Abstract

Meeting minutes are useful for efficient operations and meetings. The labor cost and time for taking the minutes can be minimized using a system that can assign a speaker to the minutes. An automatic speaker identification method using lip movements and speech data obtained by an omnidirectional camera was developed. To improve the accuracy of speaker identification, it is necessary to use the extracted data of speech sections. The proposed speech section extraction method was studied as a preprocessor for speaker identification. The proposed method consists of three processes: i) extraction of speaking frames using lip movements, ii) extraction of speaking frames using voices, and iii) discrimination of speech sections using these extraction results. In the extraction of speech sections using lip movements, the nose width was used for a threshold for automatic calculation. The speech sections can be extracted, even when the distance between the camera and the subject changes, by using a threshold based on the width of the nose. Finally, 11 sentences of speech video data (154 data) of 14 subjects were used to evaluate the usefulness of the method. The evaluation result obtained was a high F-measure of 0.96 on average. The results reveal that the proposed method can extract speech sections, even when the distance between the camera and the subject changes.

Keywords: speech section, speaker identification, lip

1. Introduction

In Japan, many companies are striving to improve the efficiency of their operations and meetings\(^1\) with work-style reform. Meeting minutes are useful for efficient operations and meetings. The labor cost and time for taking the minutes can be minimized using a system that can assign a speaker to the minutes. There are several practical speaker identification methods: one is to assign a microphone to each person, and another is to identify the speaker by voiceprint\(^2\). However, these methods require preparation in advance because they require multiple microphones and preregistered voiceprints. On the other hand, speaker identification method using image and voice data by Convolutional Neural Network was proposed\(^3\). This method needs to train the model using hundreds of hours of data. These problems can be solved by analyzing the relationship between lip movement and voice.

Fig. 1 shows an automatic speaker identification system that uses lip movements and speech captured by an omnidirectional camera\(^4\). The similarity of lip movements taken from speech and images was evaluated, and the speaker was identified. To improve the accuracy of speaker identification, it is necessary to use the extracted data of speech sections.

In this article, a speech section extraction method that uses image and voice information as a preprocessor for speaker identification is proposed. Video data (154 data points) from 14 subjects were used to evaluate the usefulness of the method.

![Fig. 1. Speaker identification system.](image-url)
2. Data

The data acquisition environment is shown in Fig. 2. In this study, videos of 11 Japanese sentences (selected from Web news articles) spoken by 14 subjects (A–N, of Asian descent) were used. The videos were acquired using an omnidirectional camera (RICOH: THETA V) and a microphone (RICOH: TA-1). The omnidirectional camera was set horizontally with the ground. The lens was in front of the subject, and the camera height was set approximately 100 cm. The data used in this investigation were acquired in accordance with the ethical regulations concerning studies involving humans at Akita University, Japan.

3. Proposed Method

3.1 Overview

An overview of the proposed method is shown in Fig. 3. The proposed method consists of three processes: i) extraction of speaking frames using lip movements, ii) extraction of speaking frames using voices, and iii) discrimination of speech sections using these extraction results. First, the face areas in the image were extracted, and the height of the inside of the mouth was acquired. Second, the speaking frames were extracted using a time-series variation of the height of the inside of the mouth. Third, the speaking frames were re-extracted. Fourth, voice features were extracted from the voice data, and speaking frames were extracted using a time-series variation of voice features. Finally, the results of the speech sections were calculated using the speaking frames that were extracted using images and voice.

3.2 Extraction of height of the inside of the mouth

The face area was detected using the open-source library Dlib\(^5\), which can detect 68 feature points on the face. Twenty feature points can be used to extract the shape of the lip. Two points of the lip feature points were used to calculate the height of the inside of the mouth. An example of the calculation results of the height of the inside of the mouth is shown in Fig. 4. A time-series variation of the height was calculated and smoothed using a moving average filter. These waveforms were used as lip movement features.

3.3 Speaking frame extraction by lip movement

The movement of the lips in the speaking frames is larger than that in the nonspeaking frames. This movement difference was used to extract the speaking frames. First, the maximum and minimum values of lip movement features were calculated from the target frame and the 10 frames before and after it (21 frames). Second, the difference \( \text{avg}_{\text{lip}} \) between the maximum and minimum values was calculated. Third, if \( \text{avg}_{\text{lip}} \) was larger than or equal to the

---

Fig. 2. Data acquisition environment.

Fluorescent light

700–900 lx (around the face)

Approx. 70 cm

Approx. 50 cm

Display

THETA V

TA-1

Fig. 2. Data acquisition environment.

i) Extraction of speaking frames using lip movements.

Extraction of height of inside of the mouth

Speaking frames extraction by lip movement

Speaking frames re-extraction

ii) Extraction of speaking frames using voices.

Voice feature extraction

Speaking frames extraction by voice

iii) Discrimination of speech sections using extraction results of i) and ii).

Speech sections extraction

Fig. 3. Overview of the proposed method.
threshold \((T_h)\), this frame was extracted as an utterance frame. The \(T_h\) is calculated using Equation (1), where \(N_w\) is the median width of the nose in the video. An example of the nose width calculation is shown in Fig. 4. Equation (1) was calculated based on the results of a preliminary experiment using subjects A–F.

\[
T_h = 0.043N_w + 0.067
\]  

\(1\)

### 3.4 Speaking frame re-extraction

The speaking frames were re-extracted because the speaking frames were not extracted correctly owing to small lip movements. Specifically, the frames that had speaking frames within the next five frames were re-extracted as speaking frames.

### 3.5 Voice feature extraction

The mel-frequency cepstrum coefficient (MFCC) is a feature value for speech recognition. The low-dimensional MFCC has features related to the acoustic properties of the vocal tract and the shape of the oral cavity. In this study, the zeroth dimension of the MFCC, as the lowest dimension, was used as the voice feature. Time series data of the voice features were smoothed using a moving average filter and used for speaking frame extraction.

### 3.6 Speaking frame extraction by voice

The voice feature values of the speaking frames are larger than those of the nonspeaking frame and maintain a large value. The speech sections were extracted by focusing on these features. First, the maximum and minimum values of the voice features were calculated from the target frame and the 20 frames before and after it (41 frames). Second, the difference \((Dm)\) between the maximum and minimum values was calculated. Third, the difference \((Vf)\) between the voice features and \(Dm\) on each frame was calculated. Finally, the case in which \(Vf\) was larger than or equal to the 0.00 threshold was extracted as the speaking frame.

### 3.7 Speech section extraction

The speech sections were defined as consecutive speaking frames extracted using lip movement features. In this process, the sections were finalized as speech sections if the speech section contained speaking frames that were extracted using the voice features. Three lip movement features and 10 speech features are equal to 0.1 s in a video. An example of speech section extraction is shown in Fig. 5.

### 4. Evaluation Process

#### 4.1 Labeling of speaking frames

The speech start and end frames were set manually. Furthermore, the frame between the speech start frame and end frame was defined as the speaking section.
4.2 Evaluation index

The F-measure was used as the evaluation index. The evaluation index was calculated using Equations (2)–(4), where $TP$ represents the number of frames from which speaking frames are correctly extracted as speaking frames, $FP$ represents the number of frames in which nonspeaking frames are accidentally extracted as speaking frames, and $FN$ represents the number of frames in which speaking frames are accidentally extracted as nonspeaking frames. The F-measure ranges from 0.00 to 1.00, and the closer the value is to 1.00, the higher the success rate.

$$
precision = \frac{TP}{TP + FP} \quad (2)
$$

$$
recall = \frac{TP}{TP + FN} \quad (3)
$$

$$
F\text{-measure} = \frac{2 \times recall \times precision}{recall + precision} \quad (4)
$$

4.3 Evaluation procedure

The automatic threshold ($Th$) setting method of the proposed method was evaluated using 11 sentence videos (154 data points) of 14 subjects. A method using a $Th$ of 1.5 was used as a comparison method. In the comparison method, $Th$ was set to the best value based on the results of a preliminary study that used the data of subjects A–F. In this evaluation process, the median nose width ($W$) and the height of the inside of the mouth were multiplied by weights (set in increments of 0.1 from 0.5 to 5.0) for pseudo-change of the distance between the camera and subject. Furthermore, the F-measure of the proposed method and the comparison method were compared for each weight value.

5. Results

The mean F-measure values for the 14 subjects in each weight value are shown in Fig. 6. In the comparison method, the average values of the F-measure decreased with an increase or decrease in the value of the weights. However, the mean of the F-measure remained higher than 0.95, even when the weight value changed, in the proposed method. The resolution is approximately 0.80 mm per pixels when the distance of between the camera and the subject is 50 cm. The resolution changes in proportion to the distance. For example, the resolution is approximately 1.70 mm per pixels when the distance of between the camera and the subject is 100 cm. The threshold ($Th$) calculation formula was considered using the weight 0.5–5.0 (the resolution of approximately 1.60–0.16 mm per pixel). Therefore, the proposed method is can be used for up to the distance of between the camera and the subject is approximately 100 cm. The result suggests that the proposed method can correctly extract speaking sections in different subjects.

| Subject | Proposed method | Comparison method |
|---------|-----------------|-------------------|
| A       | 0.96            | 0.80              |
| B       | 0.98            | 0.89              |
| C       | 0.98            | 0.86              |
| D       | 0.96            | 0.83              |
| E       | 0.97            | 0.85              |
| F       | 0.97            | 0.84              |
| G       | 0.95            | 0.86              |
| H       | 0.96            | 0.86              |
| I       | 0.97            | 0.89              |
| J       | 0.96            | 0.86              |
| K       | 0.95            | 0.81              |
| L       | 0.96            | 0.88              |
| M       | 0.97            | 0.86              |
| N       | 0.94            | 0.84              |
| Average | 0.96            | 0.85              |
6. Conclusions

A speech section extraction method using image and voice information as a preprocessor for speaker identification was proposed. Based on several experiments, it was determined that i) the proposed method can extract speech sections, even when the distance between the camera and the subject changes, and ii) the proposed method can correctly extract speaking sections in different subjects.

Acknowledgment

The authors would like to thank Japan Business Systems, Inc. for assistance in conducting this study. This study was supported by JSPS KAKENHI Grant Number JP19K12909.

References

(1) Japan Business Federation : “Survey of working hours in 2017 Aggregate results”, p. 13, 2017
(2) Microsoft Azure : “Speaker Recognition,”
https://azure.microsoft.com/ja-jp/services/cognitive-services/speaker-recognition/ (Accessed 01/12/2021)
(3) Joon Son Chung and Andrew Zisserman : “Out of time: automated lip sync in the wild” Workshop on Multi-view Lip-reading, ACCV, 2016
(4) Etsuro Nakamura, Yoichi Kageyama, Motonari Shirasu : “Speaker Identification Method Using Facial Image and Voice”, 2020 IEEE 2nd Global Conference on Life Sciences and Technologies (LifeTech 2020), pp. 411–412, RET 2.4, 2020
(5) “Dlib C++ Library”, http://dlib.net/
(Accessed 01/12/2021)
(6) Christos Sagonas, Epameinondas Antonakos, Georgios Tzimiropoulos, Stefanos Zafeiriou, and Maja Pantic : “300 Faces In-The-Wild Challenge; Database and results”, Image and Vision Computing (IMAVIS), Vol. 47, pp. 3–18, 2016