International Journal of Semantic Computing
Vol. 6, No. 2 (2012) 179–204
©World Scientific Publishing Company
DOI: 10.1142/S1793351X12400077

SPEECH SHOT EXTRACTION FROM BROADCAST NEWS VIDEOS

SHOGO KUMAGAI[*,¶,∥], KEISUKE DOMAN[*,§,**],
TOMOKAZU TAKAHASHI[†,††], DAISUKE DEGUCHI[‡,‡‡],
ICHIRO IDE[*,§§] and HIROSHI MURASE[*,¶¶]

[*] *Graduate School of Information Science, Nagoya University*
Furo-cho, Chikusa-ku, Nagoya, Aichi, 464-8601 Japan

[†] *Faculty of Economics and Information*
Gifu Shotoku Gakuen University
1-38 Nakauzura, Gifu, 500-8288 Japan

[‡] *Information and Communications Headquarters*
Nagoya University, Furo-cho, Chikusa-ku, Nagoya
Aichi, 464-8601 Japan

[§] *Japan Society for the Promotion of Science (JSPS), Japan*

[¶] *Currently at Ricoh Company, Ltd., Japan*
[∥] *skumagai@murase.m.is.nagoya-u.ac.jp*
[**] *kdoman@murase.m.is.nagoya-u.ac.jp*
[††] *ttakahashi@gifu.shotoku.ac.jp*
[‡‡] *ddeguchi@nagoya-u.jp*
[§§] *ide@is.nagoya-u.ac.jp*
[¶¶] *murase@is.nagoya-u.ac.jp*

We propose a method for discriminating between a speech shot and a narrated shot to extract genuine speech shots from a broadcast news video. Speech shots in news videos contain a wealth of multimedia information of the speaker, and could thus be considered valuable as archived material. In order to extract speech shots from news videos, there is an approach that uses the position and size of a face region. However, it is difficult to extract them with only such an approach, since news videos contain non-speech shots where the speaker is not the subject that appears in the screen, namely, narrated shots. To solve this problem, we propose a method to discriminate between a speech shot and a narrated shot in two stages. The first stage of the proposed method directly evaluates the inconsistency between a subject and a speaker based on the co-occurrence between lip motion and voice. The second stage of the proposed method evaluates based on the intra- and inter-shot features that focus on the tendency of speech shots. With the combination of both stages, the proposed method accurately discriminates between a speech shot and a narrated shot. In the experiments, the overall accuracy of speech shots extraction by the proposed method was 0.871. Therefore, we confirmed the effectiveness of the proposed method.

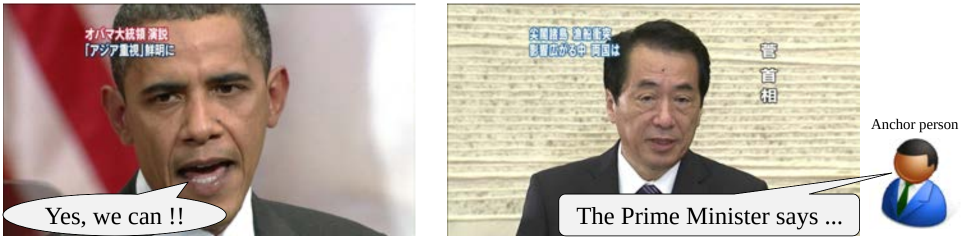
Keywords: Speech shot extraction; audio-visual integration; broadcast news videos.

1. Introduction

Recently, there is a demand for the efficient reuse of massively archived broadcast videos which consist of various genres of programs such as news, sports, dramas and so on. Especially, news videos are valuable as an archived material since they cover a wide range of real-world events that are closely related to our social lives. Accordingly, there are many researches focusing on the analysis and retrieval of broadcast news videos. Among them, there are works that focus on people that appear in news, since they attract much public attention. For example, Satoh *et al.* proposed a method for associating names and faces in news videos [1]. Ozkan and Duygulu proposed a method for extracting facial images from news videos with the name of a person [2]. Ide *et al.* proposed a method for extracting human relationships from news videos [3]. In this paper, we focus on the extraction of speech shots such as interviews, press conferences, and public speakings, from news videos. Speech shots provide a wealth of multimedia information to us, since they contain facial expressions, moods, and voice tones that are difficult to express only by text.

There is a high demand for the extraction of speech shots from news videos. Extraction of speech shots was a task in TRECVID 2002−2003 as the "news subject's monologue task" [4]. It can be used to create speech collections and summarized videos focusing on speech.

In general speech shots, as shown in Fig. 1(a), the face region of a subject appears in the center of a closeup image. Straightforwardly, the position and the size of the face region are useful for the extraction of such shots. However, as shown in Fig. 1(b), there are non-speech shots where the speaker is not the subject, namely, narrated shots. In such shots, not the subject's voice but the anchor person's voice is present in the audio. Therefore, to extract genuine speech shots from news videos, first we obtain candidate shots (hereafter called "*face shots*") by using information about the position and the size of the face region. Then, we eliminate the narrated shots from the face shots by discriminating between speech shots and narrated shots. By this way, we can obtain genuine speech shots in news videos.



(a) Speech shot (Subject = speaker)

(b) Narrated shot (Subject ≠ speaker)

Fig. 1. Examples of face shots in broadcast news videos.

Our task, speech shot extraction, is similar in some aspects to the following tasks that are different from each other.

(a) Speaker recognition: Discriminating the speaker among several subjects (active speaker detection) [5]
(b) Speaker diarization: Segment an audio stream into speaker homogeneous segments [6]
(c) Speech recognition: Recognizing what the speaker says from the voice [7]
(d) Lip reading: Recognizing what the speaker says from the lip motion [8]
(e) Lip synchronization: Synchronizing the speaker's lip motion with the voice [9].

Although our task is especially similar to task (a), it is different from discriminating the subject and an anchor person not present in the scene. That is, in task (a), it is assumed that the voice is produced by (at least) one of the subjects. On the other hand, in our task, it is assumed that the voice is produced by either the subject or the anchor person. As a matter of fact, a method for task (a) could be applied for our task by means of rejecting recognition results with low confidence. However, since it requires a reference dictionary for each subject, we consider that it is not appropriate to apply it to our task.

The basic approach for our task is, as proposed in [10, 11], to evaluate the inconsistency between the subject's lip motion and the speaker's voice. One method [10] has been proposed by Rúa *et al.* mainly for biometric identification. On the other hand, we have proposed a method in [11] for speech shot extraction in the previous work. The main applications of both methods [10, 11] are different, but their strategies are basically the same. First, several kinds of audio-visual features from the subject's lip motion and the speaker's voice are extracted. Then, the correlations between these features are evaluated. This approach can be useful in case of low level audio noise (e.g. outdoor ambient noise) and/or visual noise (e.g. face rotation, occlusion of the lip region, and various changes of lighting). However, for example, in an outdoor interview scene, the subject's face may be captured from a side view, and the voice may be recorded with high-level ambient audio noise. In such case, the methods in [10, 11] may not work well, since it is difficult to extract the audio-visual features accurately.

Another approach is to use the tendency of speech shots unique to news videos. News videos are intended to broadcast information to the viewers clearly. Especially, speech shots in broadcast news videos are captured and edited in order to lead the viewers' focus on the subject's behavior in the shot. In contrast, that is not necessarily the case in a narrated shot, since the audio in a narrated shot is replaced by an anchor person's voice. In addition, speech shots may often contain some level of audio noise such as outdoor ambient audio noise, whereas narrated shots may often contain less audio noise, because of the difference of the environment where the voices were recorded. Therefore, there is a tendency of speech shots which can be used to probabilistically (but not necessarily absolutely) discriminate between speech shots and narrated shots. Note that the tendency is not considered in previous works [10, 11].

The tendency-based approach and the inconsistency-based approach are complementary. Accordingly, we can expect to improve the accuracy of speech shot extraction with a combination of these two approaches. Focusing on this point, in this paper, we propose a two-stage framework for extracting genuine speech shots from broadcast news videos. The first stage directly evaluates the inconsistency between a subject and a speaker based on the co-occurrence between the subject's lip motion and the speaker's voice. The second stage discriminates based on the intra- and inter-shot features focusing on the tendency of speech shots. The contribution of this paper is the improvement of the accuracy of speech shot extraction by the two-stage framework.

The paper is organized as follows. Section 2 describes the proposed method to discriminate between a speech shot and a narrated shot. Section 3 reports and discusses the results of experiments to evaluate the effectiveness of the proposed method. Section 4 concludes the paper with our future work.

2. Discrimination Between a Speech Shot and a Narrated Shot

The framework of the proposed method is a two-stage cascade structure as shown in Fig. 2. The input of the proposed method is a face shot. The first stage discriminates between a speech shot and a narrated shot based on the co-occurrence between a subject's lip motion and a speaker's voice. Here, it is highly unlikely that the co-occurrence is detected by chance, even if the input shot contains sufficient audio or visual noise. Therefore, if the co-occurrence is detected, the discrimination result of the first stage would be reliable. If not, however, it requires further inspection since the co-occurrence may simply be hidden within audio or visual noise. Focusing on this point, the proposed method adds the second stage based on the tendency of speech shots. The second stage rather takes advantage of audio and visual noise, and re-evaluates the shot judged as a narrated shot by the first stage. The proposed method finally judges an input shot as a speech shot if the input shot is judged as a speech shot in either stage. Otherwise, the proposed method judges the input shot as a narrated shot. By applying the proposed method to each face shot in a broadcast

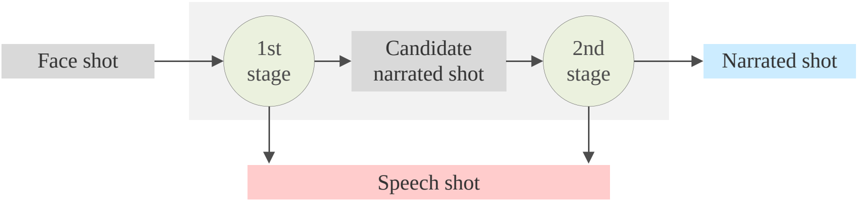


Fig. 2. The framework of the proposed method.

news video, we expect the proposed method to accurately extract genuine speech shots. Note that the proposed method assumes that face shots can be accurately obtained by existing detection and tracking techniques [12–19]. Incidentally, the first stage may not work well for face shots due to rotation or occlusion of the faces, since it may be difficult to accurately extract the lip region in such face shots. However, it is not fatal, since the second stage will try to discriminate these shots in a different approach. The details of the first stage and the second stage are described in the following sections.

2.1. *The first stage: Discrimination based on the co-occurrence between lip motion and voice*

The process flow of the first stage of the proposed method is shown in Fig. 3. First, several kinds of audio-visual features are extracted from an input shot. Next, NCCs (Normalized Correlation Coefficients) for each combination of a visual feature and an audio feature are calculated, as a representation of the co-occurrence of a subject's lip motion and a speaker's voice. Finally, based on the NCCs, a classifier constructed in advance discriminates the input shot. The details of the extraction of the audio-visual features, the calculation of NCCs, and the discrimination between a speech shot and a narrated shot are described below.

2.1.1. *Extraction of audio-visual features*

The process flow of the extraction of audio-visual features is shown in Fig. 4. First, a face shot is separated into the video stream and the audio stream. And then, visual features and audio features are extracted from each stream. The visual features represent the lip motion of a subject, whereas the audio features represent the voice of a speaker. In this paper, for each n-th input frame, visual features are denoted as

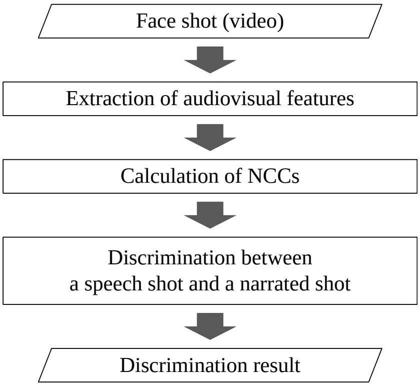

Face shot (video)
Extraction of audiovisual features
Calculation of NCCs
Discrimination between a speech shot and a narrated shot
Discrimination result

Fig. 3. Flow of the first stage of the proposed method.

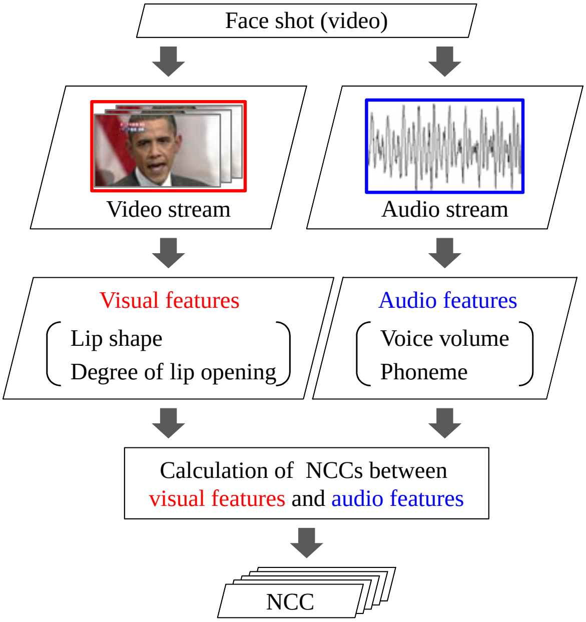


Fig. 4. Calculation of NCCs (Normalized Correlation Coefficients).

$v_i(n)$ $(i = 1, \ldots, 4)$, and audio features are denoted as $a_j(n)$ $(j = 1, \ldots, 26)$. The details of the visual features and the audio features are as follows.

Visual features $\boldsymbol{v_i(n)}$ $\boldsymbol{(i = 1, \ldots, 4)}$: A lip shape and the degree of a lip opening will differ according to the phoneme type. For example, as shown in Fig. 5, the lip shape for /a/ extends longitudinally, whereas the lip shape for /i/ extends transversally. Although it may be not necessarily so depending on language, a lip motion of a subject highly relates to his/her utterance. Focusing on this point, we extract visual features based on lip motions. Concretely, for each input frame, we extract visual features $v_i(n)$ $(i = 1, \ldots, 4)$ defined as follows.

- Lip shape: aspect ratio of lip region $v_1(n)$ and its time-derivative $v_2(n)$
- Degree of lip opening: area of lip region $v_3(n)$ and its time-derivative $v_4(n)$

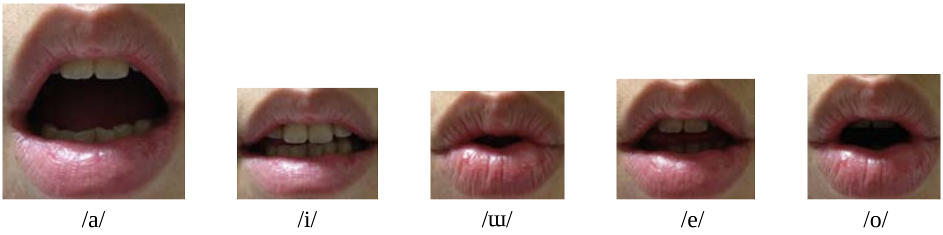


Fig. 5. Example of lip shapes for utterances (Japanese vowels represented in English phonetic symbols).

We expect that these visual features are useful for representing the lip motion of a subject, since they are used for works on lip reading and speech recognition [20, 21]. After extracting these features for all input frames, we compose visual feature vectors $\boldsymbol{v}_i$ $(i = 1, \ldots, 4)$ defined by

$$\boldsymbol{v}_i = (v_i(1), \ldots, v_i(N))^T, \tag{1}$$

where N is the number of frames in an input shot.

As for the extraction of a lip region, many methods have already been proposed. For example, there are methods which use ASM (Active Shape Model) and Snakes proposed by Jang [18], which use AAM (Active Appearance Model) proposed by Matthews *et al.* [15], and so on [16, 19]. These methods may also be applied to the extraction of lip regions from face shots in news videos.

Audio features $\boldsymbol{a_j(n)}$ $\boldsymbol{(j = 1, \ldots, 26)}$: A speaker's utterance relates to his/her lip motion. Focusing on this point, we extract audio features based on a speaker's utterance. For each input audio stream, we extract audio features $a_j(n)$ $(j = 1, \ldots, 26)$ defined as follows.

- Voice volume: audio energy $a_1(n)$ and its time-derivative $a_2(n)$
- Phoneme: 12-dimensional MFCCs (Mel-Frequency Cepstral Coefficients) $a_j(n)$ $(j = 3, \ldots, 14)$ and their time-derivatives $a_j(n)$ $(j = 15, \ldots, 26)$

Here, as shown in Fig. 6, we use the range of the audio observed from time t_n to time t_{n+1} to extract the n-th audio features. The voice volume represents the voice activity, whereas the MFCCs represent the spectrum envelope of an audio wave corresponding to the produced phoneme. We expect that these audio features are useful for representing the voice of a speaker, since they are used for speech processing works such as voice activity detection [22] and speech recognition [23]. In a similar manner with the visual features, we compose audio feature vectors

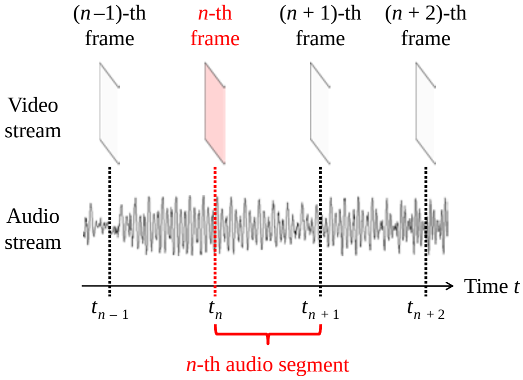


Fig. 6. The range of an audio segment corresponding to an input frame.

$\boldsymbol{a}_j$ $(j = 1, \ldots, 26)$ defined as

$$\boldsymbol{a}_j = (a_j(1), \ldots, a_j(N))^T. \tag{2}$$

2.1.2. *Calculation of normalized cross correlations*

After extracting visual feature vectors $\boldsymbol{v}_i$ $(i = 1, \ldots, 4)$ and audio feature vectors $\boldsymbol{a}_j$ $(j = 1, \ldots, 26)$, we calculate NCCs (Normalized Cross Correlations) by Eq. (3) for each combination of $\boldsymbol{v}_i$ and $\boldsymbol{a}_j$.

$$c_{i,j} = \frac{\sum_{n=1}^{N}(v_i(n) - \bar{v}_i)(a_j(n) - \bar{a}_j)}{\sqrt{\sum_{n=1}^{N}(v_i(n) - \bar{v}_i)^2}\sqrt{\sum_{n=1}^{N}(a_j(n) - \bar{a}_j)^2}}, \tag{3}$$

where

$$\bar{v}_i = \frac{1}{N}\sum_{n=1}^{N} v_i(n), \tag{4}$$

$$\bar{a}_j = \frac{1}{N}\sum_{n=1}^{N} a_j(n). \tag{5}$$

Then, using all of the NCCs $c_{i,j}$, we compose a 104-dimensional vector $\boldsymbol{c}$ defined as

$$\boldsymbol{c} = (c_{1,1}, c_{1,2}, \ldots, c_{4,25}, c_{4,26})^T. \tag{6}$$

The NCC vector $\boldsymbol{c}$ is a feature vector calculated by integrating the visual and audio features for each input shot, and represents the co-occurrence of the lip motion and the voice.

2.1.3. *Discrimination between a speech shot and a narrated shot*

The first stage discriminates between a speech shot and a narrated shot based on the NCC vector $\boldsymbol{c}$. Here, a classifier based on a SVM (Support Vector Machine) introduced by Vapnik [24] is used to discriminate an input shot. The SVM is applied in many pattern recognition applications. In SVM, a separating hyperplane is determined based on the margin maximization strategy, which enhances the generalization capability of the classification function. In addition, with the Kernel trick [25], SVM realizes a nonlinear classification with low computational cost.

In the proposed method, the classification function to discriminate an input NCC vector $\boldsymbol{c}$ is defined as

$$g(\boldsymbol{c}) = \sum_{i=1}^{l} \alpha_i y_i K(\boldsymbol{c}, \boldsymbol{c}_i) + b, \tag{7}$$

where $K(\boldsymbol{c}, \boldsymbol{c}_i)$ is a kernel function. The parameters α_i and b are trained with training NCC vectors $\boldsymbol{c}_i$ $(i = 1, \ldots, l)$ with labels y_i $(i = 1, \ldots, l)$. Here, $y_i = +1$ if the i-th training sample is a speech shot, otherwise $y_i = -1$. The parameter α_i is computed by maximizing the following quadratic problem

$$\sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{l} \alpha_i \alpha_j y_i y_j K(\boldsymbol{c}_i, \boldsymbol{c}_j) \tag{8}$$

under $\alpha_i \geq 0$ $(i = 1, \ldots, l)$, $\sum_{i=1}^{l} \alpha_i y_i = 0$. A training vector with $\alpha_i \neq 0$ is the so-called support vector. Support vectors determine the separating hyperplane, and are used to compute the parameter b.

A NCC vector $\boldsymbol{c}$ is evaluated by the trained SVM-based classifier, and discriminated by the following discrimination rule

$$f(\boldsymbol{c}) = \mathrm{sign}(g(\boldsymbol{c})), \tag{9}$$

where, $f(\boldsymbol{c}) \in \{-1, +1\}$. If $f(\boldsymbol{c}) = +1$, then the classification result is a speech shot, otherwise a narrated shot.

2.2. *The second stage: Discrimination based on the tendency of speech shots*

The process flow of the second stage of the proposed method is shown in Fig. 7. The second stage evaluates the intra- and inter-shot features based on the tendency of speech shots. First, the following feature vector $\boldsymbol{f}$ is extracted from an input shot and its neighbors.

$$\boldsymbol{f} = (f_{w_1}, f_{w_2}, f_{b_1}, f_{b_2}, f_{b_3}, f_{b_4}, f_{b_5}, f_{b_6})^T, \tag{10}$$

where f_{w_1} and f_{w_2} are intra-shot features, and f_{b_i} $(i = 1, \ldots, 6)$ are inter-shot features. Then, based on $\boldsymbol{f}$, the second stage discriminates between a speech shot and a narrated shot with a classifier. The details of the extraction of intra- and inter-shot features, and the discrimination between a speech shot and a narrated shot are described below.

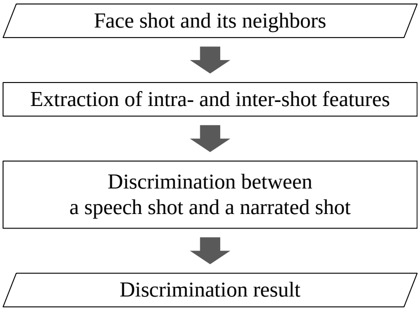

Face shot and its neighbors
Extraction of intra- and inter-shot features
Discrimination between
a speech shot and a narrated shot
Discrimination result

Fig. 7. Flow of the second stage of the proposed method.

2.2.1. *Extraction of intra- and inter-shot features*

The definitions of the intra-shot features and the inter-shot features are as follows.

Intra-shot features $\boldsymbol{f_{w_1}}$, $\boldsymbol{f_{w_2}}$: Generally, at an interview or a news conference, a subject speaks with small movement. Moreover, a speech shot is often taken with small camera motion in order to capture the behavior of the subject. Therefore, it is highly possible that the visual change in a speech shot is small. Focusing on this point, the second stage uses f_{w_1} that is the visual change between the initial frame and the last frame in an input shot. Concretely, f_{w_1} is calculated by

$$f_{w_1} = D(H^{(I)}, H^{(L)}), \tag{11}$$

where $H^{(I)}$ is the normalized RGB color histogram at the initial frame, and $H^{(L)}$ is that at the last frame of an input shot. D is the function that returns the Bhattacharyya distance between two input histograms.

Also, a speech shot is not always captured in a quiet environment. Therefore, a speech shot often contains some audio noise such as ambient noise and voices of people around. On the other hand, there are less audio noise in a narrated shot, since the voice for a narrated shot is usually recorded at a quiet environment in a broadcast station. Focusing on this point, the second stage uses f_{w_2} that is the level of audio noise in an input shot. Concretely, first, an input shot is divided into N sections. Next, the audio energy $P(n)$ $(n = 1, \ldots, N)$ is calculated by

$$P(n) = \frac{1}{T_n} \sum_{t_n \leq t \leq t_{n+1}} x^2(t), \tag{12}$$

where $x(t)$ is the sampled audio value at time t. T_n is the number of audio samples in the n-th section which is from time t_n to time t_{n+1}. Then, f_{w_2} is calculated by

$$f_{w_2} = \frac{1}{\lceil N/10 \rceil} \sum_{m=1}^{\lceil N/10 \rceil} P_{\text{low}}(m), \tag{13}$$

where $P_{\text{low}}(m)$ $(m = 1, \ldots, \lceil N/10 \rceil)$ are $\ln P(n)$ with values within the lowest ten percent of all the data in the shot, as shown in Fig. 8.

Inter-shot features $\boldsymbol{f_{b_i}}$ $(\boldsymbol{i = 1, \ldots, 6})$: In a broadcast news video, a scene where a person addresses a speech is often composed of a combination of more than one speech shot. This is a result of a concatenation of multiple speeches or excerpts from a long speech. Therefore, there is often a series of visually-similar speech shots. Focusing on this point, the second stage uses f_{b_1} and f_{b_2} that are the visual changes between a shot and its neighbor. For more details, f_{b_1} is calculated by

$$f_{b_1} = D(H_p^{(L)}, H^{(I)}), \tag{14}$$

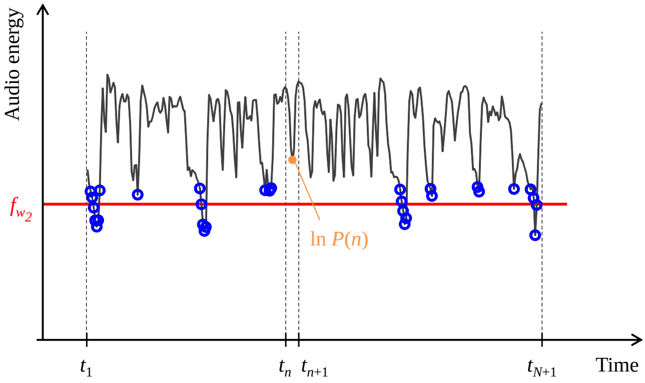


Fig. 8. Calculation of the level of audio noise f_{w_2} (circle: $P_{\text{low}}(m)$ $(m = 1, \ldots, \lceil N/10 \rceil)$).

where $H_p^{(L)}$ is the normalized RGB color histogram at the last frame of the previous shot of an input shot. Also, f_{b_2} is calculated by

$$f_{b_2} = D(H^{(L)}, H_n^{(I)}), \tag{15}$$

where $H_n^{(I)}$ is the normalized RGB color histogram at the initial frame of the next shot of an input shot.

In a speech shot, the audio is often muted at the initial frame and the last frame. This is an editorial technique so that sudden interruption of the voice when several speech shots are concatenated should not sound abrupt. In contrast, in a narrated shot, the volume at the shot change is not controlled, since an anchor person narrates continuously regardless of the shot change. Focusing on these points, the second stage uses f_{b_3} and f_{b_4} that are the volumes at the shot change. Concretely, f_{b_3} is calculated by

$$f_{b_3} = \frac{1}{T_p} \sum_{t_p^{(L)} \leq t \leq t^{(I)}} x^2(t), \tag{16}$$

where $t_p^{(L)}$ and $t^{(I)}$ are 1/30 second before and after the shot change to an input shot from the preceding shot, respectively. T_p is the number of audio samples from time $t_p^{(L)}$ to time $t^{(I)}$. Also, f_{b_4} is calculated by

$$f_{b_4} = \frac{1}{T_n} \sum_{t^{(L)} \leq t \leq t_n^{(I)}} x^2(t), \tag{17}$$

where $t_n^{(L)}$ and $t^{(I)}$ are 1/30 second before and after the shot change from an input shot to the succeeding shot, respectively. T_n is the number of audio samples from time $t^{(L)}$ to time $t_n^{(I)}$.

As described above, a speech shot may be noisier than a narrated shot because of the difference of the captured conditions. Therefore, if the level of audio noise in an input shot is much higher than that in its neighbor, it is highly possible that the input shot is a speech shot. Focusing on this point, the second stage uses f_{b_5} and f_{b_6} that are the differences of the level of audio noise between an input shot and its neighbors. Concretely, f_{b_5} and f_{b_6} are calculated by

$$f_{b_5} = f_{w_2} - f_{w_2}^{(p)}, \tag{18}$$

$$f_{b_6} = f_{w_2} - f_{w_2}^{(n)}, \tag{19}$$

where $f_{w_2}^{(p)}$ and $f_{w_2}^{(n)}$ are the levels of audio noise (Eq. (13)) in the preceding shot and the succeeding shot, respectively.

2.2.2. *Discrimination between a speech shot and a narrated shot*

The second stage discriminates between a speech shot and a narrated shot based on $\boldsymbol{f}$. Here, a SVM-based classifier is used to discriminate an input shot in a similar way to the first stage. Concretely, the SVM-based classifier is constructed with a training dataset in advance. The training dataset should contain speech shots and their neighbors, and narrated shots and their neighbors. Then, using the constructed SVM-based classifier, the second stage discriminates an input shot with $\boldsymbol{f}$ extracted from the input shot and its neighbors.

3. Experiments

We conducted mainly three experiments to evaluate the effectiveness of the proposed method. The first one was to investigate the discrimination accuracy of the first stage alone, and is described in Sec. 3.1. The second one was to investigate the discrimination accuracy of the second stage alone, and is described in Sec. 3.2. The last one was to investigate the overall accuracy of speech shot extraction by the proposed method, and is described in Sec. 3.3.

3.1. *Evaluation for the first stage*

As described in the previous section, we expect the first stage of the proposed method to accurately discriminate an input shot in case that an input shot contains a small amount of audio and visual noise. Therefore, we investigated the discrimination accuracy of the first stage with videos captured under a laboratory condition.

3.1.1. *Method*

We captured face shots (a total of 3,481 seconds) of ten males in their twenties under a laboratory condition with very low audio or visual noise. Here, each person read aloud different news articles. The specifications of the video and audio streams are shown in Tables 1 and 2. With these face shots, as shown in Fig. 9, subsequences in the shots were extracted, and then the first stage of the proposed method was applied to

Table 1. Specification of the video streams.

Frame rate	29.97 [frame/second]
Resolution	1,440 × 810 [pixel]
Lip region	Over 150 × 80 [pixel]

Table 2. Specification of the audio streams.

Sampling rate	12 [kHz]
Quantization bit rate	16 [bit]
Channel	Monaural

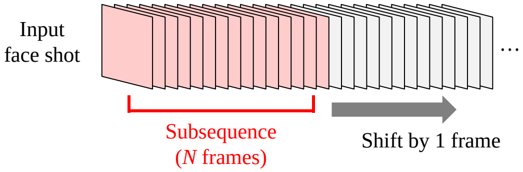


Fig. 9. Subsequences for discriminations.

each subsequence. The length of each subsequence was set to $N = 150$ frames (5 seconds) considering the length of face shots in actual broadcast news videos. We extracted a lip region in each frame of the face shots manually to avoid the influence of the extraction error.

For evaluation, five datasets were created from the face shots as shown in Table 3. We investigated the discrimination accuracy with five-fold cross validation on these datasets. That is, one dataset was used for validation while the remaining four datasets were used for training, and a total of five results for all datasets were averaged. As the evaluation criterion for each dataset, we used the discrimination accuracy defined by

$$\begin{aligned}&\text{Discrimination accuracy}\\&\quad=\frac{\text{Number of correctly-discriminated subsequences}}{\text{Total number of subsequences}}.\end{aligned} \tag{20}$$

For comparison, we also investigated the performances of a comparative method which did not use NCCs. That is, the comparative method used the following feature vector $\boldsymbol{c}'$.

$$\boldsymbol{c}' = (v_1, \ldots, v_4, a_1, \ldots, a_{26})^T. \tag{21}$$

Table 3. The datasets used for the five-fold cross validation (subject/speaker).

	Dataset				
	1	2	3	4	5
Subject = Speaker (Speech shot)	A/A B/B	C/C D/D	E/E F/F	G/G H/H	I/I J/J
Subject ≠ Speaker (Narrated shot)	A/B B/A	C/D D/C	E/F F/E	G/H H/G	I/J J/I

The difference between the proposed method and the comparative method was in only the used features. By comparing these two methods, we evaluated the performance of the first stage of the proposed method.

3.1.2. *Results*

The experimental results are shown in Table 4. The discrimination accuracy by the comparative method was 0.543, whereas that by the proposed method was 0.967. A higher accuracy was obtained by the proposed method. Therefore, we confirmed that the proposed method is effective for the discrimination between a speech shot and a narrated shot.

3.1.3. *Discussions*

We discuss (1) the effectiveness of using correlations between visual features and audio features, (2) the relation between the length of subsequences and the discrimination accuracy, (3) the effectiveness of integrating visual features and audio features, (4) the effectiveness of using time-derivative features, and (5) the robustness to audio noise.

The effectiveness of using correlations between visual features and audio features: The difference between the proposed method (only the first stage) and the comparative method was only whether NCCs between visual features and audio features were used or not. The comparative method discriminated in a space represented by the original audio-visual features. By this way, the correlations between visual features and audio features were expected to be implicitly evaluated by the SVM-based classifier. In contrast, the proposed method discriminated in a space represented by NCCs between visual features and audio features. By this way, the

Table 4. Discrimination accuracy of the proposed method (the first stage alone) and the comparative method.

	Comparative method	Proposed method
Discrimination accuracy	0.543	0.967

correlations between visual features and audio features were explicitly evaluated by the SVM-based classifier. That is, in the proposed method, the co-occurrence of a subject's lip motion and a speaker's voice was evaluated directly. As a result, the SVM-based classifier could discriminate between speech shots and narrated shots. We consider that this lead to the higher discrimination accuracy by the proposed method.

The relation between the length of subsequences and discrimination accuracy: As for the value of N; the number of frames used for calculating NCCs in the proposed method, we investigated the discrimination accuracy while changing N from 15 to 300. The result is shown in Fig. 10. As we can see in Fig. 10, the discrimination accuracy by the proposed method was higher than that by the comparative method for each N. Moreover, a larger N leads to a higher discrimination accuracy. These results make intuitive sense in view of the difference of the amount of information for discriminating between a speech shot and a narrated shot. Also, in general, the length of a speech shot in a broadcast news video is a few seconds. In this regard, since the discrimination accuracy with $N = 60$ (2 seconds) was about 0.90, and that with $N \geq 120$ (4 seconds) was over 0.95, we consider that the proposed method was very accurate. Therefore, for the application to broadcast news videos, we can set N to the length of an input face shot.

The effectiveness of integrating visual features and audio features: We compared discrimination accuracies by the proposed method and eight comparative

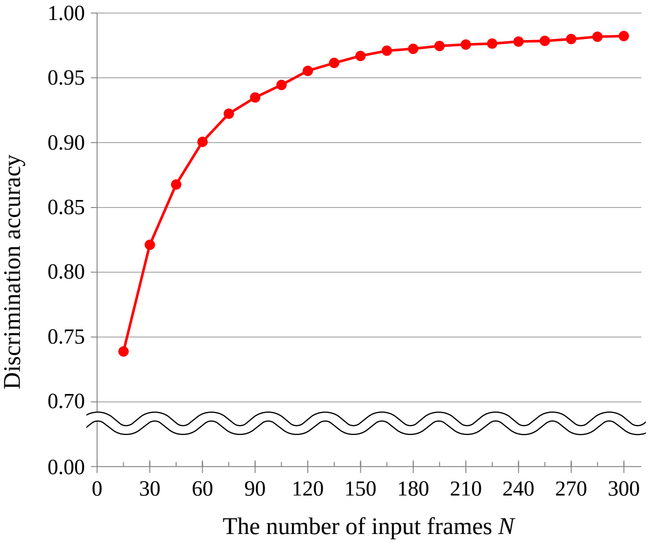


Fig. 10. Discrimination accuracies while changing the length of a subsequence.

Table 5. Comparisons to evaluate the effectiveness of integrating visual features and audio features.

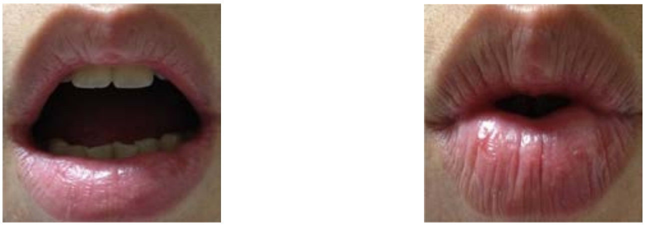
	Visual features		Audio features		
Method	Aspect ratio of lip region and its time-derivative	Area of lip region and its time-derivative	Audio energy and its time-derivative	MFCCs and their time-derivatives	Discrimination accuracy
Comparative 1A	√		√		0.883
Comparative 1B	√			√	0.930
Comparative 1C		√	√		0.892
Comparative 1D		√		√	0.951
Comparative 1E	√	√	√		0.892
Comparative 1F	√	√		√	0.955
Comparative 1G	√		√	√	0.940
Comparative 1H		√	√	√	0.962
Proposed	√	√	√	√	0.967

methods in which the used audio-visual features differed from each other. The results are shown in Table 5. Here, N (the number of input frames) was fixed to 150 frames (5 seconds). The discrimination accuracy by the proposed method was the highest of all the other methods. Also, adding any feature improved the discrimination accuracy. This indicates that the audio-visual features used in the proposed method are effective for measuring the co-occurrence between a lip motion and a voice.

Especially, the improvement by adding MFCCs and their time-derivatives was relatively-large. For example, only with audio energy, it would have been difficult to discriminate between an utterance of /a/ and an utterance of /i/ in case where the voice volumes were equal. In fact, there were many shots where the voice volumes were equal although the actual phonemes were different in the experimental datasets. MFCCs can discriminate the difference of utterances even if voice volumes are equal. Thus, by using MFCCs and their derivatives as well as the audio energy, the speaker's voice was expressed more accurately.

Similarly, there were many shots where the aspect ratios or the areas of a subject's lip region were equal in the experimental datasets. It is difficult to discriminate between the subject's lip shapes shown in Fig. 11 without using the area of the lip region, because different utterances may yield a similar aspect ratio. Thus, by using not only the aspect ratio of a lip region but also the area of the region, the subject's lip shape was expressed more accurately.

The effectiveness of using time-derivative features: We compared the performance of three methods: (1) the proposed method, (2) comparative method 1I without time-derivative features, and (3) comparative method 1J only with time-derivative features. The results are shown in Table 6. Here, N (the number of input frames) was fixed to 150 frames (5 seconds). The proposed method outperformed both comparative methods. In comparative method 1I, NCCs between absolute



/a/ /ɯ/

Fig. 11. Lip shapes with different utterances (Japanese vowels represented in English phonetic symbols).

Table 6. Comparisons to evaluate the effectiveness of time-derivative features.

	Visual features		Audio features		
Method	Aspect ratio and area of lip region	Time-derivatives	Audio energy and MFCCs	Time-derivatives	Discrimination accuracy
Comparative 1I	√		√		0.935
Comparative 1J		√		√	0.956
Proposed	√	√	√	√	0.967

states of a lip region and the voice were evaluated. In comparative method 1J, NCCs between their relative states were evaluated. Compared to these comparative methods, in the proposed method, both absolute and relative states were integrated and evaluated to discriminate a speech shot and a narrated shot. Thus, it is considered that this feature integration enabled the proposed method to achieve the higher performance.

The robustness to audio noise: The discrimination accuracy was 0.967 under a laboratory condition without any audio-visual noise. Thus, as expected, the first stage of the proposed method could accurately discriminate between a speech shot and a narrated shot in case that an input shot contains a small amount of audio and visual noise.

On the other hand, we also evaluated the discrimination accuracy of the first stage with 20 speech shots extracted from actual broadcast news videos (NHK News7). These speech shots were composed of 10 indoor shots and 10 outdoor shots, and varied from 8 to 12 seconds in length. The variation of age groups and genders in these shots is shown in Table 7. Note that the specifications of the video and audio streams were the same as those shown in Tables 1 and 2. The experimental results are shown in Table 8. Here, we manually extracted a lip region in each frame of the face shots to avoid the influence of the extraction error, and an SVM-based classifier was constructed with all datasets shown in Table 3. The number of correctly-discriminated shots was 9 out of 10 for indoor shots and 5 out of 10 for

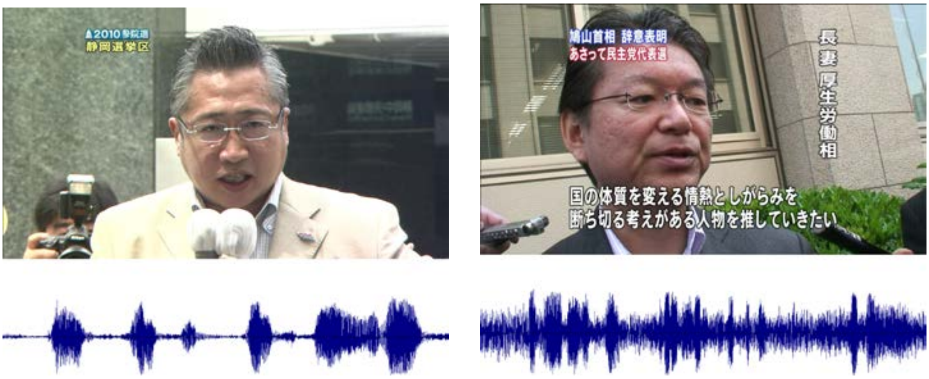
Table 7. Variation of age groups and genders in the shots.

	Indoor		Outdoor	
Age group	Male	Female	Male	Female
20	2	0	1	0
30	0	0	2	1
40	3	0	1	0
50	2	0	3	0
60	3	0	1	0
70	0	0	0	1
Total	10	0	8	2

Table 8. Discrimination accuracy for actually-broadcast speech shots.

Location	Discrimination accuracy
Indoor	0.9 (9/10)
Outdoor	0.5 (5/10)

outdoor shots. We consider that this can be explained mainly by the difference in the level of audio noise. Examples of a correctly-discriminated shot and a misdiscriminated shot are shown in Fig. 12. We can see that there is a small amount of audio noise in the correctly-discriminated shot. In contrast, most of the misdiscriminated shots were outdoor shots with a huge amount of audio noise. It is difficult to extract audio features only from a speaker's voice in such noisy shots. Accordingly, it should also be difficult to exactly measure the co-occurrence between a subject's lip motion and a speaker's voice. From the above results, we confirmed that the first stage of the proposed method may not correctly discriminate shots with



(a) Correctly-discriminated shot
(with a small amount of audio noise)

(b) Misdiscriminated shot
(with a huge amount of audio noise)

Fig. 12. Examples of a correctly-discriminated shot and a misdiscriminated shot.

a huge amount of audio noise, but can do so for shots with a small amount of noise accurately. Incidentally, although the first stage used the SVM-based classifier constructed with the face shots of ten persons only in their twenties, the first stage could correctly discriminate shots with a wide range of age groups in the indoor shots. Therefore, we expect the first stage to be able to deal with various age groups. However, we need a more in-depth investigation about the influence of the variation of genders, since the shots used in the experiment were almost of males (only two females).

3.2. *Evaluation for the second stage*

We investigated the discrimination accuracy of the second stage of the proposed method with actual broadcast news videos.

3.2.1. *Method*

We extracted 459 face shots from seven day's actual broadcast news videos (NHK News7) manually. We used these face shots as the experimental datasets as shown in Table 9. Here, each dataset was composed of one day's news video. We investigated the discrimination accuracy with seven-fold cross validation on these datasets. That is, one dataset was used for validation while the remaining six datasets were used for training, and a total of seven results for all datasets were averaged.

For comparison, we investigated the following five comparative methods in addition to the proposed method.

- Comparative method 2A:
  Without the intra-shot feature f_{w_1} (visual change in a shot)
- Comparative method 2B:
  Without the intra-shot feature f_{w_2} (audio noise in a shot)
- Comparative method 2C:
  Without the inter-shot feature f_{b_1} and f_{b_2} (visual change at a shot change)
- Comparative method 2D:
  Without the inter-shot feature f_{b_3} and f_{b_4} (volume at a shot change)

Table 9. Details of the face shots used in the experiment.

Dataset (Date)	Subject = speaker	Subject ≠ speaker	Total
Jan 24, 2010	48	13	61
Apr 10, 2010	37	25	62
Apr 19, 2010	49	13	62
Apr 4, 2010	55	19	74
May 6, 2010	48	18	66
May 12, 2010	41	31	72
May 24, 2010	43	19	62
Total	321	138	459

- Comparative method 2E:
  Without the inter-shot feature f_{b_5} and f_{b_6} (difference of audio noise between neighbor shots)
- Proposed method:
  With all the intra- and inter-shot features f_{w_1}, f_{w_2}, and f_{b_i} $(i = 1, \ldots, 6)$.

The differences between all methods were only in the used features. By comparing these six methods, we evaluated the performance of the second stage of the proposed method.

3.2.2. *Results*

The experimental results are shown in Table 10. The proposed method outperformed each comparative method. Also, adding any feature improved the discrimination accuracy. Therefore, this indicates that the intra- and inter-shot features used in the second stage of the proposed method are effective for discriminating a speech shot and a narrated shot.

3.2.3. *Discussion*

We discuss (1) the effectiveness of each feature used in the second stage of the proposed method, and (2) a room for improvement of discrimination accuracy.

The effectiveness of using features based on the tendency of speech shots: The discrimination accuracy was improved by adding any feature. Therefore, we confirmed that each feature was somewhat effective for the discrimination between a speech shot and a narrated shot. Especially, the inter-shot features f_{b_3} and f_{b_4} (volume at a shot change) gave the largest impact to the improvement. An example correctly-discriminated by using these features is shown in Figs. 13 and 14. We can see that there is no sound around (b) and (d) in Fig. 14. Also, such tendency was observed in many other speech shots used in the experiments, but not in most narrated shots used in the experiments. Thus, this would be one of the reasons why

Table 10. The features used in the proposed method and the comparative methods.

	Intra-shot feature		Inter-shot feature			
Method	Visual change f_{w_1}	Audio noise f_{w_2}	Visual change f_{b_1},f_{b_2}	Volume f_{b_3},f_{b_4}	Audio noise f_{b_5},f_{b_6}	Discrimination accuracy
Comparative 2A		√	√	√	√	0.760
Comparative 2B	√		√	√	√	0.797
Comparative 2C	√	√		√	√	0.771
Comparative 2D	√	√	√		√	0.758
Comparative 2E	√	√	√	√		0.780
Proposed	√	√	√	√	√	0.808

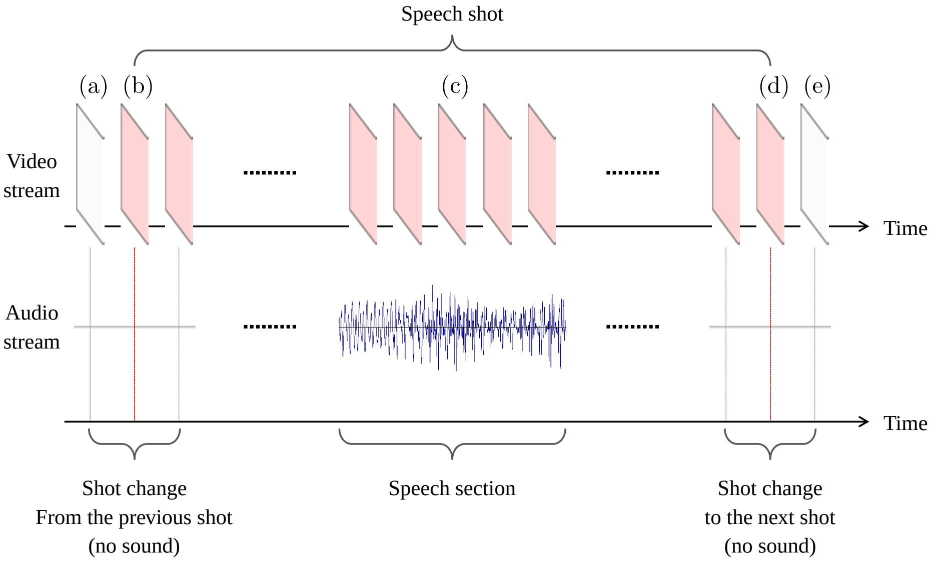


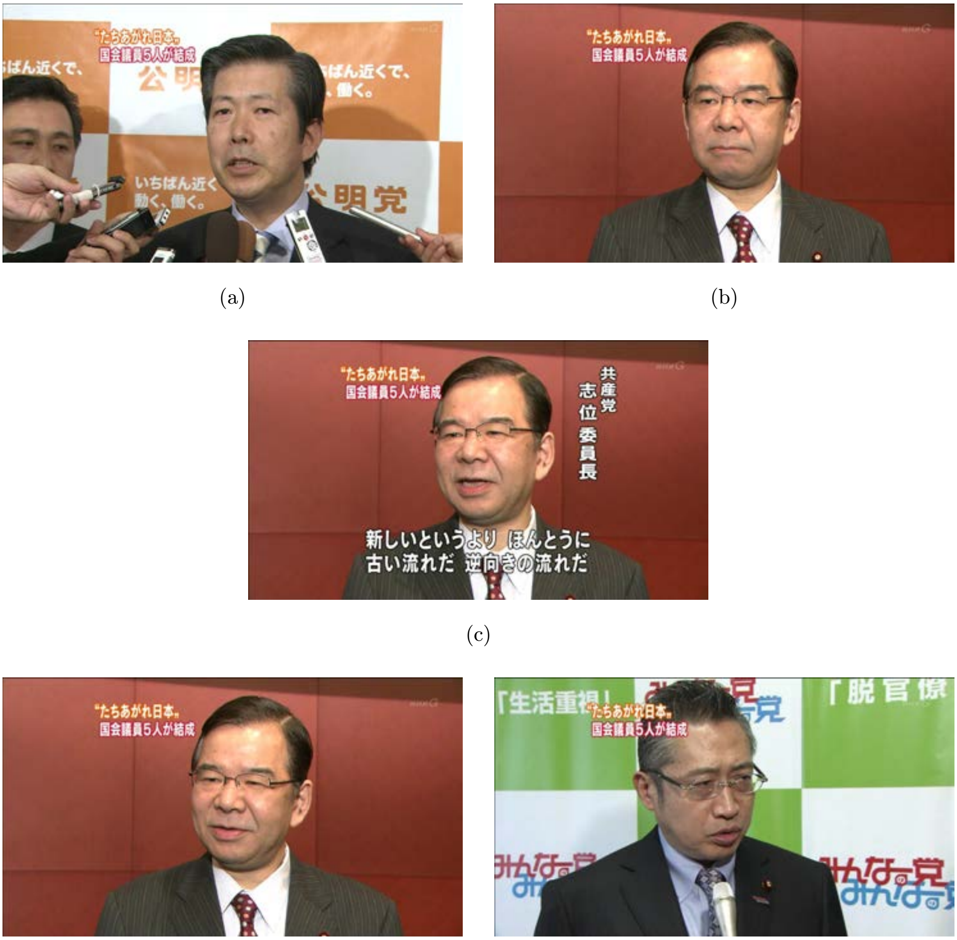
Fig. 13. An example correctly-discriminated by using the inter-shot features f_{b_3} and f_{b_4} (volume at a shot change).

higher discrimination accuracy was obtained by the second stage of the proposed method.

A room for improvement of discrimination accuracy: The discrimination accuracy by the proposed method was 0.808, whereas that by the comparative method 2B was 0.797. Note that the difference between these methods was only 0.011. We consider that there is a better way to extract features based on the tendency of speech shots. For example, the proposed method in this paper calculates the level of audio noise (f_{w_2}) from audio samples when a speaker seems not to be speaking. Here, we assumed that there are pauses while a person is speaking. Therefore, if there was no pause, the proposed method would not be able to calculate the level of audio noise. Thus, we need to study a better way to precisely capture the features (including f_{w_2}) that we focus on. Also, to improve the discrimination accuracy, we need to use the actual speech content in an input shot and its neighbor, and also whether a subject and a speaker are the same or not between neighbor shots.

3.3. *Overall system evaluation of the proposed method*

The experiments described in Secs. 3.1 and 3.2 were to investigate the discrimination accuracy of each stage of the proposed method. In this section, we present an experiment that aims to investigate the overall accuracy of speech shot extraction by the proposed method.



(a) (b)

(c)

(d) (e)

Fig. 14. Frames corresponding to Figs. 13(a)–13(e).

3.3.1. *Method*

We applied both of the two stages to the datasets (459 face shots extracted from seven day's broadcast news videos) shown in Table 9. The details are as follows. First, we applied the first stage of the proposed method to these face shots. Here, we judged an input shot as a speech shot if more than one third of subsequences were judged as a speech shot. The length of each subsequence was $N = 150$ frames (5 seconds), and the lip region in each frame of the face shots was automatically extracted by simply thresholding based on intensity and color. Next, we applied the second stage of the proposed method to the shots judged as narrated shots by the first stage. Finally, we extracted only the shots judged as speech shots in either stage.

As for the first stage, we trained the SVM-based classifier with the datasets in Table 3. As for the second stage, we trained the SVM-based classifier with the datasets in Table 9 except for those judged as speech shots in the first stage. Here, we evaluated the discrimination accuracy of the second stage with seven-fold cross validation on the datasets. That is, one dataset was used for validation while the remaining six datasets were used for training, and a total of seven results for all datasets were averaged. Note that training datasets and test datasets were completely separated in the evaluation for each stage.

As for the evaluation criteria, we used precision, recall, and F-measure which are commonly used for the evaluation in detection task. Each criterion is calculated by

$$\text{Precision} = \frac{\text{Number of correctly-extracted shots}}{\text{Number of extracted shots}}, \tag{22}$$

$$\text{Recall} = \frac{\text{Number of correctly-extracted shots}}{\text{Number of to-be-extracted shots}}, \tag{23}$$

$$\text{F-measure} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} \cdot \text{Recall}}, \tag{24}$$

respectively.

3.3.2. *Results*

The experimental results are shown in Table 11. F-measure for the first stage alone was 0.294, that for the second stage alone was 0.860, and that for both stages was 0.871. The highest F-measure was obtained by applying both stages; the proposed method. Therefore, we confirmed the effectiveness of the proposed method.

3.3.3. *Discussion*

We discuss (1) the effectiveness of the two-stage discrimination, and (2) a room for improvement of the extraction accuracy.

The effectiveness of two-stage discrimination: The highest F-measure was obtained by the combination of the first stage and the second stage. We consider that this owes to the fact that different types of features were used in the two stages. The first stage discriminates by directly-focusing on the co-occurrence between a lip motion and a voice. Therefore, as described in Sec. 3.1, the first stage can accurately discriminate in the case of a small amount of audio and visual noise but not in the

Table 11. Extraction accuracy of the proposed method.

	Precision	Recall	F-measure
First stage alone	0.949 (56/59)	0.174 (56/321)	0.294
Second stage alone	0.879 (270/307)	0.841 (270/321)	0.860
Overall	0.882 (276/313)	0.860 (276/321)	0.871

case of a huge amount of audio and visual noise. In contrast, the second stage rather takes advantage of such noise which reduced the discrimination accuracy in the first stage. However, the second stage cannot extract shots which do not follow the assumed tendency of speech shots. That is, the discrimination of the second stage is probabilistic (not absolute). In this regard, the proposed method combines both stages with such strong points and shortcomings, and covers the shortcomings of each stage. We consider that this is the reason the proposed two-stage discrimination method improved the extraction accuracy.

A room for improvement of extraction accuracy: Although the F-measure by the proposed method was 0.871 and high, there were some errors. One way to improve the extraction accuracy is to improve the discrimination accuracy of the classifier in each stage. By the first stage, the precision was 0.949, whereas the recall was 0.174. That is, the first stage could not extract many speech shots. This was mainly because of the error of extracting a lip region from a face shot. In this experiment, a lip region was extracted automatically by a simple and fast method; thresholding of intensity and color. It is difficult to accurately extract a lip region with a precision of few pixels, since the luminance in a lip region may drastically change by a flash of a camera or a shadow. The error of extracting a lip region affects the co-occurrence measuring between a lip motion and a voice, which leads to the error of speech shot extraction. To extract a lip region accurately, there are methods which use ASM (Active Shape Model) and Snakes proposed by Jang [18], which use AAM (Active Appearance Model) proposed by Matthews *et al.* [15], and so on [16, 19]. By applying these methods, we expect to improve the accuracy of the lip region extraction, and subsequently, the discrimination accuracy of the first stage. In addition, we can improve the extraction accuracy of the second stage by the way discussed in 3.2.3. By these improvements, we expect to improve the overall extraction accuracy by the proposed method.

4. Conclusion

In this paper, we proposed a method for discriminating between a speech shot and a narrated shot to extract genuine speech shots from a broadcast news video. The proposed method is composed of two stages. The first stage directly evaluates the inconsistency between a subject and a speaker based on the co-occurrence between lip motion and voice. The second stage evaluates based on the intra- and inter-shot features focusing on the tendency of speech shots. By combining the two stages, the proposed method accurately discriminates between a speech shot and a narrated shot. In the experiments, the overall accuracy of speech shot extraction by the proposed method was 0.871. Therefore, we confirmed that the proposed method is effective for the discrimination between a speech shot and a narrated shot.

In the future, for the first stage, we will study the improvement of the accuracy of lip region extraction. In addition, for in-depth investigation, we will evaluate the discrimination performance by comparing state-of-the-art methods (e.g. [10]), and analyze the influence of individual differences (e.g. age groups and genders of subjects in face shots). For the second stage, we will study on a better way to extract features based on the tendency of speech shots, and the use of the speech content in an input shot and its neighbor.

Acknowledgments

Parts of the work presented in this paper were supported by the Grants-in-Aid for Scientific Research and a joint-research project with National Institute of Informatics, Japan.

References

[1] S. Satoh, Y. Nakamura and T. Kanade, Name-it: Naming and detecting faces in news videos, *IEEE Trans. Multimedia* **6**(1) (1999) 22–35.

[2] D. Ozkan and P. Duygulu, Finding people frequently appearing in news, in *Image and Video Retrieval*, H. Sundaram *et al.* (Eds.), Lecture Notes in Computer Science, Vol. 4071 (2000), pp. 173–182.

[3] I. Ide, T. Kinoshita, H. Mo, N. Katayama and S. Satoh, TrackThem: Exploring a large-scale news video archive by tracking human relations, in *Information Retrieval Technology*, G. G. Lee, A. Yamada, H. Meng and S.-H. Myaeng (Eds.), Lecture Notes in Computer Science, Vol. 3689 (2005), pp. 510–515.

[4] A. F. Smeaton, P. Over and W. Kraaij, High-level feature detection from video in TRECVid: A 5-year retrospective of achievements, in *Multimedia Content Analysis, Theory and Applications*, A. Divakaran (Ed.), Signals and Communication Technology Series (Springer-Verlag, 2009), pp. 151–174.

[5] H. J. Nock, G. Iyengar and C. Neti, Speaker localisation using audio-visual synchrony: An empirical study, in *Image and Video Retrieval*, E. M. Bakker, M. S. Lew, T. S. Huang, N. Sebe and X. S. Zhou (Eds.), Lecture Notes in Computer Science, Vol. 2728 (2003), pp. 565–570.

[6] S. E. Tranter and D. A. Reynolds, An overview of automatic speaker diarization systems, *IEEE Trans. on Audio, Speech and Language Processing* **14**(5) (2006) 1557–1565.

[7] J. M. Baker, L. Deng, J. Glass, S. Khudanpur, C.-H. Lee, N. Morgan and D. O'Shaughnessy, Research developments and directions in speech recognition and understanding, Part 1, *IEEE Signal Processing Magazine* **26**(3) (2009) 75–80.

[8] G. Zhao, M. Barnard and M. Pietikäinen, Lipreading with local spatiotemporal descriptors, *IEEE Trans. Multimedia* **11**(7) (2009) 1254–1265.

[9] J. Lewis, Automated lip-sync: Background and techniques, *Journal of Visualization and Computer Animation* **2**(4) (1991) 118–122.

[10] E. A. Rúa, H. Bredin, C. G. Mateo, G. Chollet and D. G. Jiménez, Audio-visual speech asynchrony detection using co-inertia analysis and coupled hidden Markov models, *Pattern Analysis and Applications* **12**(3) (2009) 271–284.

[11] S. Kumagai, K. Doman, T. Takahashi, D. Deguchi, I. Ide and H. Murase, Detection of inconsistency between subject and speaker based on the co-occurrence of lip motion and

voice towards speech scene extraction from news videos, in *Proc. of 2011 IEEE Int. Symp. on Multimedia*, 2011, pp. 311–318.

[12] P. Viola and M. Jones, Rapid object detection using a boosted cascade of simple features, in *Proc. of 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition* **1** (2001) 511–518.

[13] D. Decarlo and D. Metaxas, Optical flow constraints on deformable models with applications to face tracking, *Int. J. of Computer Vision* **38** (2000) 99–127.

[14] T. F. Cootes, G. J. Edwards and C. J. Taylor, Active appearance models, *IEEE Trans. on Pattern Analysis and Machine Intelligence* **23**(6) (2001) 681–685.

[15] I. Matthews, T. F. Cootes, J. A. Bangham, S. Cox and R. Harvey, Extraction of visual features for lipreading, *IEEE Trans. on Pattern Analysis and Machine Intelligence* **24**(2) (2002) 198–213.

[16] A. W.-C. Liew, S. H. Leung and W. H. Lau, Lip contour extraction from color images using a deformable model, *J. of Pattern Recognition* **35**(12) (2002) 2949–2962.

[17] R. C. Verma, C. Schmid and K. Mikolajczyk, Face detection and tracking in a video by propagating detection probabilities, *IEEE Trans. on Pattern Analysis and Machine Intelligence* **25**(10) (2003) 1215–1228.

[18] K. S. Jang, Lip contour extraction based on active shape model and snakes, *Int. J. Computer Science and Network Security* **7**(10) (2007) 148–153.

[19] U. Saeed and J.-L. Dugelay, Combining edge detection and region segmentation for lip contour extraction, in *Proc. of 6th Intl. Conf. on Articulated Motion and Deformable Objects*, 2010, pp. 11–20.

[20] M. J. Lyons, C.-H. Chan and N. Tetsutani, Mouthtype: Text entry by hand and mouth, in *Proc. of Conf. on Human Factors in Computing Systems*, 2004, pp. 1383–1386.

[21] G. Potamianos, C. Neti, J. Luettin and I. Matthews, Audio-visual automatic speech recognition: An overview, in *Visual and Audio-Visual Speech Processing*, G. Bailly, E. Vatikiotis-Bateson and P. Perrier (Eds.), (MIT Press, 2004), pp. 1–30.

[22] S. G. Tanyer and H. Özer, Voice activity detection in nonstationary noise, *IEEE Trans. on Speech and Audio Processing* **8**(4) (2000) 478–482.

[23] G. Potamianos and C. Neti, Audio-visual speech recognition in challenging environments, in *Proc. of 8th European Conf. on Speech Communication and Technology*, (2003), pp. 1293–1296.

[24] V. N. Vapnik, *The Nature of Statistical Learning Theory*, 2nd edn. (Springer-Verlag, 1999).

[25] M. A. Aizerman, E. M. Braverman and L. I. Rozonoer, Theoretical foundations of the potential function method in pattern recognition learning, *Automation and Remote Control* **25** (1964) 821–837.