Multimed Tools Appl (2013) 63:485–500
DOI 10.1007/s11042-011-0921-z

An enhanced fuzzy c-means algorithm for audio segmentation and classification

Mohammad A. Haque · Jong-Myon Kim

Published online: 18 November 2011

Abstract Automated audio segmentation and classification play important roles in multimedia content analysis. In this paper, we propose an enhanced approach, called the correlation intensive fuzzy c-means (CIFCM) algorithm, to audio segmentation and classification that is based on audio content analysis. While conventional methods work by considering the attributes of only the current frame or segment, the proposed CIFCM algorithm efficiently incorporates the influence of neighboring frames or segments in the audio stream. With this method, audio-cuts can be detected efficiently even when the signal contains audio effects such as fade-in, fade-out, and cross-fade. A number of audio features are analyzed in this paper to explore the differences between various types of audio data. The proposed CIFCM algorithm works by detecting the boundaries between different kinds of sounds and classifying them into clusters such as silence, speech, music, speech with music, and speech with noise. Our experimental results indicate that the proposed method outperforms the state-of-the-art FCM approach in terms of audio segmentation and classification.

Keywords Audio segmentation and classification · Fuzzy c-means algorithm · Multimedia · Database retrieval

1 Introduction

Multimedia databases usually store thousands of audio recordings such as music, speech, and other sounds. The immense amounts of audio data in these domains necessitate the development of computerized methods for efficient, automated, content-based segmentation and classification of audio data [14]. Such methods have important applications in professional media production, audiovisual archive management, education, entertainment, and surveillance [20]. There are two major applications of audio content analysis. One is to segment an audio stream into a number of constituent audio streams; the other is to classify

M. A. Haque · J.-M. Kim (✉)
School of Electrical Engineering, University of Ulsan, Ulsan, South Korea 689-749
e-mail: jongmyon.kim@gmail.com

audio streams into different sound classes such as speech, music, environmental sound, and silence. A number of methods have been proposed to address issues inherent in audio segmentation and classification, such as detecting audio-cuts at which the audio signal changes, selecting audio-segments using the audio-cuts, and classifying segments into different types of audio groups [6, 8–10, 12–15, 17, 18, 20]. These methods are based on both perceptual and acoustical features.

Previously proposed audio segmentation methods can be categorized into two groups depending on how they detect and segment audio-cuts: (1) threshold processing based approaches [6, 9, 10, 20] and (2) fuzzy based approaches [12]. Threshold processing based approaches detect audio-cuts by applying threshold comparisons either to audio features or to differences between audio features at two different times. These methods are therefore subject to performance degradation when segmenting audio streams that contain audio effects such as fade-in, fade-out, and cross-fade. In previous studies, Zhang et al. [20] proposed a heuristic rule-based procedure for audio classification along with threshold processing based segmentation, which is based on morphological and statistical analysis of the time-varying functions of simple audio features including the energy function, average zero-crossing rate, fundamental frequency, and spectral peak tracks to ensure the feasibility of real-time processing while scarifying the accuracy of audio classification. Liu et al. [9] also used the threshold processing based method in audio segmentation, proposing a set of low-level audio features extracted from the time and frequency domains and then using these features as input for a neural net classifier. However, only non-speech and non-music audio data from TV programs were considered, and therefore the generality of their system was not demonstrated. Lu et al. [10] proposed a step by step classification method for classifying audio streams into different categories. They introduced a set of new audio features in both the energy and frequency domains and proposed a segmentation algorithm that is based on quasi-GMM and line spectral pair (LSP) correlation analysis. This method can segment audio streams of open-set speakers in real time without a priori knowledge about particular speakers' speech characteristics. Gang et al. [6] proposed a classification-independent segmentation (CIS) method that calculates the similarities between audio feature vectors. All of these methods are based on threshold processing in audio-cut detection and are consequently vulnerable to the aforementioned problem of performance degradation when audio streams contain sound effects. To overcome this problem, a fuzzy method was proposed by Noaki et al. [12]. This is a soft-segmentation method that utilizes the fuzzy c-means (FCM) algorithm for audio-cut detection.

A number of other methods have been also proposed, which simply consider issues related to audio stream classification. Wold et al. [18] presented an audio retrieval system, and their study is treated as a milestone because it presented a method of content based audio analysis that distinguished it from previous works [17]. Statistical values for several time and frequency domain measurements are used in Wold et al.'s method to represent perceptual features like loudness, brightness, bandwidth, and pitch. Since this method considers only statistical values, it is only suitable for classifying sounds with a single timbre. Audio classification by support vector machine (SVM) methods was proposed in [17], in which Mel-frequency cepstral coefficients (MFCC) are taken as features. Since MFCCs do not accurately represent the timbres of sounds, this method fails to distinguish music and environmental sounds with different timbre characters. Audio classification by a Hidden Markov Model (HMM) approach was proposed in [8]. However, this method requires prior training data, which decreases the robustness of the system. Park et al. proposed three different fuzzy methods for classification of audio in [13–15]. These methods utilize the Gradient-Based FCM algorithm (GBFCM) for soft classification of

audio contents. However, the FCM-based methods proposed in [12–15] do not consider the temporal relationships of audio features between audio frames or segments in segmentation or classification, respectively. Therefore, further research is warranted with the goal of improving the performance of GBFCM methods.

Through our intensive study of different methods of audio segmentation and classification, we observed that the FCM algorithm can effectively detect audio-cuts even if the audio signal contains fade-in, fade-out and cross-fade. The FCM algorithm interprets the existence of audio-cuts as real values between *0* and *1* and thereby detects segments. Further, it subdivides audio-segments into different audio classes such as silence, speech, music, speech with music background, and speech with noise. To improve the performance of audio segmentation and classification based on audio content analysis; this paper proposes a correlation intensive FCM (CIFCM) algorithm that employs the features of audio and incorporates the concepts of FCM based methods proposed in [12–15]. Unlike conventional FCM clustering approaches, the CIFCM algorithm utilizes temporal correlation information between neighboring frames and segments in the context of the current frame and segment for audio-cut detection and classification. We assess recall rate and precision rate to evaluate the performance of the CIFCM method for audio-cut detection and segmentation [12]. We analyze the classification performance of the proposed method on a labeled audio-segment dataset of five broad audio genres. In addition, we employ the four eminent cluster validity functions that are summarized in [11] to evaluate the performance of CIFCM for audio classification. Our experimental results indicate that the proposed CIFCM algorithm outperforms the conventional FCM-based method [12] for audio segmentation and classification.

The rest of this paper is organized as follows. Section 2 describes the conventional FCM algorithm and cluster validity functions. Section 3 introduces the proposed CIFCM algorithm for audio segmentation and classification, and Section 4 presents experimental results. Section 5 concludes the paper.

2 Background information

2.1 Fuzzy c-means algorithm

The FCM algorithm is an unsupervised clustering method. It was developed by Dunn in 1973, and revised by Bezdek in 1981 [1]. The FCM algorithm has been successfully applied to feature analysis, clustering, and especially pattern recognition [2, 7]. The effects of correlations between the features of the current experimental data and those of neighboring data in conventional FCM clustering were intensively analyzed in [11].

The conventional FCM algorithm is an iterative method of clustering that allows one piece of data to belong to two or more clusters. Let an unlabelled data set $X=(x_1, x_2, x_3, \ldots, x_n)$ represent the features of the n items. The FCM algorithm sorts the data set X into c clusters. The standard FCM objective function with the Euclidian distance metric is defined as follows:

$$\mathrm{J_m}(U,V)=\sum_{i=1}^{c}\sum_{k=1}^{n}u_{ik}^{m}d^2(x_k,v_i) \tag{1}$$

where $d^2(x_k, v_i)$ represents the Euclidian distance between the data point x_k and the center v_i of the i-th cluster, and u_{ik} is the degree of membership of the data x_k to the k-th cluster, along with the constraint $\sum_{i=1}^{c}u_{ik}=1$. The parameter m controls the fuzziness of the resulting partition, with $m\geq 1$, and c is the total number of clusters. Local minimization of the objective

function $J_m(U,V)$ is accomplished by repeatedly adjusting the values of u_{ik} and v_i according to the following equations:

$$u_{ik} = \left[\sum_{j=1}^{c} \left(\frac{d^2(x_k, v_i)}{d^2(x_k, v_j)} \right)^{\frac{1}{m-1}} \right]^{-1} \tag{2}$$

$$v_i = \frac{\sum_{k=1}^{n} u_{ik}^m x_k}{\sum_{k=1}^{n} u_{ik}^m}, 1 \le i \le c \tag{3}$$

As J_m is iteratively minimized, v_i becomes more stable. The iteration of the FCM algorithm is terminated when the ending condition $\max_{1 \le i \le c} \left\{ abs\left(v_i^t - v_i^{t-1}\right) \right\} < \varepsilon$ is satisfied, where $v^{(t-1)}$ are the centers of the previous iteration, *abs()* stands for the absolute value, and ε is the predefined termination threshold. Finally, all data points are distributed into clusters according to the maximum membership u_{ik}. In addition, the fuzzy partition matrix U is congregated for further operations to evaluate the efficiency of the clustering.

2.2 Cluster validity functions

Two important types of cluster validity functions are used for the quantitative evaluation of cluster performance; they are based on the fuzzy partitions [3] and the feature structure of the data set [5, 19]. The fuzzy partitions use two parameters: Bezdek's partition coefficient v_{pc} and the partition entropy v_{pe} [3], which are defined as follows:

$$v_{pc}(U) = \sum_{j=1}^{n} \sum_{i=1}^{c} u_{ij}^2 \tag{4}$$

$$v_{pe}(U) = -\frac{1}{n} \left[\sum_{j=1}^{n} \sum_{i=1}^{c} \left(u_{ij} \log\left(u_{ij}\right) \right) \right] \tag{5}$$

When v_{pc} is maximal or v_{pe} is minimal, optimal clustering is achieved. However, these two parameters depend only upon the membership values of data in the clusters, not the data themselves. To overcome this shortcoming, two other validity functions based on the feature structure of data set have been proposed: the Fukuyama-Sugeno function v_{fs} [5] and the Xie-Beni function v_{xb} [19]. These are defined as follows:

$$v_{fs}(U, V, X) = \sum_{i=1}^{c} \sum_{j=1}^{n} u_{ij}^m \left(\left\| x_j - v_i \right\|^2 - \left\| v_i - \overline{v} \right\|^2 \right) \tag{6}$$

$$v_{xb}(U) = \frac{\sum_{i=1}^{c} \sum_{j=1}^{n} u_{ij}^2 \left\| x_j - v_i \right\|^2}{n \left(\min_{i \ne k} \left(\left\| v_i - v_k \right\|^2 \right) \right)} \tag{7}$$

where $\overline{v} = \frac{1}{c} \sum_{i=1}^{c} v_i$. The smaller the values of v_{fs} or v_{xb}, the better the clustering results.

The aforementioned four cluster validity functions are used as the bases for comparing the performance of the proposed CIFCM and the conventional FCM [12] for audio segmentation and classification.

3 The correlation intensive fuzzy c-means algorithm (CIFCM) for audio segmentation and classification

3.1 The proposed CIFCM algorithm

The traditional FCM algorithm for audio segmentation operates by detecting audio-cuts in a frame using the attributes of only that frame [12]. It classifies each segment in an audio stream using only the features of that segment [12–15]. However, the general aspects of an audio frame or segment are highly correlated with those of neighboring frames or segments due to the similarities of the temporal features. This leads to accuracy degradation in the segmentation and classification procedures. This aspect of FCM was comprehensively explored by Luong et al. in the image segmentation domain [11]. In audio segmentation, when frames contain sound effects such as fade-in, fade-out, or cross-fade; then abrupt changes (i.e., audio-cuts) in the signal cannot be detected from the differences between two consecutive frames. Therefore, it is important to consider the impact of changes of neighboring frames within a specified window length. Similar conditions also occur when classifying audio segments. To solve this problem, which is inherent in audio segmentation and classification, we propose the CIFCM algorithm that utilizes not only attributes of the current frame or segment but also considers the memberships of its neighboring frames or segments, respectively, by modifying the membership function of the traditional FCM algorithm. The membership of each data element is calculated as a weighted sum of the current element membership and the memberships of the previous and following neighboring elements in a window length of w_f, where the center element x_k is the current element.

In CIFCM, a neighboring impact factor, called P_{ik}, is used to consider temporal information about the neighbors to determine the fit of the data element x_k in the cluster i. The smaller the distances between the center element and its neighbors, the higher the probability that a given element and its neighbors are in the same cluster. The neighboring impact factor is defined as follows:

$$p_{ik} = \sum_{j=k-\frac{w_f}{2}}^{k+\frac{w_f}{2}} h\left(x_k, x_j\right) u_{ij} \tag{8}$$

where the function $h(x_k, x_j)$ is the distance coefficient between the center element x_k and the neighbor x_j; and u_{ij} is the membership value of the neighbor x_j in the cluster i as described in Section 2 for conventional FCM. To assign an appropriate function of $h(x_k, x_j)$ as shown in (9), we define hypotheses as follows:

1. The neighbor impact factor P_{ik} ranges from [0:1] with j in the range of $k - \frac{w_f}{2} : k + \frac{w_f}{2}$, indicates the neighbor elements.
2. If all elements in the range of w_f belong completely to cluster i, then the impact factor value P_{ik}=1. This implies that this segment is mostly impacted by its neighbors.

To determine the function $h(x_k, x_j)$, it is assumed that u_{ij} =1. As a result $\sum_{j=k-\frac{w_f}{2}}^{k+\frac{w_f}{2}} h\left(x_k, x_j\right) = 1$ when the neighbor impact factor P_{ik} =1. The function $h(x_k, x_j)$ is defined as follows:

$$h\left(x_k, x_j\right) = \left[\sum_{l=k-\frac{w_f}{2}}^{k+\frac{w_f}{2}} \left(\frac{d^2\left(x_k, x_j\right)}{d^2\left(x_k, x_l\right)}\right)\right]^{-1} \tag{9}$$

where longer distances between x_k and x_j generates smaller values of $h(x_k, x_j)$.

Subsequently, the function p_{ik} is defined as follows:

$$p_{ik} = \left(\sum_{l=k-\frac{w_f}{2}}^{k+\frac{w_f}{2}} \frac{1}{d^2(x_k, x_l)}\right)^{-1} \left(\sum_{j=k-\frac{w_f}{2}}^{k+\frac{w_f}{2}} \frac{u_{ij}}{d^2(x_k, x_j)}\right) \tag{10}$$

To make use of this impact factor, we include it in the distance measurement of the conventional FCM in (2), and generate a new distance function as follows, in place of simple Euclidian distance:

$$d^2_{new}(x_k, v_i) = d^2(x_k, v_i)(f(p_{ik}))^{-1} \tag{11}$$

where $f(P_{ik})$ indicates the function of P_{ik} which is $(1/P_{ik})$in this study. Thus, the new membership function using the new distance metric is calculated as:

$$w_{ik} = \left[\sum_{j=1}^{c} \left(\frac{d^2_{new}(x_k, v_i)}{d^2_{new}(x_k, v_j)}\right)^{\frac{1}{m-1}}\right]^{-1} \tag{12}$$

By simplifying (12), we obtain the membership function for CIFCM in (13) and the center of clusters in (14) in place of (2) and (3), respectively, in the conventional FCM algorithm:

$$w_{ik} = \frac{u_{ik}(f(p_{ik}))^{\frac{1}{m-1}}}{\sum_{j=1}^{c} u_{jk}(f(p_{ik}))^{\frac{1}{m-1}}} \tag{13}$$

$$v_i = \frac{\sum_{k=1}^{n} w_{ik}^m x_k}{\sum_{k=1}^{n} w_{ik}^m}, 1 \le i \le c \tag{14}$$

The steps of CIFCM for audio segmentation and classification are summarized as follows:

1. Distribute the data elements of the audio stream into data set X and initiate the center values $V^0 = (v_1^0, v_2^0, ..., v_c^0)$.
2. Calculate the membership values u_{ik} from (2) and the impact factor P_{ik} from (10).
3. Compute the new membership values w_{ik} from (13).
4. Calculate the new center values of the clusters using (14).
5. Evaluate the termination condition $\max_{1 \le i \le c} \{abs(v_i^t - v_i^{t-1})\} < \varepsilon$ The process is finished if this condition is satisfied, otherwise repeat the process starting with step 2.
6. Assign each data element to clusters according to their maximum memberships in the clusters as derived from the membership matrix $U_{c \times N}$.

The proposed CIFCM is further used for audio segmentation and classification.

3.2 Audio segmentation and classification using CIFCM

In order to segment and classify an audio stream, we must initially extract the audio features. After rigorous studies of different audio features as described in [12, 20] and [16], we calculate some characteristic parameters of audio signals as follows:

1. The power of the audio signal $E(n)$ in a frame of w_l samples is given as follows:

$$E(n) = \frac{1}{w_l} \sum_{k=1}^{w_l} \left(\frac{sig_n(k)}{\max(abs(sig_n))}\right)^2 \tag{15}$$

where n is the index of the frames in the signal, $sig_n(k)$ is the amplitude of the k^{th} sample in the n^{th} frame, and sig_n is the array of all sample values within the n^{th} frame. This provides a convenient representation of the amplitude variation in the signal over time [20].

2. The parameter sequence $C(n)$ is defined as:

$$C(n) = \frac{\sum_{j=0}^{w_c-1} E(n+j)E(n-w_c+j)}{\sqrt{\sum_{j=0}^{w_c-1} E(n+j)^2 \sum_{j=0}^{w_c-1} E(n-w_c+j)^2}} \tag{16}$$

where w_c indicates the number of frames in a pre-specified window. This feature is useful for identifying abrupt changes in the signal [12]. When the value of $C(n)$is closer to 0, the possibility of the existence of an audio-cut in the n^{th} frame is increased. The length w_c must be set in such a way that multiple audio-cuts do not occur within a window.

3. The mean μ_E and the standard deviation σ_E of the power sequence $E(n)$. The patterns of variation of these two values help to classify audio signals into different groups.
4. The center of gravity $G(n)$ is a parameter that observes alterations of signal in a low frequency domain [12] and is computed as follows:

$$G(n) = \frac{\sum_{k=1}^{w_l} k \times \{F_n(k)\}^2}{\sum_{k=1}^{w_l} \{F_n(k)\}^2} \tag{17}$$

where $F_n(k)$ is the Fourier transform coefficient of the k^{th} sample in the n^{th} frame.

5. The mean μ_G and the standard deviation σ_G of the center of gravity $G(n)$ facilitate the analysis of the spectral shape of the audio data [16].
6. If successive samples have different signs, a zero-crossing occurs in a discrete time signal. The rate at which zero crossing occurs is a simple measure of the frequency content of a signal. The zero-crossing rate $Z(n)$ is calculated as follows [20]:

$$Z(n) = \frac{1}{w_l} \sum_{k=1}^{w_l} \frac{1}{2} \{\mathrm{sgn}[sig_n(k)] - \mathrm{sgn}[sig_n(k-1)]\} \tag{18}$$

where $\mathrm{sgn}[sig_n(k)] = \begin{cases} -1, & sig_n(k)<0 \\ 1, & sig_n(k)\geq 0 \end{cases}$

7. The zero ratio Z_R, which is defined as the ratio of the number of zero indices to the total number of indices in a signal, plays an important role in measuring the noisiness in audio signals of different classes [16]. It can be derived from (18) as shown below:

$$Z_R = \frac{1}{N} \sum_{n=1}^{N} Z(n), \tag{19}$$

where N is the number of frames in the signal.

Audio-cut detection and segmentation: Audio-cuts can be detected efficiently by observing the parameter sequence $C(n)$ [12]. Therefore, the proposed CIFCM utilizes $C(n)$ to detect audio-cuts. In audio signal segmentation, three vectors defined in (20), (21) and (22) are grouped into two clusters by applying the proposed CIFCM algorithm:

$$\mathbf{P}_n = [C(n), ..., C(n+w_x-1)]^T \tag{20}$$

$$\mathbf{P}_{n-\nabla} = [C(n-\nabla), ..., C(n-\nabla+w_x-1)]^T \tag{21}$$

$$\mathbf{Z} = [0, ..., 0]^T \tag{22}$$

where w_x is the number of frames in a predefined window, and ∇ and T represent the step size of the window and the transpose of the matrix, respectively. We determine the values of w_x and ∇ analytically so that audio-cuts do not occur simultaneously within the periods from n to $(n + w_x - 1)$ and from $(n–\nabla)$ to $(n - \nabla + w_x - 1)$ as shown by the windows L_1 and R_1 in Fig. 1.

Audio-cuts can be detected by the membership of $\mathbf{P}_n$ in the cluster of z. When an audio-cut exists in the time interval from n to $(n + w_x - 1)$, for instance as illustrated by the window R_1 in Fig. 1, the element $C(n)$ of vector $\mathbf{P}_n$ closes to *0*, and all elements of vector $\mathbf{P}_{n–\nabla}$ become sufficiently high compared to *0*. Since the elements of $\mathbf{P}_n$ are closer to *0* than those of $\mathbf{P}_{n–\nabla}$, the distance between $\mathbf{P}_n$ and z becomes smaller and the membership of $\mathbf{P}_n$ in the cluster of z is increased. Therefore they are assigned to the same cluster, and an audio-cut can be identified in the n^{th} frame as shown in Fig. 2, which plots the signal amplitude and the membership value of $\mathbf{P}_n$ in the cluster containing z, against the time. For example, the membership values become very high at 9 s (app.) and 11 s (app.) as indicated by the red dotted-lines, since two audio-cuts are detected in these two points of the signal and these audio-cuts act as the delimiters of a speech audio-segment. On the other hand, when no audio-cut exists in either of the two time intervals from n to $(n + w_x - 1)$ and from $(n–\nabla)$ to $(n - \nabla + w_x - 1)$, as illustrated by the windows L_2 and R_2 in Fig. 1, the elements of both $\mathbf{P}_n$ and $\mathbf{P}_{n–\nabla}$ become significantly greater than *0*. Thus, the distance between $\mathbf{P}_n$ and $\mathbf{P}_{n–\nabla}$ becomes shorter, and they move away from z.

Audio-segment classification: The audio-cuts indicate the segment boundaries in the signal, which allows us to identify the segments. After segmenting a long audio stream that includes different classes of sounds, each segment is classified into one of five groups such as silence, speech, music, speech with music background, and speech with noise background. The selected feature vector for classifying audio signals is shown in (23):

$$\mathbf{V}_f = [\mu_E, \sigma_E, \mu_G, \sigma_G, Z_R] \tag{23}$$

This includes five features from the aforementioned characteristic parameters of audio signals defined in (15)–(19). The distributions of these features for the five audio-classes are depicted in [12]. The rule based approach of segment-classification into five audio classes using the feature vector V_f is summarized below:

- Silence: This type of signal only contains quasi-stationary background noise and has relatively low values for μ_E and σ_E, but high values for Z_R.

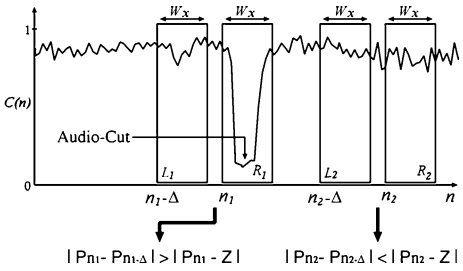


Fig. 1 Transition of the parameter at the C (n) audio-cut [12]

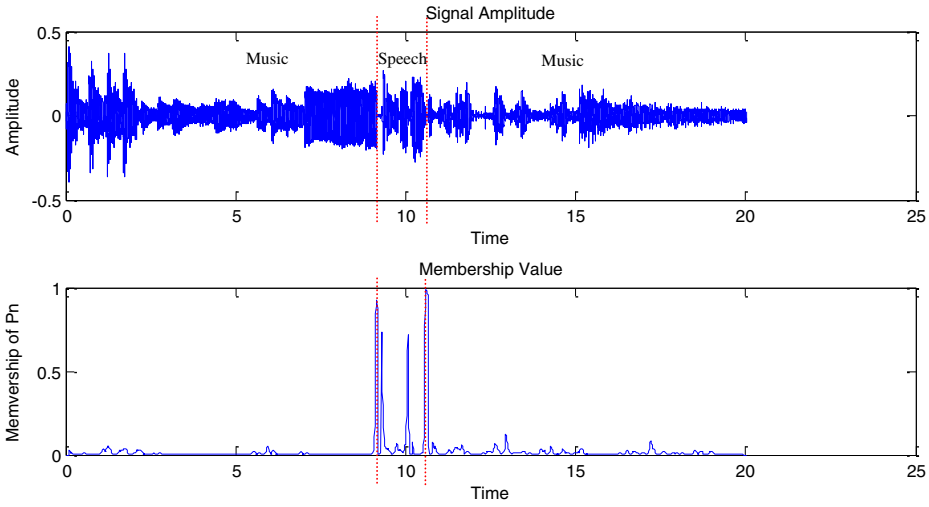


Fig. 2 An example of the signal amplitude and membership function values of the vector P_n in the cluster of z, where arrows indicate the existence of an audio-cut with corresponding higher membership

- Speech: An audio signal that contains the voices of human beings, such as the sound of conversation, and has relatively high values for μ_E and σ_E, and low values for Z_R compared to silence and music.
- Music: These are audio signals that contain sounds made by musical instruments. These sounds have relatively low values for Z_R, μ_G and σ_G.
- Speech with music background: These audio signals contain speech in an environment with music in the background. These can be discriminated from pure music by differences in σ_G.
- Speech with noise background: This type of audio signal contains speech in an environment with noise in the background. These sounds have relatively higher values for μ_G and σ_G than those of other audio classes.

We employ the proposed CIFCM approach to identify the specific class of sound represented by each audio segment. We determine the audio class of a segment according to its highest class membership. The steps of the heuristic rule-based fuzzy clustering procedure are listed below:

- Step 1: Acquisition of the audio signals.
- Step 2: Extraction of the feature C (n) by calculating E (n) from (15) and (16); and defining the vectors $\mathbf{P}_n$, $\mathbf{P}_{n-\nabla}$, and $\mathbf{z}$ by using (20)–(22).
- Step 3: Application of the CIFCM for segmentation.
- Step 4: Detection of segments from the audio-cuts.
- Step 5: Extraction of the feature vector $\mathbf{V}_f$ in (23) from each segment by applying (15) and (17)–(19).
- Step 6: Application of the CIFCM to classify segments into five audio classes by the values in $\mathbf{V}_f$.
- Step 7: Determination of the specific class of each segment depending upon its highest class membership.

4 Experimental results

This section evaluates the performance of the proposed CIFCM algorithm in audio segmentation and classification experiments. This section also compares the performance of the proposed approach and the conventional FCM algorithm [12].

4.1 Experimental environment

To evaluate the performance of the proposed CIFCM algorithm, we developed a Graphical User Interface (GUI) using MATLAB7.6, shown in Fig. 3. Here we included two frames, a signal browsing frame and a classification distribution frame. The signal browsing frame presents the signal being processed as well as segmentation results while the classification distribution frame shows audio-segment classification results. A "Detect Audio-cuts" button is used to determine audio-cuts and "Classify Audio-segs" button is used for classification.

In this study, we used a number of audio samples obtained from TV programs, including music and drama programs, as the input signals of the proposed algorithm. We used the following empirical parameters: the fuzzy weighting exponent $m=2.0$, the convergence tolerance $\varepsilon=0.001$, the maximum number of iterations in CIFCM=100, the number of samples in each frame $w_l=400$, the step size $\nabla=10$, and the pre-defined window length $w_c=w_x=10$.

In this study, two kinds of errors were generated and evaluated for audio-cut detection: (1) misdetection, in which the algorithm fails to detect existing audio-cuts and (2) over-detection, in which it incorrectly detects an audio-cut even where no audio-cut exists. In addition, we used two metrics to measure the effectiveness of

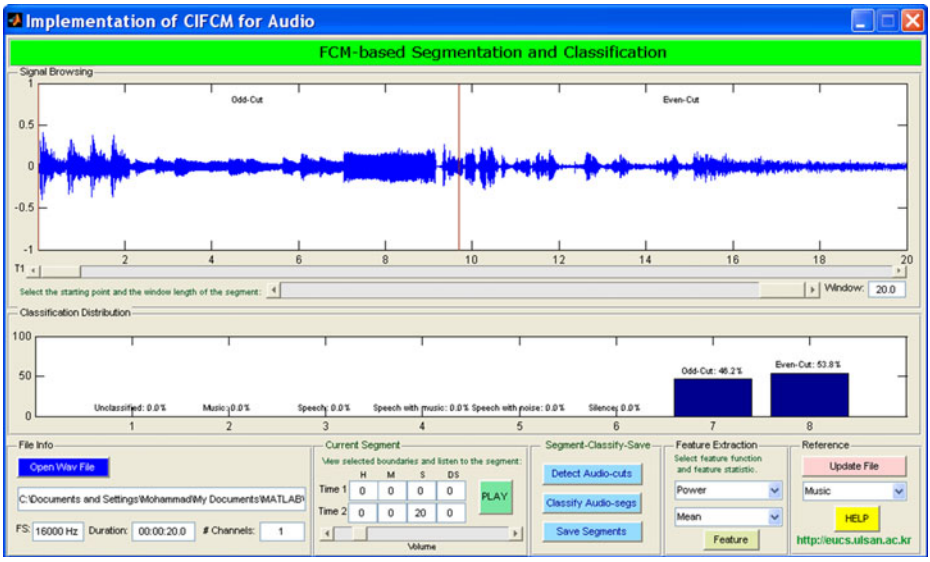


Fig. 3 Screenshot of GUI system implemented for CIFCM-based audio segmentation and classification

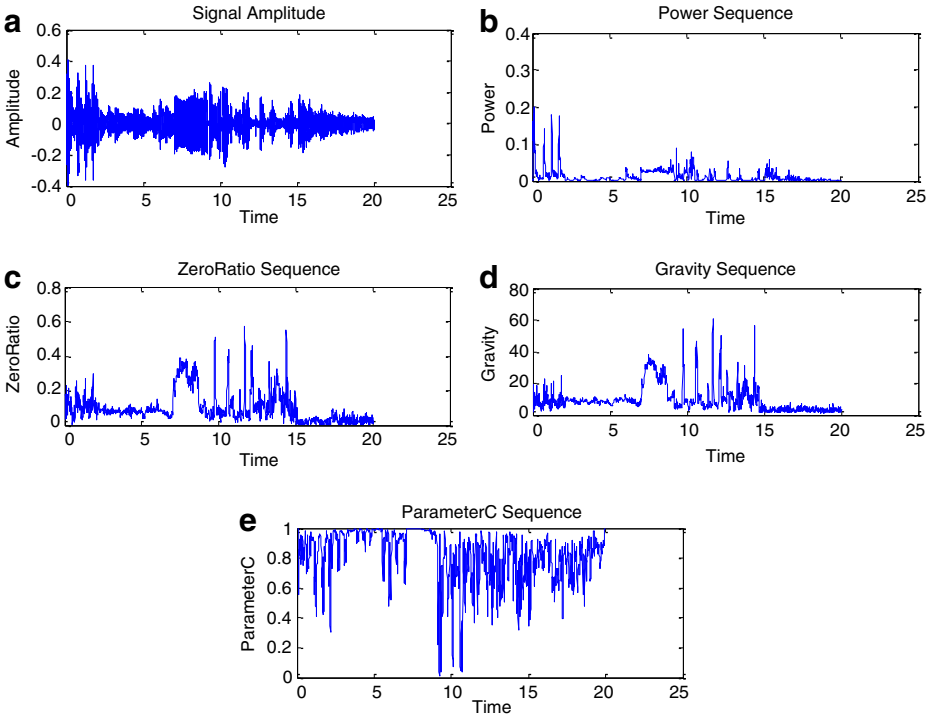


Fig. 4 An example of audio features extracted from a selected audio file, *DEMO*; (**a**) input signal, (**b**) power sequence $E(n)$, (**c**) gravity $G(n)$, (**d**) zero-crossing rate $Z(n)$, and (**e**) parameter sequence $C(n)$

audio-cut detection using the proposed approach: (1) recall rate and (2) precision rate [12], which are as follows:

$$\text{Recall rate} = \frac{\text{Number of correctly detected audio} - \text{cuts}}{\text{Number of manually detected audio} - \text{cuts}} \tag{24}$$

$$\text{Precision rate} = \frac{\text{Number of correctly detected audio} - \text{cuts}}{\text{Number of all detected audio} - \text{cuts}} \tag{25}$$

Table 1 Audio-cut detection results

	CIFCM		FCM	
	TEST1	*TEST2*	*TEST1*	*TEST2*
Number of all audio-cuts	103	138	103	138
Number of correct detections	98	132	85	117
Number of misdetections	5	6	18	21
Number of over-detections	8	16	15	26
Recall rate	0.951	0.957	0.825	0.848
Precision rate	0.925	0.892	0.850	0.818

Table 2 Resulting confusion matrix of genres from applying the proposed CIFCM for classification of audio-segments

Genres	Speech	Speech with music	Silence	Speech with noise	Music
1. Speech	129	20	0	5	0
2. Speech with Music	8	115	0	0	4
3. Silence	0	0	92	0	0
4. Speech with Noise	6	0	0	75	6
5. Music	11	0	0	0	102

The values of both rates are within the range [0:1]. If the recall and precision rates approach 1, there are few misdetections and over-detections, respectively. We use a well-known metric to measure the effectiveness of classification [12], which is given as follows:

$$\text{Classification Precision rate, CPR} = \frac{\text{Number of correctly classified audio segments}}{\text{Number of all audio segments}} \times 100\% \tag{26}$$

The value of *CPR* is within the range [0:100]%. Higher values indicate higher accuracy in classification.

The clustering performance of the proposed algorithm for audio classification is also measured in terms of four cluster validity functions described in Section 2.2.

4.2 Analysis and results

Figure 4 (a, b, c, d, e) depicts the results of the power sequence $E(n)$, the gravity $G(n)$, the zero-crossing rate $Z(n)$, and the parameter sequence $C(n)$, respectively, extracted from a selected audio sample using the proposed algorithm. These values depend on the types of sounds existing in the audio signal, as stated in Section 3.2 and analyzed in [12]. $E(n)$ is used to discriminate speech signals from the others. $G(n)$ and $Z(n)$ help to classify music, speech-with-music and speech-with-noise signals. The parameter sequence $C(n)$ is used to detect audio-cuts. For instance, the signal depicted in Fig. 4(a) contains speech-with-music in the 10–15 s (app.) time interval and merely music in the 15–20 s (app.) time interval. If we observe $E(n)$ of the signal depicted in Fig. 4(b), we do not get sufficient difference in these values and cannot discriminate speech-with-music and music signals. However we can discriminate them by using the values of $G(n)$, since the mean and variance parameters of $G(n)$ vary sufficiently within the time intervals 10–15 s and 15–20 s, as shown in Fig. 4(d).

Table 1 shows the audio-cut detection results of the proposed CIFCM and conventional FCM algorithm [12]. These results were generated from two tests, namely *TEST1* and *TEST2,*

Table 3 Confusion matrix results of genres from applying the FCM for classification of audio-segments

Genres	Speech	Speech with music	Silence	Speech with noise	Music
1. Speech	96	27	0	4	27
2. Speech with Music	13	107	0	0	7
3. Silence	0	0	92	0	0
4. Speech with Noise	7	0	0	74	6
5. Music	21	18	0	0	74

Table 4 Comparison results of the audio-segments classification

Approach	CPR
Conventional FCM [12]	77.31%
One-against-all Multiclass SVM	86.21%
Proposed CIFCM	89.53%

which were conducted on two different audio signals with a total of 241 audio-cuts. Here 'the number of all audio-cuts' indicates the number of manually detected audio-cuts in the audio signals. Manually detected audio-cuts are identified by listening to the audio signals and by observing the waveforms in the frame level. We observe that, the proposed CIFCM algorithm outperforms the conventional FCM algorithm for audio-cut detection in terms of both recall rate and precision rate. However, the proposed CIFCM algorithm generates a number of over-detections of audio-cuts due to the relatively long periods of silence between consecutive sentences in the speech segments, yet it is still better than the conventional FCM algorithm. We can reduce the inaccuracy of these over-detections by merging those three consecutive segments, where two segments of same audio-group are separated by a small duration of silence segment. Misdetection was also observed in the results due to the gradient way of changing signals at transition points. However this type of error has been reduced considerably in the proposed CIFCM in compare with conventional FCM algorithm.

There is no standard dataset on broad audio genres of the sort we are interested in, for analyzing the classification performance [12]. Thus we used a number of audio signals obtained from TV programs and the Internet (including the signals used in audio-cuts detection) to create the dataset. We summarized an experimental dataset of 573 audio-segments from all five broad categories. To ensure robustness in the experimental analysis, we included male speech, female speech, conversation of both male and female in the speech audio-segments, and instrumental as well as vocal songs of different music genres in the musical audio-segments. In addition, we included speech-with-noise with different levels of environmental noises. The confusion matrixes using the proposed CIFCM algorithm and the conventional FCM are shown in Tables 2 and 3, respectively. The rows represent the manually classified results and the columns represent the classification results of these algorithms. We found misclassifications for different groups in the results of Table 2, especially for speech and speech-with-music groups. Misclassifications occurred in the speech group (1st column) due to a relatively lower fraction of music or noise in speech with music or noise, and in some music-segments with lack of continuous strings. Similarly, in the speech-with-music group, misclassification occurred mostly due to fast speech-segments. Classification in the other groups is remarkably efficient.

Table 4 presents the performance comparison among the proposed CIFCM, one-against-all multiclass SVM-based approach [4], and conventional FCM [12] in

Table 5 Comparison of clustering performances in terms of fuzzy-cluster validity functions in audio-segments classification

Experiments	Technique	V_{pc}	V_{pe}	V_{xb}	V_{fs}
TEST1	FCM	0.7530	0.2154	0.3245	−6.6759
	CIFCM	0.8472	0.1177	0.1489	−9.1918
TEST2	FCM	0.8542	0.1359	0.1844	−0.5844
	CIFCM	0.9612	0.0350	0.0626	−0.7826

classification of audio-segments into five broad audio genres. We used the same feature vector for all three classification approaches. In addition, as SVM is a supervised learning method, we used a small and manually labeled training set from our dataset in order to train the SVM. From the comparison results, we observed that by considering temporal correlations of neighboring audio data, we could achieve better classification performance (89.53%) by using the proposed CIFCM algorithm. In addition, experimental results indicate that the CIFCM algorithm outperforms the conventional FCM algorithm and one-against-all multiclass SVM-based approach [4] in audio-segments classification. Quantitative evaluation using cluster validity functions is also used, to analyze the effectiveness of evolving better clusters in the proposed CIFCM over the traditional FCM algorithm [12]. As described in Section 2.2, superior clustering performance, as assessed by compactness, is achieved if V_{pc} is maximized and V_{pe}, V_{fs}, V_{xb} are minimized. Table 5 is a comparison of the clustering performance of audio-segments classification by using the proposed CIFCM and the traditional FCM in terms of the cluster validity function. It is observed from the results that, the proposed CIFCM algorithm achieves better clustering performance than the conventional FCM algorithm in the classification experiments (namely, *TEST1* and *TEST2*) due to the increased compactness of fuzzy clusters of segments that is achieved by considering neighboring impact factors in the proposed CIFCM.

Overall, the proposed CIFCM algorithm outperforms the conventional FCM in both audio-cut detection and audio-segment classification.

5 Conclusions

This paper presents a CIFCM algorithm that was designed to improve audio segmentation and classification performance. Unlike the conventional FCM approach, the CIFCM algorithm efficiently incorporates the influence of neighboring frames and segments from the audio stream for improved segmentation and classification of the current frame and segment, respectively. In addition, we analyzed a number of characteristic audio features to explore the differences among different types of audio data. Our experimental results indicate that the proposed CIFCM algorithm outperforms the conventional FCM algorithm and one-against-all multiclass SVM-based approach in audio segmentation and classification.

In the future, we will explore audio feature extraction from compressed audio data, since most digital audio data that are currently available are in compressed formats such as WMA, MP3, AAC, etc. Numerous different types of audio signals will be studied as well for segmentation and classification. In addition, different statistical classifiers will be investigated with the goal of producing a fully featured automatic audio retrieval system.

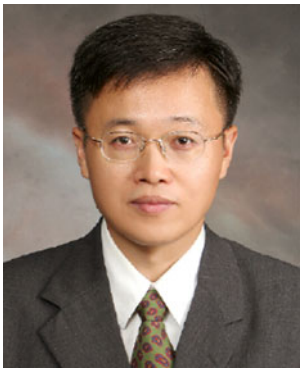
Acknowledgements This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MEST) (No. 2011-0017941)

References

1. Bezdek JC (1981) Pattern recognition with fuzzy objective function algorithms. Pleum Press, New York
2. Bezdek JC, Keller J, Krisnapuram R, Pal NR (2005) Fuzzy models and algorithms for pattern recognition and image processing. Kluwer Academic Publishers, Norwell

3. Bezdek JC, Nikhil R (1995) On cluster validity for the fuzzy c-means model. IEEE Trans on Fuzzy Syst 3(3):370–379
4. Chen L, Gunduz S, and Ozsu MT. (2006) Mixed type audio classification with support vector machine. IEEE Int. Conf. on Multimedia and Expo, 781–784
5. Fukuyama Y and Sugeno M. (1989) A new method for Fuzzy clustering. 5th Fuzzy System Symp., 247–250
6. Gang C, Hui T, Xin-meng C (2005) Audio segmentation via the similarity measure of audio feature vectors. Wuhan Univ J Natur Sci 10(5):833–837
7. Krinidis S, Chatzis V (2010) A Robust fuzzy local information c-means clustering algorithm. IEEE Trans Image Process 19(5):1328–1337
8. Liu Z, Huang J, and Yang Y. (1998) Classification of TV programs based on audio information using hidden markov model. IEEE 2nd Workshop on Multimedia Sig Process, 27–32
9. Liu Z, Wang Y (1998) Audio feature extraction and analysis for scene segmentation and classification. J VLSI Sign Process 20:61–79
10. Lu L, Zhang HJ, Jiang H (2002) Content analysis for audio classification and segmentation. IEEE Trans Speech Audio Process 10(7):504–516
11. Luong HV and Kim J-M. (2009) A Generalized spatial fuzzy C-means algorithm for medical image segmentation. IEEE Int. Conf. on Fuzzy Systems, 409–414
12. Nitanda N, Haseyama M, Kitajima H (2006) Audio signal segmentation and classification using fuzzy c-means clustering. Syst Comput Jpn 37(4):23–34
13. Park DC (2009) Classification of audio signals using fuzzy c-means with Divergence-based Kernel. Pattern Recognit Lett 30(9):794–798
14. Park DC, Nguyen DH, Beack SH, Park S (2005) Classification of audio signals using Gradient-based fuzzy c-means algorithm with divergence measure. Adv Multimedia Inf Process PCM 2005:698–708
15. Park DC, Tran CN, Min BJ, and Park S. (2006) Modeling and classification of audio signals using Gradient-based fuzzy C-means algorithm with a Mercer Kernel. In 9th Pacific Rim International Conference on Artificial Intelligence: 1104–1108
16. Tzanetakis G, Cook P (2002) Music genre classification of audio signals. IEEE Trans on Speech Audio Process 10(5):293–302
17. Wang JC, Wang JF, Lin CB, Jian KT, Kuok WH (2006) Content-based audio classification using support vector machines and independent component analysis. 18th Int. Conf. on. Pattern Recognit 4:157–160
18. Wold E, Blum T, Keislar D, Wheaton J (1996) Content-based classification search and retrieval of audio. IEEE Multimedia Mag 3:27–36
19. Xie XL, Beni GA (1991) A validity measure for fuzzy clustering. IEEE Trans on Pattern Anal Mach Intell 13(8):841–847
20. Zhang T, Kuo C-CJ (2001) Audio content analysis for online audiovisual data segmetation and classification. IEEE Trans on Speech Audio Process 9(4):441–457

Mohammad A. Haque Received his BS degree in Computer Science and Engineering from University of Chittagong, Bangladesh, in 2008, and is currently studying in MS at Computer and Information Technology in University of Ulsan, South Korea. He is a lecturer (now in study leave) of Computer and Communication Engineering at International Islamic University Chittagong, Bangladesh. His research interest includes audio signal processing, embedded system design, and biometrics and image processing.



Jong-Myon Kim Received the BS degree in electrical engineering from the Myongji University, Yongin, Korea, in 1995, the MS degree in electrical and computer engineering from University of Florida, Gainesville, in 2000, and the PhD degree in electrical and computer engineering from the Georgia Institute of Technology, Atlanta, in 2005. He is an assistant professor of Computer Engineering and Information Technology at University of Ulsan, Korea. His research interests include multimedia processing, multimedia specific processor architecture, parallel processing, and embedded system. He is a member of the IEEE and the IEEE Computer Society.