Detection of Inconsistency between Subject and Speaker based on the Co-occurrence of Lip Motion and Voice Towards Speech Scene Extraction from News Videos

Shogo Kumagai†, Keisuke Doman†, Tomokazu Takahashi‡,†
Daisuke Deguchi†, Ichiro Ide†, and Hiroshi Murase†

† Graduate School of Information Science, Nagoya University, Nagoya, Japan
Email: {skumagai, kdoman}@murase.m.is.nagoya-u.ac.jp, {ddeguchi, ide, murase}@is.nagoya-u.ac.jp
‡ Faculty of Economics and Information, Gifu Shotoku Gakuen University, Gifu, Japan
Email: ttakahashi@gifu.shotoku.ac.jp

***Abstract*—We propose a method to detect the inconsistency between a subject and the speaker for extracting speech scenes from news videos. Speech scenes in news videos contain a wealth of multimedia information, and are valuable as archived material. In order to extract speech scenes 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 approach, since news videos contain non-speech scenes where the speaker is not the subject, such as narrated scenes. To solve this problem, we propose a method to discriminate between speech scenes and narrated scenes based on the co-occurrence between a subject's lip motion and the speaker's voice. The proposed method uses lip shape and degree of lip opening as visual features representing a subject's lip motion, and uses voice volume and phoneme as audio feature representing a speaker's voice. Then, the proposed method discriminates between speech scenes and narrated scenes based on the correlations of these features. We report the results of experiments on videos captured in a laboratory condition and also on actual broadcast news videos. Their results showed the effectiveness of our method and the feasibility of our research goal.**

***Keywords*-speech scene extraction; audiovisual integration; news videos; lip motion; correlation;**

I. INTRODUCTION

Recently, there is a demand for the efficient reuse of massively archived broadcast videos which consist of various programs such as news, sports, dramas and so on. Especially, news video is valuable as an archived material, since they involve a wide range of real-world events that are closely related to our daily lives. Accordingly, there are many researches focussing on the reuse of broadcast news videos. 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 scenes such as interviews, press conferences, and public speakings, from news videos. Speech scenes provide a wealth of multimedia information to us, since they contain facial expressions, moods, and voice tones that are difficult to express only in text.

There is a high demand for the extraction of speech scenes from news videos. Extraction of speech scenes 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 focussing on speech.

As shown in Figure 1(a), in general speech scenes, the face region of a subject will appear in the center of a closeup image. Straightforwardly, the position and size of the face region is useful for the extraction of such scenes. However, as shown in Figure 1(b), there are non-speech scenes where the speaker is not the subject, such as narrated scenes. In such scenes, not the subject's voice but the anchor person's voice is present in the audio. Therefore, to extract genuine speech scenes 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 scenes from the face shots. By this way, we can obtain genuine speech scenes in news videos. Thus, we focus on the detection of the inconsistency between a subject and a speaker based on the co-occurrence between lip motion and voice.

To detect the inconsistency, it is necessary to discriminate between speech scenes and narrated scenes. To do that, some methods have been proposed. For example, the method proposed by Nock et al. is based on the score of the mutual information between visual and audio features [5]. This method works well for identifying the speaker among several people in a shot. However, this method is not adequate for discriminating between speech scenes and narrated scenes, because the distributions of audiovisual features extracted from each scene overlaps, which makes it difficult to obtain a decision boundary. Another method proposed by Rúa et al. [6] is based on the co-inertia analysis and coupled hidden Markov models. This method is for a biometric identification for which some kinds of test words (e.g. name, address) are input to an identification system under a specific condition. In news videos, there is an infinite variation in a number

(a) Speech scene (Subject = speaker) (b) Narrated scene (Subject ≠ speaker)

Figure 1. Examples of face shots in news videos.

of words spoken under various conditions. Therefore, it is difficult to simply apply this method to news videos.

In this paper, we propose a method to discriminate between speech scenes and narrated scenes in news videos based on the co-occurrence between lip motion and voice. The proposed method calculates the co-occurrence by integrating several visual features and audio features. By this way, we extract only speech scenes from the input news videos.

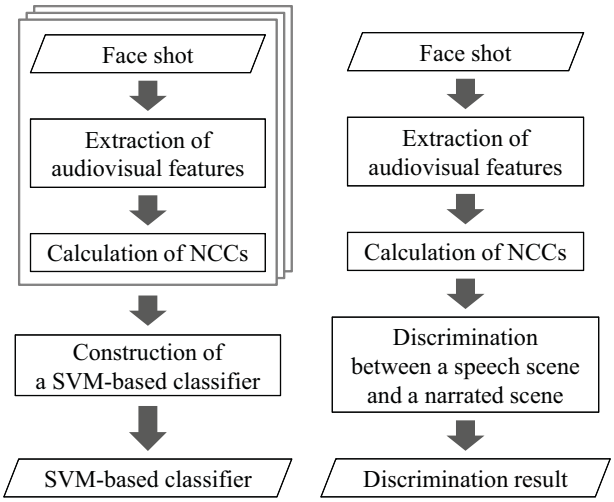
The paper is organized as follows. First, Section II describes the proposed method to discriminate between speech scenes and narrated scenes. Next, Section III reports and discusses the results of experiments on videos captured in a laboratory condition to evaluate the effectiveness of the proposed method. Then, Section IV reports and discusses the results of experiments on actual broadcast news videos to explore the feasibility of our research goal. Finally, Section V concludes the paper with our future work.

II. DISCRIMINATION BETWEEN A SPEECH SCENE AND A NARRATED SCENE

The process flow of the proposed method is shown in Figure 2. The proposed method is composed of mainly two stages: the training stage and the discrimination stage. In the training stage, first, visual features and audio features are extracted from training face shots. Next, NCCs (Normalized Correlation Coefficients) for each combination of a visual feature and an audio feature are calculated. Then, a classifier based on an SVM (Support Vector Machine) is constructed by using the NCCs. At the discrimination stage, first, visual features and audio features are extracted from an input face shot in the same way as the training stage. Then, to discriminate a speech scene from a narrated scene, the SVM-based classifier detects the inconsistency between a subject and a speaker. The details for each step are described below.

A. Extraction of audiovisual features

The process flow of the extraction of audiovisual features is shown in Figure 3. 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.



(a) Training stage (b) Discrimination stage

Figure 2. Flow of the proposed method.

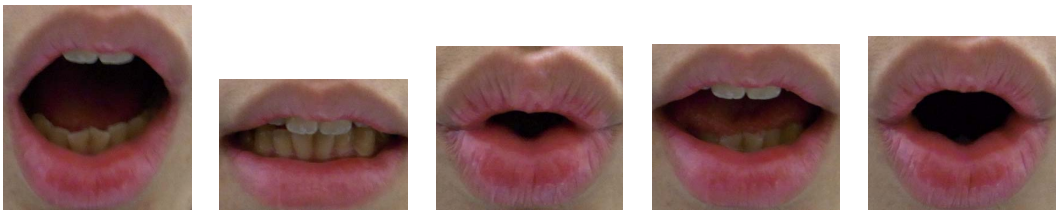
The visual features represent the lip motion of a subject, whereas audio features represent the voice of a speaker. In this paper, for each n-th input frame, visual features are denoted by $v_i(n)$ $(i = 1, \ldots, 4)$, and audio features are denoted by $a_j(n)$ $(j = 1, \ldots, 26)$. The details of the visual features and the audio features are as follows.

1) Visual features $v_i(n)$ $(i = 1, \ldots, 4)$: A lip shape and the degree of a lip opening will differ according to the phoneme type. For example, as shown in Figure 4, the lip shape for vowel "a" extends longitudinally, whereas the lip shape for vowel "i" extends transversally. That is, a lip motion of a subject highly is related to his/her utterance.

Focussing on this point, we extract visual features based on a lip motion. For more details, for each input frame, we extract visual features $v_i(n)$ $(i = 1, \ldots, 4)$ defined below.

- 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 $v_4(n)$

We expect that these visual features are useful for repre-



(a) “a” (b) “i” (c) “u” (d) “e” (e) “o”

Figure 4. Example of lip regions for each vowel.

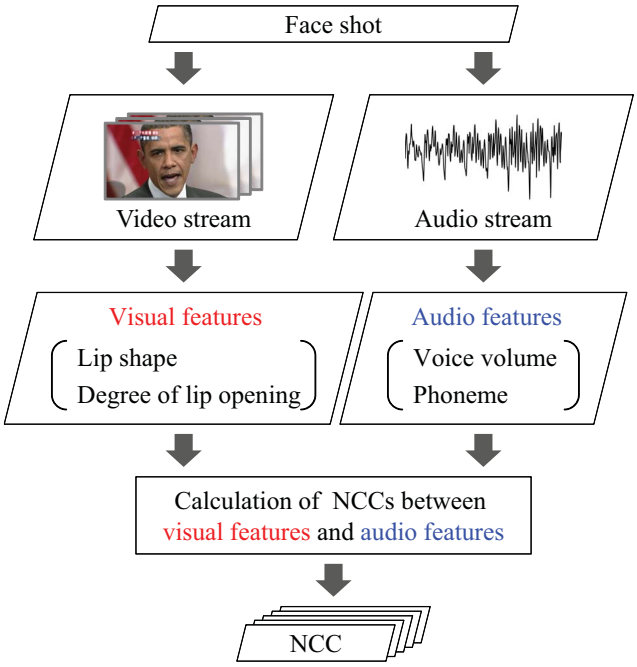


Figure 3. Calculation of Normalized Correlation Coefficients (NCCs).

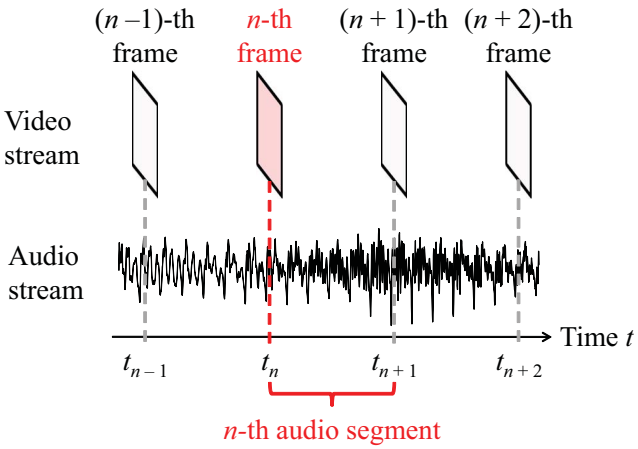


Figure 5. The range of an audio segment corresponding to an input frame.

senting the lip motion of a subject, since they are used for works on lip reading and speech recognition [7], [8]. After extracting these features for all input frames, we compose visual feature vectors $\boldsymbol{v}_i$ $(i = 1, \ldots, 4)$ defined by

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

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

As for the extraction of a lip region, many techniques have already been proposed. For example, there are methods with active shape model (ASM) and Snakes proposed by Jang [9], with active appearance model (AAM) proposed by Matthews et al. [10], and so on [11], [12]. These techniques may be applied to the extraction of lip regions from face shots in news videos.

2) Audio features $a_j(n)$ $(j = 1, \ldots, 26)$: A speaker's utterance differs according to his/her lip motion. Hence, as mentioned above, a speaker's utterance is related to his/her lip motion.

Focussing on this point, we extract audio features based on a speaker's utterance. For each input audio, we extract audio features $a_j(n)$ $(j = 1, \ldots, 26)$ defined below.

- 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 $a_j(n)$ $(j = 15, \ldots, 26)$

A voice volume represents a voice activity, whereas MFCCs represent the spectrum envelope of an audio wave corresponding to a 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 [13] and speech recognition [14]. In a similar way to the visual features, we compose audio feature vectors $\boldsymbol{a}_j$ $(j = 1, \ldots, 26)$ defined by

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

The range of an audio segment corresponding to an input frame is shown in Figure 5. We use the range of the audio observed from time t_n to time t_{n+1} to extract the n-th audio features. For example, if the video frame rate is 30 fps, the length of an audio segment is $\frac{1}{30}$ sec.

B. 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 $\boldsymbol{a}_j$.

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

where $\bar{v_i} = \frac{1}{N} \sum_{n=1}^{N} v_i(n)$, and $\bar{a_j} = \frac{1}{N} \sum_{n=1}^{N} a_j(n)$. Then, using all of the NCCs $c_{i,j}$, we compose a 104-dimensional vector $\boldsymbol{c}$ defined by

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

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

C. Construction of an SVM-based classifier

SVM (Support Vector Machine) was introduced by Vapnik [15], and is used in many pattern recognition applications. In SVM, a separating hyperplane is determined based on the margin maximization, which enhances the generalization capability of the classification function. In addition, with the Kernel trick [16], SVM achieves a nonlinear classification with low computational cost.

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

$$g(\boldsymbol{c}) = \sum_{i=1}^{l} \alpha_i y_i K(\boldsymbol{c}, \boldsymbol{c}_i) + b, \quad (5)$$

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 label y_i $(i = 1, \ldots, l)$. Here, $y_i = +1$ if the i-th training sample is a speech scene, otherwise $y_i = -1$. The parameter α_i is computed by maximizing the following quadratic problem

$$\sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{l} \alpha_i \alpha_j y_i y_j K(\boldsymbol{c}_i, \boldsymbol{c}_j) \quad (6)$$

under $\alpha_i \geq 0$ $(i = 1, \cdots, l)$, $\sum_{i=1}^{l} \alpha_i y_i = 0$. 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.

Table I
SPECIFICATION OF THE VIDEO AND AUDIO STREAMS.

Frame size	1,440 × 810 pixels
Frame rate	30 fps
Sampling rate	16 kHz

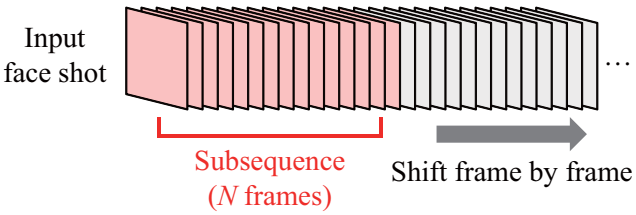


Figure 6. Subsequences for discriminations. More than one subsequence is extracted from a face shot by shifting frame by frame.

D. Discrimination between a speech scene and a narrated scene

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

$$f(\boldsymbol{c}) = \mathrm{sign}(g(\boldsymbol{c})), \quad (7)$$

where, $f(\boldsymbol{c}) \in \{-1, +1\}$. If $f(\boldsymbol{c}) = +1$ then the classification result is a speech scene, otherwise a narrated scene. By this way, we expect to discriminate between a speech scene and a narrated scene.

III. EXPERIMENT ON VIDEOS CAPTURED IN A LABORATORY CONDITION

This section reports and discusses the experimental results on videos captured in a laboratory condition to evaluate the effectiveness of the proposed method.

A. Experimental method

We captured face shots of ten persons under a laboratory condition. Each shot vary in length from 259 to 398 sec. (a total of 3,481 sec.). The specification of the video and audio streams is shown in Table I. Here, each person read aloud different news articles. Using these face shots, as shown in Figure 6, subsequences in the face shots were extracted, and then the proposed method was applied to each subsequence. In this experiment, the length of each subsequence was $N = 150$ frames (5 sec.) considering the length of face shots in actual broadcast news videos. Also, an RBF kernel function was used for the SVM-based classifier. As for the extraction of a lip region, we extracted a lip region in each frame of the face shots manually to avoid the influence of the extraction error.

To evaluate the performance of the proposed method, as shown in Table II, five datasets were created from the face shots. We evaluated the discrimination accuracy with 5-fold cross validation on these datasets; one dataset was used for

Table II
THE DATASETS USED FOR 5-FOLD CROSS VALIDATION (SUBJECT / SPEAKER).

	Data set				
	1	2	3	4	5
Subject = speaker (Speech scene)	A / A B / B	C / C D / D	E / E F / F	G / G H / H	I / I J / J
Subject ≠ speaker (Narrated scene)	A / B B / A	C / D D / C	E / F F / E	G / H H / G	I / J J / I

Table III
EXPERIMENTAL RESULTS ON VIDEOS CAPTURED IN A LABORATORY CONDITION.

	Proposed method	Comparative method
Discrimination accuracy	0.967	0.543

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

$$\text{Discrimination accuracy} = \frac{N_c}{N_t}, \tag{8}$$

where N_c is the number of correctly-discriminated subsequences, and N_t is the total number of subsequences.

For comparison, we investigated the performances of the proposed method and a comparative method. The comparative method used an SVM-based classifier without NCCs. In the comparative method, the following feature vector $\boldsymbol{c}'$ was extracted from each input frame.

$$\boldsymbol{c}' = (v_1, \ldots, v_4, a_1, \ldots, a_{26})^{\mathrm{T}}. \tag{9}$$

B. Results

Table III shows the experimental results. The discrimination accuracy by the proposed method was 0.967, whereas that by the comparative method was 0.543. A higher accuracy was obtained by the proposed method. Therefore, we confirmed that the proposed method is effective for the discrimination between a speech scene (Subject = speaker) and a narrated scene (Subject ≠ speaker).

C. Discussions

We discuss the effectiveness of 1) using correlations between visual features and audio features, 2) integrating visual features and audio features, and 3) using time-derivative features.

1) The effectiveness of using correlations between visual features and audio features: The difference between the proposed method and the comparative method was only whether NCCs between visual features and audio features were used. The comparative method discriminated in a space represented by original audiovisual features. By this way, the correlations between visual features and audio features would 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 correlations between visual features and audio features should be 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 scenes and narrated scenes. We consider that this lead to the higher discrimination accuracy by the proposed method.

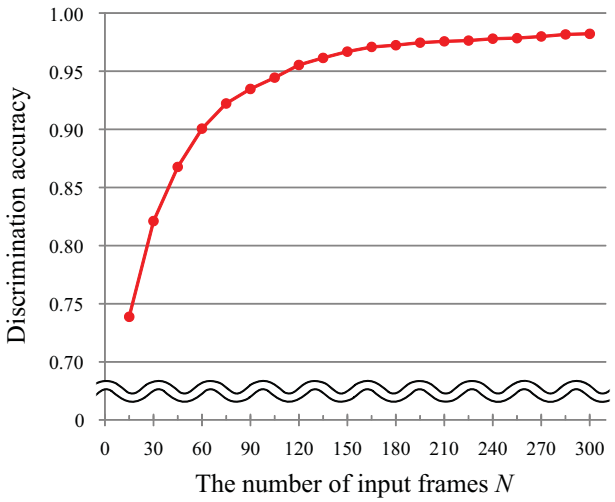


Figure 7. Discrimination accuracies while changing the length of a subsequence.

(a) "a" (b) "u"

Figure 8. Lip region with different utterances. Aspect ratios of both lip regions are nearly equal, however, their areas are quite different.

As for the choice 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 Figure 7. As we can see in Figure 7, a larger N lead to a higher discrimination accuracy. In addition, for each N, the discrimination accuracy by the proposed method was higher than that by the comparative method. These results make intuitive sense in view of the difference of the amount of information for discriminating between a speech scene and a narrated scene. Also, these results show that, for the application to broadcast news videos, we can choose N depending on the length of an input face shot.

Table IV
COMPARISONS TO EVALUATE THE EFFECTIVENESS OF INTEGRATING VISUAL FEATURES AND AUDIO FEATURES.

Method	Visual features		Audio features		Discrimination accuracy
	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	
Proposed	✓	✓	✓	✓	0.967
Comparative (A)	✓		✓		0.883
Comparative (B)	✓			✓	0.930
Comparative (C)		✓	✓		0.892
Comparative (D)		✓		✓	0.951
Comparative (E)	✓	✓	✓		0.892
Comparative (F)	✓	✓		✓	0.955
Comparative (G)	✓		✓	✓	0.940
Comparative (H)		✓	✓	✓	0.962

Table V
COMPARISONS TO EVALUATE THE EFFECTIVENESS OF TIME-DERIVATIVE FEATURES.

Method	Visual features		Audio features		Discrimination accuracy
	Aspect ratio and area of lip region	Time-derivatives of aspect ratio and area of lip region	Audio energy and MFCCs	Time-derivatives of audio energy and MFCCs	
Proposed	✓	✓	✓	✓	0.967
Comparative (I)	✓		✓		0.935
Comparative (J)		✓		✓	0.956

2) The effectiveness of integrating visual features and audio features: Table IV shows the comparison of discrimination accuracies by the proposed method and eight comparative methods in which the used audiovisual features differed from each other. Note that every comparative method used NCCs and an SVM-based classifier. Here, N (the number of input frames) was fixed to 150 frames (5 sec.). As seen from Table IV, the discrimination accuracy by the proposed method was the highest of all the other methods. Adding any feature improved the discrimination accuracy. Therefore, this indicates that the audiovisual 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 be difficult to discriminate between an utterance of "a" and an utterance of "i" in case where the voice volumes are equal. In fact, there were many scenes 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 scenes 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 Figure 8 without using the area of the lip region because different utterances may have a close 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.

3) The effectiveness of using time-derivative features: To investigate the effectiveness of using time-derivative features, we compared the performance of three methods: 1) the proposed method, 2) the comparative method (I) without time-derivative features, and 3) the comparative method (J) only with time-derivative features. The results are shown in Table IV. The proposed method outperformed both comparative methods. In the comparative method (I), NCCs between absolute states of a lip region and the voice were evaluated. In the comparative method (J), NCCs between relative states of them were evaluated. Compared to these comparative methods, in the proposed method, both absolute and relative states were integrated and evaluated to discriminate a speech scene and a narrated scene. Thus, it is considered that this feature integration enabled the proposed method to achieve the higher performance.

IV. EXPERIMENT ON ACTUAL BROADCAST NEWS VIDEOS

This section reports and discusses the experimental results on actual broadcast news videos to explore the feasibility of our research goal.

A. Experimental method

We used 20 speech scenes (Subject = speaker) and 20 narrated scenes (Subject $\neq$ speaker) to evaluate the performance of the proposed method. These scenes varied from 8 to 12 sec. in length, and the specification of the video and audio streams was the same as that shown in Table I. The speech scenes were extracted from actual broadcast news videos (NHK News7). The narrated scenes were built artificially by combining video streams of the speech scenes with the

Table VI
EXPERIMNETAL RESULTS ON ACTUAL BROADCAST NEWS VIDEOS.

	Scene type		Total (Average)
	Speech scene (Subject = speaker)	Narrated scene (Subject $\neq$ speaker)	
Discrimination accuracy	0.533	0.997	0.765

anchor person's voice in the other broadcast news videos. As for the extraction of a lip region, we extracted a lip region in each frame of the face shots manually to avoid the influence of the extraction error. An SVM-based classifier with an RBF kernel function was constructed with all datasets shown in Table II. Then these speech scenes and narrated scenes were discriminated by the classifier. Here, N (the number of input frames) was fixed to 150 frames (5 sec.).

B. Results

Table VI shows the experimental results. The discrimination accuracy of speech scenes was 0.533, and that of narrated scenes was 0.997. The average discrimination accuracy was 0.765.

C. Discussions

We discuss 1) the cause of discrimination errors and 2) the relation between the extraction accuracy of a lip region and discrimination accuracy.

1) The cause of discrimination errors: There were measurable audio noises in many of the speech scenes, which were discriminated as narrated scenes (i.e. miss-discriminated). In such scenes, audio features are extracted from not only speaker's voices but also ambient audio noises. Therefore, it would be difficult to measure the co-occurrence between a lip motion and a voice. To solve this problem, there are three approaches: 1) the reduction of the audio noises as a pre-processing, 2) the use of robust audio features to the audio noises, and 3) the training of a classifier with noisy samples.

As for the feasibility of our research goal, we aim at automatically generating speech collections, summarized videos focussing on speech, etc. To realize this, we consider that it is necessary to discriminate both speech scenes and narrated scenes accurately. As seen from Table VI, the proposed method could work for narrated scenes, but not for speech scenes. Therefore, in the future work, the discrimination accuracy of speech scenes should be improved.

2) The relation between the extraction accuracy of lip regions and discrimination accuracy: In this experiment, lip regions were extracted manually. In the future, this process should be done automatically with active shape model (ASM) and Snakes proposed by Jang [9], with active appearance model (AAM) proposed by Matthews et al. [10], and so on [11], [12]. However, it is difficult to accurately extract a lip region with the error range in a 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 causes the error of measuring the co-occurrence between a lip motion and a voice. We will study on the accurate extraction of a lip region from a news video, and training a classifier using training samples with the errors of lip region extractions, in the future.

V. CONCLUSION

In this paper, we proposed a method to discriminate between speech scenes and narrated scenes based on the correlations between visual features representing a subject's lip motion and audio features representing the speaker's voice. In the experiment applied to videos captured in a laboratory condition, the discrimination accuracy by the proposed method was 0.967, and the effectiveness of our method was shown. Also, in the experiment applied to actual broadcast news videos, the discrimination accuracy of speech scenes was 0.533, and that of narrated scenes was 0.997. In the future, to obtain higher discrimination accuracy, we will study on the reduction of audio noises in pre-processing, and the accurate extraction of lip regions. Additionally, we will study on using the structure of news videos and also refer to closed-caption for the accurate extraction of speech scenes.

ACKNOWLEDGMENT

Parts of the work presented in this paper were supported by the Grants-in-Aid for Scientific Research.

REFERENCES

[1] S. Satoh, Y. Nakamura, and T. Kanade, "Name-It: Naming and detecting faces in news videos," *IEEE Multimedia*, vol. 6, no. 1, pp. 22–35, January-March 1999.

[2] D. Ozkan and P. Duygulu, "Finding people frequently appearing in news," in *Image and Video Retrieval*, ser. Lecture Notes in Computer Science, H. Sundaram, M. Naphade, J. R. Smith, and Y. Rui, Eds. Springer, July 2006, vol. 4071, 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*, ser. Lecture Notes in Computer Science, G. G. Lee, A. Yamada, H. Meng, and S.-H. Myaeng, Eds. Springer, October 2005, vol. 3689, 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*, ser. Signals and Communication Technology, A. Divakaran, Ed. Springer, 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*, ser. Lecture Notes in Computer Science, E. M. Bakker, M. S. Lew, T. S. Huang, N. Sebe, and X. S. Zhou, Eds. Springer, July 2003, vol. 2728, pp. 565–570.

[6] 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*, vol. 12, no. 3, pp. 271–284, September 2009.

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

[8] 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.

[9] K. S. Jang, "Lip contour extraction based on active shape model and snakes," *Intl. J. of Computer Science and Network Security*, vol. 7, no. 10, pp. 148–153, October 2007.

[10] I. Matthews, T. F. Cootes, J. A. Bangham, S. Cox, and R. Harvey, "Extraction of visual features for lipreading," in *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 24, no. 2, pp. 198–213, February 2002.

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

[12] 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*, vol. 35, no. 12, pp. 2949–2962, December 2002.

[13] S. G. Tanyer and H. Özer, "Voice activity detection in nonstationary noise," in *IEEE Trans. on Speech and Audio Processing*, vol. 8, no. 4, pp. 478–482, July 2000.

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

[15] V. N. Vapnik, *The nature of statistical learning theory*, 2nd ed. Springer, 1999.

[16] 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*, vol. 25, pp. 821–837, June 1964.