4th International Conference on
Analysis and Modeling of Faces and Gestures (F. J. Perales and R. B. Fisher (Eds.): AMDO 2006, LNCS 4069, Springer), pp. 453-463, 2006

Modeling Timing Structure in Multimedia Signals

Hiroaki Kawashima, Kimitaka Tsutsumi, and Takashi Matsuyama

Kyoto University, Yoshida-Honmachi Sakyo, Kyoto 6068501, JAPAN,
{kawashima,tm}@i.kyoto-u.ac.jp,
http://vision.kuee.kyoto-u.ac.jp/

Abstract. Modeling and describing temporal structure in multimedia signals, which are captured simultaneously by multiple sensors, is important for realizing human machine interaction and motion generation. This paper proposes a method for modeling temporal structure in multimedia signals based on temporal intervals of primitive signal patterns. Using temporal difference between beginning points and the difference between ending points of the intervals, we can explicitly express *timing structure*; that is, synchronization and mutual dependency among media signals. We applied the model to video signal generation from an audio signal to verify the effectiveness.

1 Introduction

Measuring dynamic behavior such as speech, musical performances, and sport actions with multiple sensors, we obtain media signals across different modalities. We often exploit the temporal structure of co-occurrence, synchronization, and temporal difference among temporal patterns in these signals. For example, it is well-known fact that the simultaneity between auditory and visual patterns influences human perception (e.g., the McGurk effect [9]). On the other hand, modeling the cross-modal structure is important to realize the multimedia systems of human computer interaction; for example, audio-visual speech recognition [11] and media signal generation from another related signal (e.g., motion from audio signal)[2]. Motion modeling also exploits this kind of temporal structure, because motion timing among different parts plays an important role in natural motion generation.

State based co-occurrence models, such as coupled hidden Markov models (HMMs) [3], are strong methods for media integration [11]. These models describe a relation between adjacent or co-occurred states that exist in the different media signals (Fig. 1(a)). Although this frame-wise representation enables us to model short term relations or interaction among multiple processes, it is ill-suited to systems in which the features of synchronization and temporal difference between media signal patterns become significant. For example, an opening lip motion is strongly synchronized with an explosive sound /p/; on the other hand, the lip motion is loosely synchronized with a vowel sound /e/, and the motion always precedes the sound. We can see such an organized temporal difference

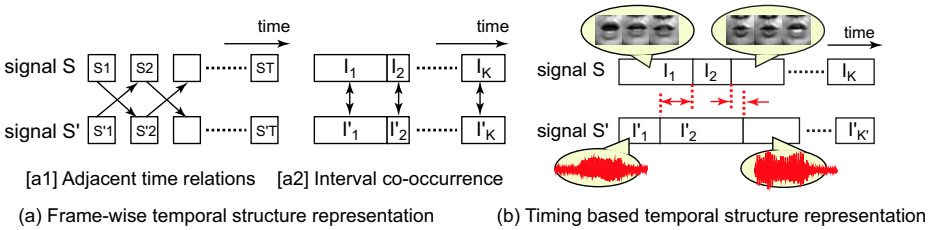


Fig. 1. Temporal structure representation in multimedia signals.

in music performances also; performers often make preceding motion before the actual sound.

In this paper, we propose a novel model that directly represents this important aspect of temporal relations, what we refer to as *timing structure*, such as synchronization and mutual dependency with organized temporal difference among multiple media signals (Fig. 1(b)).

First, we assume that each media signal is described by a finite set of *modes*: primitive temporal patterns. Segment models [13], which are the generalization of segmental HMMs [7], become popular models in the speech recognition community to describe audio signals based on this assumption. A number of similar models are widely proposed in different communities, for example, hybrid systems [4, 6] in the computer vision, and the motion texture [8] in the graphics. These models describe complex temporal variations not only by a physical-time based state transition but also by an event-based state transition that is free from temporal metric space (i.e., it models just the order of events). We refer to these models as *interval models*, because every model provides an interval representation of media signals, where each interval is a temporal region labeled by one of the modes.

Then, we introduce a *timing structure model*, which is a stochastic model for describing temporal structure among intervals in different media signals. Because the model explicitly represents temporal difference between beginning and ending points of intervals, it provides a framework of integrating multiple interval models across modalities. Consequently, we can exploit the model to human machine interaction systems in which media synchronization plays an important role. In the experiments, we verify the effectiveness of the method by applying it to media signal conversion that generate a media signal from another media signal.

2 Modeling Timing Structure in Multimedia Signals

2.1 Temporal Interval Representation of Media Signals

To define timing structure, we assume that each media signal is represented by a single interval model, and the parameters of the interval model are estimated in advance (see [13, 8], for example). Then, each media signal is described by an interval sequence. In the following paragraphs, we introduce some terms and notations for the structure and the model definition.

Media signals: Multimedia signals are obtained by measuring dynamic event with N_s sensors simultaneously. Let S_c be a single media signal. Then, multimedia signals become $\mathcal{S} = \{S_1, \cdots, S_{N_s}\}$. We assume that S_c is a discrete signal that is sampled by rate ΔT_c.

Modes and Mode sets: Mode $M_i^{(c)}$ is the property of temporal variation occurred in signal S_c (e.g., "opening mouth" and "closing mouth" in a facial video signal). We define a mode set of S_c as a finite set: $\mathcal{M}^{(c)} = \{M_1^{(c)}, \cdots, M_{N_c}^{(c)}\}$. Each mode is modeled by a sub model of the interval models. For example, hybrid systems, which we use in our experiments, use linear dynamical systems for the mode models.

Intervals: Interval $I_k^{(c)}$ is a temporal region that a single mode represents. Index k denotes a temporal order that the interval appeared in signal S_c. Interval $I_k^{(c)}$ has properties of beginning and ending time $b_k^{(c)}, e_k^{(c)} \in \mathbb{N}$ (the natural number set), and mode label $m_k^{(c)} \in \mathcal{M}^{(c)}$. Note that, we simply refer to the indices of sampled order as "time". We assume signal S_c is partitioned into interval sequence $\mathcal{I}^{(c)} = \{I_1^{(c)}, ..., I_{K_c}^{(c)}\}$ by the interval model, where the intervals have no overlaps or gaps (i.e., $b_{k+1}^{(c)} = e_k^{(c)} + 1$ and $m_k^{(c)} \neq m_{k+1}^{(c)}$).

Interval representation of media signals: Interval representation of multimedia signals is a set of interval sequences: $\{\mathcal{I}^{(1)}, ..., \mathcal{I}^{(N_s)}\}$.

2.2 Definition of Timing Structure

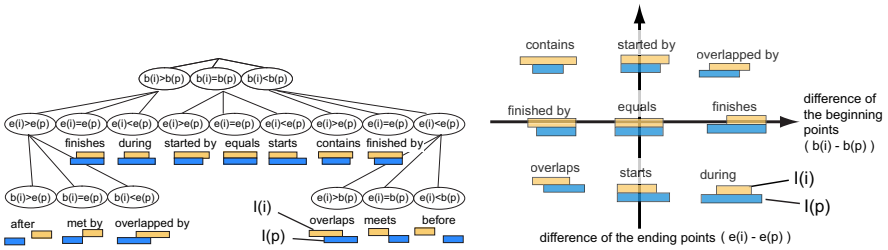
In this paper, we concentrate on modeling timing structure between two media signals S and S'. (We use the mark " ' " to discriminate between the two signals.)

Let us use notation $I_{(i)}$ for an interval I_k that has mode $M_i \in \mathcal{M}$ in signal S (i.e.,$m_k = M_i$), and let $b_{(i)}, e_{(i)}$ be its beginning and ending points, respectively. (We omit index k, which denotes the order of the interval.) Similarly, let $I'_{(p)}$ be an interval that has mode $M'_p \in \mathcal{M}'$ in the range $[b'_{(p)}, e'_{(p)}]$ of signal S'. Then, the temporal relation of two modes becomes the quaternary relation of the four temporal points $R(b_{(i)}, e_{(i)}, b'_{(p)}, e'_{(p)})$. If signal S and S' has different sampling rate, we have to consider the relation of continuous time such as $b_{(i)}\Delta T$ on behalf of $b_{(i)}$. In this subsection, we just use $b_{(i)} \in \mathbb{R}$(the real number set) for both continuous time and the indices of discrete time to simplify the notation.

Let us define timing structure as the relation R that can be determined by four binary relations $R_{bb}(b_{(i)}, b'_{(p)})$, $R_{be}(b_{(i)}, e'_{(p)})$, $R_{eb}(e_{(i)}, b'_{(p)})$, $R_{ee}(e_{(i)}, e'_{(p)})$. In the following, we specify the four binary relations that we focus on this paper.

Considering temporal ordering relations $R_<, R_=, R_>$, which are often used in temporal logic [1], for these binary relations, we get 3^4 relations for R. Because of $b_{(i)} \leq e_{(i)}$ and $b'_{(p)} \leq e'_{(p)}$, it can be reduced to 13 relations as shown in Fig. 2(a). However, temporal metric information is omitted in these 13 relations, which often becomes significant for modeling human behavior with temporal structure (e.g., temporal difference between sound and motion).

We therefore introduce metric relations for R_{bb} and R_{ee} by assuming that R_{be} and R_{eb} is $R_\leq$ and $R_\geq$, respectively (i.e., the two modes have overlaps). This assumption is natural when the influence of one mode to the other modes with



(a) Temporal ordering relations. (b) Temporal difference relation.

Fig. 2. Temporal relations of two intervals. (a) The temporal order of beginning and ending time provides 13 relations of the two intervals. (b) The horizontal and vertical axes denote the difference between beginning points $b_{(i)} - b'_{(p)}$ and the difference between ending points $e_{(i)} - e'_{(p)}$, respectively.

long temporal distance can be ignored. For the metric of R_{bb} and R_{ee}, we use temporal difference $b_{(i)} - b'_{(p)}$ and $e_{(i)} - e'_{(p)}$, respectively; the relation is represented by a point $(D_b, D_e) \in \mathbb{R}^2$ (see also Fig. 2(b)). In the next subsection, we model this type of temporal metric relation using two-dimensional distributions.

2.3 Modeling Timing Structure

Temporal difference distribution of overlapped mode pairs: To model the metric relations that described in the previous subsection, we introduce the following distribution for every mode pair $(M_i, M'_p) \in \mathcal{M} \times \mathcal{M}'$:

$$P(b_k - b'_{k'} = D_b, e_k - e'_{k'} = D_e | m_k = M_i, m'_{k'} = M'_p, [b_k, e_k] \cap [b'_{k'}, e'_{k'}] \neq \phi). \quad (1)$$

We refer to this distribution as a *temporal difference distribution*. Because the distribution explicitly represent the frequency of the metric relation between two modes (i.e., temporal difference between beginning points and the difference between ending points), it provides significant temporal structure for two media signals. For example, if the peak of the distribution comes to the origin, the two modes tend to be synchronized in their beginning and ending points; on the other hand, if $b_k - b_{k'}$ has large variance, the two modes loosely synchronized in their beginning.

As we described in Subsection 2.2, the domain of the distribution is $\mathbb{R}^2$. To estimate the distribution from a finite number of samples (i.e., overlapped mode pairs), we fit a density function such as Gaussian or its mixture models to the samples when we use the model in real applications.

Co-occurrence distribution of mode pairs: As we see in Eq. (1), the temporal difference distribution is a probability distribution under the condition of the given mode pair. To represent frequency that each mode pair appears in the overlapped interval pairs, we introduce the following distribution:

$$P(m_k = M_i, m_{k'} = M'_p \,|\, [b_k, e_k] \cap [b'_{k'}, e'_{k'}] \neq \phi). \quad (2)$$

We refer to this distribution as *co-occurrence distribution* of mode pairs. The distribution can be easily estimated by calculating a mode pair histogram from every overlapped interval pairs.

Mode transition probability: Using Eq. (1) and (2), we can represent timing structure that is defined in Subsection 2.2. Although timing structure models temporal metric relations between media signals, temporal relation in each media signal is also important. Therefore, similar to previously introduced interval models, we use the following transition probability of adjacent modes in each signal:

$$P(m_k = M_j | m_{k-1} = M_i)\,(M_i, M_j \in \mathcal{M}). \tag{3}$$

3 Media Signal Conversion Based on Timing Structure

Once we estimate the timing structure model that introduced in Section 2 from simultaneously captured multimedia signals, we can exploit the model for generating a media signal from another related signal. We refer to the application as *media signal conversion*, and introduce the algorithm in this section.

The overall flow of media signal conversion from signal S' to S is as follows: (1) a reference (input) media signal S' is partitioned into an interval sequence $\mathcal{I}' = \{I'_1, ..., I'_{K'}\}$, (2) a media interval sequence $\mathcal{I} = \{I_1, ..., I_K\}$ is generated from a reference interval sequence $\mathcal{I}'$, (3) a media signal S is generated from $\mathcal{I}$. (K and K' is the number of intervals in $\mathcal{I}$ and $\mathcal{I}'$, and $K \neq K'$ in general.)

Since the methods of (1) and (3) have been already introduced in some literatures of interval models (see [8, 6], for example), we focus on (2), and propose a novel method that generates a media interval sequence from another related media interval sequence based on the timing structure model. In the following subsections, we assume that the two media signals S, S' have the same sampling rate to simplify the algorithm.

3.1 Formulation of Media Signal Conversion Problem

Let Φ be the timing structure model that is estimated in advance. Then, the problem of generating an interval sequence $\mathcal{I}$ from a reference interval sequence $\mathcal{I}'$ can be formulated by the following optimization:

$$\hat{\mathcal{I}} = \arg\max_{\mathcal{I}} P(\mathcal{I}|\mathcal{I}', \Phi) \tag{4}$$

In the equation above, we have to determine the number of intervals K and triples (b_k, e_k, m_k) for all the intervals $I_k\,(k = 1, ..., K)$, where $b_k, e_k \in [1, T]$ and $m_k \in \mathcal{M}$. T is the length of signal S', and the mode set $\mathcal{M}$ is estimated simultaneously with the signal segmentation. If we search for all the possible interval sequences $\{\mathcal{I}\}$, the calculation order increases exponentially in the increase of T. We therefore use a dynamic programming method, which is similar to the Viterbi algorithm in HMMs, to solve Eq. (4) (see Subsection 3.2).

We currently do not consider online media signal conversion, because it requires trace back mechanism. If online processing is necessary, one of the simplest

method is dividing input stream comparatively longer range than the sampling rate and apply the following method repeatedly.

3.2 Interval Sequence Conversion via Dynamic Programming

To simplify the notation, we omit the model parameter variable Φ in the following equations. Let us use notation $f_t = 1$ that denotes an interval "finishes" at time t, which follows Murphy's notation that is used in a research note about segment models [10]. Then, $P(m_t = M_j, f_t = 1|\mathcal{I}')$, which is the probability when an interval finishes at time t and the mode of time t becomes M_j in the condition of the given interval sequence $\mathcal{I}'$, can be calculated by the following recursive equation:

$$\begin{aligned} &P(m_t = M_j, f_t = 1|\mathcal{I}') \\ &= \sum_{\tau} \sum_{i\,(\neq j)} \left\{ \begin{array}{r} P(m_t = M_j, f_t = 1, l_t = \tau | m_{t-\tau} = M_i, f_{t-\tau} = 1, \mathcal{I}') \\ \times P(m_{t-\tau} = M_i, f_{t-\tau} = 1|\mathcal{I}') \end{array} \right\}, \end{aligned} \tag{5}$$

where l_t is a duration length of an interval (i.e., it continues l_t at time t) and m_t is a mode label at time t. The lattice in Fig. 3 depicts the path of the above recursive calculation. Each pair of arrows from each circle denotes whether the interval "continues" or "finishes", and every bottom circle sums up all the finishing interval probabilities.

The following dynamic programming algorithm is deduced directly from the recursive equation (5):

$$\begin{aligned} E_t(j) = \max_{\tau} \max_{i\,(\neq j)} \underline{P(m_t = M_j, f_t = 1, l_t = \tau | m_{t-\tau} = M_i, f_{t-\tau} = 1, \mathcal{I}')} E_{t-\tau}(i), \\ \text{where} \quad E_t(j) \triangleq \max_{m_1^{t-1}} P(m_1^{t-1}, m_t = M_j, f_t = 1|\mathcal{I}'). \end{aligned} \tag{6}$$

$E_t(j)$ denotes the maximum probability when the interval of mode M_j finishes at time t, and is optimized for the mode sequence from time 1 to $t-1$ under the condition of given $\mathcal{I}'$. The probability with underline denotes that interval I_k with a triple $(b_k = t-\tau+1, e_k = t, m_k = M_j)$ occurs just after the interval I_{k-1} that has mode $m_{k-1} = M_i$ and ends at $e_{k-1} = t-\tau$. We refer to this probability as an *interval transition probability*.

We recursively calculate the maximum probability for every mode that finishes at time $t (t = 1, ..., T)$ using Eq. (6). After the recursive calculation, we find the mode index $j^* = \arg\max_j E_T(j)$. Then, we can get the duration length of the interval that finishes at time T with mode label M_{j*}, if we preserve τ that gives the maximum value at each recursion of Eq. (6). Repeating this trace back, we finally obtain the optimized interval sequence and the number of intervals.

The remaining problem for the algorithm is the method of calculating the interval transition probability. As we see in the next subsection, this probability can be estimated from a trained timing structure model.

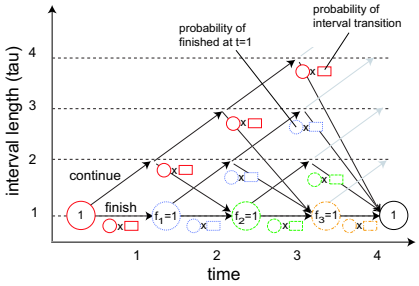


Fig. 3. Lattice to search optimal interval sequence (num. of mode =2). We assume that $\sum_j P(m_T = M_j, f_T = 1|\mathcal{I}') = 1$

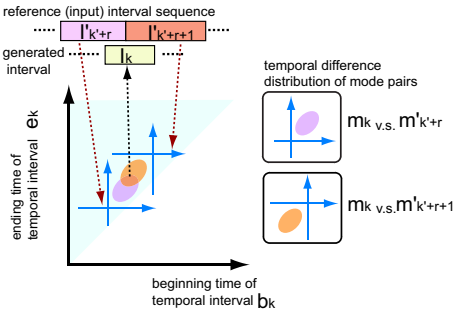


Fig. 4. An interval probability calculation from the trained timing structure model.

3.3 Calculation of Interval Transition Probability

As we described in previous subsection, the interval transition probability appeared Eq. (6) is the transition from interval I_{k-1} to I_k. To simplify the notation, let us replace $t-\tau+1$ with B_k. Let $e_{\min} = B_k$ and $e_{\max} = \min(T, B_k + l_{\max} - 1)$ be the minimum and maximum values of e_k, where $l_{\max}$ is the maximum length of the intervals. Let $I'_{k'}, ..., I'_{k'+R} \in \mathcal{I}'$ be reference intervals that are possible to overlap with I_k. Assuming that the reference intervals are independent of each other (this assumption empirically works well), the interval transition probability can be calculated by the following equation:

$$\begin{aligned}
&P(m_t = M_j, f_t = 1, l_t = \tau | m_{t-\tau} = M_i, f_{t-\tau} = 1, \mathcal{I}') \\
&= P(m_k = M_j, e_k, e_k \in [e_{\min}, e_{\max}] | m_{k-1} = M_i, b_k = B_k, I'_{k'}, ..., I'_{k'+R}) \\
&= \prod_{r=0}^{R} \{ \mathrm{Rect}(e_k, e_k \in [e_{\min}, b'_{k'+r} - 1]) \\
&\quad + \kappa_r P(m_k = M_j, e_k, e_k \in [b'_{k'+r}, e_{\max}] | m_{k-1} = M_i, b_k = B_k, I'_{k'+r}) \} ,
\end{aligned} \tag{7}$$

where $\mathrm{Rect}(e, e \in [a, b]) = 1$ in the range $[a, b]$; else 0. Since the domain of e_k is $[e_{\min}, e_{\max}]$, Rect is out of range when $r = 0$, and $b'_{k'} = e_{\min}$. κ is a normalizing factor: $\kappa_r = 1\,(r = 0)$ and

$$\kappa_r = P(m_k = M_j, e_k, e_k \in [b'_{k'+r}, e_{\max}] | b_k = B_k, m_{k-1} = M_i)^{-1} \quad (r = 1, ..., R).$$

In the experiments, we assume κ_r is uniform for (m_k, e_k); thus, $\kappa_r = N(e_{\max} - e_{\min} + 1)$ (N is the number of modes).

Using some assumptions that we will describe later, we can decompose the probability in Eq. (7) as follows:

$$\begin{aligned}
&P(m_k = M_j, e_k, e_k \in [b'_{k'+r}, e_{\max}] \,|\, m_{k-1} = M_i, b_k = B_k, I'_{k'+r}) \\
=&P(e_k \,|\, e_k \in [b'_{k'+r}, e_{\max}], m_k = M_j, b_k = B_k, I'_{k'+r}) \\
&\times P(m_k = M_j | e_k \in [b'_{k'+r}, e_{\max}], m_{k-1} = M_i, b_k = B_k, I'_{k'+r}) \\
&\times P(e_k \in [b'_{k'+r}, e_{\max}] \,|\, m_{k-1} = M_i, b_k = B_k)
\end{aligned}$$

The first term is the probability of e_k under the condition that I_k overlaps with $I'_{k'+r}$. We assume that it conditionally independent of m_{k-1}. This proba-

bility can be calculated from Eq. (1). Here, we omit the details of the deduction, and just make an intuitive explanation using Fig. 4. First, an overlapped mode pair in I_k and $I'_{k'+r}$ provides a relative distribution of $(b_k - b'_{k'+r}, e_k - e'_{k'+r})$. Since $I'_{k'+r}$ is given, the relative distribution is mapped to the absolute time domain (the upper triangle region). Normalizing this distribution of (b_k, e_k) for $e_k \in [b'_{k'+r}, e_{\max}]$, we obtain the probability of the first term. The second term can be calculated using Eq. (2) and (3). For the third term, we assume that the probability of $e_k \geq b'_{k'+r}$ is independent of $I'_{k'+r}$. Then, this term can be calculated by modeling temporal duration length l_t. In the experiments, we assumed uniform distribution of e_k and used $(e_{\max} - b'_{k'+r})/(e_{\max} - e_{\min} + 1)$.

The calculation cost strongly depends on the maximum interval length $l_{\max}$. If we successfully estimate the modes, $l_{\max}$ becomes comparatively small (i.e., balanced among modes); thus, the cost will be reasonable.

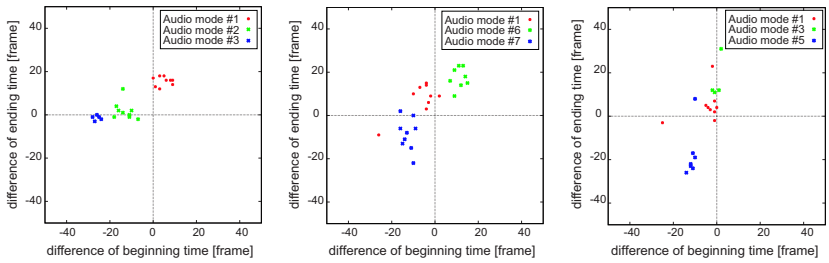
4 Experiments

We applied the media conversion method described in Section 3 to the application that generates image sequences from an audio signal.

Feature extraction: First, we captured continuous utterance of five vowels /a/,/i/,/u/,/e/,/o/ (in this order) using a pair of camera and microphone. This utterance was repeated nine times (18 sec.). The resolution of the video data was 720×480 and the frame rate was 60fps. The sampling rate of the audio signal was 48kHz (downsampled to 16kHz in the analysis). Then, we applied short-term Fourier transform to the audio data with the window step of 1/60msec; thus, the frame rate corresponds to the video data. Using filter bank analysis, we obtained 1134 frames of audio feature vectors (dimensionality was 25). For the video feature, we extracted lip region exploiting the Active Appearance Model [5]. Then, we downsampled the lip region to 32×32 pixels and applied principal component analysis (PCA) to the extracted lip image sequence. Finally, we obtained 1134 frames of video feature vectors (dimensionality was 27).

Segmentation and mode estimation of each media signal: Considering the extracted audio and visual feature vector sequences as signal S' and S, we estimated the number of modes, parameters of each mode, and the temporal partitioning of each signal. We used linear dynamical systems for the models of modes. To estimate the parameters, we exploited hierarchical clustering of the dynamical systems based on eigenvalue constraints [6]. The estimated number of modes was 13 and 8 for audio and visual modes, respectively. The segmentation results are shown in Fig. 6 (the first and second rows). Because of the noise, some vowel sounds were divided into several audio modes.

Training of the timing structure model between audio and video: Using the two interval sequences obtained by the segmentation, we estimated distributions of Eq. (1), (2), and (3). Figure 5 is the scattered plots of the samples that are temporal difference between beginning points and ending points of overlapped modes. Each chart shows samples of one visual mode to typical (two or



(a) visual mode #1 (b) visual mode #5 (c) visual mode #7

Fig. 5. Scattering plots of temporal difference between overlapped audio and visual modes. Visual mode #1, #5, and #7 corresponds to lip motion /o/ → /a/, /e/ → /o/, and /a/ → /i/, respectively

three) audio modes. We see that the beginning motion from /a/ to /i/ synchronized with the actual sound (right chart) compared to the motion from /o/ to /a/ (left) and from /e/ to /o/ (middle). Applying Gaussian mixture models to these distributions, we estimated the temporal difference distributions.

Lip image sequence generation from an audio signal: Using the trained timing structure model, we applied the method in Section 3 to the audio signal. To verify the ability of the timing structure model, we input the audio interval sequence that we used in the parameter estimation. First, we generated a visual interval sequence from the input audio interval sequence. Figure 6 (the third row) shows the generated visual interval sequence. We see that the sequence is almost the same as the training data shown in the second row.

Then, we generated visual feature vector sequences using the parameters of modes (linear dynamical systems) estimated in the segmentation process. Finally, we obtained an image sequence by calculating linear combination of principal axes (eigenvectors of PCA). The result of frame 140 to 250 was shown in the fifth row in Fig. 6. The lip motion in the sequence almost corresponds to the original motion (in the sixth row), and we also see the visual motion precedes the actual sound by comparing to the wave data (in the bottom row).

5 Conclusion

We present a timing structure model that explicitly represents temporal metric relations in a multimedia signal. The experiment shows that the model can be applied to generate lip motion from speech signal across the modalities. We also applied the method to generate the silhouette motion of piano performance from audio signal. Although the current results is in the stage of the verification of the model, its basic ability for representing temporal synchronization is expected to be useful for wide variety of human machine interaction systems including speaker tracking and audio-visual speech recognition. Moreover, the model provide general framework to integrate variety of signals such as motion

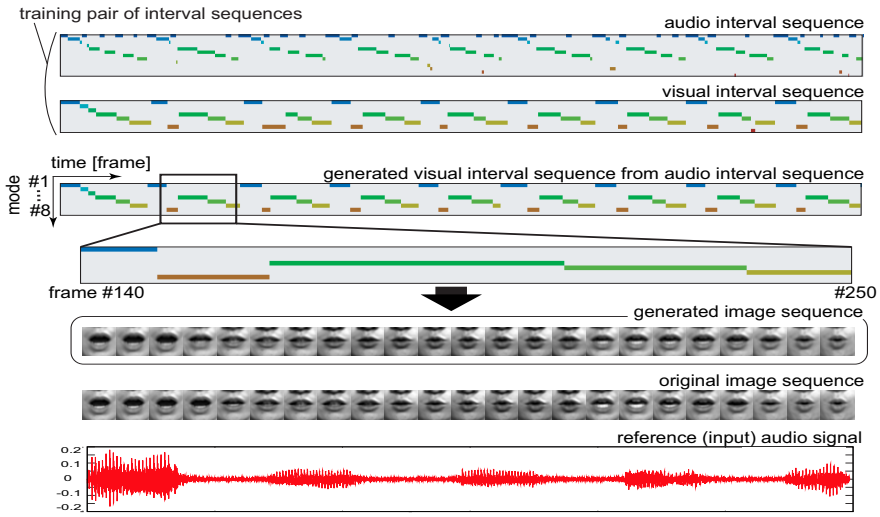


Fig. 6. Generated visual interval sequence and an image sequence from the audio signal.

in each part of facial deformation [12]. Our future work is to extend the current framework to realize interaction systems that share a sense of time with human.

Acknowledgment: This work is in part supported by Grant-in-Aid for Scientific Research of the Ministry of Education, Culture, Sports, Science and Technology of Japan under the contract of 18049046.

References

1. J. F. Allen. Maintaining knowledge about temporal interval. *Commun. of the ACM*, 26(11):832–843, 1983.
2. M. Brand. Voice puppetry. *Proc. SIGGRAPH*, pages 21–28, 1999.
3. M. Brand, N. Oliver, and A. Pentland. Coupled hidden Markov models for complex action recognition. *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 994–999, 1997.
4. C. Bregler. Learning and recognizing human dynamics in video sequences. *Proc. Int. Conference on Computer Vision and Pattern Recognition*, pages 568–574, 1997.
5. T. F. Cootes, G. J. Edwards, and C. J. Taylor. Active appearance model. *Proc. European Conference on Computer Vision*, 2:484–498, 1998.
6. H. Kawashima and T. Matsuyama. Multiphase learning for an interval-based hybrid dynamical system. *IEICE Trans. Fundamentals*, E88-A(11):3022–3035, 2005.
7. S. E. Levinson. Continuously variable duration hidden Markov models for automatic speech recognition. *Computer Speech and Language*, 1:29–45, 1986.
8. Y. Li, T. Wang, and H.-Y. Shum. Motion texture: A two-level statistical model for character motion synthesis. *Proc. SIGGRAPH*, pages 465–472, 2002.
9. H. McGurk and J. MacDonald. Hearing lips and seeing voices. *Nature*, pages 746–748, 1976.
10. K. P. Murphy. Hidden semi-Markov models (HSMMs). *Informal Notes*, 2002.

11. A. V. Nefian, L. Liang, X. Pi, X. Liu, and K. Murphy. Dynamic Bayesian networks for audio-visual speech recognition. *EURASIP Journal on Applied Signal Processing*, 2002(11):1–15, 2002.
12. M. Nishiyama, H. Kawashima, T. Hirayama, and T. Matsuyama. Facial expression representation based on timing structures in faces. *IEEE Int. Workshop on Analysis and Modeling of Faces and Gestures (LNCS 3723)*, pages 140–154, 2005.
13. M. Ostendorf, V. Digalakis, and O. A. Kimball. From HMMs to segment models: A unified view of stochastic modeling for speech recognition. *IEEE Trans. Speech and Audio Process*, 4(5):360–378, 1996.