The visual parameters were extracted and modeled by 12 lip parameters. The acoustic parameters were 12 cepstral coefficients. In the figure, the V line shows the percentage of correct responses of the video recognizer alone. The A curve shows the scores of the audio recognizer alone. The dashed curves represent the averages of the test scores while the line segments show the standard deviations for the individual test conditions. The equal weights are set to audio and visual modalities.

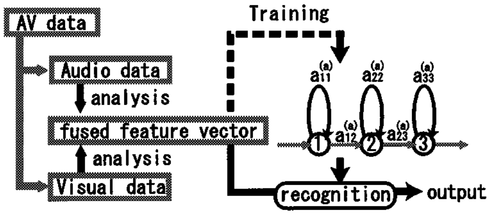


Fig. 3. Block diagram for the early integration model.

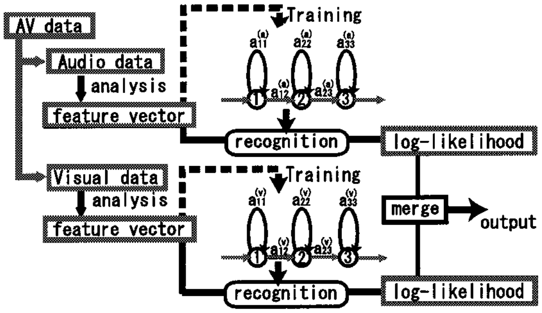


Fig. 4. Block diagram for the late integration model.

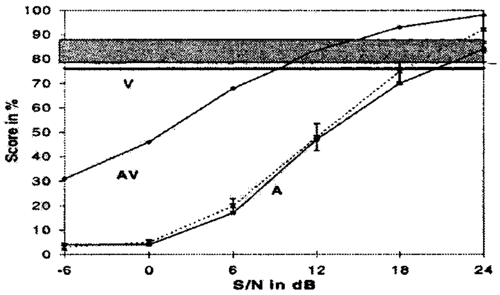


Fig. 5. Word recognition results by early integration models [28].

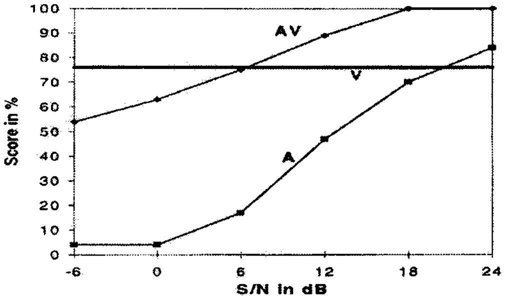


Fig. 6. Word recognition results by late integration models [29].

These figures suggest that the late integration model performs better than the early integration model in automatic audio–visual speech recognition. The reason is that the early integration is trained considering strong time synchronicity and modality reliability in the training data. Therefore, if one of the modalities is corrupted, the overall performance easily suffers from serious degradation.

The characteristics of the models can be summarized as follows:

- The early integration model assumes synchronicity between the events of two modalities and the initial modality reliability. This model seems to be suitable for audio–visual speech processing, but requires a huge amount of data to train the joint audio–visual events accurately, because the events between the audio speech and visual face movements are not rigidly synchronized. Or, it can be said that this relationship is synchronized but loosely.

The early integration model also has a drawback when one of the modalities is corrupted by some reason, and this results in a situation where the overall performance suffers from serious degradation.

- The late integration model ignores synchronization between audio and visual information. However, each model can be well trained by a sufficient amount of data. The performance is relatively robust even when one of the modalities is damaged.

On the other hand, human perception works better if visual modalities are presented so as to add to audio modalities. This fact insists on the development of a new model for automatic audio–visual speech recognition.

Therefore, the model for audio–visual integration requires the following functions.

- Representability of loose synchronicity between audio and visual modalities.
- Controllability of the reliability weight on each modality.

C. *Audio–Visual Product Model*

Next, we describe related works in the development of an intermediate model able to represent loose audio–visual synchronicity. The model introduced in this paper is called a product HMM. A product HMM is defined so as to produce a product code (AV) from a product set $A \times V$ of multiple information source sets, A and V. Fig. 7 shows the training algorithm. First, in order to create an audio–visual phoneme HMM, audio and visual features are extracted from audio–visual data. In general, the frame rate of audio features is higher than that of visual features. Accordingly, the extracted visual features are incorporated such that the audio and visual features have the same frame rate. Second, the audio and visual features are modeled individually into two HMMs by the EM algorithm. The audio–visual phoneme HMM is composed as the product of these two HMMs. The output probability at state ij of the audio–visual HMM is

$$b_{ij}(O_t) = b_i^A(O_t^A)^{\alpha_A} \times b_j^V(O_t^V)^{\alpha_V} \tag{5}$$

which is defined as the product of the output probabilities of the audio and visual streams. Here, $b_i^A(O_t^A)^{\alpha_A}$ is the output probability of the audio feature vector at time instance t in state i, $b_j^V(O_t^V)^{\alpha_V}$ is the output probability of the visual feature vector at time instance t in state j, and α_A and α_V are the audio stream weight and visual stream weight, respectively.

In general, since (5) does not represent a probability mass function, it is improper to estimate the stream weights by the maximum likelihood principle [39], [45]. In a similar manner, the transition probability from state ij to state kl in the audio–visual HMM is defined as follows:

$$a_{ij,kl} = a_{ik}^A \times a_{jl}^V \tag{6}$$

where a_{ik}^A is the transition probability from state i to state k in the audio HMM, and a_{kl}^V is the transition probability from state j to state l in the visual HMM. This composition is performed for all phonemes.

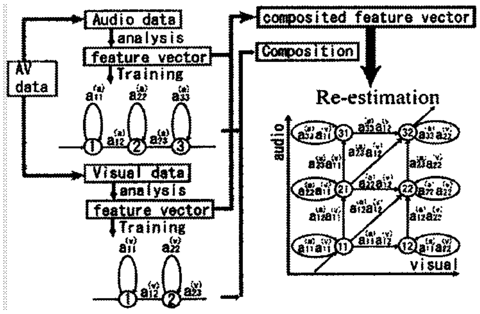


Fig. 7. Training for a product HMM.

This methods includes the following two problems.

1) The product HMMs can not represent the loose synchronicity within phonemes.
2) The product HMMs force a strict synchronization at the phoneme boundary.

This paper introduces our approaches, already published in [48], [49], to solve the two problems. First, we describe the reestimation of the product HMMs by using a small amount of audio–visual synchronous adaptation data. Second, we describe a structure for the product HMMs. This new structure includes loose state synchronicity beyond the phoneme boundary.

D. *State Asynchronous Modeling*

1) Asynchronicity Within a Phoneme: The first problem is from the inability of the conventional product HMMs to represent loose state synchronicity within a phoneme. This problem is caused by the fact that the transition probabilities and output probabilities are obtained by the multiplication of probabilities from independent states of audio and visual HMMs. This section describes our approach in which product HMMs' parameters are reestimated using audio–visual synchronous adaptation data [48]. The advantage of performing reestimation is that the reestimation is able to introduce the loose state synchronicity of the states of two modalities into the product HMMs.

The reestimation procedure is carried out using a small amount of audio–visual synchronous data. After the composition of two HMMs, the product HMMs can be reestimated based on the Baum–Welch algorithm for multistream HMMs. Fig. 8 shows results comparing audio HMMs, visual HMMs, early integration, late integration, and product HMMs with and without reestimation [48]. The experimental conditions are the same as those in a later section except that the audio HMMs are trained using clean audio speech data. The figure shows that the product HMMs with reestimation achieve the best performance, while the product HMMs without reestimation are worse than those of the early and late integration schemes.

2) Asynchronicity Beyond a Phoneme Boundary: The second problem is that the conventional product HMMs force a strict synchronization at every phoneme boundary. This is because the human speech organs normally move earlier than the audio speech to be produced. Sometimes, the speech organs have already articulated the previous audio phoneme utterance. Accordingly, we have to consider state synchronous modeling beyond the phoneme boundary.

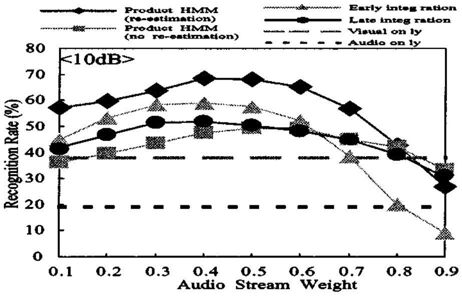


Fig. 8. Results of product HMMs.

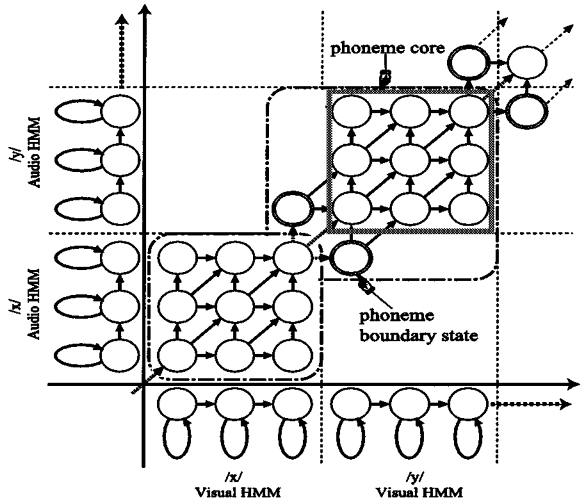


Fig. 9. Pseudobiphone product HMMs.

Considering this fact, we have proposed a new product HMM that includes extra asynchronous states at phoneme boundaries [49] as indicated in Fig. 9. The core states of the phoneme HMMs are the same as those of context independent phoneme product HMMs. In addition, the new product HMMs have two extra HMM states aiming to work similarly to word unit-based HMMs. The first extra state is composed of the initial audio state and final visual state of the preceding phoneme HMM. The second extra state is composed of the initial visual state and final audio state of the preceding phoneme HMM. Since these extra states are dependent on the preceding phoneme, they can only be reestimated in a manner similar to the biphone HMMs. Therefore, we call these HMMs pseudobiphone product HMMs. The proposed HMMs can tolerate one state asynchronicity beyond the phoneme boundary.

E. Evaluation Experiments

The audio signal is sampled at 12 kHz (down-sampled) and analyzed with a frame length of 32 ms every 8 ms. The audio features are 16-dimensional MFCC and 16-dimensional delta MFCC. On the other hand, the visual image signal is sampled at 30 Hz with 256 gray-scale levels from RGB. Then, the image level and location are normalized by a histogram and template matching. Next, the normalized images are analyzed by two-dimensional (2-D) fast Fourier transform (FFT) to extract 6×6 log power 2-D spectra for audio–visual speech recognition. Finally, 35-dimensional 2-D log power spectra and their delta features are extracted. For each modality, the basic coefficients and the delta coefficients are collectively merged into one stream. Since the frame rate of the video images is 1/30, we insert the same images so as to synchronize the face image frame rate to the audio speech frame rate. For the HMMs, we use a two-mixture Gaussian distribution and assign three states for the audio stream and three states for the visual stream in the late integration HMMs and the baseline product HMMs. In this research, we perform word recognition evaluations using a bimodal database [35]. We use 4740 words for HMM training and two sets of 200 words for testing. These 200 words are different from the words used in the training.

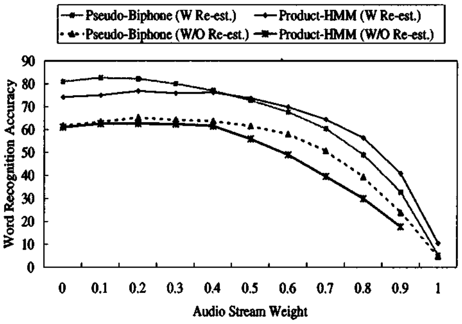


Fig. 10. Audio–visual speech recognition performance (SNR $= -5$ dB).

Fig. 10 shows word accuracies for acoustic SNR $= -5$ dB. We compared the processed product HMMs without reestimation [Product-HMM(W/O Re-est.)], the product HMMs with reestimation [(Product-HMM(W Re-est.)], the pseudobiphone product HMMs without reestimation [(Pseudo-Biphon(W/O Re-est.)], and the pseudo-biphone product HMMs with reestimation [Pseudo-Biphon(W Re-est.)]. White noise was used to reduce the acoustic SNR in this experiment. The audio HMMs were trained using SNR $=$ 15 dB data. The results can be summarized as follows.

- The reestimation of the product HMMs quite effectively improves the performance. The reestimation is able to introduce the loose state synchronicity of the states of two modalities into the product HMMs. The reestimation also produces a consistent state alignment to the multiple input modalities. There has been proposed a product HMM [57], [33], [32]. However, the reestimation performed in our experiments indicates that audio–visual synchronicity within a product HMM should be taken into account.
- State synchronous modeling beyond the phoneme boundary results in significant improvements to the product HMMs. This finding indicates the importance of considering the loose synchronicity of audio speech and speech organs over the phoneme boundary.
- The optimal stream weights for the maximum performance vary according to each method and acoustic SNR. Further investigations are necessary to adjust the optimal weights for modalities.

F. Modality Weight Optimization

Audio–visual speech recognition systems need to be optimized adaptively for their respective environments since the phonetic discrimination accuracy of visual information alone is too poor to obtain a sufficient performance of an speech recognition system. This should be done in the audio–visual speech recognition system itself; if there is no acoustically noisy environment, the audio features are more important than the visual ones, otherwise, they are not.

To optimize reliability modality weights, there are many reported methods based on the dispersion rate [28], [63], acoustic SNR [64], maximum likelihood estimation [65], maximum classification estimation [44], and generalized probabilistic decent (GPD) algorithm [66], [39], [67] and their extensions [68].

However, the maximum likelihood based methods have a serious estimation drawback because the scales of two probabilities are normally very different and so the weights can not be estimated optimally. The GPD based methods have a substantial possibility for optimizing the weights. However, one serious problem is that these methods require a lot of adaptation data for the weight estimation procedure. In this paper, we introduce our adaptive estimation of stream weights based on the GPD algorithm for new noisy acoustic conditions.

The approach by the GPD training defines a misclassification measure, which provides distance information concerning the correct class and all other competing classes. The misclassification measure is formulated as a smoothed loss function. This loss function is minimized by the GPD algorithm. Here, let $L_c^{(x)}(\Lambda)$ be the log-likelihood score in recognizing input data x for adaptation using the correct word HMM, where $\Lambda = \{\lambda_A, \lambda_V\}$.

In a similar way, let $L_n^{(x)}(\Lambda)$ be the score in recognizing data x using the n-th best candidate among the mistaken word HMMs.

The misclassification measure is defined as

$$d^{(x)} = -L_c^{(x)}(\Lambda) + \log\left(\left[\frac{1}{N}\sum_{n=1}^{N}\exp\{\eta L_n^{(x)}(\Lambda)\}\right]^{\frac{1}{\eta}}\right) \quad (7)$$

where η is a positive number, and N is the total number of candidates. The smoothed loss function for each data is defined as

$$l^{(x)} = \left[1 + \exp\{-\alpha d^{(x)}(\Lambda)\}\right]^{-1} \quad (8)$$

where α is a positive number. In order to stabilize the gradient, the loss function for the entire data is defined as

$$l(\Lambda) = \sum_{x=1}^{x} l^{(x)}(\Lambda) \quad (9)$$

where X is the total amount of data. The minimization of the loss function expressed by (9) is directly linked to the minimization of the error. The GPD algorithm adjusts the stream weights recursively according to

$$\Lambda_{k+1} = \Lambda_k - \varepsilon_k E_k \nabla l(\lambda), k = 1, \ldots \quad (10)$$

where $\varepsilon_k > 0$, $\sum_{k=1}^{\infty}\varepsilon_k = \infty$, $\sum_{k=1}^{\infty}\varepsilon_k^2 < \infty$, and E is a unit matrix. The algorithm converges to a local minimum as $k \to \infty$ [66].

We introduce our experimental results for stream weight optimization of a product HMM [48]. In an experiment to optimize stream weights adaptively, we used 100 words as the adaptation data, excluding the training and test data. We also performed experiments using 15, 25, and 50 words, which were extracted from the 100 words. The context of the data for the adaptation differed from that of the test data. The size of the vocabulary in the dictionary was 500 words during the recognition of the adaptation data. We set $N = 1$ in (7), $\alpha = 0.1$ in (8), $N = 100/k$, and the maximum iteration count $= 8$. This was because the GPD algorithm convergence pattern is known to greatly depend on the choice of parameters.

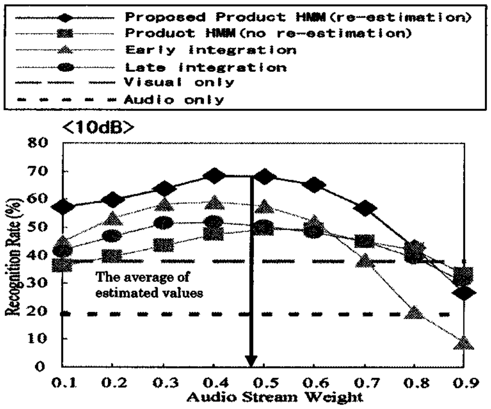


Fig. 11. Recognition results for various stream weights [10 dB].

Fig. 11 shows recognition rates for stream weights for audio only, visual only, early integration, late integration, and the product HMM based integration (reestimated/no reestimation). In the figures, the stream weights change under $\lambda_A + \lambda_V = 1$ in the acoustically noisy audio SNR 10 dB environment.

Generally, audio–visual speech recognition systems have their peak recognition rates by stream weights (See Fig. 11). Therefore, if the optimal stream weights can be estimated by a small amount of adaptation data, audio–visual speech recognition systems can achieve higher recognition rates in various environments. Fig. 12 shows an average of estimated stream weights from 25 word data by the GPD algorithm. Note that the performance of the audio–visual speech recognition system with optimized stream weights is better than that of the unimodal one, even in the noise free case.

These experiments show that the GPD-based estimation provides optimal weights for the modalities by using a relatively small amount of adaptation data. However, weights should be different for different phonemes. Accordingly, these weights are expected to be determined for each phoneme and each environment considering the amount of adaptation data.

III. Audio–Visual Speech Synthesis

This section describes transmodal processing between audio speech and visual speech. As mentioned, audio speech and visual speech have a strong correlation since these modalities originate from the same information source. Transmodal mapping between audio speech and visual speech can be achieved using the correlation.

We investigate synthesis methods for achieving human-like visual lip movements by mapping from audio speech. The employed audio speech includes more information than text to synthesize real lip movements, such as the pitch frequency and phoneme duration.

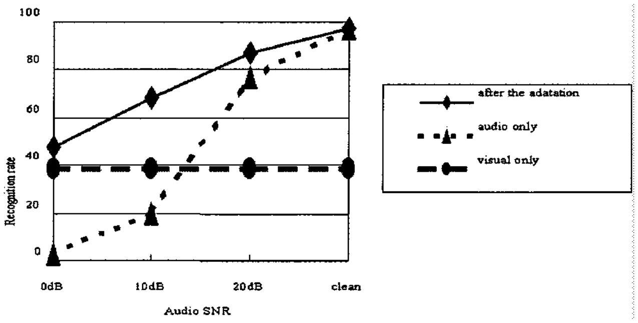


Fig. 12. Recognition results with optimized weights.

The lip movement synthesis from audio speech also permits lip-synchronization between audio speech and a synthesized visual lip image sequence. Lip-synchronization is required for human-like computer agents in interactive communication systems [10]. If lip movements can be synthesized well enough to allow lip-reading, hearing impaired people can benefit by comprehending the auditory information given by computer agents.

Mapping algorithms from audio speech to lip movements have been reported based on: articulatory model [69]–[71], vector quantization [72], [73], ANNs [74], [75], and Gaussian mixtures [76]. These methods are based on frame-by-frame or frames-by-frames mapping from audio speech parameters to visual lip parameters. However, they have problems, e.g., 1) the mapping is fundamentally complicated many-to-many mapping and 2) extensive training is required to take account of context information. The required audio–visual database increases in proportion to the length over the preceding or succeeding frames. On the other hand, there is another approach using techniques of audio speech recognition, such as phonetic segmentation [77] and HMMs [78]–[92]. These techniques convert audio speech into visual lip parameters based on such information as a phonetic segment, a word, a phoneme, an acoustic event, or so on. This HMM-based approach has an advantage that explicit phonetic information is available to allow the consideration of coarticulation effects caused by surrounding phoneme contexts.

This section introduces two audio–visual mapping methods based on HMMs for lip movement synthesis. These two methods require no transcription of the utterance beforehand. If transcription is available, other methods like waveform and image sequence concatenation methods can be utilized. However, these methods can not be applied to the case of no transcription. This paper introduces our approaches based on HMMs, already published in [81], [84], [85], and [91].

A. Mapping Method Using Viterbi Algorithm

The first method is a baseline method, *Map*ping based on *V*iterbi alignment (MAP-V), which is composed of two processes. These processes are composed of a decoding process that converts the input audio speech to the most likely HMM state sequence and a lookup table process that converts an HMM state to its corresponding visual parameters [81].

In the decoding process, the likelihood of the input audio speech by the kth audio HMM, M_k^A, is defined as (11)

$$P(O^A|M_k^A) \approx \max\nolimits_Q \{\pi_{q_1}(M_k^A) \times \prod_{t=1}^{T} a_{q_{t-1}q_t}(M_k^A) b_{q_t}(o^A(t)|M_k^A)\} \quad (11)$$

where $Q = q_1 \cdots q_T$ denotes the audio HMM state sequence, $O^A = o^A(1) \cdots o^A(T)$ is the sequence of input audio parameters, $\pi_j(M_k^A)$ is the initial state probability, $a_{ij}(M_k^A)$ is the transition probability from state i to j of M_k^A, and $b_j(o^A(t)|M_k^A)$ is the output probability at state j.

In (11), the optimal state sequence for the observation is obtained by Viterbi alignment. Along the alignment, the correspondence between each audio HMM state and the visual parameters is then calculated and stored in the look-up table in the training step. The visual parameters per audio HMM state are obtained by taking the expectation for all visual parameters assigned to the same audio HMM state.

B. Mapping Method Using EM Algorithm

The MAP-V method converts audio parameters to visual parameters through a deterministic single HMM state sequence. The quality of the visual parameters by the MAP-V method depends on the accuracy of the Viterbi alignment. The deterministic process involves a substantial problem, which may give rise to incorrect visual parameters out of an incorrect HMM state sequence. For example, if a bilabial consonant were to be decoded to other categories classified by place of articulation, the synthesized lip movement would generate a sense of incompatibility among audience members. To solve the problem, we extend the MAP-V method to an undeterministic process [93], [91].

The extended method, *Map*ping based on the *E*xpectation-*M*aximization algorithm (MAP-EM), reestimates visual parameters for the given audio parameters by the EM algorithm using audio–visual HMMs. Although visual parameters do not exist initially, the required visual parameters are synthesized iteratively from the initial parameters by a reestimation procedure that maximizes the audio–visual joint probability likelihood of the above audio–visual HMMs. The reestimating operation is regarded as the auto-association of a complete pattern out of an incomplete pattern over a time series. In experiments, the MAP-EM method is compared to the MAP-V method

$$\hat{o}^V(t) = \arg\max\nolimits_{o^V(t)} P(O^{A,V}|o^A(t), M^{AV}) \quad (12)$$

where $\hat{o}^V(t)$ denotes the estimated visual parameter. The likelihood of this method is derived by considering all HMM models and states at a time. To obtain the expectation over all HMM models and states, the likelihood of the audio–visual joint probability is defined as follows:

$$\begin{aligned} &\sum_{Q(all\ k)} P(M_k^{AV}) P(O^{A,V}|Q, M_k^{AV}) \\ &= \sum_{Q(all\ k)} P(M_k^{AV}) \pi_{q_1}(M_k^{AV}) \\ &\quad \times \prod_{t=1}^{T} a_{q_{t-1}q_t}(M_k^{AV}) b_{q_t}(o^{A,V}(t)|M_k^{AV}) \end{aligned} \quad (13)$$

where $P(M_k^{AV})$ is the probability of the kth HMM, and $\pi_j(M_k^{AV})$, $a_{ij}(M_k^{AV})$, and $b_j(o^{A,V}(t)|M_k^{AV})$ are the joint initial state probability, the joint transition probability, and the joint output probability of the audio and visual parameters, respectively. The summation of $Q(all\ k)$ considers all models M_k^{AV} at a time. The reestimation formula is defined to maximize the auxiliary function $A(\hat{o}^V(t), o^V(t))$ over estimated visual parameter $\hat{o}^V(t)$

$$\begin{aligned} &A(\hat{o}^V(t), o^V(t)) \\ &= \sum_{Q(all\ k)} P(M_k^{AV}) P(O^{A,V}|Q, M_k^{AV}) \\ &\quad \times \log P(M_k^{AV}) P(\hat{O}^{A,V}|Q, M_k^{AV}). \end{aligned} \tag{14}$$

The maximization of the auxiliary function is equivalent to increasing the likelihood to the input data. The reestimation formula of the visual parameters is derived by differentiating the auxiliary function by the mth visual parameter $\hat{o}_m^V(t)$. Let the output probability density function be mixed Gaussian distributions with a mean vector with $\mu_n^A(M_k^{AV}, j)$, $\mu_m^V(M_k^{AV}, j)$, and covariance matrix Σ with components $\sigma_{nn'}^{A,A}(M_k^{AV}, j)$, $\sigma_{mm'}^{V,V}(M_k^{AV}, j)$, $\sigma_{nm}^{A,V}(M_k^{AV}, j)$. n, m are the numbers of dimensions of the audio and visual parameters. The reestimation formula is derived as follows:

$$\begin{aligned} o_m^V(t) =& \frac{1}{\sum_k \sum_j P(M_k^{AV}) \gamma(t; M_k^{AV}, j) \frac{\Sigma'^{\,V,V}_{mm}(M_k^{AV}, j)}{|\Sigma(M_k^{AV}, j)|}} \\ & \times \sum_k \sum_j P(M_k^{AV}) \gamma(t; M_k^{AV}, j) \\ & \times \frac{1}{|\Sigma(M_k^{AV}, j)|} (\mu_m^V(M_k^{AV}, j) \Sigma'^{\,V,V}_{mm}(M_k^{AV}, j) \\ & - \sum_n (o_n^A(t) - \mu_n^A(M_k^{AV}, j)) \Sigma'^{\,A,V}_{nm}(M_k^{AV}, j)) \end{aligned} \tag{15}$$

where $\gamma(t; M_k^{AV}, j)$ is the state occupation probability in state j of M_k^{AV} at time t. $\Sigma'^{\,V,V}_{mm'}(M_k^{AV}, j)$ means the adjoint of $\Sigma_{mm'}^{V,V}(M_k^{AV}, j)$. Formula (15) is under the constraint that covariance $\sigma_{nn'}^{A,A}(M_k^{AV}, j) = 0$ at $n \neq n'$ and $\sigma_{mm'}^{V,V}(M_k^{AV}, j) = 0$ at $m \neq m'$. Furthermore, the reestimation formula is simplified as follows if the covariance matrix is diagonal

$$\hat{o}_m^V(t) = \frac{\sum_k \sum_j P(M_k^{AV}) \gamma(t; M_k^{AV}, j) \frac{\mu_m^V(M_k^{AV}, j)}{\sigma_{mm}^{V,V}(M_k^{AV}, j)}}{\sum_k \sum_j P(M_k^{AV}) \gamma(t; M_k^{AV}, j) \frac{1}{\sigma_{mm}^{V,V}(M_k^{AV}, j)}}. \tag{16}$$

The algorithm for the visual parameter reestimation can be summarized as follows:

```
Step 1 Set the initial value for visual parameter $o_m^V(t)$.
Step 2 Calculate $\gamma(t; M_k^{AV}, j)$ for all frames under the forward–backward algorithm (EM algorithm for HMM).
Step 3 Re-estimate $\hat{o}_m^V(t)$ at each frame.
Step 4 If a convergence condition is satisfied, go to the end, otherwise return to step 2).
```

C. Lip Synthesis Experiments

The audio speech and visual image data in our lip synthesis experiments are synchronously recorded at 125 Hz. The visual parameters are height (X) and width (Y) of the outer lip contour and protrusion (Z) of the lip sides from an original point, where the three parameters are used to construct a lip shape with three-dimensional (3–D) lip creation software [94]. The audio parameter has 33 dimensions of 16-order mel-cepstral coefficients, their delta coefficients, and the delta log power. Fifty-four monophone and two pause models are used for the HMMs in both the MAP-V method and the MAP-EM method. The pause models are prepared separately for the word beginning and the word ending. Triphone HMMs are not adopted, because they require huge amounts of time synchronous training data. Each audio HMM and audio–visual HMM has a left-to-right structure with three states, where the output probability on each state has 256 tied-mixture Gaussian distributions. The HMMs are trained by an audio or audio–visual synchronous database composed of 326 Japanese words, all phonetically balanced. Another 100 words are prepared for testing. The measure to evaluate the synthesized lip movements is Euclidian error distance E between the synthesized visual parameters and the original parameters extracted from human movements.

In the MAP-EM method, the state occupation probabilities $\gamma(t; M_k^{AV}, j)$ are updated after all visual parameters for the utterance are reestimated. In this paper, formula (16) is used in the experiments.

To verify the effect of the MAP-EM method, the five synthesis methods in Table I were compared in experiments. The MAP-EM method could be implemented by various conditions. We tried to make the number of parameter vectors fluctuate taking account of the dependency on the quality of the HMMs. In the MAP-EM-2 method, the visual parameter vector consisted of six parameters of three lip parameters and their time differential parameters. Likewise, the MAP-EM-3 method contained the acceleration part of the lip parameters in addition to the parameters of the MAP-EM-2 method. Note that in all of the MAP-EM methods, the number of tied-mixture distributions was fixed at 256 like in the MAP-V method. For the initial values of the visual parameters, the MAP-EM-1,2,3 methods used visual parameters synthesized by the MAP-V method and the MAP-EM-4 method used visual parameters of the lip closure shape during a pause.

The results of an objective evaluation of the five methods are shown in Table II. Each column in Table II indicates the error distances averaged by all frames or correctly decoded, incorrectly decoded, and incorrectly decoded /p//b//m/ frames by the MAP-V method. The results show that the use of time differential and acceleration parameters reduces error distances. The MAP-EM-4 method gives a large error due to the flat start of the lip closure.

We investigated errors of the MAP-EM-1,2,3 methods under incorrectly decoded frames by the MAP-V method. The errors are compared in three detailed categories of palatal, dental, and bilabial consonants in Fig. 13. The MAP-V method shows large errors under the bilabial consonant category of incorrectly decoded frames. Bilabial consonants /p//b//m/ are known to be quite sensitive to audience members. For these phonemes, the errors of the MAP-EM-3 method are reduced by 26% com-

TABLE I
COMPARISON OF SYNTHESIS METHODS

method	#params HMM training		#params Mapping		Initial visual parameters
	A	V	A	V	
MAP-V	33	-	33	-	-
MAP-EM1	33	3	33	3	MAP-V
MAP-EM2	33	6	33	3	MAP-V
MAP-EM3	33	9	33	3	MAP-V
MAP-EM4	33	9	33	3	pause lip

TABLE II
ERROR DISTANCES OF SYNTHESIS METHODS

	E cm			
	All frames	Correct de-coded	Incorrect de-coded	Incorrect de-coded /p//b//m/
MAP-V	1.066	1.062	1.075	1.701
MAP-EM-1	1.093	1.106	1.051	1.370
MAP-EM-2	1.063	1.077	1.021	1.392
MAP-EM-3	1.052	1.072	0.989	1.254
MAP-EM-4	1.207	1.231	1.134	1.061

pared to the errors of the MAP-V method at incorrectly decoded frames. The table also indicates that the MAP-EM method has a dependency on the initial conditions for iterative estimation. This dependency is caused by the fact that the reestimation of the MAP-EM method takes the expectation over not only the HMM states but also over all HMM models. This expectation may involve many local optima.

An effect of the MAP-EM method is illustrated in Figs. 14 and 15. The figures show a test Japanese word "kuchibiru." The horizontal axis means the number of frames corresponding to time. The vertical axis means visual parameters. The solid lines in the figures are synthesized visual parameters, and the dotted lines are visual parameters by the original recorded human movements. The two vertical lines show the beginning and ending times of the utterance. The synthesized height visual parameter of the MAP-V method does not form a valley of the lip closure of /b/ because of the incorrect Viterbi alignment in Fig. 14. However, the MAP-EM-3 method of Fig. 15 shows the correct articulation.

It is interesting to see that the MAP-EM method can be interpreted a generalization of the MAP-V method. The MAP-EM method estimates time domain feature parameters by maximizing the audio–visual joint probability. Although this method does not assume any transcription of utterances beforehand, if the transcription is given, the mapping accuracy can be improved. Furthermore, if a large amount of audio–visual synchronous data becomes available, the MAP-EM method can be extended to the case of using full covariance estimation. Full covariance estimation can provide a precise joint relationship between every pair of audio–visual parameters.

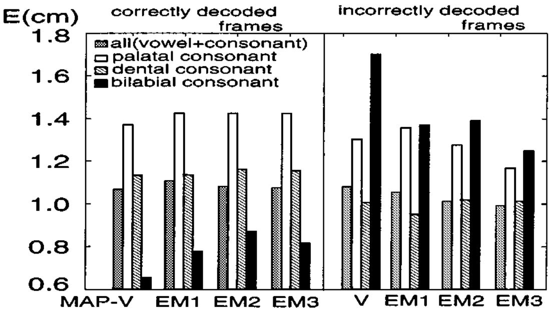


Fig. 13. Errors by consonant category.

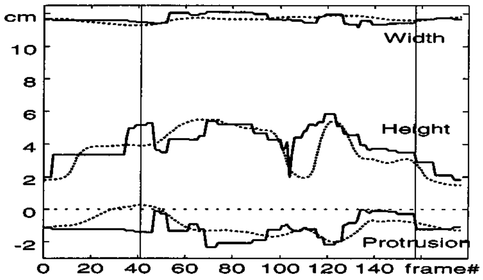


Fig. 14. Visual parameters synthesized by the MAP-V method.

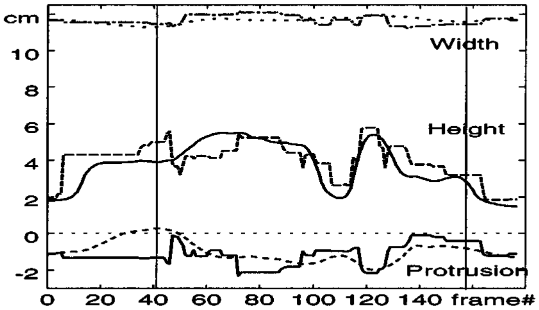


Fig. 15. Visual parameters synthesized by the MAP-EM-3 method.

IV. AUDIO–VISUAL SPEECH TRANSLATION

Audio-visual speech recognition and speech synthesis can be applied to audio–visual interaction. Recently, focus has been on interaction through human-like computer agents, or Avatars. Audio-visual speech recognition and speech synthesis are essential technologies for achieving natural and accurate movements of such Avatars. The Avatars can then be utilized to improve the degree of friendliness in human-machine interaction. The Avatars can also be applied to human-human interaction. By transmitting the essential parameters of the Avatars, the bit rate for video transmission can be reduced drastically. Other types of Avatars can also be extended to speech-to-speech translation [95], [96]. One of the dreams of human beings is to achieve automatic speech translation allowing people of different languages to communicate with each other. Speech translation systems have mainly been studied for verbal information. However, both verbal information and nonverbal information are indispensable for natural human communications. Facial expression plays an important role in transmitting both verbal and nonverbal information in face-to-face telecommunications. Lip movements transmit visual speech information along audio speech.

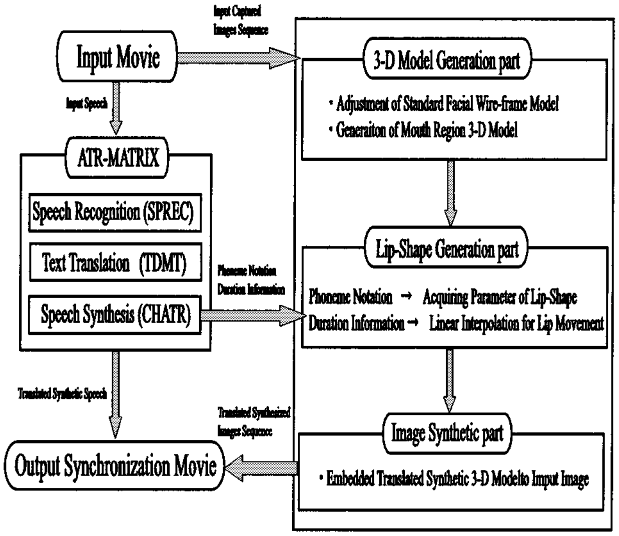

Input Captured Images Sequence
Input Movie
Input Speech
3-D Model Generation part
• Adjustment of Standard Facial Wire-frame Model
• Generaiton of Mouth Region 3-D Model
ATR-MATRIX
Speech Recognition (SPREC)
Text Translation (TDMT)
Speech Synthesis (CHATR)
Phoneme Notation
Duration Information
Lip-Shape Generation part
Phoneme Notation → Acquiring Parameter of Lip-Shape
Duration Information → Linear Interpolation for Lip Movement
Translated Synthetic Speech
Translated Synthesized Images Sequence
Image Synthetic part
Output Synchronization Movie
• Embedded Translated Synthetic 3-D Modelto Imput Image

Fig. 16. System overview of audio–visual speech translation.

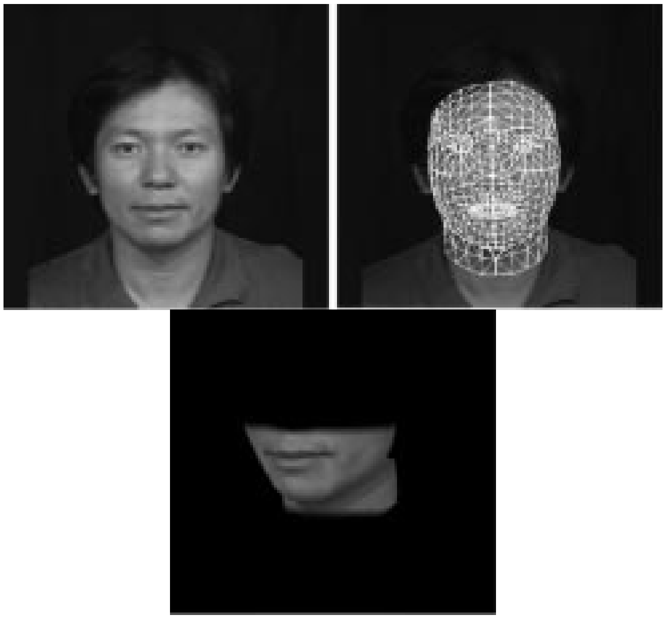


Fig. 17. 3-D Personal face synthesis process.

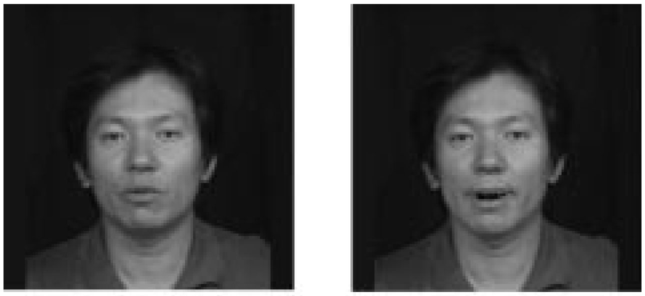


Fig. 18. Translated speaking face image. (a) Original image. (b) Translated synthetic image.

For instance, stand-in speech in movies does not match the lip movements of the facial images. However, face movements are necessary to transmit the nonverbal information of the speaker. If we choose to create all facial images by computer graphics, it might be difficult to send messages of nonverbal information. However, if we were to develop a technology capable of translating original facial speaking motion synchronized to translated speech, we could achieve natural multilingual multimodal speech translation.

In this paper, we introduce a work on an audio–visual translation system 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 a video camera for the other regions for natural speaking face generation [95].

V. System Overview

Fig. 16 shows an overview of the system in this work. The system is divided into two parts: one is a speech translation part and the other is an image translation part. The speech translation part is composed of ATR-MATRIX [97], which was developed at ATR. ATR-MATRIX is composed of a audio speech recognizer, dialog translator, and audio speech synthesizer, CHATR [98], to generate synthesized audio speech. In the audio–visual speech translation process, the two parameters of phoneme notation and duration information, which are output from CHATR, are applied to the facial image translation procedure. The first step of the image translation part is to create 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 the 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 in the audio speech synthesis stage. Duration information is also applied and used for linear interpolation for smooth lip movements. Here, the lip-shape parameters are defined by a vector derived from the natural face at lattice points on the wire-frame model 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 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/s.

VI. Generation of a 3-D Personal Face Model

Fig. 17 shows a personal face model synthesis process. The original image in Fig. 17 is used as the original personal face image, fitting shows the fitting result of a generic face model [99], and mouth model denotes a mouth part model constructed by a personal model used for mouth synthesis for lip synchronization. Fig. 18 is an example of a translated face. To control the mouth region's 3-D model, we use seven control points on the model [100]. These points are controlled by geometric movement rules based on the bone and muscle structure. Here, reference lip-shape images are prepared from the front and side. Then, we transform the wire-frame model to approximate the reference images. In this process, we acquire vectors of lattice points on the wire-frame model. Then, we store these vectors in the lip-shape database. This database is normalized by the mouth region's size, so we do not need speaker adaptation. As a result, this system achieves talking face generation with a small database. A viseme is a word created from a phoneme, which is the smallest linguistic sound unit. Visemes are defined for lip movement information like [au] and [ei] of the phonetic alphabet, but in this work, 22 kinds for English and additional five kinds for Japanese are used. A silence interval viseme is also prepared. The lip-shape database of this system is defined by only vectors of lattice points on a wire-frame. However, there is no data among the standard lip-shapes. Linear interpolation is applied for lip movements by using duration information from CHATR.

As a result of this research, the developed 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 multimodal English-to-Japanese and Japanese-to-English translation, has been achieved by applying parameters from CHATR. For further improvement of this system, we need to model the tongue in the 3-D model. The tongue model also needs to be controlled by parameters. This should help the entire 3-D model express images more precisely. At the same time, we must retrace the linear interpolation method of the lip-shape. The automatic model match-move algorithm [101] clearly needs to be applied to complete the system. Furthermore, this system only works off-line in this research. To operate the system on-line, a greater speed is necessary. Of last note, because of the different durations between the original audio speech and translated audio speech, a method that can control the duration information from the image synthesis part to the audio speech synthesis part needs to be developed.

VII. Conclusion

This paper introduced attempts at audio–visual multimodal integration for speech recognition, audio–visual trans-modal mapping for audio–visual speech synthesis, and audio–visual speech translation, all of which are applied by audio–visual complementary correlation and synchronicity. For audio–visual speech recognition, the paper showed a strong correlation with loose synchronicity between audio speech and mouth movement images. This fact necessitates a sophisticated model able to represent audio–visual loose synchronicity even beyond the phoneme boundary. Furthermore, adaptive weight control for each modality is also necessary considering modality reliability in the target environment. For audio–visual speech synthesis, the paper described benefits of an approach based on HMMs. Although the model does not require the transcription of the utterance, the performance can be further improved if the transcription is provided. This paper also introduced a sophisticated method that can estimate visual image parameters using the EM algorithm based on the audio–visual joint probability. This algorithm basically can avoid the misalignment problem of the Viterbi algorithm. Finally, this paper introduced an attempt at audio–visual speech translation. This was an example applying audio–visual integration for natural telecommunications. There indeed are many problems in terms of the speech recognition performance, speech translation quality, face image quality, and processing time by image processing. However, this audio–visual translation will be one of the most important applications of audio–visual integration for globalized natural and friendly telecommunications among human beings in the future.

References

[1] H. McGurk and J. MacDonald, "Hearing lips and seeing voices," *Nature*, pp. 746–748, Dec. 1976.

[2] E. Vatikiotis-Bateson, I. M. Eigsti, S. Yano, and K. Munhall, "Eye movement of perceivers during audiovisual speech perception," *Perception Psychophys.*, 1998.

[3] W. H. Sumby and I. Pollack, "Visual contribution to speech intelligibility in noise," *J. Acoust. Soc. Amer.*, vol. 26, pp. 212–215, Mar. 1954.

[4] C. Benoit, T. Guiard-Marigny, B. LeGoff, and A. Adjoudani, "Which components of the face do humans and machines best speechread?," in *Speechreading by Humans and Machines: Models, Systems, and Applications*, ser. NATO ASI. New York: Springer-Verlag, 1996, pp. 315–328.

[5] *Hearing by Eye: The Psychology of Lipreading*, B. Dodd and R. Campbell, Eds., Lawrence Erlbaum, Hillsdale, NJ, 1987.

[6] J. Robert-Ribes, J. L. Schwartz, and P. Escudier, "A comparison of models for fusion of the auditory and visual sensors in speech perception," *Artificial Intell. Rev. J.*, vol. 9, pp. 323–345, 1995.

[7] D. G. Stork and M. E. Hennecke, Eds., *Speechreading by Humans and Machines*: Springer, 1996.

[8] R. Campbell, B. Dodd, and D. Burnham, Eds., *Hearing by Eye II: Advances in the Psychology of Speechreading and Auditory–Visual Speech*: Psychlogy Press, 1999.

[9] B. P. Yuhas, M. H. Goldstein, and T. Sejnowski, "Integration of acoustic and visual speech signals using neural networks," *IEEE Commun. Mag.*, vol. 27, pp. 65–71, 1989.

[10] T. Chen, "Audiovisual speech processing," *IEEE Signal Processing Mag.*, pp. 9–21, Jan. 2001.

[11] J.-L. Schwartz, J. Robert-Ribes, and P. Escudier, "Ten Years After Summerfield: a Taxonomy of Models for Audio-visual Fusion in Speech Perception," in *Hearing by Eye II: Advances in the Psychology of Speechreading and Auditory-visual Speech*: Psychology Press, 1999, pp. 85–108.

[12] N. M. Brooke and E. D. Petajan, "Seeing speech: Investigations into the synthesis and recognition of visible speech movements using automatic image processing and computer graphics," in *Proc. Int. Conf. Speech Imput/Output, Techniques, Applicat.*, 1986, pp. 104–109.

[13] E. K. Finn and A. A. Montgomery, "Automatic optically based recognition of speech," *Pattern Recognition Lett.*, vol. 8, no. 3, pp. 159–164, 1988.

[14] B. P. Yuhas, M. H. Goldstein, T. J. Sejnowski, and R. E. Jenkins, "Neural network models of sensory integration for improved vowel recognition," *Proc. IEEE*, vol. 78, no. 10, pp. 1658–1668, 1990.

[15] T. Watanabe and M. Kohda, "Lip-reading of Japanese vowels using neural networks," in *Proc. Int. Conf. Spoken Language Processing*, 1990, pp. 1373–1376.

[16] J. Wu, S. Tamura, H. Mitsumoto, H. Kawai, K. Kurosu, and K. Okazaki, "Neural network vowel-recognition jointly using voice features and mouth shape image," *Pattern Recognition*, vol. 24, no. 10, pp. 921–927, 1991.

[17] D. G. Stprl, "Sources of structure in neural networks for speech and language," *Int. J. Neural Syst.*, vol. 2, no. 3, pp. 159–167, 1991.

[18] D. G. Stork, G. J. Wolff, and E. P. Levine, "Neural network lipreading system for improved speech recogniton," in *Proc. IJCNN'92*, vol. 2, 1992, pp. 285–295.

[19] P. Silsbee, "Computer Lipreading for Improved Accuracy in Automatic Speech Recognition," Ph.D. dissertation, Univ. Texas, —AUTHOR, PLEASE LIST CITY-, 1993.

[20] A. J. Goldschen, "Continuous Automatic Speech Recognition by Lipreading," Ph.D. dissertation, George Washington Univ., Washington, DC, 1993.

[21] C. Bregler, H. Hild, S. Manke, and A. Waibel, "Improving connected letter recognition by lipreading," in *Proc. Int. Conf. Acoust., Speech, Signal Processing*, vol. 1, 1993, pp. 557–560.

[22] C. Bregler and Y. Konig, "'Eigenlips' for robust speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP'94)*, 1994, pp. 669–672.

[23] N. M. Brooke, M. J. Tomlinson, and R. K. Moore, "Automatic speech recognition that includes visual speech cues," in *Proc. Inst. Acoust.*, vol. 16, 1994, pp. 15–22.

[24] P. Duchnowski, M. Hunke, D. Busching, U. Meier, and A. Waibel, "Toward movement-invariant automatic lip-reading and speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP'95)*, vol. 1, 1995, pp. 109–112.

[25] A. Adjoudani and C. Benoit, "Audio-visual speech recognition compared across two architectures," in *Proc. Eurospeech'95*, vol. 2, 1995, pp. 1563–1566.

[26] N. M. Brooke, "Talking heads and speech recognizers that can see:the computer processing of visual speech signals," *Speechreading Man Machine*, pp. 351–373, 1996.

[27] M. E. Hennecke, D. G. Stork, and K. V. Prasad, "Visionary Speech: Looking Ahead to Practical Speechreading Systems," in *Speechreading by Humans and Machines*. New York: Springer-Verlag, 1996, pp. 331–350.

[28] A. Adjoudani and C. Benoit, "On the integration of auditory and visual parameters in an HMM-based asr," in *Proc. Auditory–Visual Speech Processing*, 1996, pp. 461–471.

[29] ——, "On the integration of auditory and visual parameters in an HMM-based asr," in *Speechreading by Humans and Machines*. New York: Springer-Verlag, 1996, pp. 461–472.

[30] P. L. Silsbee and Q. Su, "Audio-Visual Sensory Integration Using Hidden Markov Models," in *Speechreading by Humans and Machines: Models, Systems, and Applications*. New York: Springer-Verlag, 1996, pp. 489–496.

[31] P. Deleglise, A. Rogozan, and M. Alissali, "Asynchronous integration of audio and visual sources in bimodal automatic speech recognition," in *Proc. EUSIPCO,'96*, 1996.

[32] M. J. Tomlinson, M. J. Russell, and N. M. Brooke, "Integrating audio and visual information to provide highly robust speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing(ICASSP'96)*, 1996, pp. 821–824.

[33] P. Jourlin, "Handling disynchronization phenomena with HMM in connected speech," in *Proc. EUSIPCO '96*, 1996.

[34] G. Potamianos, E. Cosatto, H. P. Graf, and D. B. Roe, "Speaker independent audiovisual database for bimodal asr ," in *Proc. European Tutorial Workshop Audiovisual Speech Processing*, 1997.

[35] S. Nakamura, R. Nagai, and K. Shikano, "Improved bimodal speech recognition using tied-mixture HMMs and 5000 word audio–visual synchronous database," in *Proc. Eurospeech'97*, 1997, pp. 1623–1626.

[36] S. Cox, I. Matthews, and A. Bangham, "Combining noise compensation with visual information in speech recognition," in *Proc. Auditory–Visual Speech Processing*, 1997, pp. 53–56.

[37] B. Andre-obrecht, B. Jacob, and N. Parlangeau, "Audio visual speech recognition and segmental master slave HMM," in *Proc. Auditory–Visual Speech Processing*, 1997, pp. 49–52.

[38] A. Rogozan, P. Deleglise, and M. Alissali, "Adaptive determination of audio–visual weights for automatic speech recognition," in *Proc. Auditory–Visual Speech Processing*, 1997, pp. 67–64.

[39] G. Potamianos and H. P. Graf, "Discriminative training of HMM stream exponents for audio–visual speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP'98)*, vol. 6, May 1998, pp. 3733–3736.

[40] G. Potamianos and A. Potamianos, "Speaker adaptation for audio–visual speech recognition," in *Proc. EUROSPEECH'99*, 1999, pp. 1291–1294.

[41] Y. Nankaku, K. Tokuda, and T. Kitamura, "Intensity-and location-normalized training for HMM-based visual speech recognition," in *Proc. EUROSPEECH'99*, 1999, pp. 1287–1290.

[42] C. Neti, G. Iyengar, G. Potamianos, A. Senior, and B. Maison, "Perceptual interfaces for information interaction: Joint processing of audio and visual information for human-computer interaction," in *Proc. Int. Conf. Spoken Language Processing (ICSLP'00)*, vol. 3, Oct. 2000, pp. 11–14.

[43] M. Heckmann, F. Berthommier, C. Savario, and K. Kroschel, "Labeling audio–visual speech corpora and training an ANN/HMM audio–visual speech recognition system," in *Proc. Int. Conf. Spoken Language Processing (ICSLP'00)*, vol. 4, Oct. 2000, pp. 9–12.

[44] S. Nakamura, H. Ito, and K. Shikano, "Stream weight optimization of speech and lip image sequence for audio–visual speech recognition," in *Proc. Int. Conf. Spoken Language Processing (ICSLP'00)*, vol. 3, 2000, pp. 20–23.

[45] C. Miyajima, K. Tokuda, and T. Kitamura, "Audio-visual speech recognition using mce-based HMMs and model-dependent stream weights," in *Proc. Int. Conf. Spoken Language Processing (ICSLP'00)*, Oct. 2000, pp. 1023–1026.

[46] G. Potamianos, A. Verma, C. Neti, G. Iyengar, and S. Basu, "A cascade image transform for speaker independent automatic speechreading," in *Proc. IEEE Int. Conf. Multimedia*, Aug. 2000.

[47] F. J. Huang and T. Chen, "Tracking of multiple faces for huan-computer interfaces and virtual environments," in *Proc. IEEE Int. Conf. Multimedia*, 2000.

[48] K. Kumatani, S. Nakamura, and K. Shikano, "An adaptive integration based on product HMM for audio–visual speech recognition," in *Proc. IEEE Int. Conf. Multimedia Expo. (ICME'01)*, vol. 1, Aug, 2001.

[49] S. Nakamura, K. Kumatani, and S. Tamura, "State synchroonous modeling of audio–visual information for bi-modal speech recognition," in *Proc. IEEE Workshop Automatic Speech Recognition Understanding*, vol. 1, Dec. 2001.

[50] E. D. Petajan, "Automatic Lipreading to Enhance Speech Recognition," Ph.D. dissertation, Univ. Illinois, 1984.

[51] C. Bregler and S. Omohundro, "Nonlinear manifold learning for visual speech recognition," in *Proc. IEEE ICCV*, 1995, pp. 494–499.

[52] P. L. Silsbee, "Sensory integration in audiovisual automatic speech recognition," in *28th Annu. Asilomar Conf. Signals, Syst., Comput.*, vol. 1, 1994, pp. 561–565.

[53] S. M. Peeling, R. K. Moore, and M. J. Tomlinson, "The multilayer perceptron as a tool for speech pattern processing research," in *Proc. Inst. Acoust.*, vol. 8, 1986, pp. 307–314.

[54] C. Bregler, H. Hild, S. Manke, and A. Waibel, "Improving connected letter recognition by lipreading," in *Proc. Int. Joint Conf. Speech Signal Processing*, vol. 1, 1993, pp. 557–560.

[55] M. E. Hennecke, K. V. Prasad, and D. G. Stork, "Using deformable templates to infer visual speech dynamics," in *Proc. 28th Annu. Asilomar Conf. Signals, Syst., Comput.*, vol. 1, 1994, pp. 578–582.

[56] P. Cosi, E. Magno, K. Caldognetto, K. Vagges, A. Mian, and M. Contolini, "Bimodal recognition experiments with recurrent neural networks," in *Proc. Int. Conf. Acoust., Speech, Signal Processing*, vol. 2, 1994, pp. 553–556.

[57] P. Jourlin, "Automatic bimodal speech recognition," in *Proc. ICPhS*, 1995.

[58] Y. Zhang, S. Levinson, and T. Huang, "Speaker independent audio–visual speech recognition," in *Proc. IEEE Int. Conf. Multimedia Expo. (ICME'00)*, 2000, TP8.01.

[59] S. M. Chu and T. S. Huang, "Bimodal speech recognition using coupled hidden Markov models," in *Proc. Int. Conf. Spoken Language Processing (ICSLP'00)*, vol. 2, Oct. 2000, pp. 747–750.

[60] H. Pan, Z. P. Liang, and T. S. Huang, "A new approach to integrate audio and visual features of speech," in *Proc. IEEE Int. Conf. Multimedia Expo. (ICME'00)*, 2000, TP8.06.

[61] J. Luettin, G. Potamianos, and C. Neti, "Asynchronous stream modeling for large vocabulary audio–visual speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP'01)*, 2001, pp. 165–168.

[62] M. R. Naphade, A. Garg, and T. S. Huang, "Duration dependent input output Markov models for audio–visual event detection," in *Proc. IEEE Int. Conf. Multimedia Expo. (ICME'01)*, 2001, pp. 369–372.

[63] M. Heckmann, T. Wild, F. Berthommier, and K. Kroschel, "Comparing audio- and a posteriori-probability-based stream confidence measure for audio–visual speech recognition," in *Proc. EUROSPEECH'01*, 2001, pp. 1023–1026.

[64] P. Teissier, A. Guerin-Dugue, and J.-L. Schwartz, "Models for audio-visual fusion in a noisy-vowel recognition task," *J. VLSI Signal Processing*, vol. 20, no. 1/2, pp. 25–44, 1998.

[65] J. Hernando, "Maximum likelihood weighting of dynamic speech features for CDHMM speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP'97)*, 1997, pp. 1267–1270.

[66] S. Katagiri, B.-H. Juang, and C.-H. Lee, "Pattern recognition using a family of design algorithms based upon the generalized probabilistic descent method," *Proc. IEEE*, vol. 86, no. 11, pp. 2345–2373, 1998.

[67] G. Potamianos and C. Neti, "Stream confidence estimation for audio–visual speech recognition," in *Proc. Int. Conf. Spoken Language Processing (ICSLP'00)*, vol. 3, Oct. 2000, pp. 746–749.

[68] H. Glotin, D. Vergyri, C. Neti, G. Potamianos, and J. Luettin, "Weighting schemes for audio–visual fusion in speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP'01)*, 2001, pp. 173–176.

[69] T. Kuratate, K. G. Munhall, P. E. Rubin, E. Vatikiotis-Bateson, and H. Yehia, "Audio-visual synthesis of talking faces from speech production correlates," in *Proc. EUROSPEECH'99*, 1999, pp. 1279–1282.

[70] Z. Wen, P. Hong, and T. S. Huang, "Real time speech driven facial animation using formant analysis," in *Proc. IEEE Int. Conf. Multimedia Expo.(ICME'01)*, 2001, pp. 1024–1027.

[71] S. Kshirsagar and N. Magnena-Thalmann, "Lip synchronization using linear predictive analysis," in *Proc. IEEE Int. Conf. Multimedia Expo. (ICME'00)*, 2000, TP8.02.

[72] S. Morishima, K. Aizawa, and H. Harashima, "An intelligent facial image coding driven by speech and phoneme," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP'89)*, 1989, pp. 1795–1798.

[73] L. M. Arslan and D. Talkin, "Codebook based face point trajectory synthesis algorithm using speech input," *Speech Commun.*, vol. 27, no. 2, pp. 81–93, 1999.

[74] S. Morishima and H. Harashima, "A media conversion from speech to facial image for intelligent man-machine interface," *IEEE J. Select. Areas Commun.*, vol. 9, pp. 594–600, 1991.

[75] F. Lavagetto, "Converting speech into lip movements: A multimedia telephone for hard of hearing people," *IEEE Trans. Rehab. Eng.*, vol. 3, pp. 90–102, Mar. 1995.

[76] R. R. Rao and T. Chen, "Cross-modal prediction in audio–visual communication," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP'96)*, vol. 4, 1996, pp. 2056–2059.

[77] W. Goldenthal, "Driving synthetic mouth gestures: Phonetic recognition for faceme!," in *Proc. Eurospeech'97*, 1997, pp. 1995–1998.

[78] A. D. Simons and S. J. Cox, "Generation of mouthshape for a synthetic talking head," in *Proc. Inst. Acoust.*, vol. 12, 1990, pp. 475–482.

[79] W. Chou and H. Chen, "Speech recognition for image animation and coding," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP'95)*, 1995, pp. 2253–2256.

[80] T. Chen, Y. Wang, H. P. Graf, and C. Swain, "A new frame interpolation scheme for talking head sequences," in *Proc. IEEE Int. Symp. Multimedia*, vol. 2, 1995, pp. 591–594.

[81] E. Yamamoto, S. Nakamura, and K. Shikano, "Speech-to-lip movement synthesis by HMM," in *Proc. Auditory–Visual Speech Processing*, 1997, pp. 137–140.

[82] N. M. Brooke and S. D. Scott, "Two- and three-dimensional audio–visual speech synthesis," in *Proc. Auditory–Visual Speech Processing*, 1998.

[83] T. Masuko, T. Kobayashi, M. Tamura, J. Masubuchi, and K. Tokuda, "Text-to-visual speech synthesis based on parameter generation from HMM," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP'98)*, 1998, pp. 3745–3748.

[84] E. Yamamoto, S. Nakamura, and K. Shikano, "Subjective evaluation for HMM-based speech-to-lip movement synthesis," in *Proc. Auditory–Visual Speech Processing*, 1998.

[85] ——, "Lip movement synthesis from speech based on hidden Markov models," *Speech Commun.*, pp. 105–115, 1998.

[86] F. J. Huang and T. Chen, "Real-time lip-synch face animation driven by human voice," in *Proc. IEEE Workshop Multimedia Signal Processing (MMSP'98)*, 1998.

[87] M. Tamura, S. Kondo, T. Masuko, and T. Kobayashi, "Text–audio–visual speech synthesis based on parameter generation from HMM," in *Proc. EUROSPEECH'99*, 1999, pp. 959–962.

[88] S. Sako, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "HMM-based text-to-audio–visual speech synthesis," in *Proc. Int. Conf. Spoken Language Processing (ICSLP'00)*, vol. 3, 2000, pp. 25–28.

[89] T. A. Faruquie, C. Neti, N. Rajput, L. V. Subramaniam, and A. Verma, "Translingual visual speech synthesis," in *Proc. IEEE Int. Conf. Multimedia Expo. (ICME'00)*, 2000, pp. 2393–2396.

[90] K. Kakihara, S. Nakamura, and K. Shikano, "Speech-to-face movement synthesis based on HMMs," in *Proc. IEEE Int. Conf. Multimedia Expo. (ICME'00)*, 2000, MP7.07.

[91] S. Nakamura and E. Yamamoto, "Speech-to-lip movement synthesis by maximizing audio–visual joint probability based on the em algorithm," *J. VLSI Signal Processing*, vol. 27, no. 1/2, pp. 119–126, 2001.

[92] K. Choi, Y. Luo, and J.-N. Hwang, "Hidden Markov model inversion for audio-to-visual conversion in an mpeg-4 facial animation system," *J. VLSI Signal Processing*, vol. 29, no. 1/2, pp. 51–61, 2001.

[93] E. Yamamoto, S. Nakamura, and K. Shikano, "Speech-to-lip movement synthesis based on em algorithm using audio–visual HMMs," in *Proc. Int. Conf. Spoken Language Processing (ICSLP'98)*, 1998, pp. 1275–1278.

[94] T. Guiard-Marigny, T. Adjoudani, and C. Benoit, "A 3-D model of the lips for visual speech synthesis," in *Proc. 2nd ESCA/IEEE Workshop Speech Synthesis*, Sept. 1994, pp. 49–52.

[95] S. Ogata, K. Murai, S. Nakamura, and S. Morishima, "Model-based lip synchronization with a automatically translated synthetic voice toward a multimodal translation system," in *Proc. IEEE Int. Conf. Multimedia Expo. (ICME'01)*, 2001.

[96] J. Yang, J. Xiao, and M. Ritter, "Automatic selection of visemes for image-based visual speech synthesis," in *Proc. IEEE Int. Conf. Multimedia Expo. (ICME'00)*, 2000, pp. 1081–1084.

[97] T. Morimoto, Y. Sagisaka, N. Campbell, H. Iida, F. Sugaya, A. Yokoo, and S. Yamamoto, "Japanese-to-English speech translation system:atr-matrix," in *Proc. Int. Conf. Spoken Language Processing (ICSLP'98)*, 1998, pp. 957–960.

[98] N. Campbell and A. W. Black, "CHATR: A multilingual speech re-sequencing synthesis system,", IEICE, 1995.

[99] H. Harashima and S. Morishima. (1997) Facial image processing system for human-like 'kansei' agent. [Online]. Available: http://www.tokyo.image-lab.or.jp/aa/ipa

[100] K. Ito, T. Misawa, J. Muto, and S. Morishima, "3-D lip expression generation by using new lip parameters,", IEICE Rep., vol. A16-24, 2000.

[101] T. Misawa, S. Nakamura, and S. Morishima, "Automatic face tracking and model match move in video sequence using 3D face model," in *Proc. IEEE Int. Conf. Multimedia and Expo. (ICME'01)*, 2001, TP7.1.

Satoshi Nakamura received the B.S. degree in electronics engineering from Kyoto Institute of Technology, Kyoto, Japan, in 1981 and the Ph.D. degree in information science from Kyoto University in 1992.

From 1981 to 1986 and 1990 to 1993, he was with the Central Research Laboratory, Sharp Corporation, Nara, Japan, where he was engaged in speech recognition research. From 1986 to 1989, he was a Researcher in the Speech Processing Department at ATR Interpreting Telephony Research Laboratories. From 1994 to 2000, he was an Associate Professor of 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. He is currently the head of Department 1 at ATR Spoken Language Translation Laboratories, Japan. His current research interests include speech recognition, speech translation, spoken dialogue systems, stochastic modeling of speech, and microphone arrays.

Dr. Nakamura received the Awaya Award from the Acoustical Society of Japan in 1992. He is a Member of the Acoustical Society of Japan, Institute of Electrical, Information, and Electronics Engineers (IEICE), Information Processing Society of Japan, and IEEE. He is currently a Member of the Speech Technical Committee of the IEEE Signal Processing Society and an Editor for the *Journal of the IEICE Information and System Society*.