Speaker Detection Using the Timing Structure of Lip Motion and Sound

Yu Horii Hiroaki Kawashima Takashi Matsuyama
Graduate School of Informatics, Kyoto University
Yoshida-Honmachi, Sakyo, Kyoto 6068501, Japan
horii@vision.kuee.kyoto-u.ac.jp {kawashima,tm}@i.kyoto-u.ac.jp

Abstract

In this paper, we propose a novel approach to speaker detection by an integration of audio-visual information using the cue of timing structure. We first extract feature sequences of lip motion and sound, and segment each of them into temporal intervals. Then, we construct a cross-media timing-structure model of human speech by learning the temporal relations of overlapping intervals. Based on the learned model, we realize speaker detection by evaluating the timing structure of the observed video and audio. Our experimental result shows the effectiveness of using temporal relations of intervals for speaker detection.

1. Introduction

In human speech communication, we recognize the other's utterance state (i.e. who is speaking at that time) using not only audio information (e.g. direction of sound and phonetic characteristics) but also visual information (e.g. face position and lip motion). Similarly, for a speech understanding system, visual information may play an important role to improve the accuracy of speaker detection. For example, some methods of merging visual and audio information have been proposed for automatic video recording of teleconferences or archiving lectures [5, 8].

Most existing methods for speaker detection are realized by combining techniques of sound localization via a microphone array and human tracking via background subtraction by using coupled Hidden Markov Models (HMMs) or Dynamic Bayesian Networks (DBNs) [11, 2]. However, because of the spatial resolution of the microphone array, these methods can become ineffective in situations where speakers are physically close to each other.

To achieve high accuracy in such situations, we can use the cue of co-occurrence patterns between lip motion and speech sound. For example, when we produce a plosive (e.g. /pa/), the starting of lip motion and that of sound are almost coherent. In contrast, in utterance of a vowel (e.g. /a/), the lip motion precedes the speech sound. That is, they are not necessarily synchronized. However, we human consider that these temporal differences are perfectly normal. In fact, it is known that some temporal variance is allowed in our speech perception [13]. Frame-wise integration methods are often used in speech recognition, however, they sometimes fail to describe such loose synchronization.

From these consideration, in this paper, we propose a novel approach to speaker detection by using a model that directly represents the specific temporal relations between lip motion and sound; We describe these relations as *timing structure*. As a related research, Nishiyama et al. applied the idea of timing structure in modeling the temporal relations among partial movements in facial image sequences, and made a success in separation of intentional smiles and spontaneous smiles [9]. In contrast with this, we deal in the timing structure of multimedia signal.

Our goal is to detect the speaker in a scene in which there are multiple persons. This technique is also applicable to content analysis of archived video data, because our method is realizable with only one camera and one microphone.

2. Problem Definition and Our Approach

The goal of this paper is to correctly detect the speaker, when either one of participants is speaking. In order to concentrate on verifying the effectiveness of using the timing structure, we consider the following situation:

- There are multiple persons in the scene.
- Only one camera and one microphone are used.
- Frontal faces of all persons are captured.
- Lip motion of all persons can occur at a time.
- Ordinary background noise is included in audio data.
- There is no overlap in utterance.

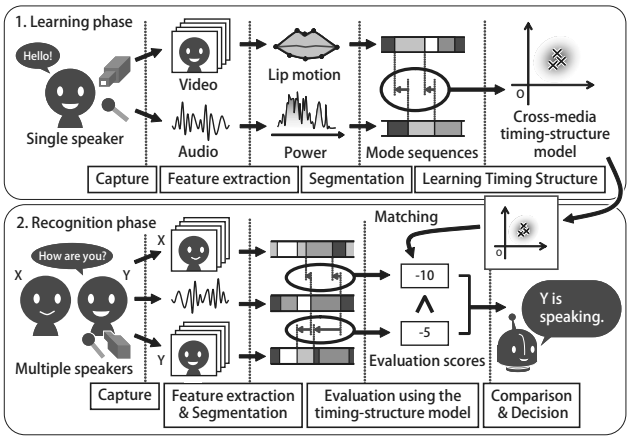

Figure 1. Flow of speaker detection using the timing structure.

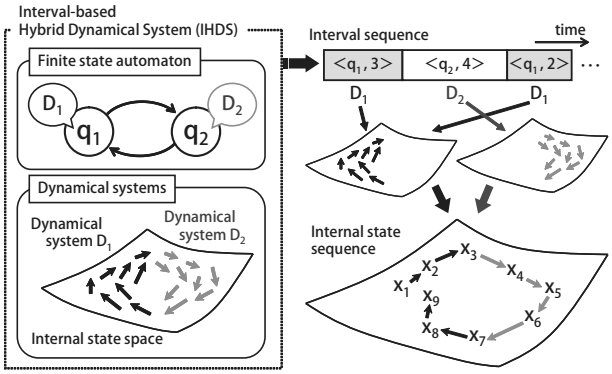

Figure 2. Interval-based Hybrid Dynamical System (IHDS).

The flow of our method is the followings (See Figure 1): This is a two-step approach which consists of the learning phase and the recognition phase.

In the learning phase, at first, we extract a feature sequence of mouth shape and that of sound power level from captured data, and segment each of them into temporal intervals that can be described in linear dynamical systems. Then, we construct a cross-media timing-structure model of human speech by learning the temporal relations of overlapping temporal intervals. This model describes the allowable range of the temporal fluctuation in human speech.

Secondly, in the recognition phase, we get interval sequences of newly-observed data in the same way. Then, based on the learned timing-structure model, we calculate scores of the timing structure of lip motion of each person and sound. Finally, by comparison between the evaluation scores, we detect the speaker.

In Section 3, we describe modeling of a single media signal. In Section 4, we propose a method for learning timing structure, and for evaluating newly-observed data. In Section 5, we evaluate our proposed method using real data.

3. Modeling a Single Media Signal

In this paper, we utilize the Interval-based Hybrid Dynamical System (IHDS), which was introduced by Kawashima et al. [6, 7], to describe single-media signals based on the structure of intervals. The IHDS consists of a discrete-event system and a dynamical system which is described by differential equations. This structure is similar to the Switching Linear Dynamical System (SLDS) [12]; however, the IHDS is more like Segment Models [10] because the two types of systems are integrated based on intervals (segments). By exploiting the operation on the intervals, the IHDS provides efficient learning techniques including the hierarchical clustering of linear dynamical systems. In addition, the IHDS is a stochastic model that can generate multivariate vector sequences with interval-based structures.

3.1. Interval-Based Hybrid Dynamical System

System architecture. The IHDS has a two-layer architecture (Figure 2). The first layer has a finite state automaton that models stochastic transitions between intervals. The second layer consists of a set of multiple linear dynamical systems, $\mathcal{D} = \{D_1, \cdots, D_N\}$. To integrate these two layers, intervals are introduced; each interval is described by $\langle q_i, \tau \rangle$ where q_i denotes a discrete state in the automaton and τ denotes the physical temporal duration length of the interval. It is assumed that each state, q_i, in the automaton corresponds to a unique linear dynamical system, D_i.

Linear dynamical system. The state transition of dynamical system D_i is modeled by the following equation:

$$\boldsymbol{x}_t \quad = \quad F^{(i)}\boldsymbol{x}_{t-1} + \boldsymbol{g}^{(i)} + \boldsymbol{\omega}_t^{(i)}, \qquad (1)$$

where $\boldsymbol{x}_t$ is the internal state vector at time t. $F^{(i)}$ is a transition matrix and $\boldsymbol{g}^{(i)}$ is a bias vector. $\boldsymbol{\omega}_t^{(i)}$ is the process noise which is modeled by a Gaussian distribution. Note that each dynamical systems has $F^{(i)}$, $\boldsymbol{g}^{(i)}$, and $\boldsymbol{\omega}_t^{(i)}$ individually, and that all the dynamical systems share a single internal state space.

Interval-based state transition. Let $\mathcal{I} = [I_1, \cdots, I_K]$ be an interval sequence generated by the automaton of the IHDS. Here, they are assumed that the first-order Markov property for the generated intervals and that the adjacent intervals have no temporal gaps or overlaps. Then, the state transition process can be modeled by the conditional probability, $P(I_k = \langle q_j, \tau \rangle | I_{k-1} = \langle q_i, \tau_p \rangle)$, where it denotes that the interval $\langle q_j, \tau \rangle$ occurs after the interval $\langle q_i, \tau_p \rangle$.

3.2. Learning and Segmentation Method

Learning method for the IHDS. The goal of the IHDS identification is to estimate the number of linear dynamical systems, N, and the parameter set of all the systems D_i. The estimation process is divided into two steps: a clustering process of dynamical systems using a typical training data set, and a refinement process for all the parameters based on the EM (Expectation-Maximization) algorithm [4] using all the training data. At the same time, we can segment all the training data into temporal interval sequences. The details of the learning algorithms are described in [6].

Segmentation of newly-observed data. We can segment newly-observed signal data using the learned IHDS. When a observed sequence is given, the IHDS finds an optimal interval sequence to describe the observed data based on a likelihood calculation.

4. Modeling the Cross-Media Timing Structure

Applying the learned IHDSs to each of video and audio signals, we obtain a set of interval sequences. In this section, we concentrate on modeling the timing structure between two media signals, S and S'.

We use the term *mode* to describe the primitive event of motion or sound observed in media signals (e.g. "opening mouth" in video), and we assume that each mode, M_i, uniquely corresponds to a linear dynamical system D_i in the IHDS. Let I_k be an interval that has mode $m_k \in \{M_1, \cdots, M_N\}$ in signal S and let b_k and e_k be its starting and ending timing points , respectively. Similarly, let $I'_{k'}$ be an interval that has mode $m'_{k'} \in \{M'_1, \cdots, M'_{N'}\}$ in the range $[b'_{k'}, e'_{k'}]$ of signal S'.

Here, we define the *timing structure* between S and S' as temporal relations of overlapping interval pairs. Especially, to describe a particular timing structure, we introduce the following distribution for every mode pair:

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

We refer to this distribution as a *temporal difference distribution* (Figure 3). For example, if the peak of the distribution appears at the origin, the two modes tend to be synchronized to each other at both starting and ending points.

4.1. Learning the Timing Structure

We consider that two interval sequences, $\mathcal{I}$ and $\mathcal{I}'$, in media signals, S and S', are given as a training data set. To estimate a temporal difference distribution of the mode pair (M_i, M'_p), we collect all pairs of overlapping intervals that have the mode pair (M_i, M'_p) from the training data. Since the training data is usually finite in real applications, we fit

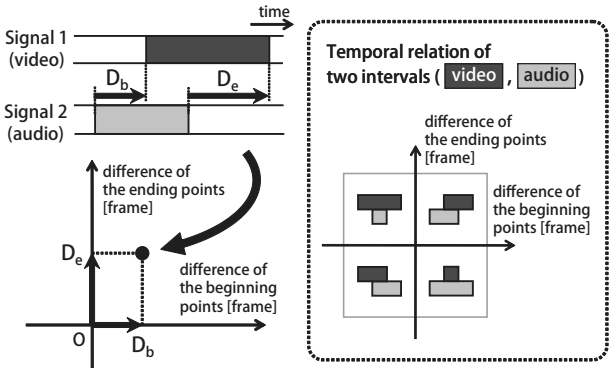

Figure 3. Expression of temporal relations of two intervals.

a density function such as Gaussian mixture models to the samples.

By computing the distributions of temporal differences for all possible mode pairs as above, we can obtain fundamental characteristics of the cross-media timing structure of a given data set. That is, we can obtain the following function F of an interval pair $(I_k, I'_{k'})$:

$$F(I_k, I'_{k'}) = P(b_k - b'_{k'} = D_\mathrm{b}, e_k - e'_{k'} = D_\mathrm{e} | \\ m_k = M_i, m'_{k'} = M'_p, [b_k, e_k] \cap [b'_{k'}, e'_{k'}] \neq \emptyset). \quad (3)$$

4.2. Evaluation of the Timing Structure

Let us assume that the temporal interval sequences $\mathcal{I} = [I_1, \cdots, I_K]$ and $\mathcal{I}' = [I'_1, \cdots, I'_{K'}]$ are obtained by segmenting newly observed media signals S and S', respectively. We label the set of the overlapping interval pairs included in $\mathcal{I}$ and $\mathcal{I}'$ as set $\mathcal{P}$.

To evaluate the timing structure between $\mathcal{I}$ and $\mathcal{I}'$, we calculate the score of set $\mathcal{P}$ based on joint probabilities as follows:

$$\hat{F}(\mathcal{P}) = \left[\prod_{(I_k, I'_{k'}) \in \mathcal{P}} F(I_k, I'_{k'}) \right]^{\frac{1}{n(\mathcal{P})}}, \quad (4)$$

where F is defined in Eq. (3). Note that the score is normalized for the size of set $\mathcal{P}$, $n(\mathcal{P})$. If all the pairs of overlapping intervals in $\mathcal{P}$ have the similar temporal differences of the starting and ending points, the score becomes close to 1. In practice, we take logarithm of $\hat{F}(\mathcal{P})$ to prevent underflow. Thus, we define the score of a media signal pair (S, S') as follows:

$$E(S, S') = \log \hat{F}(\mathcal{P}) = \frac{1}{n(\mathcal{P})} \sum_{(I_k, I'_{k'}) \in \mathcal{P}} \log F(I_k, I'_{k'}). \quad (5)$$

For example, let us consider that two signals, $S^{(\mathrm{X})}$ and $S^{(\mathrm{Y})}$, are observed in one media (e.g. lip motion of two

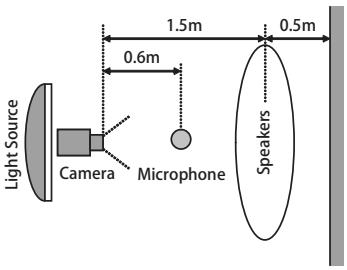

Figure 4. The layout of the recording room

Figure 5. An example of the captured video image (training data).

persons in video) and a signal S' from another media (e.g. speech sound). Comparing the scores $E(S^{(\mathrm{X})}, S')$ and $E(S^{(\mathrm{Y})}, S')$, we can estimate which one of $S^{(\mathrm{X})}$ and $S^{(\mathrm{Y})}$ has more likely timing structure between S'.

5. Experimental Evaluations

To evaluate our proposed method, we conducted two experiments. At first, we obtained the timing-structure model of human speech from real data of two persons (see the top layer of Figure 1). Note that we trained a single model by using the learning data of two speakers. Secondly, as the first experiment, we evaluated the accuracy of our proposed method using the data of speech scenes of two persons who are the same persons in the training data. As the second experiment, we used the speech data of another five persons and tried to detect the speaker based on the timing-structure model learned in the first experiment.

5.1. Learning the Timing-Structure Model

Data capture. As a training sample set, we used speech data spoken by two persons. The data recordings were made in a room with one camera and one microphone (Figure 4). There were no significant noise sources other than an air conditioner and PC fans. The video data were captured in an image of 640 $\times$ 480 pixels at 60 fps (a mouth is about 40 $\times$ 20 pixels). The sampling rate of the audio signal was 48kHz. Figure 5 shows an example of the captured image.

The total length of training data set was 17010 frames ($\simeq$ 4.7 minutes). In the data, the speakers took a turn after each utterance of one or two sentences (about 20 or 30 seconds).

Feature extraction. Our final goal is to realize person-independent speaker detection. Therefore, we need to obtain the timing-structure model for unspecified speakers. That is, we should use features that are consistent from person to person. As for the video signal, the lip shapes and the horizontal motions are highly individual. Hence, we used the frame difference of the vertical-coordinates of the bottom lip (described by 5 feature points) as the feature vector of lip motion. We used the Active Appearance Model

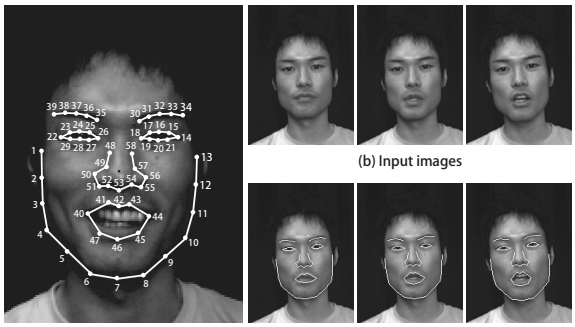

Figure 6. Examples of (a) a training image to build AAMs, (b) a captured face images, and (c) face images with tracked feature points.

(AAM) [3] to extract the feature points of the face in each image. The AAM is a statistical model that can represent both the shape and texture variability in a training set, and can be matched to a target image robustly. Then, we assumed that all the feature points were on the same plane, and normalized the translation, the rotation, and the scale of the feature shapes based on singular value decomposition [1]. Finally, we obtained 5D feature vector sequences.

As for the audio signal, we used the sound power level as the feature of sound, because power-level patterns are less affected by the difference of individuals. The time interval of analysis frame length was 33.3 ms, and the frame shift was 16.6 ms (it is equal to the video frame rate). We used HTK Ver.3.4 [14] as an extraction tool, and obtained 1D feature sequences.

Segmentation of the feature vector sequences. We trimmed away the parts without utterance in the extracted feature sequences manually. The total length was 13533 frames, that consisted of 6761 frames of one person's utterance and 6772 frames of the other's. Then, we estimated the parameter of IHDSs and segmented each signal into an interval sequence by the method described in Section 3. The estimated number of video modes was five, and that of au-

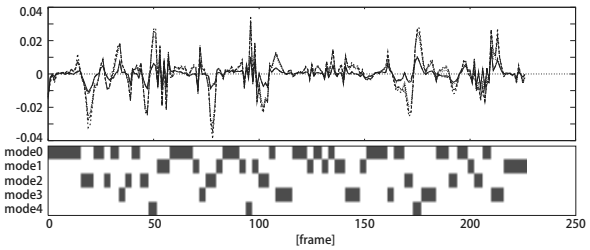

Figure 7. A result of segmentation of a video signal.

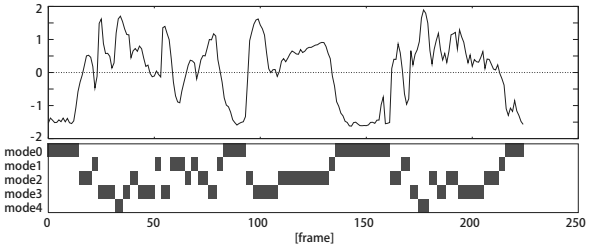

Figure 8. A result of segmentation of a audio signal.

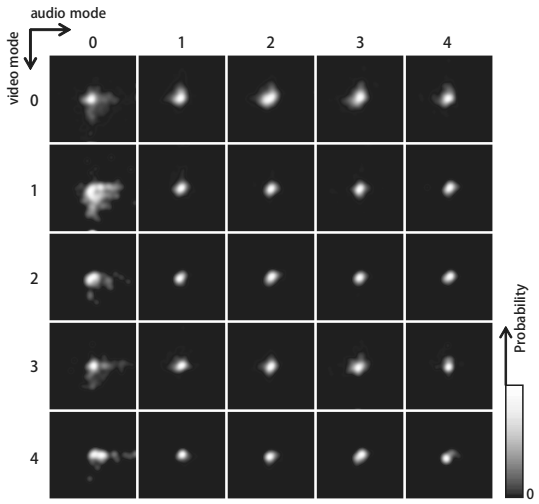

Figure 9. A table of temporal difference distributions. Horizontal (vertical) axis is the starting (ending) points in each distribution. Both of their ranges are from -50 frames to 50 frames.

dio modes was also five. Examples of the segmentation results are shown in Figure 7 and 8. Thus, the training data was converted into interval sequences that were labeled by modes which described the elemental patterns.

Learning the cross-media timing-structure model. Using the interval sequences of lip motion and sound as signal S and S', we computed the distributions of temporal difference according to the method described in Section 4.1. To obtain the whole density distributions, we convoluted sample points with a Gaussian distribution whose standard deviation was 3 frames [1]. The result is shown in Figure 9.

As an interpretation of the distributions, for example, we can find that these interval pairs tend to synchronize at the ending points in the distribution of video-mode 4 and audio-mode 0. Video-mode 4 corresponds to "opening one's mouth", and audio-mode 0 corresponds to "no sound". Therefore, this distribution indicates that the lip motion tends to precede the occurrence of speech sound in human speech.

5.2. Speaker Detection Using the Timing Structure

Using the learned timing-structure model, we evaluated our method of speaker detection. We newly captured speech scenes of two persons, X and Y (the same persons in the training data), and segmented them into interval sequences by the method described in Section 3.2. We obtained 12 sequences (each had about 2000 frames, the total length was 23831 frames). Note that this test data was recorded in the same situation with the training data.

Here, let $S^{(\mathrm{v,X})}$ ($S^{(\mathrm{v,Y})}$) be video signal of person X (Y), and $S^{(\mathrm{a})}$ be audio signal. In addition, we define 3 terms to describe the states of lip motion as the followings:

- *Utterance* : Lip motion of the actual speaker.
- *Silence* : No lip motion.
- *Fake lip motion* : Lip motion not related to sound. (e.g. a change in facial expression or whispering.)

5.2.1 Evaluation Method

Let $E_i(S^{(\mathrm{v})}, S^{(\mathrm{a})})$ be an evaluation function of the timing structure between the signals, $S^{(\mathrm{v})}$ and $S^{(\mathrm{a})}$. We estimate the correct speaker based on a comparison of the scores. That is, if $E_i(S^{(\mathrm{v,X})}, S^{(\mathrm{a})}) > E_i(S^{(\mathrm{v,Y})}, S^{(\mathrm{a})})$, we decide that the speaker is X.

To compare with other methods, we define the following three functions E_1, E_2, and E_3 (see also Figure 10). Note that E_3 is the proposed method, and we calculate the score in a time window with T frames.

(a) Mode pair co-occurrence in the same frame. In the learning phase, we obtain the probability $P(m_t^{(\mathrm{v})} = v, m_t^{(\mathrm{a})} = a)$ from the training data set, where (v, a) denotes the mode pair of video and audio signal at a frame index t. We assume that $P(m_t^{(\mathrm{v})} = v, m_t^{(\mathrm{a})} = a)$ is constant with time t. In the recognition phase, we use evaluation function E_1

[1] According to the research on the over all timing tolerance between video and audio by A. Peregudov et al. [13], the thresholds of acceptability are about $+90$ ms (sound leading) to -185 ms (sound delayed). For these reason, we decided the standard deviation of the Gaussian, σ, as $2\sigma \simeq 100$ ms (the frame rate was 60 fps, therefore $\sigma = 3$ frames $\simeq 50$ ms).

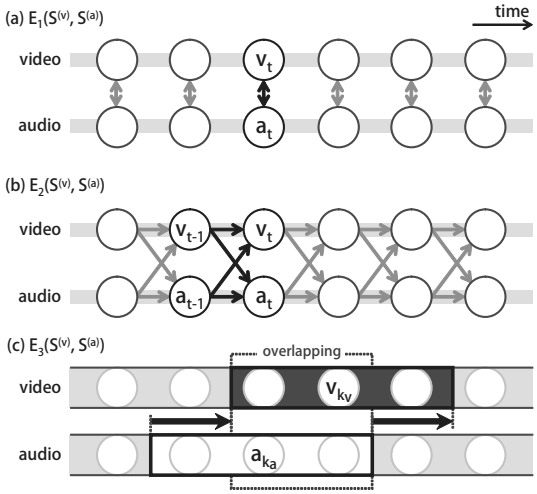

Figure 10. Temporal relations used in each of the evaluation functions. (a) Mode pair co-occurrence in the same frame. (b) Mode transition probability between the adjacent frames. (c) Temporal difference of overlapping interval pairs (proposed method).

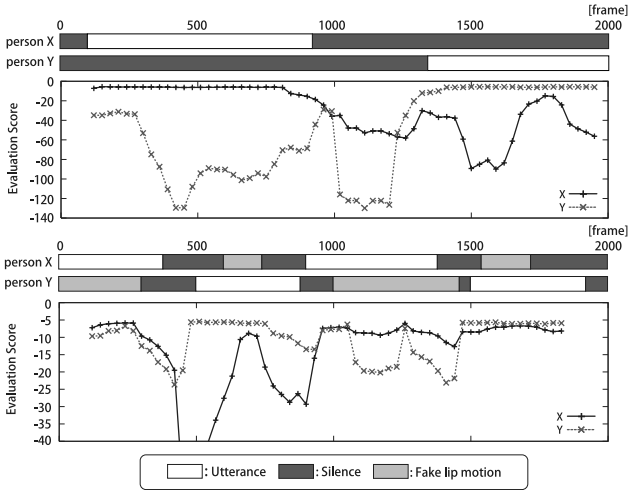

Figure 11. Examples of temporal change in the evaluation scores. The top part of each graph shows the state of lip motion. The upper graph is the case that no fake lip motion occurs, and the lower one is the case that fake lip motion is included in most frames.

defined by the following equation:

$$E_1(S^{(\mathrm{v})}, S^{(\mathrm{a})}) = \frac{1}{T}\sum_{t=0}^{T-1} \log P\Big(m_t^{(\mathrm{v})} = v_t, m_t^{(\mathrm{a})} = a_t\Big). \quad (6)$$

(b) Mode transition probability between the adjacent frames. Evaluation function E_2 uses the mode transition probability between the adjacent frames in a similar manner as Coupled HMMs. We learn the probability $P(m_t^{(\mathrm{v})} = v, m_t^{(\mathrm{a})} = a | m_{t-1}^{(\mathrm{v})} = v_{\mathrm{p}}, m_{t-1}^{(\mathrm{a})} = a_{\mathrm{p}})$ from the training data, where $(v_{\mathrm{p}}, a_{\mathrm{p}})$ is the mode pair at the previous frame $t-1$. Then, in the recognition phase, we use function E_2 defined by the following equation:

$$E_2(S^{(\mathrm{v})}, S^{(\mathrm{a})}) = \frac{1}{T-1}\sum_{t=1}^{T-1} \log P\Big(m_t^{(\mathrm{v})} = v_t, m_t^{(\mathrm{a})} = a_t \Big| m_{t-1}^{(\mathrm{v})} = v_{t-1}, m_{t-1}^{(\mathrm{a})} = a_{t-1}\Big). \quad (7)$$

(c) Temporal difference distribution of overlapping interval pairs (proposed method). Using the function shown in Eq. (5), we again define E_3 by follows:

$$\mathcal{P} = \Big\{(I_{k_{\mathrm{v}}}^{(\mathrm{v})}, I_{k_{\mathrm{a}}}^{(\mathrm{a})}) \Big| I_{k_{\mathrm{v}}}^{(\mathrm{v})} \in \mathcal{I}^{(\mathrm{v})}, I_{k_{\mathrm{a}}}^{(\mathrm{a})} \in \mathcal{I}^{(\mathrm{a})}, [b_{k_{\mathrm{v}}}^{(\mathrm{v})}, e_{k_{\mathrm{v}}}^{(\mathrm{v})}] \cap [b_{k_{\mathrm{a}}}^{(\mathrm{a})}, e_{k_{\mathrm{a}}}^{(\mathrm{a})}] \neq \emptyset\Big\}, \quad (8)$$

$$E_3(S^{(\mathrm{v})}, S^{(\mathrm{a})}) = \frac{1}{n(\mathcal{P})} \sum_{(I_{k_{\mathrm{v}}}^{(\mathrm{v})}, I_{k_{\mathrm{a}}}^{(\mathrm{a})}) \in \mathcal{P}} \log F(I_{k_{\mathrm{v}}}^{(\mathrm{v})}, I_{k_{\mathrm{a}}}^{(\mathrm{a})}), \quad (9)$$

where $F(I_{k_{\mathrm{v}}}^{(\mathrm{v})}, I_{k_{\mathrm{a}}}^{(\mathrm{a})})$ is calculated by the learned temporal difference distributions in Eq. (3).

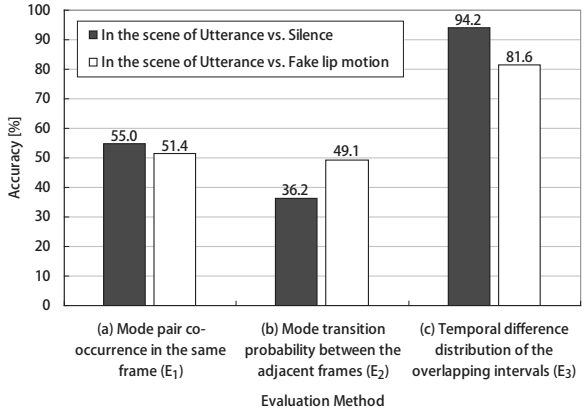

Figure 12. The accuracy by using each evaluation function on the same persons in the learning data. The black bar depicts R_{sil}, and the white bar does R_{fak}.

5.2.2 Experimental Results

At first, we calculated the evaluation scores of each person by shifting a time window. We set the window size T 180 frames (3 seconds), and the window shift 30 frames. Figure 11 shows the examples of temporal change in the scores evaluated by function E_3. We can observe that the score of the speaker was larger than that of the other in most parts. When the non-speaker was in silence, the difference between the evaluation scores was particularly large (the upper graph in Figure 11).

To evaluate the accuracy of the speaker detection, we obtained the ratio of the correctly detected time windows to the

Table 1. The accuracy by using each evaluation function (see Section 5.2.1). The figure in parenthesis shows the number of data points. In the evaluated data, the same persons in the learning data (X and Y) were captured. R_{sil} (R_{fak}) = the accuracy in the scene that the non-speaker's state was silence (fake lip motion). R_{X} (R_{Y}) = the accuracy in the scene that the speaker was person X (Y).

Evaluation Function	R_{sil} [%]		R_{fak} [%]		R_{X} [%]		R_{Y} [%]		R [%]	
E_1	55.0	(170)	51.4	(179)	86.1	(309)	13.4	(40)	53.1	(349)
E_2	36.2	(112)	49.1	(171)	77.7	(279)	1.3	(4)	43.1	(283)
E_3 (proposed method)	94.2	(291)	81.6	(284)	82.2	(295)	94.0	(280)	87.5	(575)

total windows in which either person was speaking. Here, we describe the accuracy as R. Especially, R_{sil} (R_{fak}) denotes the accuracy in the scene that the non-speaker's state was silence (fake lip motion). R_{X} (R_{Y}) denotes the accuracy in the scene that the speaker was person X (Y). The calculated results of R_{sil} and R_{fak} are shown in Figure 12, and $R_{\text{sil}}, R_{\text{fak}}, R_{\text{X}}, R_{\text{Y}}$, and R are shown in Table 1.

In Figure 12, we can see that the accuracy of the proposed method (E_3) is higher than those of the other methods, regardless of whether the non-speaker was in silent or with fake lip motion. However, R_{sil} is over 10 points larger than R_{fak}. Furthermore, the average of the score differences was 34.4 in the scene that the non-speaker was in silent. In contrast, when the non-speaker was with fake lip motion, that was 2.2. The reason of such a sharp contrast is that the interval lengths in silence tend to be long, and the temporal relations are far from the peak of the learned distributions.

From the accuracy of E_1 and E_2 in Table 1, we see that R_{Y} was significantly smaller than R_{X}, and the total accuracy was about 50%. On the other hand, we can correctly detect both persons using E_3.

5.3. Evaluation on the Generalization Capability

To evaluate whether the learned timing-structure model can apply to unspecified persons, we additionally captured speech data of another five persons (person 1, $\cdots$, person 5). The recording situation was the same in Figure 4. The number of the captured sequences was eight, and the total length was 12837 frames. In the most parts of the sequences, at least one person was in measurable fake lip motion. An example of the image is shown in Figure 13.

Then, we evaluated these data using the timing-structure model obtained in the preceding experiment (learned from the data of person X and Y). The accuracy by using each evaluation function is shown in Table 2. Here, R_i denotes the accuracy in the scene that the speaker was person i ($i = 1, \cdots, 5$), and R denotes the total accuracy.

In Table 2, we can find that our proposed method (E_3) detected the speaker more correctly than the other evaluation functions (E_1, E_2). In the result by using E_1 and E_2, there are significant differences between the accuracies of each person, and thus the total accuracy was very low.

In the result of the proposed method, the accuracies of person 2 and 5 were both nearly 80 % ($\simeq R_{\text{fak}}$ in the preceding experiment), however, the others were lower (around 50 or 60 %). This result shows that the difference of speech style may have affected the detection accuracy. For example, the average of speech rates of person 2 and 5 were faster than those of other persons, and were close to that of the training data.

Figure 13. An example of the captured video image for evaluation on the generalization capability (the different five persons from the ones in the training data).

6. Discussion and Conclusions

We proposed the method of speaker detection by evaluating observed data based on the timing-structure model, which directly describes the co-occurrence and specific timing differences between lip motion and sound. Although the current results are still preliminary, we see that the proposed method detects a speaker with higher accuracy than frame-wised methods.

We used simple features and a straightforward estimation method in the experiments to concentrate on evaluating the effectiveness of the timing structure. The accuracy therefore would be improved by considering additional factors, for example, temporal constraints among time windows and the difference of speech style (e.g. speech rate, habit of lip motion).

For practical applications, we however have to investigate more natural and diverse situations where all persons are in silent, or the cases including temporary lip occlusion and overlapping utterance. In order to obtain better generalization ability, it will also be worth seeking the extension of the model learning; for example, the use of larger train-

Table 2. The accuracy by using each evaluation function (see Section 5.2.1). The figure in parenthesis shows the number of data points. In the evaluated data, the new five persons (person 1, $\cdots$, 5) were captured. R_i = the accuracy in the scene that the speaker was person i.

Evaluation Function	R_1 [%]		R_2 [%]		R_3 [%]		R_4 [%]		R_5 [%]		R [%]	
E_1	61.8	(21)	6.3	(4)	53.1	(34)	6.0	(6)	3.5	(2)	21.0	(67)
E_2	47.1	(16)	0.0	(0)	32.8	(21)	0.0	(0)	3.5	(2)	12.2	(39)
E_3 (proposed method)	51.5	(17)	78.1	(50)	67.2	(43)	64.0	(64)	84.2	(48)	69.8	(222)

ing data with respect to the number of persons and speech styles would be necessary for practical implementation.

Acknowledgement

This study is supported by Grant-in-Aid for Scientific Research No.18049046 of the Ministry of Education, Culture, Sports, Science and Technology.

References

[1] K. S. Arun, T. S. Huang, and S. D. Blostein. Least-squares fitting of two 3-d point sets. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 9(5):698–700, 1987. 4

[2] F. Asano, K. Yamamoto, I. Hara, J. Ogata, T. Yoshimura, Y. Motomura, N. Ichimura, and H. Asoh. Detection and separation of speech event using audio and video information fusion and its application to robust speech interface. *EURASIP Journal on Applied Signal Processing*, 2004(11):1727–1738, 2004. 1

[3] T. F. Cootes, G. J. Edwards, and C. J. Taylor. Active appearance model. *Proc. European Conference on Computer Vision*, pages 484–498, 1998. 4

[4] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *J. R. Statist. Soc. B*, 39:1–38, 1977. 3

[5] D. Gatica-Perez, G. Lathoud, I. McCowan, J. Odobez, and D. Moore. Audio-visual speaker tracking with importance particle filters. In *IEEE International Conference on Image Processing (ICIP)*, 2003. 1

[6] H. Kawashima and T. Matsuyama. Hierarchical clustering of dynamical systems based on eigenvalue constraints. *3rd International Conference on Advances in Pattern Recognition (S. Singh et al. (Eds.): ICAPR 2005, LNCS 3686)*, pages 229–238, 2005. 2, 3

[7] H. Kawashima and T. Matsuyama. Interval-based linear hybrid dynamical system for modeling cross-media timing structures in multimedia signals. *International Conference on Image Analysis and Processing*, pages 789–794, 2007. 2

[8] S. Nishiguchi, Y.Kameda, K. Kakusho, and M. Minoh. Automatic video recording of lecture's audience with activity analysis and equalization of scale for students observation. *Journal of Advanced Computational Intelligence and Intelligent Informatics*, 8(2):181–189, 2004. 1

[9] M. Nishiyama, H. Kawashima, T. Hirayama, and T. Matsuyama. Facial expression representation based on timing structures in faces. *IEEE International Workshop on Analysis and Modeling of Faces and Gestures (W. Zhao et al. (Eds.): AMFG 2005, LNCS 3723)*, pages 140–154, 2005. 1

[10] 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. 2

[11] V. Pavlović, A. Garg, J. Rehg, and T. Huang. Multimodal speaker detection using error feedback dynamic Bayesian networks. *Proc. Computer Vision and Pattern Recognition*, pages 34–43, 2000. 1

[12] V. Pavlović, J. M. Rehg, and J. MacCormick. Learning switching linear models of human motion. *Proc. Neural Information Processing Systems*, 2000. 2

[13] A. Peregudov, K. Glasman, and A. Logunov. Relative timing of sound and vision: evaluation and correction. *Proceedings of the Ninth International Symposium on Consumer Electronics*, pages 198–202, 2005. 1, 5

[14] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland. *The HTK Book (for HTK Version 3.4)*. Cambridge University Engineering Department, 2006. 4