International Journal of Computer and Organization Trends
ISSN: 2249 – 2593 / https://doi.org/10.14445/22492593/IJCOT-V15I2P304

Volume 15 Issue 2, 35-39, May-Aug 2025

Original Article

Eliminating Video Music Separator Using K-Means Algorithm and MFCC

Eman Hato

Computer Science Department, College of Sciences, Mustansiriyah University, Baghdad, Iraq.

Corresponding Author : emanhato@uomustansiriyah.edu.iq

Received: 02 June 2025 Revised: 04 July 2025 Accepted: 25 July 2025 Published: 13 August 2025

Abstract - *The news theme serves as the first separator in the news video's well-defined framework. This separator significantly raises the false detection rates during video temporal segmentation since it comprises a sequence of rapidly moving interlaced pictures with a particular musical accompaniment. The separator frames are eliminated before the video segmentation process starts to minimize extraneous frames and lower false detections. To effectively and efficiently eliminate unnecessary video frames, this paper proposes an automatic technique for separating the music portion of a news video using Mel-frequency Cepstral Coefficients (MFCC) and the K-means clustering algorithm. There are two steps in the suggested approach. The audio signal taken from the input video is used to calculate the MFCC features in the first stage. To do this, the audio stream is divided into overlapping windows, and each window is processed separately. The result is a matrix of MFCC coefficients. In the second stage, the k-means algorithm is employed to initially cluster centers from a predefined matrix, making them more closely related to each cluster, specifically music and speech in this case. The algorithm then classifies the MFCC features into music and speech clusters. To locate the intervals of consecutive music clusters, the sequences of the same cluster are determined and removed from the input video. The results demonstrated the effectiveness of the proposed method, achieving a clustering accuracy of 99%. Its efficiency was further evidenced by a reduction in errors during the segmentation process and the elimination of irrelevant information.*

Keywords - *Hamming windows, K-Means, Feature extraction, MFCC, Audio clustering.*

1. Introduction

The most sophisticated and powerful multimedia format, video, encompasses a variety of information types. Video analysis is challenging due to the unstructured information format [1]. Large-scale video data analysis is a practical and efficient method that may be applied to a variety of video applications, including ranking, indexing, retrieval, and summarization. The fundamental challenge of video analysis is to bridge the gap between the user's interpretation of higher-level data and the low-level information that is extracted [2]. A critical first stage in the indexing process is video segmentation, which allows the needed information to be extracted from a massive amount of video data. By breaking the video up into meaningful segments, video segmentation aims to reveal the temporal structure of the video [3].

Features refer to descriptive parameters that are extracted from video data. It is important to extract various features, including those based on audio, color, shape, texture, or combinations of these elements [4]. Audio features are particularly significant for distinguishing between types of sounds, such as speech and music, as well as identifying audio events and spoken text. In general, audio features are divided into two categories: frequency domain features, which include spectrograms, cepstral coefficients, and MFCC and time domain features, which include amplitudes, pitch, and zero-crossing rates [5, 6].

This paper aims to separate the musical part that appears at the beginning of news videos from the part in which the speech appears, given the subsequent negative impact of this part on the process of segmenting the video's visual content.

2. Mel Frequency Cepstral Coefficients (MFCC)

MFCC is a technique used for extracting features from stationary and pseudo-stationary signals. It is highly effective for classifying speech and music signals and for speaker recognition. To create a feature vector that captures all the important information from an audio signal, MFCC emulates the way the human auditory system processes sound. The power spectrum is adjusted according to the mel scale, which allows the frequency resolution to align with the characteristics of human hearing [7].

The following procedures are used to calculate the MFCC technique's coefficients [8, 9]:

2.1. Pre-Emphasis

To highlight the higher frequencies, the passing signal is filtered. Balancing the signal spectrum with a steep roll-off in the high-frequency band is the aim of this step.

2.2. Windowing and Frame Blocking

The audio signal is split up into frames, each of which has N samples. To provide seamless transitions devoid of sharp edges, a window function is applied to each frame to reduce the amplitude of discontinuities at the boundaries. For this, Hamming windows—which are described as in Equations 1and 2:

$$X(n) = x(n) \times Hw(n) \quad (1)$$

$$Hw(n) = a\_0 - a\_1 \cos(2\pi n/(N-1)) \quad (2)$$

Where n is $0 < n \leq N\text{-}1$, a0 and a1 are hamming window coefficients, and the ideal values are a0 = 0.53836 and a1 = 0.46164. N is the number of samples in each frame, and x(n) is the input signal, X(n) is the output signal, and Hw (n) is the Hamming window.

2.3. Discrete Fourier Transform (DFT)

DFT is used to transform each windowed frame from the time domain into the frequency domain (spectrum), which is described as in Equation 3:.

$$Y(k) = \sum_{n=0}^{N-1} X(n).\, e^{\frac{-I_u 2\pi nk}{N}} \quad 0 \leq k \leq N-1 \quad (3)$$

2.4. Me-Spectrum

To determine the mel spectrum, the converted signal is run through a bank of band-pass filters called a mel filter bank. Based on how the human ear interprets frequency, a "mel" is a unit of measurement. A particular formula can be used to estimate the relationship between mel and physical frequency, as in Equation 4:

$$f\_mel = 2595 \times \log(1 + f/700) \quad (4)$$

Where fmel represents the perceived frequency in mel and f is the physical frequency in Hertz. The magnitude spectrum Y (k) is multiplied by each of the triangular mel weighting filters to determine the Mel Spectrum MS (m) as in Equation 5:

$$MS(m) = \sum_{K=0}^{N-1} (|Y(k)|^2 H_m(k)) \quad 0 \leq m \leq M-1 \quad (5)$$

Where Hm(k) is the weight assigned to the kth energy spectrum, and M is the total number of triangular mel weighting filters.

2.5. Discrete Cosine Transform (DCT)

It is used to create a set of cepstral coefficients by converting the log mel spectrum into the time domain. The MFCC coefficients are computed as in Equation 6:

$$c(n) = \sum_{m=0}^{M-1} \log(MS(m)) \cos\left(\frac{\pi n(m-0.5)}{M}\right) \quad n=0,1,\ldots,C-1 \quad (6)$$

Where C is the number of MFCCs and c (n) is the cepstral coefficient.

3. K-Means Clustering Algorithm

One of the most basic methods for unsupervised learning is the k-means clustering algorithm. As seen in Figure 1, this approach provides a simple means of classifying a given dataset [10]. The following steps make up the algorithm [11, 12]:

- Insert k points into the search area space that is represented by the clustered samples. These points represent the initial cluster centroids.
- Assign every sample to the cluster whose centroid is nearest.
- Recalculate the k centroids' locations after all samples have been assigned.
- Until the centroids remain the same, repeat steps two and three.

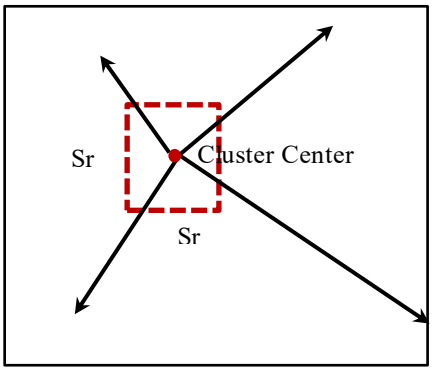

Fig. 1 The search area of the K-means algorithm

4. The Proposed Method

The structure of news videos is mostly set; news segments typically start with a pause. A break is a sequence of frames used to divide narratives. The comparable break pattern—a sequence of rapidly moving overlapping images accompanied by particular background music—significantly raises the false detection rate of temporal video segmentation. To avoid unnecessary frames and false detection, frames from the break segment that do not contain any useful information are removed before the video segmentation process. The proposed method involves two stages: first, extracting MFCC features from the audio signal obtained from the video, and second, classifying the audio signal into either music or speech clips while also identifying the frames associated with these clips.

4.1. MFCC Features Extraction

MFCC analysis is very useful in distinguishing between speech signals and music, where intervals contain music without spoken words. Following the extraction of the audio

signal from the input video, MFCC features are computed for every video. By splitting the audio stream into overlapping windows and processing each window independently, MFCC features are retrieved. Making sure that audio samples from the input sequence are roughly centred in a specified window is the aim of overlapping analysis. Therefore, some or possibly all of the signal information that was lost during the segmentation process is preserved by overlapping windows. Algorithm 1 presents the steps of MFCC extraction.

Algorithm 1: MFCC Features Extraction
Input: Input video.
Output: MFCC features Matrix.
Start
Step 1: Extract the audio signal from the input video.
Step 2: Blocking signal into short windows (frames).
Step 3: For I: 1 to window number
Temp← Holds single window.
WTemp← Apply hamming on Temp window based on Eq. 1.
FTemp← Apply DFT using WTemp based on Eq. 3.
STemp← Compute mel spectrum from FTemp based on Eq. 5.
Coeffs ← Apply DCT of Stemp based on Equation 6.
Store DCT coefficients in the MFCCFeat matrix and discard the rest.
End For
End.

4.2. K-means Clustering

Every time the k-means algorithm is executed, it converges to a local minimum because it randomly selects the starting cluster centers. Cluster centers are derived from a predetermined matrix CMat to generate clusters with more accuracy, ensuring that the outcomes are consistent across algorithm runs. The CMat has MFCC capabilities for speech taken from a news video that only includes speech snippets and MFCC features for music taken from the BBC news start-up theme video. The initial centers are always closer and more connected to each identified cluster as a result of this procedure. Using the k-means technique, the audio signal is separated into music and speech clusters once the matrix of cluster centers and MFCC features has been initialized. To find the interval for each series of music clusters, the subsequent series of the same cluster (either speech or music cluster) are identified. Algorithm 2 illustrates the operation sequences of audio clustering.

Algorithm 2: Audio Clustering
Input: Input video, Clusters centers matrix (CMat), Clusters number (No_{Clust}).
Output: Music frames index.
Start
Step 1: Load the input video.
Step 2: Calculate the MFCC features as in Algorithm 1.
Step 3: Extract the number of iterations and the number of MFCC features.
Step 4: Cluster the MFCC features of the audio signal and save the cluster labels into a DT using the K-means algorithm and the Cluster Center Matrix (CMat).
Step 5: Identify the successive series of the same DT cluster and save them into a Cluster.
Step 6: Identify the frames within each music cluster.
End.

5. Experimental Results

The ten news videos that were collected from the British Broadcasting Corporation's (BBC) news channel archive make up the dataset. The videos' audio signal was separated into clusters for speaking and music. The trials were carried out on an Intel Core i7, 64-bit operating system, 2.40 GHz processor, and 12 GB RAM. The suggested system was developed using the programming language MATLAB 2017a.

Music separators were excluded from all ten test files, achieving an Accuracy of up to 99% to eliminate unhelpful information represented by music separator frames. Table 1 presents the details of the results.

Table 1. The evaluation of the proposed method.

Video Files	Frames Number	Video Time Duration in Sec.	Music Time Duration in Sec.	Music Frames		False Frame Detection	Accuracy
				Start Frame	End Frame		
V01	3622	108660	46.1	30	1413	4	99.99
V02	435	14.5	1.16	0	35	0	100
V03	5453	181.766	1.46	0	44	16	99.98
V04	2993	99.766	46.2	28	1415	0	100
V05	2286	68.580	20.9	80	707	0	100
V06	5998	199.933	23.5	0	705	0	100
V07	6704	223.466	47.16	80	1415	0	100
V08	5135	171.166	20.93	80	708	0	100
V09	4401	146.7	44.43	80	1413	6	99.99
V10	3694	110.82	20.93	50	678	0	100
Average	**4072**					**2.6**	**99.996**

The excellent performance of the proposed method is demonstrated in Table 1, achieving very high accuracy and a very low error rate, averaging about two frames out of 4,000. Therefore, this method can be used in practical applications for video analysis. In addition, music breaks confuse the temporal segmentation process, increasing the false detection rate and affecting the performance of video segmentation. A video track is used as an example to illustrate the effect of music breaks on frames. The video (V01) contained these frames in the first 6 seconds, which lacked any valuable information, as shown in **Figure 2**. First, it is known that the music interruption was considered as only one segment, while the result indicated that there were four segments in the V01 video, namely 90, 140, 171, and 195. Second, although it is a single segment, if visual interruptions are considered to detect temporal segmentation by comparing the difference between consecutive frames, then segment 195 was a false detection, and segment 200 was also a missed detection, as shown in Figure 3.

Upon examining Figure 3, it is evident that frames 195 and 196 belong to the same segment, while frames 199 and 200 are associated with different segments. A mistake was made in detecting the shot boundaries, which occurred within the first 200 frames of the video. The experimental test demonstrated the exclusion of music segments, which helped improve the performance of the temporal segmentation and reduced processing time. Consequently, the frames that fell within the duration of the music segments were disregarded for further video processing.

6. Conclusion

This paper introduces a new method for distinguishing between musical segments and speech segments at the beginning of news videos. This differentiation is crucial since it significantly impacts subsequent video segmentation. The presence of music at the start of news segments has complicated and hindered the performance of temporal video segmentation. To address this issue, the music frames have been removed to decrease the incidence of false positives and missed detections in video segmentation.

The method relies on MFCC features, known for their accuracy and distinctiveness, along with the K-means algorithm to effectively separate music from speech. The results demonstrate a high level of efficiency in distinguishing between music and speech frames in news videos. This allows for the exclusion of frames that do not contain useful information, enabling a more focused analysis of the relevant content in the remaining frames.

Funding Statement

The development and publication of this work were entirely funded by the authors, without any external financial support or sponsorship.

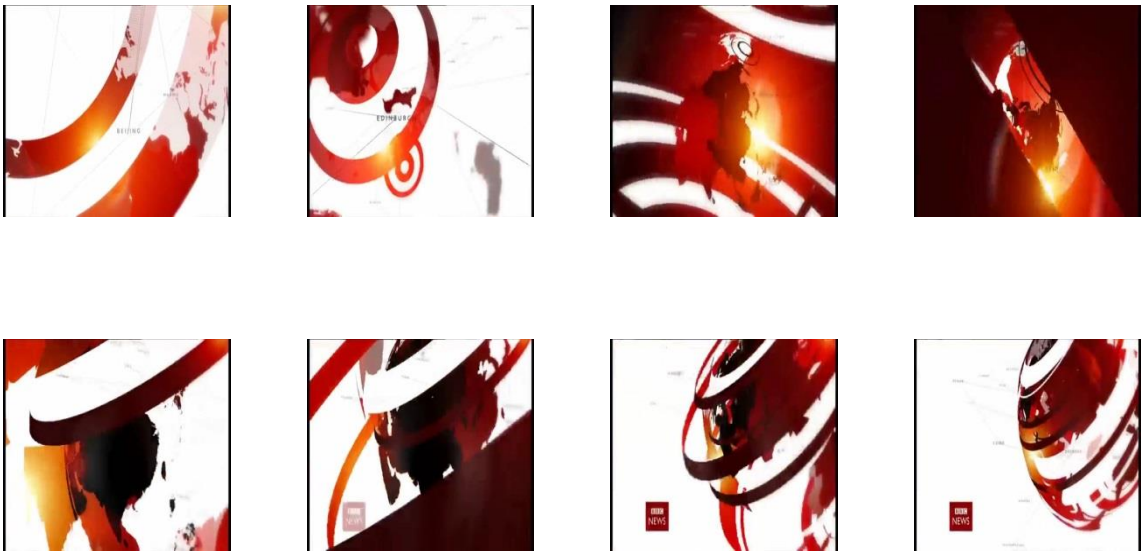

Fig. 2 Example music separator frames in Video 1 (V01).

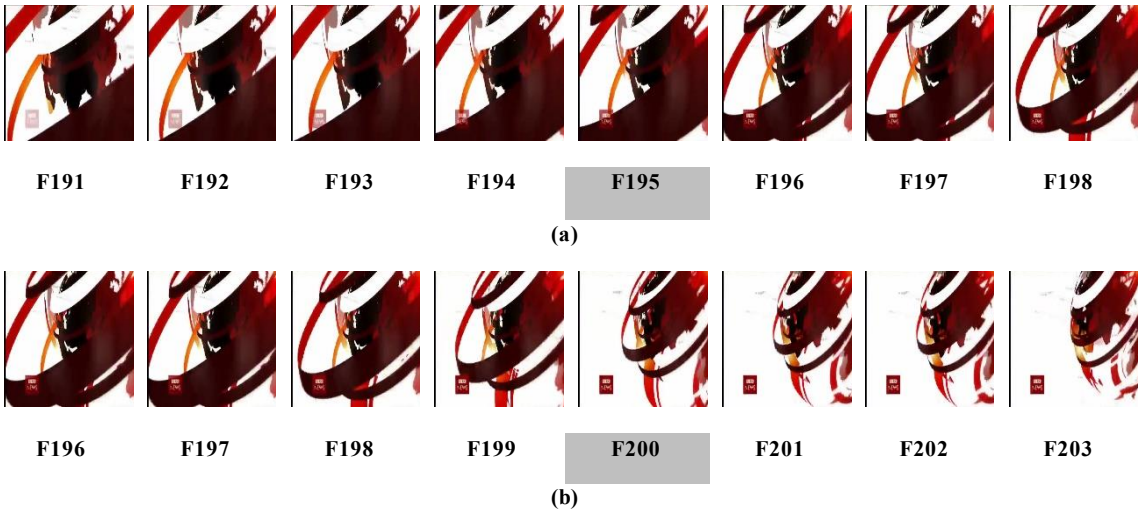

Fig. 3 False detection example. (a) False segment detection. (b) Missing segment detection.

References

[1] R. Priya, and T. N. Shanmugam, "A Comprehensive Review of Significant Researches on Contentbased Indexing and Retrieval of Visual Information," *Frontiers of Computer Science*, vol. 7, pp. 782-799, 2013. [CrossRef] [Google Scholar] [Publisher Link]

[2] Nitin J. Janwe, and Kishor K. Bhoyar, "Multi-Label Semantic Concept Detection in Videos Using Fusion of Asymmetrically Trained Deep Convolutional Neural Networks and Foreground Driven Concept Co-Occurrence Matrix," *Applied Intelligence*, vol. 48, no. 8, pp. 2047-2066, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[3] Arbind Agrahari Baniya et al., "Frame Selection Using Spatiotemporal Dynamics and Key Features as Input Pre-Processing for Video Super-Resolution Models," *SN Computer Science*, vol. 5, no. 3, pp. 1-15, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[4] T. Kar et al., "Video Shot-Boundary Detection: Issues, Challenges and Solutions," *Artificial Intelligence Review*, vol. 57, no. 4, pp. 1-38, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[5] M. Rhevanth et al., "Deep Learning Framework Based on Audio–Visual Features for Video Summarization," *Proceedings of the Springer Conference on Advanced Machine Intelligence and Signal Processing*, vol. 858, pp. 229-243, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[6] Yunzuo Zhang et al., "Key Frame Extraction Method for Lecture Videosbased on Spatio-Temporal Subtitles," *Multimedia Tools and Applications*, vol. 83, no. 2, pp. 5437-5450, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[7] Tsung-Han Tsai, Ping-Cheng Hao, and Chiao-Li Wang, "Self-Defined Text-Dependent Wake-Up-Words Speaker Recognition System," *IEEE Access*, vol. 9, pp. 138668-138676, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[8] Adal A. Alashban et al., "Spoken Language Identification System Using Convolutional Recurrent Neural Network," *Applied Sciences*, vol. 12, no. 18, pp. 1-17, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[9] Young-Long Chen et al.., "Combined Bidirectional Long Short-Term Memory with Mel-Frequency Cepstral Coefficients Using Autoencoder for Speaker Recognition," *Applied Sciences*, vol. 13, no. 12, pp. 1-19, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[10] Šárka Brodinová et al., "Robust and Sparse K-Means Clustering for High-Dimensional Data," *Advances in Data Analysis and Classification*, vol. 13, pp. 905-932, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[11] Mehta, V., Bawa, S. and Singh, J., "Analytical Review of Clustering Techniques and Proximity Measures," *Artificial Intelligence Review*, vol. 53, pp. 5995-6023, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[12] Caroline X. Gao et al., "An Overview of Clustering Methods with Guidelines for Application in Mental Health Research," *Psychiatry Research*, vol. 327, p.115265, 2023. [CrossRef] [Google Scholar] [Publisher Link]