Structural and Semantic Modeling of Audio for Content-Based Querying and Browsing

Mustafa Sert[1,2], Buyurman Baykal[3], and Adnan Yazıcı[4]

[1] Başkent University, Department of Computer Engineering, 06530 Ankara, Turkey
msert@baskent.edu.tr
[2] Gazi University, Faculty of Technical Education, Department of Electronics and Computer Education, 06500 Ankara, Turkey
[3] Middle East Technical University, Department of Electrical and Electronics Engineering, 06531 Ankara, Turkey
buyurman@metu.edu.tr
[4] Middle East Technical University, Department of Computer Engineering, 06531 Ankara, Turkey
yazici@ceng.metu.edu.tr

Abstract. A typical content-based audio management system deals with three aspects namely audio segmentation and classification, audio analysis, and content-based retrieval of audio. In this paper, we integrate the three aspects of content-based audio management into a single framework and propose an efficient method for flexible querying and browsing of auditory data. More specifically, we utilize two robust feature sets namely MPEG-7 Audio Spectrum Flatness (ASF) and Mel Frequency Cepstral Coefficients (MFCC) as the underlying features in order to improve the content-based retrieval accuracy, since both features have some advantages for distinct types of audio (e.g., music and speech). The proposed system provides a wide range of opportunities to query and browse an audio data by content, such as querying and browsing for a chorus section, sound effects, and query-by-example. In addition, the clients can express their queries in the form of *point*, *range*, and k-*nearest neighbor*, which are particularly significant in the multimedia domain.

1 Introduction

Traditional information retrieval (IR) systems provide access to data stored in the form of relational tables through forms-like interfaces. The clients specify their queries by filling the forms and supplying strict conditions on the attributes stored in the database. In the case of multimedia information retrieval (MIR), this approach can lead to unexpected results since the clients might not have prior knowledge about the structure of information stored in the database. A solution approach to this problem may be to allow the use of approximate queries through well-designed user interfaces. In order to enable such a capability, efficient storage and retrieval models should be designed and developed since multimedia content have complex structures. Various studies have emphasized

H. Legind Larsen et al. (Eds.): FQAS 2006, LNAI 4027, pp. 319–330, 2006.

on image and video indexing and retrieval [1, 4, 5, 6]. Auditory data (e.g., music, speech, sound effects), on the other hand, has not been considered as much as the other media types. We can explore the existing research on content-based audio management in three categories: (1) segmentation and classification of audio data into predefined classes, such as speech, music, environmental sound, and silence; (2) content-based audio retrieval (CBAR); (3) audio analysis. Therefore, an audio content management system should support all of these issues.

In the first category, most of the audio segmentation methods deal with feature extraction. These studies [10, 11, 12] can be examined in three groups. One of them is a segmentation method which involves with the classification of audio data segments into predefined classes (e.g., music, speech, and environmental sound) [10]. The other group is to detect the abrupt changes in feature values of audio data[11]. The last one is another segmentation method which consists of relative silences or pauses in an audio data[12]. Our segmentation method relies on the second group, namely detecting abrupt changes in feature values, and advances it by realizing a structural similarity analysis technique.

In the CBAR category, a widely used approach is to extract a set of acoustic features for each sound and query item. One attempt based on this motivation is the work carried out by the Muscle Fish [7]. In this work, they made use of statistical values, which are obtained from the mean, variance, and autocorrelation parameters of five features namely loudness, pitch, brightness, bandwidth, and harmonicity. However, merely statistical values are not suitable for sounds that have multiple timbre [9]. One specific technique is used in [8], where the Mel Frequency Cepstral Coefficients (MFCC) are used as features, and sounds are characterized by templates that are obtained from a tree-based vector quantizer. However, this method fails to distinguish music and environmental sounds that have distinct timbres in general, since the MFCC does not represent the timbre of music, as well.

In the audio analysis category, the main aim is to obtain structural descriptions of auditory data for efficient indexing and retrieval. Audio analysis studies, such as semantic audio segmentation [13], music thumbnailing [2, 3, 13], music summarization [14, 19], and chorus detection [15, 20], although carrying different titles, all share the same goal of facilitating an efficient browsing and searching of audio files. Indeed, they are all built upon the identification of important audio excerpts that are adequate to represent the entire audio file. However, current studies have some drawbacks in terms of detecting all repetitive structures and recognition accuracy, and hence need some improvements.

In this paper, we integrate the three considered aspects of content-based audio management into a single framework and propose an efficient method for flexible querying and browsing of auditory data. More specifically, we utilize two robust feature sets namely MPEG-7 Audio Spectrum Flatness (ASF) [16] and MFCC as the underlying features, in order to improve the content-based retrieval accuracy, since both features have some advantages for distinct types of audio data (e.g., music and speech). The proposed system provides a wide range of opportunities to query and browse an audio content, such as querying and browsing for a