AUDIO-CUT DETECTION AND AUDIO-SEGMENT CLASSIFICATION USING FUZZY C-MEANS CLUSTERING

Naoki Nitanda, Miki Haseyama, and Hideo Kitajima

School of Engineering, Hokkaido University
N-13 W-8 Kita-ku Sapporo 060-8628, JAPAN
{nitanda, mikich, kitajima}@media.eng.hokudai.ac.jp

ABSTRACT

This paper proposes an audio-cut detection and audio-segment classification method using fuzzy c-means clustering. In the proposed method, the boundaries between two different audio signals, which are called *audio-cuts*, can be detected by the fuzzy c-means clustering. In the fuzzy c-means clustering, the fuzzy number represents the possibility that the audio-cut exists. Therefore, according to the possibility, qualified candidates for audio-cuts can be obtained even if audio effects such as fade-in, fade-out, etc. are included in the audio signal. The audio signal is segmented at the detected audio-cuts, and these segments are classified into the following five classes: silence, music, speech, speech with music background, and speech with noise background. This classification simultaneously deletes the wrongly detected audio-cuts. Consequently, we can obtain the accurate audio-cuts and the classification results.

1. INTRODUCTION

The popular use of the Internet, higher bandwidth access to the network, and widespread of digital recording and storage have created needs for segmentation and classification techniques of audio-visual materials. For the accurate audio-visual segmentation and classification, some methods which utilize the audio signal have been proposed [1]-[6]. They segment an audio signal into different audio signals at their boundaries, which are called *audio-cuts*, and classify the segments, which are called *audio-segments*, into basic audio classes such as silence, speech, music, etc. However, they cannot provide enough accuracy in the audio-cut detection, which is caused by audio processing for several effects such as fade-in, fade-out, etc.

Therefore, we propose an accurate audio-cut detection and audio-segment classification method using fuzzy c-means clustering. In the proposed method, the fuzzy c-means clustering is applied to the audio-cut detection so that the possibility that the audio-cut exists can be represented by the fuzzy number, while the previous work can only represent that the audio-cut exists or not. Therefore, we obtain not only highly reliable results but also possible candidates for the audio-cuts. After the audio-cut detection, the audio-segments are classified into the following five classes: silence, music, speech, speech with music background, and speech with noise background. Since this classification also deletes the wrongly detected audio-cuts, we can obtain the accurate audio-cuts and the classification results. Further, since the proposed method can directly process the MPEG audio signal without any decoding procedures, it can be easily applied to the audio-visual indexing which is compressed by MPEG.

2. AUDIO-CUT DETECTION

The proposed method processes an audio signal which are coded by MPEG Audio Layer III (MP3). Therefore, the MDCT coefficients in the MP3 codes are utilized for our audio-cut detection.

First, we compute the power of the audio signal $E(n)$ as follows:

$$E(n) = \sum_{i=0}^{31} \sum_{j=0}^{17} \{F_n(i,j)\}^2, \tag{1}$$

where n, i, and j denote the granule number, the sub-band number, and the sample number in the MP3 codes, respectively. $F_n(i,j)$ denotes the MDCT coefficient of nth granule, ith sub-band, and jth sample.

Second, we define the parameter sequence $C(n)$ by using the power sequence $E(n)$. The parameter sequence $C(n)$ is computed as follows:

$$C(n) = \frac{\sum_{k=0}^{W_1-1} E(n+k)E(n+k-W_1-1)}{\sqrt{\sum_{k=0}^{W_1-1}\{E(n+k)\}^2}\sqrt{\sum_{k=0}^{W_1-1}\{E(n+k-W_1-1)\}^2}}, \tag{2}$$

where W_1 is a predefined window length.

The existence of the audio-cut can be detected by observing the above sequence $C(n)$. Let us explain why it is possible. By the definition of Eq. (2), the sequence $C(n)$ is computed by using the power sequence $E(n)$ in two adjoining sliding windows which are illustrated in Fig. 1. The case that an audio-cut exists in either of the window is considered as follows: An example of this case is shown in the window L_1 and the window R_1 of Fig. 1. As shown in Fig. 1, the power sequence $E(n)$ in the window R_1 changes at the audio-cut abruptly, while the power sequence $E(n)$ in the window L_1 does not change abruptly. Therefore, the numerator of $C(n)$ is much smaller than its denominator; and thereby $C(n)$ is close to 0. As opposed to this, the case that an audio-cut exists in neither of the window is considered as follows: An example of this case is shown in the window L_2 and the window R_2 of Fig. 1. In Fig. 1, the power sequences $E(n)$ in the both windows do not change abruptly. Therefore, the numerator of $C(n)$ is close to its denominator; and then the number of $C(n)$ is close to 1. This shows that the existence of the audio-cut can be detected by observing the sequence $C(n)$.

Finally, the audio-cuts are detected by applying the fuzzy c-means clustering to the sequence $C(n)$ $(n = 1, 2, \cdots)$ obtained

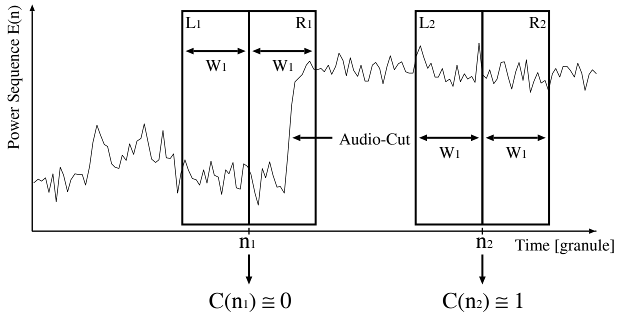


Fig. 1. The parameter $C(n)$: if either of the two adjacent sliding windows contains an audio-cut, $C(n)$ is close to 0; and if neither of them contains an audio-cut, $C(n)$ is close to 1.

above. In this clustering, we define the following three target vectors:

$$\boldsymbol{P_n} = [C(n), \cdots, C(n+W_2-1)]^T,$$
$$\boldsymbol{P_{n-\Delta}} = [C(n-\Delta), \cdots, C(n-\Delta+W_2-1)]^T,$$
$$\boldsymbol{Z} = [0, \cdots, 0]^T,$$

where W_2 is a predefined window length, and Δ is a step size. Character T represents the transpose of a matrix. These vectors are classified into two clusters by applying fuzzy c-means clustering.

If an audio-cut is included in the time interval form (n) to $(n+W_2-1)$, each element of $\boldsymbol{P_n}$ is close to 0. Therefore, the Euclidean distance between $\boldsymbol{P_n}$ and $\boldsymbol{Z}$ becomes shorter, and thereby $\boldsymbol{P_n}$ and $\boldsymbol{Z}$ are classified into the same cluster. On the contrary, if any audio-cuts are not included in the time interval from (n) to $(n+W_2-1)$, the elements of $\boldsymbol{P_n}$ are close to those of $\boldsymbol{P_{n-\Delta}}$. Therefore, the Euclidean distance between $\boldsymbol{P_n}$ and $\boldsymbol{P_{n-\Delta}}$ becomes shorter, and thereby $\boldsymbol{P_n}$ and $\boldsymbol{P_{n-\Delta}}$ are classified into the same class. By using these properties, qualified candidates for audio-cuts can be obtained.

An example of the audio-cut detection results is shown in Fig. 2. The audio signal captured from a TV news program at 44.1 kHz is used. A fade-out exists between two different audio signals whose contents are music and speech, respectively. As shown in Fig. 2, since the membership value becomes higher at the beginning and the end of the fade-out, the audio-cut can be detected by the proposed method.

3. AUDIO-SEGMENT CLASSIFICATION

Our proposed method includes an audio-segment classification into the following five classes:

- Silence: This class contains only a quasi-stationary background noise.
- Music: This class contains the sound made by the musical instrument.
- Speech: This class contains the voice of human beings such as the sound of conversation.
- Speech with music background: This class contains the speech under the environment where music exists in a background.
- Speech with noise background: This class contains the speech under the environment where noise exists in a background.

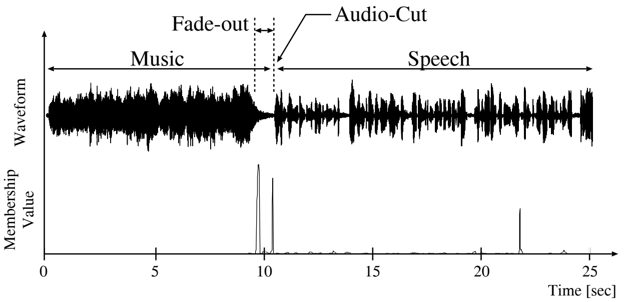


Fig. 2. An example of the audio-cut detection results: The membership value becomes higher at the beginning and the end of the fade-out.

The above five audio classes have the following features, which are utilized for our audio-segment classification.

(i) The average of the power of the audio signal (μ_E):
The average μ_E of the power sequence $E(n)$ defined in Eq. (1) is utilized for the audio-segment classification.

(ii) The variance of the power of the audio signal (σ_E^2):
The variance σ_E^2 of the power sequence $E(n)$ defined in Eq. (1) is utilized for the audio-segment classification.

(iii) The average of the center of gravity of the 0th sub-band (μ_G):
In order to observe alteration of a low frequency domain, the center of gravity in the 0th sub-band $G(n)$ is utilized. $G(n)$ is defined as follows:

$$G(n) = \frac{\sum_{j=0}^{17} j\{F_n(0,j)\}^2}{\sum_{j=0}^{17}\{F_n(0,j)\}^2}. \tag{3}$$

The average μ_G of the sequence $G(n)$ is utilized for the audio-segment classification.

(iv) The variance of the center of gravity of the 0th sub-band (σ_G^2):
The variance σ_G^2 of the sequence $G(n)$ defined in Eq. (3) is utilized for the audio-segment classification.

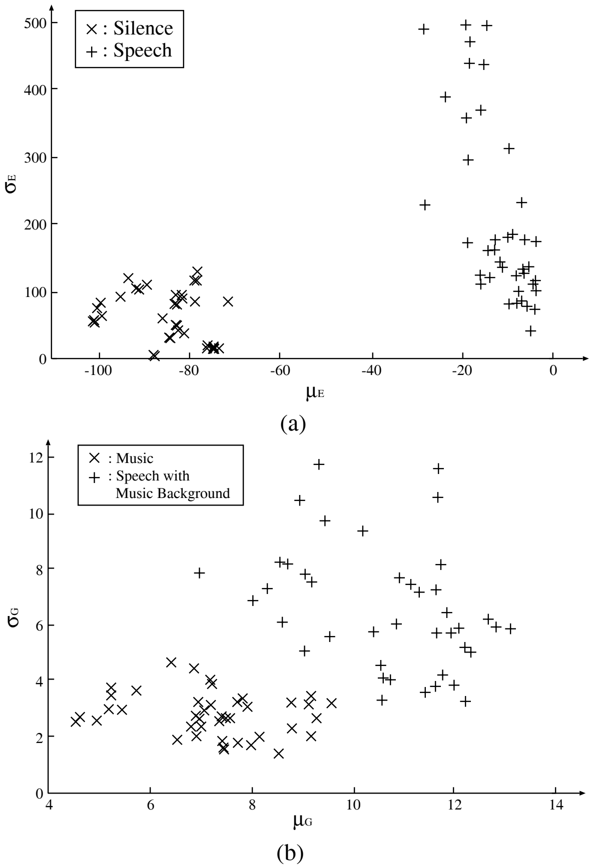
(v) The zero ratio (Z_R):
Reference [6] proposed the zero ratio Z_R in order to ascertain whether the audio-segment contains the music components or not. We utilize this feature for the audio-segment classification. Though [6] computes the zero ratio from the power spectrum by using AR coefficients; the zero ratio is computed by using MDCT coefficients in the MP3 codes because the proposed method utilizes MP3 compressed audio signal. Computing process is described below.

1) Compute the sequence $F_n'(i,j)$ which is defined as follows:

$$F_n'(i,j) = 10\log_{10}\{F_n(i,j)\}^2. \tag{4}$$

2) Smoothing is performed in each granule of $F_n'(i,j)$.

3) If there are peaks detected in consecutive sequence $F_n'(i,j)$ which stay at the same frequency level for a certain period of time, this time period is indexed as 1. Otherwise, the index value is set to 0.



(a)

(b)

Fig. 3. An example of distribution of features: (a) μ_E and σ_E^2, (b) μ_G and σ_G^2.

4) The zero ratio is computed as the ratio between the number of zeros in the indexes and the total number of the indexes, and this feature is utilized for the audio-segment classification.

An example of μ_E, σ_E^2, μ_G, and σ_G^2 is shown in Fig. 3. These features are computed from 200 clips of audio signals captured from several TV programs such as news, music, CM, etc. Each clip is 5 seconds long. As shown in Fig. 3(a), μ_E and σ_E^2 of the silence class and these of the speech class are clearly separated. This indicates that the silence class has the low level of energy compared with the speech class, and it is well-known. Thereby these two classes can be classified by watching these two parameters. Further, it is also well-known that the energy of the speech class is concentrated on the low frequency level. Therefore, as shown in Fig. 3(b), the music class and the speech with noise background class can be classified by watching μ_G and σ_G^2. The proposed method also introduces Z_R to the audio-segment classification, and thereby the speech with music background class can be isolated from the speech with noise background class.

According to the above properties, the fuzzy c-means clustering is applied to the following feature vector $\boldsymbol{V_f}$ for the audio-segment classification into the five classes:

$$\boldsymbol{V_f} = [\mu_E, \sigma_E^2, \mu_G, \sigma_G^2, Z_R]^T. \qquad (5)$$

Table 1. Experimental Results.

	Recall	Precision	Correctness
Proposed Method	98.2 [%]	93.7 [%]	92.1 [%]
Reference [6]	89.5 [%]	91.3 [%]	94.4 [%]

The feature vectors obtained from all the audio-segments are utilized for the fuzzy c-means clustering, and they are classified into the audio class whose membership value is the highest. Then if the adjacent audio-segments are classified into the same audio class, these two audio-segments are merged into one. According to the process, the wrongly detected audio-cuts can be removed.

4. EXPERIMENTAL RESULTS

In this section, we show the effectiveness of our proposed method with some simulations. In these simulations, the parameters W_1, W_2, and Δ are 40, 10, and 30, respectively.

If an audio-segment which is shorter than 5 seconds is between the two audio-segments classified into the same class, the three audio-segments are merged into one. Such a process which merges the audio-segment of such a short time is also adopted in [6].

For evaluation, we define recall, precision, and correctness as follows:

$$\text{Recall} = \frac{\text{Num. of correctly detected audio-cuts}}{\text{Num. of correct audio-cuts}} \times 100[\%]$$

$$\text{Precision} = \frac{\text{Num. of correctly detected audio-cuts}}{\text{Num. of all detected audio-cuts}} \times 100[\%]$$

$$\text{Correctness} = \frac{\text{Num. of correctly classified audio-segments}}{\text{Num. of all audio-segmnets}} \times 100[\%]$$

For the experiments, the audio signal captured from TV news program (CNN World News) at 44.1 kHz is used. This signal is 30 minutes long. The experimental results of the first five minutes of the signal are shown in Fig. 4. As shown in Fig. 4, all the audio-cuts are successfully detected and the whole audio-segments are correctly classified. Furthermore, there are two cross-fades in the audio signal illustrated in Fig. 4 and the proposed method could detect both of them. This indicates that the proposed method is robust to the audio effect.

The whole experimental results are summarized in Table 1. The audio-cut detection results by [6] are also shown in Table 1 for comparison. As shown in Table 1, the recall, precision, and correctness of the proposed method are all above 90%. This indicates that the proposed method can detect the audio-cuts and classify the audio-segments more accurately than [6].

5. CONCLUSIONS

This paper has proposed an accurate audio-cut detection and audio-segment classification method using fuzzy c-means clustering. The proposed method can effectively detect the audio-cut even if several audio effects such as fade-in, fade-out, etc. are included.

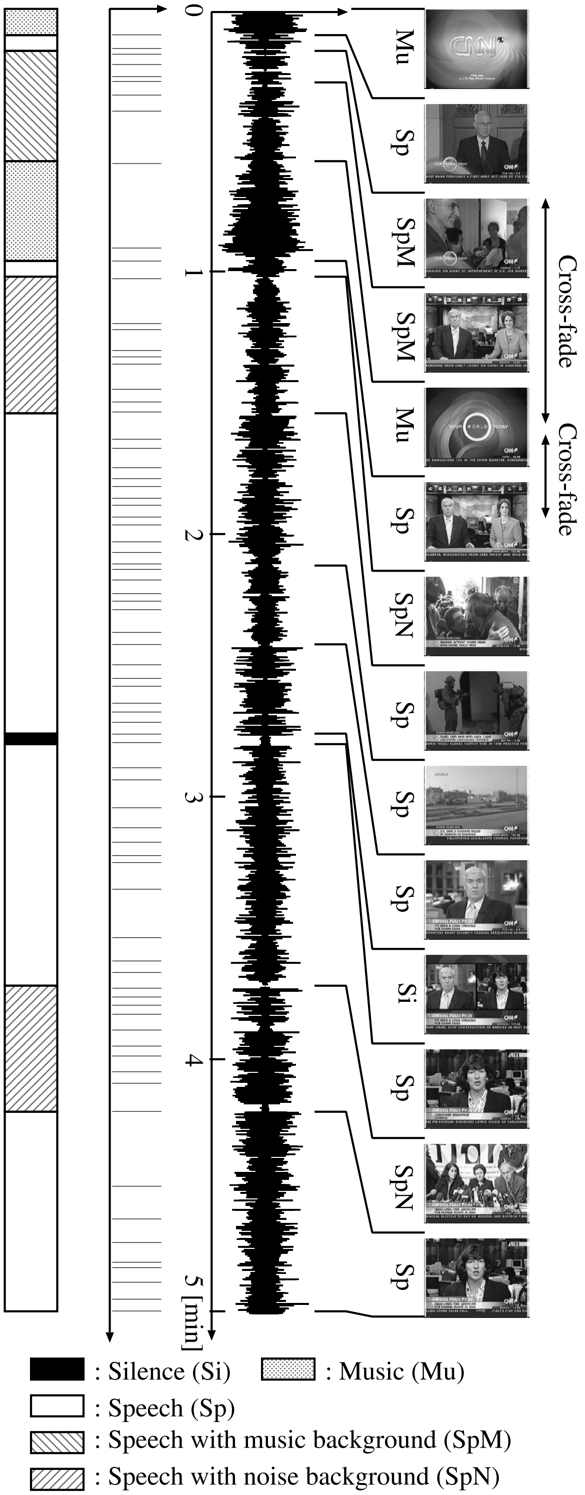


Fig. 4. The experimental results of the first five minutes: they are the contents, the wave form, the results of audio-cut detection, and the results of audio-segment classification from the right.

Since this method can directly process the MPEG audio signal, it can be easily applied to the audio-visual indexing which is compressed by MPEG.

6. REFERENCES

[1] N.V. Patel and I.K. Sethi, "Audio Characterization for Video Indexing," Storage and Retrieval for Image and Video Databases (SPIE), vol. 2670, pp. 373-384, 1996.

[2] J. Huang, Z. Liu, and Y. Wang, "Integration of Audio and Visual Information for Content-Based Video Segmentation," IEEE Int. Conf. Image Processing, 1998.

[3] G. Lu and T. Hankinson, "A Technique towards Automatic Audio Classification and Retrieval," 4th Int. Conf. Signal Processing, 1998.

[4] G. Tzanetakis and P. Cook, "Sound Analysis Using MPEG Compressed Audio," IEEE Int. Conf. Acoustics, Speech and Signal Processing, vol. 2, pp. 761-764, 2000.

[5] C. Huang and B. Liao, "A Robust Scene-Change Detection Method for Video Segmentation," IEEE Trans. Circuits and Systems for Video Technology, vol. 11, no. 12, pp. 1281-1288, 2001.

[6] T. Zhang and C.-C.J. Kuo, "Audio Content Analysis for Online Audiovisual Data Segmentation and Classification," IEEE Trans. Speech and Audio Processing, vol. 9, No. 4, 2001.