ACCURATE AUDIO-SEGMENT CLASSIFICATION USING FEATURE EXTRACTION MATRIX

Naoki Nitanda, Miki Haseyama, and Hideo Kitajima

Graduate School of Information Science and Technology, Hokkaido University
N-14 W-9 Kita-ku Sapporo 060-0814, JAPAN
{nitanda, mikich, kitajima}@md.ist.hokudai.ac.jp

ABSTRACT

This paper proposes an accurate audio signal classification method using feature extraction matrix. The proposed method classifies the segments of the audio signal into the following five audio classes: silence, speech, music, speech with music background, and speech with noise background. In this classification, a diagonal matrix, which is called feature extraction matrix, is utilized in order to extract the effective audio features for the classification. By using this feature extraction matrix, the five audio classes are clearly separated each other in the feature space, and thereby highly precise classification can be attained. Experimental results performed by applying the proposed method to real audio signals are shown to verify its high performance.

1. INTRODUCTION

Automatic segmentation and classification technique of audio signal is required for the audiovisual indexing, and some methods have been proposed[1]-[5]. They divide an audio signal into segments, which are called audio-segments in this paper, and classify them into basic audio classes, such as silence, speech, music, etc. For the audio signal segmentation, it is required to detect the boundaries between two adjacent audio-segments, which are called audio-cuts in this paper. Though these methods can detect abruptly changed audio-cuts accurately, gradually changed audio-cuts, which are caused by audio processing for several effects such as fade-in, fade-out, cross-fade, etc., cannot be detected accurately. In order to overcome this problem, we proposed the audio-cut detection and audio-segment classification method using fuzzy c-means clustering[6]. In this 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. According to this fuzzy number, not only highly reliable but also possible candidates for the audio-cuts can be obtained, and thereby both abruptly and gradually changed audio-cuts can be detected.

By the way, the audio-segment classification method proposed in [6] classifies the audio-segments into the following five audio classes: silence, speech, music, speech with music background, and speech with noise background. In this clustering, five kinds of audio features, which are the average and the variance of the power of the signal, the average and the variance of the frequency centroid, and the zero ratio proposed in [3], are utilized. Though these audio features well represent the characteristic of the audio classes, the classification error can be occurred. This is because the audio-segment classification of [6] is attained by using all the audio features in single classification procedure, and thereby five audio classes cannot be clearly separated each other in the feature space.

Therefore, we propose a new accurate audio-segment classification method. The proposed method classifies the audio-segments by using sequential classification procedures. In each classification procedure, a diagonal matrix, which is called feature extraction matrix, is utilized in order to extract the effective audio features for the classification. By using this feature extraction matrix, the five audio classes are clearly separated each other in the feature space, and thereby highly precise classification can be attained.

This paper is organized as follows. In Section 2, the audio-cut detection method proposed in [6] is summarized. In Section 3, the audio-segment classification using feature extraction matrix is proposed. Finally, in Section 4, experimental results performed by applying the proposed method to real audio signals are shown to verify its high performance.

2. AUDIO-CUT DETECTION USING FUZZY C-MEANS CLUSTERING

Since the audio-cuts must be detected before audio segmentation, we detect them by the audio-cut detection in [6] before applying our method. This method processes an audio signal that is coded by MPEG Audio Layer III (MP3), and the MDCT coefficients in the MP3 codes are utilized for the audio-cut detection. First, the power of the audio signal $E(n)$ is computed by the following equation:

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

where $F_n(i,j)$ denotes the MDCT coefficient of nth granule, ith sub-band, and jth sample. Then, the parameter sequence $C(n)$ is computed by using the power sequence $E(n)$ 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. This parameter sequence $C(n)$ is close to 0 at audio-cut, because the power of the audio signal $E(n)$ changes before and after the audio-cut. Therefore, the audio-cuts can be detected as the time when the parameter sequence $C(n)$ is close to 0. In order to judge whether the parameter sequence $C(n)$ is close to 0, we utilize the fuzzy c-means clustering, which classifies the following three vectors into

two clusters.

$$\boldsymbol{P}_n = [C(n), \cdots, C(n+W_2-1)]^{\mathrm{T}}$$
$$\boldsymbol{P}_{n-\Delta} = [C(n-\Delta), \cdots, C(n-\Delta+W_2-1)]^{\mathrm{T}}, \quad (3)$$
$$\boldsymbol{Z} = [0, \cdots, 0]^{\mathrm{T}}$$

where W_2 is a predefined window length, and T represents the transpose of a matrix. According to the clustering, the audio-cuts can be obtained as the time when the vector $\boldsymbol{P}_n$ and $\boldsymbol{Z}$ are classified into the same cluster, and then the audio-segment, whose boundaries are audio-cuts, can be obtained.

3. AUDIO-SEGMENT CLASSIFICATION USING FEATURE EXTRACTION MATRIX

Our proposed method classifies the audio-segment into silence, speech, music, speech with music background, and speech with noise background, which are the same audio classes as [6]. These five audio classes are defined as below.

- Silence: This class contains only a quasi-stationary background noise.
- Speech: This class contains the voice of human beings such as the sound of conversation.
- Music: This class contains the sound made by the musical instrument.
- 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.

For the audio-segment classification, the proposed method first computes the seven kinds of audio features, and then applies the fuzzy c-means clustering to the audio features. Each part is described in the following subsections.

3.1. Computation of Audio Features

For the audio-segment classification, we first compute the following seven kinds of audio features, which include the same audio features as [6] and newly added ones.

(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 average of the log scaled power of the audio signal μ_{LE}: The average μ_{LE} of the log scaled power sequence $LE(n)$ is utilized for the audio-segment classification. $LE(n)$ is defined as follows:

$$LE(n) = 10\log_{10} E(n). \quad (4)$$

(iii) The variance of the log scaled power of the audio signal σ^2_{LE}: The variance σ^2_{LE} of the log scaled power sequence $LE(n)$ defined in Eq. (4) is utilized for the audio-segment classification.

(iv) The variance of the frequency centroid σ^2_{FC}: The variance σ^2_{FC} of the frequency centroid sequence $FC(n)$ is utilized for the audio-segment classification. $FC(n)$ is defined as follows:

$$FC(n) = \frac{\sum_{k=0}^{575} kF'_n(k)}{\sum_{k=0}^{575} F'_n(k)}, \quad (5)$$

where

$$F'_n(18i+j) = 10\log_{10}\{F_n(i,j)\}^2. \quad (6)$$

Table 1. The diagonal elements a^i_k of the feature extraction matrix $\boldsymbol{M}_i$.

	a^i_1	a^i_2	a^i_3	a^i_4	a^i_5	a^i_6	a^i_7
$\boldsymbol{M}_1$ $(i=1)$	1	1	0	0	0	0	0
$\boldsymbol{M}_2$ $(i=2)$	0	0	1	1	0	0	1
$\boldsymbol{M}_3$ $(i=3)$	0	0	1	0	0	1	0
$\boldsymbol{M}_4$ $(i=4)$	0	0	0	1	1	0	0

(v) The variance of the frequency bandwidth σ^2_{BW}: The variance σ^2_{BW} of the frequency bandwidth sequence $BW(n)$ is utilized for the audio-segment classification. $BW(n)$ is defined as follows:

$$BW(n) = \sqrt{\frac{\sum_{k=0}^{575}\{k - FC(n)\}^2 F'_n(k)}{\sum_{k=0}^{575} F'_n(k)}}. \quad (7)$$

(vi) The variance of the fuzzy number σ^2_{FN}: The variance σ^2_{FN} of the fuzzy number obtained in Section 2 is utilized for the audio-segment classification.

(vii) The zero ratio Z_R: Reference [3] 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 [3] 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) Smoothing is performed in each granule of $F'_n(k)$ defined in Eq. (6).
2) If there are peaks detected in consecutive sequence $F'_n(k)$, 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.
3) The zero ratio is computed as the ratio between the number of zeros in the indexes and the total number of the indexes.

3.2. Audio-Segment Classification

The proposed method classifies the audio-segments by using the sequential classification procedures, which are illustrated in Fig. 1. For the clustering, the feature vector $\boldsymbol{f}$ is defined as follows:

$$\boldsymbol{f} = [\mu_E, \mu_{LE}, \sigma^2_{LE}, \sigma^2_{FC}, \sigma^2_{BW}, \sigma^2_{FN}, Z_R]^{\mathrm{T}}. \quad (8)$$

Since both effective and ineffective audio features for each classification are contained in $\boldsymbol{f}$, the proposed method utilizes the feature extraction matrix $\boldsymbol{M}_i \triangleq diag[a^i_1, \cdots, a^i_k]$, which is indicated in Table 1, in order to control the selection of the audio features for the classification. By using this feature extraction matrix, the effective audio features for the clustering are extracted, and thereby highly precise classification can be attained. The details of each classification procedure are described below.

(i) Preparing the model feature vector: Before classifying the audio-segments, feature vectors are computed from the training data, whose audio classes are already known. These

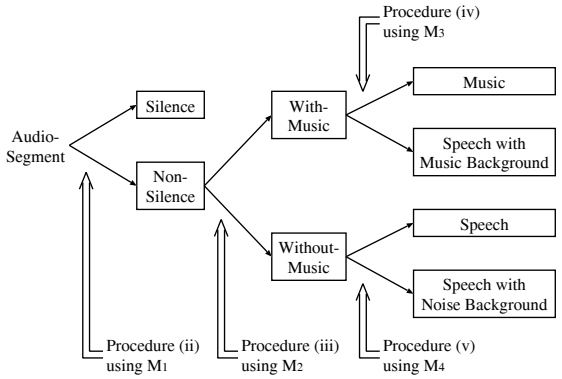

Fig. 1. The procedures of the audio-segment classification.

feature vectors of silence, speech, music, speech with music background, and speech with noise background are represented as $\boldsymbol{f}_{Si}$, $\boldsymbol{f}_{Sp}$, $\boldsymbol{f}_{Mu}$, $\boldsymbol{f}_{SpMu}$, and $\boldsymbol{f}_{SpNo}$, respectively.

(ii) Separating silence and non-silence using feature extraction matrix $\boldsymbol{M}_1$: The first step is to separate the audio-segments into silence and non-silence by using the feature extraction matrix $\boldsymbol{M}_1$. As shown in Table 1, this matrix $\boldsymbol{M}_1$ extracts μ_E and μ_{LE} from $\boldsymbol{f}$, and these two features well represent the characteristic of the silence class, that is, the silence class has low power level compared to the other audio classes. By using the matrix $\boldsymbol{M}_1$, the following three vectors are computed: $\boldsymbol{V}_1$ $(= \boldsymbol{M}_1\boldsymbol{f})$, $\boldsymbol{V}_{Si}$ $(= \boldsymbol{M}_1\boldsymbol{f}_{Si})$, and $\boldsymbol{V}_{Non\text{–}Si}$ $(= \frac{1}{4}\boldsymbol{M}_1(\boldsymbol{f}_{Sp}+\boldsymbol{f}_{Mu}+\boldsymbol{f}_{SpMu}+\boldsymbol{f}_{SpNo}))$. These three vectors are classified into two clusters by using fuzzy c-means clustering. When $\boldsymbol{V}_1$ and $\boldsymbol{V}_{Si}$ are classified into the same cluster, we judge that the audio-segment is the silence. Otherwise we judge that the audio-segment is the non-silence. According to this procedure, the audio-segments are separated into silence and non-silence.

(iii) Separating sounds with/without music using feature extraction matrix $\boldsymbol{M}_2$: The second step is to separate the non-silence audio-segments into the audio-segments which contain music and which do not contain music by using the feature extraction matrix $\boldsymbol{M}_2$. As shown in Table 1, this matrix $\boldsymbol{M}_2$ extracts σ^2_{LE}, σ^2_{FC}, and Z_R from $\boldsymbol{f}$, and these three features well represent the characteristic of the music, that is, the sound made by the musical instruments usually exists continuously and the peaks in consecutive power spectra are inclined to stay at the same frequency level for a certain period of time. By using the matrix $\boldsymbol{M}_2$, the following three vectors are computed: $\boldsymbol{V}_2$ $(= \boldsymbol{M}_2\boldsymbol{f})$, $\boldsymbol{V}_{with}$ $(= \frac{1}{2}\boldsymbol{M}_2(\boldsymbol{f}_{Mu}+\boldsymbol{f}_{SpMu}))$, and $\boldsymbol{V}_{without}$ $(= \frac{1}{2}\boldsymbol{M}_2(\boldsymbol{f}_{Sp}+\boldsymbol{f}_{SpNo}))$. These three vectors are classified into two clusters by using fuzzy c-means clustering. When $\boldsymbol{V}_2$ and $\boldsymbol{V}_{with}$ are classified into the same cluster, we judge that the audio-segment contains music. Otherwise we judge that the audio-segment does not contain music. According to this procedure, the non-silence audio-segments are separated into the audio-segments which contain music and which do not contain music.

(iv) Separating music and speech with music background using feature extraction matrix $\boldsymbol{M}_3$: The third step is to separate the audio-segments, which contain music, into the music and the speech with music background by using the feature extraction matrix $\boldsymbol{M}_3$. As shown in Table 1, this matrix $\boldsymbol{M}_3$ extracts σ^2_{LE} and σ^2_{FN} from $\boldsymbol{f}$, and these two features well represent the characteristic of the speech class, that is, the speech class has large fluctuation in the power of the audio signal. By using the matrix $\boldsymbol{M}_3$, the following three vectors are computed: $\boldsymbol{V}_3$ $(= \boldsymbol{M}_3\boldsymbol{f})$, $\boldsymbol{V}_{Mu}$ $(= \boldsymbol{M}_3\boldsymbol{f}_{Mu})$, and $\boldsymbol{V}_{SpMu}$ $(= \boldsymbol{M}_3\boldsymbol{f}_{SpMu})$. These three vectors are classified into two clusters by using fuzzy c-means clustering. When $\boldsymbol{V}_3$ and $\boldsymbol{V}_{Mu}$ are classified into the same cluster, we judge that the audio-segment is the music. Otherwise we judge that the audio-segment is the speech with music background. According to this procedure, the audio-segments, which contain music, are separated into music and speech with music background.

(v) Separating speech and speech with noise background using feature extraction matrix $\boldsymbol{M}_4$: The last step is to separate the audio-segments, which do not contain music, into the speech and the speech with noise background by using feature extraction matrix $\boldsymbol{M}_4$. As shown in Table 1, this matrix $\boldsymbol{M}_4$ extracts σ^2_{FC} and σ^2_{BW} from $\boldsymbol{f}$, and these two features well represent the characteristic of the speech with noise background, that is, the frequency centroid and bandwidth change intricately compared to the pure speech. By using the matrix $\boldsymbol{M}_4$, the following three vectors are computed: $\boldsymbol{V}_4$ $(= \boldsymbol{M}_4\boldsymbol{f})$, $\boldsymbol{V}_{Sp}$ $(= \boldsymbol{M}_4\boldsymbol{f}_{Sp})$, and $\boldsymbol{V}_{SpNo}$ $(= \boldsymbol{M}_4\boldsymbol{f}_{SpNo})$. These three vectors are classified into two clusters by using fuzzy c-means clustering. When $\boldsymbol{V}_4$ and $\boldsymbol{V}_{No}$ are classified into the same cluster, we judge that the audio-segment is the speech. Otherwise we judge that the audio-segment is the speech with noise background. According to this procedure, the audio-segments, which do not contain music, are separated into speech and speech with noise background.

By applying the above-mentioned four separations, the audio-segments are certainly classified into one of the five audio classes, and thereby the audio-segment classification is attained completely.

4. EXPERIMENTAL RESULTS

In this section, we show the effectiveness of our proposed method with some simulations. For the experiments, the three kinds of audio signals captured from TV programs, which are news, music, and drama program, are used. Each signal is 10 minutes long, and 30 minutes in all. Experimental results of the first two minutes of the news program are shown in Fig. 2. As shown in Fig. 2, a classification error is occurred in the classification results of [6]. On the other hand, the proposed method classifies the all audio-segments into the correct audio class successfully.

The whole experimental results of the proposed method and [6] are summarized in Table 2. For evaluation, we define recall and precision rates as follows:

$$\text{Recall} = \frac{\text{Num. of correctly classified audio-segments}}{\text{Num. of manually classified audio-segments}}, \quad (9)$$

$$\text{Precision} = \frac{\text{Num. of correctly classified audio-segments}}{\text{Num. of all audio-segments}}. \quad (10)$$

By the way, it is required that both recall and precision rates to be high for the accurate audio-segment classification. Therefore, we

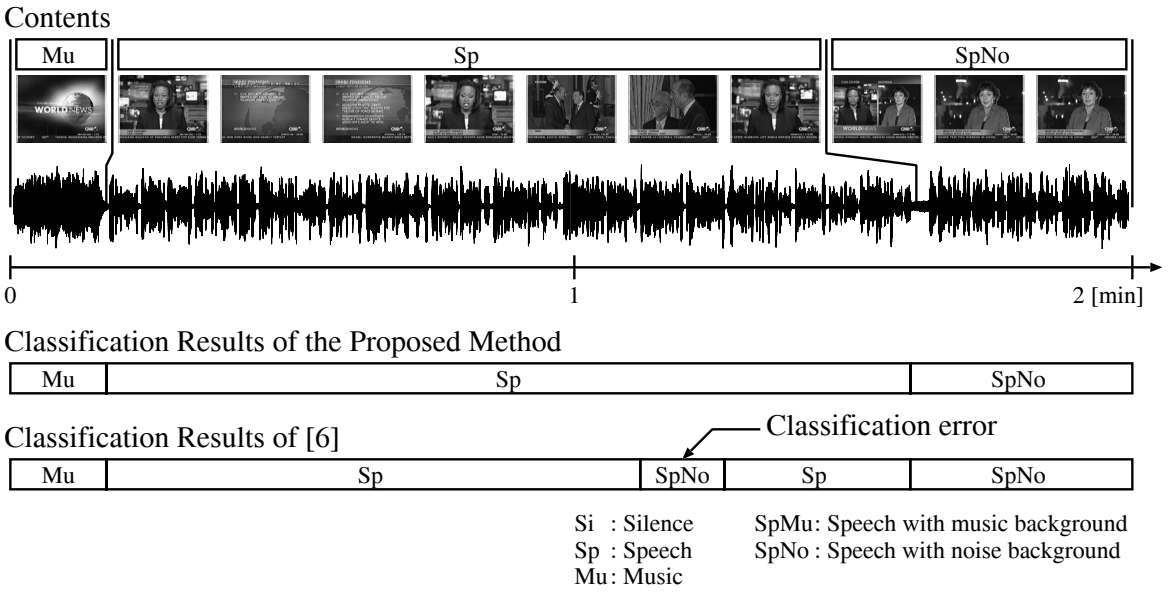

Fig. 2. The experimental results of the first two minutes of the news program: they are the contents, the waveform, the results of the audio-segment classification by using the proposed method, and the results of the audio-segment classification by using [6] from the top.

Table 2. Experimental results of whole audio signals.

	Precision		Recall		F-measure	
	Proposed	Reference [6]	Proposed	Reference [6]	Proposed	Reference [6]
Silence	1.000	1.000	1.000	0.900	1.000	0.947
Speech	0.927	0.824	0.905	0.875	0.916	0.848
Music	0.961	0.929	0.881	0.867	0.919	0.896
Speech with Music Background	0.913	0.818	0.913	0.818	0.913	0.818
Speech with Noise Background	0.931	0.905	0.947	0.950	0.939	0.927
Average	0.946	0.895	0.929	0.882	0.937	0.887

utilize not only recall and precision rates, but also the F-measure for the evaluation. F-measure is defined as follows:

$$\text{F-measure} = \frac{2}{\frac{1}{\text{Recall}} + \frac{1}{\text{Precision}}}. \quad (11)$$

As shown in Table 2, the proposed method improves the accuracy of the audio-segment classification in all the audio classes compared to the method of [6]. Moreover, the F-measure of the proposed method is all above 0.90. This indicates that the proposed method can provide enough accuracy for the audio signal classification.

5. CONCLUSIONS

This paper has proposed an accurate audio-segment classification method. The proposed method utilizes the feature extraction matrix in order to extract the effective audio features for the classification. By using this feature extraction matrix, the five audio classes are clearly separated each other in the feature space, and thereby highly precise classification can be attained.

6. REFERENCES

[1] Z. Liu, Y. Wang, and T. Chen, "Audio Feature Extraction and Analysis for Scene Segmentation and Classification, " Journal of VLSI Signal Processing, vol. 20, no. 1–2, pp.61–69, 1998.

[2] D. Li, I.K. Sethi, N. Dimitrova, and T. McGee, "Classification of General Audio Data for Content-Based Retrieval," Pattern Recognition Letters, vol. 22, no. 5, pp. 533–544, 2001.

[3] 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, pp. 441–457, 2001.

[4] L. Lu, H. Zhang, and H. Jiang, "Content Analysis for Audio Classification and Segmentation, " IEEE Trans. Speech and Audio Processing, vol. 10, no. 7, pp. 504–516, 2002.

[5] L. Lu, H. Zhang, and S.Z. Li, "Content-Based Audio Classification and Segmentation by using Support Vector Machines, " Multimedia Systems, vol. 8, no. 6, pp. 482–492, 2003.

[6] N. Nitanda, M. Haseyama, and H. Kitajima, "Audio-Cut Detection and Audio-Segment Classification Using Fuzzy C-Means Clustering, " Proc. ICASSP, vol. IV, pp. 325–328, 2004.