Content-Based Searching in Group Communication Systems

Gábor Richly
Budapest University of Technology and Economics, Hungary

Gábor Hosszú
Budapest University of Technology and Economics, Hungary

Ferenc Kovács
Budapest University of Technology and Economics, Hungary

INTRODUCTION

The importance of real-time pattern recognition in streaming media is rapidly growing (Liu, Wang, & Chen, 1998). Extracting information from audio and video streams in an e-collaboration scenario is getting increasing relevance as the networking infrastructure develops. This development enables the use of rich media content. Shared archives of this kind of knowledge need tools for exploration, navigation and searching. As an example, to filter out redundant copies of an audio record, added by different members of an e-community, helps to keep the knowledge-base clean and compact.

The work presented here focuses on algorithms used for content-sensitive searching in audio. The main problem of such algorithms is the optimal selection of the reference patterns applied in the recognition procedure. The proposed method is based on distance maximization. It is able to quickly choose the reference pattern to be used by the pattern recognition algorithm (Richly, Kozma, Kovács, & Hosszú, 2001).

The article presents a novel approach of searching patterns in shared audio file storages such as peer-to-peer (P2P) based systems. The proposed method is based on the recognition of specific patterns in the audio contents extending the searching possibility from the description based model to the content based model.

The presented method identified as EMESE (experimental media-stream recognizer) is an important component of a light-weight content-searching system which is suitable for the investigation of network-wide shared file storages. The efficiency of the proposed procedure is demonstrated in the article.

BACKGROUND

From the introduction of Napster (Parker, 2004), Internet based communication has been developing toward the application level networks (ALN). Hosts are getting more and more powerful enabling various collaborative applications to run and create mutual logical connections (Hosszú, 2005). They establish a virtual overlay and, as an alternative to the older client/server model, they use the P2P communication. The majority of such system deals with file sharing, requiring the important task of searching in large, distributed shared file storages (Cohen, 2003; Qiu & Srikant, 2004).

Up to this time, searches have been based on the various attributes of the media contents (Yang & Garcia-Molina, 2002). The attributes called metadata can be the name of the media file, the name of authors, date of recording, type of media (genre) or some other keywords of descriptive attributes. However, if incorrect metadata were accidentally recorded, the media file may become invisible due to misleading descriptions.

Currently, powerful computers provide the possibility to implement and widely use pattern recognition methods. Due to the large amount of media files and their very rich content, limited pattern identification should be reached as a realistic goal. This article introduces the problem of media identification based on the recognition of well-defined patterns.

Another problem is introduced, if the pattern-based identification method must be extended from media files to real-time media streams. The hardness of this problem is the requirement that the pattern identification system should work real-time even in less powerful computing environment as well. For this purpose, full-featured media monitoring methods are not applicable since they require large processing power in order to run their full-featured pattern recognition algorithms.

The novel system called EMESE is dedicated for solving the special problem where a small but significant pattern must be found in a large voice stream or bulk voice data file in order to identify known sections of audio. Since this work is limited to sound files, the pattern is referred to as the soundprint of the specific media**.** It serves for uniquely identifying the media. This kind of patterns have also became known as the *audio fingerprint* (Haitsma, Jaap, & Kalker, 2002). The developed method is light-weight, meaning that its design goals were fast operation and the relatively small computing power. In order to reach these goals, the length of the pattern to be recognized should be very limited and the total score is not required.

This article deals mainly with the heart of the EMESE system, the pattern recognition algorithm, especially with the selection of the soundprints, the process called *reference selection*.

THE PROBLEM OF THE PATTERN RECOGNITION

In the field of sound recognition, there are many different methods and applications for specific tasks (Kondoz, 1994; Coen, 1995).

The demand for working efficiently with streaming media on the Internet increases rapidly. These audio streams may contain artificial sound effects besides the mix of music and human speech. The effects furthermore may contain signal fragments that are not audible by the ear. As a consequence, processing of this kind of *audio signal* is rather different from the already developed methods. For example, the short-term predictability of the signal is not applicable.

The representation of digital audio signal as individual sample values lacks any semantic structure to help automatic identification. For this reason, the audio signal is transformed into several different orthogonal or quasiorthogonal bases to enable detecting certain properties.

Already, there are solutions for classifying the type of broadcast on radio or television using the audio signal. An application (Akihito et al., 1998) makes basically a speech/music decision by examining the spectrum for harmonic content, and the temporal behavior of the spectral-peek distribution. Although it was applied successfully to that decision problem, it cannot be used directly for generic recognition purposes. Liu et al. (1998) also describe a scheme classifying method where the extracted features are based on short-time spectral distribution represented by a bandwidth and a central frequency value. Several other features, for example the volume distribution and the pitch contour along the sound clip are also calculated. The main difficulty of these methods is their high computation-time demand. Therefore their application for real-time or fast monitoring is hardly possible, when taking the great number of references to be monitored into account.

A similar monitoring problem has already been introduced (Lourens, 1990) where the energy envelope is used as the signal feature and its section on the record signal as the *reference.* That was correlated with the input (*test*) signal. The demand on real-time execution drove the development of a novel recognition scheme (Richly et al., 2000) that is capable of recognizing a pattern of transformed audio signal in an input stream, even in the presence of level-limited noise. This algorithm selects a short segment of the signals in the case of each record in the set of records to be monitored (Richly, Kozma, Kovács, & Hosszú, 2001).

Tests carried out on live audio broadcasts showed that the success of identification process depends on the proper selection of the representative short segment. The position where this soundprint can be extracted is determined by the recognition algorithm of the proposed system EMESE. The selected references must be non-correlated to avoid false alarms.

The method applied in EMESE is analyzed in the following section in order to explain the synchronization of the monitoring system to the test stream under various conditions. The measured results are also presented.

THE SOUND IDENTIFICATION IN EMESE

The audio signal, sampled at $f_s = 16kHz$ is transformed into a spectral description. It is a block of data, where the columns are feature vectors of the sound corresponding to a *frame* of time-domain data ($N_f = 256$ samples, $T_f = 16ms$ long). First, the amplitude of the Fourier spectrum is computed from the frame. Then averaging is adapted to the neighboring frequency lines to project the spectrum onto the Bark-scale. The reason