PERFORMANCE IMPROVEMENT of an AUTOMATED MEDIA CONTENT RECOGNITION SYSTEM

Berk Üstündağ[1] Celal Ergün[2]
bustundag@itu.edu.tr celale@formulsoft.com

[1] *Istanbul Technical University, Faculty of Electricity & Electronics, Maslak, 34469, Istanbul, Türkiye,*
[2]*ITU-KOSGEB Technology Development Center, Formulsoft, Maslak 34469, Istanbul, Türkiye*

Keywords: pattern recognition, media content indexing, audio fingerprint

ABSTRACT

There are several algorithms proposed for media content recognition and indexing but operation for more than thousand media channels with millions of audio tracks requires special considerations. In this study, we've explained modification of an audio fingerprint recognition algorithm with respect to hardware topology of the system in order to get optimal performance.

I. INTRODUCTION

Content monitoring and reporting of broadcast media are important for advertisement tracking, legal issues and commercial analysis. Due to rapid increment of media channels, recent tendency is automated recognition and analysis of the media content. Audio fingerprint based recognition and analysis appears as an effective solution against need of high amount of human resource in conventional fully operator dependent solutions. Although there are several algorithms proposed for content recognition and analysis in reporting systems, performing for more than thousand channels with millions of audio tracks requires special considerations. Performance optimization criteria includes storage & memory requirement, processing power requirement, operational costs, applicability, reliability, fault tolerance, error rate and minimum human support for meta data coupling.

Audio fingerprinting method for signal processing based media content monitoring has become popular in recent years. The main philosophy of the audio fingerprinting is extraction of a comparable image data file that represents distinguishable features of the sound track. An audio fingerprint based recognition procedure was proposed by Haitsma and Kalker [1][2] where hash functions are used for extraction and recognition. If two data streams X and Y are similar than their hash files as output of the hash functions H(X) and H(Y) have to be similar. Then recognition is defined as finding the matching X in a set of hash records $\{X_1,X_2,..,X_n\}$ from an indexed file. This seems as relatively a simple solution but there are several variants of possible algorithms and their configurations dramatically affect the need of CPU (MIPS in total) & memory (RAM) need besides false positive & false negative error rate. When the amount of parallel media channels to be recognized is high as thousand and such defined content size as advertisements and songs exceeds million then performance improvement becomes important as automated recognition success. Modification proposals for this purpose take place in section 2. Real performance on 1 year of 34 channel data takes place in the 3rd section as the sample application. Besides these hardware-software dependency related optimization, providing the robustness against deflection from the origin due to noise [3], [4] and codec effects [5] must also be taken into consideration. Audio fingerprinting is quite sensitive to change in audio play speed (approximately ±2%) and some solutions were proposed in past studies [6].

In this study, we've explained improvement of an audio fingerprint recognition algorithm with respect to hardware topology of the system in order to get optimal performance. System is currently being used for content reporting of local and national radio and TV channels in Turkey. If we consider that amount of all these real time audio stream resources is approximately 1300 stations, then importance of optimizing the media content reporting system can be more clearly understood. The most critical case occurs for advertisements since they dramatically increase the amount of meta-data to be coupled due to high renewal rates and version management requirements (figure 8). On the other hand they set the limits for minimum recognition time interval that is around 2.5 seconds. This limit is important not only for recognition of short audio tracks but also for catching the missing or non-broadcasted seconds of the advertisements.

MEDIA CONTENT REPORTING SYSTEM

Our first fully automated media content monitoring system was established in 2003 for experimental purposes (ITU-KOSGEB Technology Development Center) and enlarged it for recording and reporting purposes of all 34 national radio channels in 2004. In 2005 we've begun integration of regional information centers. Recognition of national TV channels was integrated to the system in the beginning of 2006. Each channel is received through preferably 3 methods for the data redundancy. Cable, satellite, air broadcast and Internet are the probable connection mediums. A multiplexer for each channel connects 2 of these received channels to the recording system. Each channel has two different records as a precaution against temporary loss of single channel record. On the other hand if signal is totally lost in one of these channels then system switches the lost line input to the 3rd signal receiver.

Hence reliability increases on digital record. Unfortunately this scheme cannot be applied to some local radios and TV channels since their broadcast is made through single way. On the other regional channels don't have transmission failure probability since most of them broadcasts over a specific region. Raw content record through the receivers is more important than doing the same with direct connection to stations since one of the goals is controlling the real air time and schedule of the ordered advertisements.

Scalability is an important feature of the system since memory and processing power requirement continuously increases as long as new audio tracks are loaded to the system. A grid architecture consisting of standard personal computers and servers are used in the system. PCs where operators input new meta-data are also used for recognition. Only three operators and a technician used to work for the automated operations of all 34 national channels. Totally more than 70 persons were working in 3 shifts for the same purpose in the previous manual monitoring equivalent system. A flexible reporting and data mining software runs incorporation with the database. People can externally reach to specified reports through an IIS in real time. Some basic system structures for audio signal recognition based radio & TV monitoring, reporting and data mining has been proposed in other studies [7], [8]. Grid architecture that runs on Gigabit Ethernet provides memory and CPU load balancing between the personal computers. Hence staff can handle 24hours process within 8 hours in the work day by using ordinary PC equipment. Increment of total channels to be monitored with local radio and TV connections caused raise in operator population within the proportion of 1 person/8 channel in time.

II. AUDIO FINGERPRINT BASED RECOGNITION

Using audio fingerprint requires special considerations in software with respect to target hardware infrastructure. Received sound stream is separated into frames as input of Fast Fourier Transform (Figure 1). FFT provides spectral distribution in p bands. Energy of the signals in each band is compared with the previous and the neighboring value as shown in figure 1.

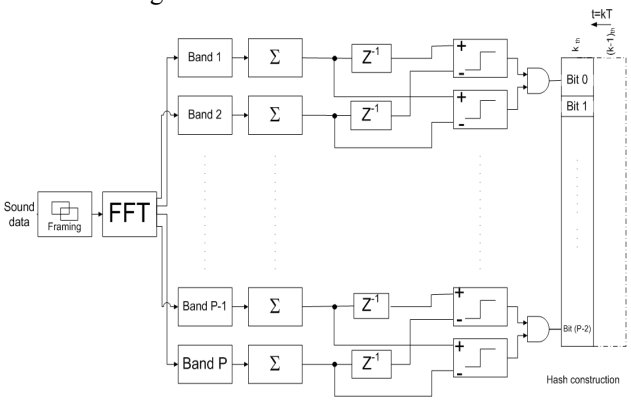

Figure 1. Hash file construction from the audio stream

Set of (p-1) bits are read out as the current data of the hash file due to audio stream. Most of the computers reach their optimal integer-arithmetic and logic calculation performance in their CPU bit size. So if the computer uses 64Bits CPU then p can be 65 so that 64bits handled as each framing interval of the system. Former studies had suggested 33 spectral bands [1] those yield 32bits-wide hash file and logarithmic spectrum distribution due to nature of human sensation. They had later improvements in amount of used energy bands to increase redundancy against play speed variations [6].

FFT has the highest CPU load portion with respect to frame size at a specified sampling frequency and the frame shift rate. Gaussian mixture models (GMM) using short time Fourier Transform was also proposed for audio classification and content based retrieval [9].

Finding the optimal frame size and shift rate in time affects the over all system performance both in hash file extraction and recognition phases. Audio fingerprint of 27.4 seconds length (11025Hz sampling) audio track with 2048 samples width frame size and 128bits frame shift rate, 4096 samples width frame size and 256bits frame shift rate, 8192 samples width frame size and 512bits frame shift rate are shown in figure 2. Wider shift rates cause smaller files but they also require larger frames. Frame size is the main factor that affects the hash extraction speed because FFT calculation load of CPU is proportional to sample amount in a frame due to number of MAC operations.

Figure 2. Audio finger print of 27.4 seconds 11025Hz audio track for a) 2048 samples of frame size – 128 samples of frame movement rate b) 4096 samples of frame size – 256 samples of frame movement a) 8192 samples of frame size – 512 samples of frame movement

Hash construction time of 1 hour audio record with respect to different frame widths are shown in figure 3. For the sake of objective comparison load sharing feature over the Intranet was cancelled during these tests. The test platform was a commercially available PC having an Intel Pentium IV CPU@2GHz and 2GByte RAM. It must be considered that only unrecognized records require new hash generation and recognition time is less than 20% of the generation time.

An important problem is high memory bank requirement of indexing phase contrary to the cost effective solution. Data indexing provides us finding the matching sub-fingerprints within the previously constructed hash file.

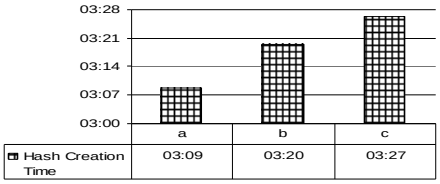

Figure 3. Comparison of hash creation time(min:sec) for 1 hour fingerprint of the audio stream sampled at 11025Hz a) 2048 samples of frame size – 128 samples of frame movement rate b) 4096 samples of frame size – 256 samples of frame movement c) 8192 samples of frame size – 512 samples of frame movement

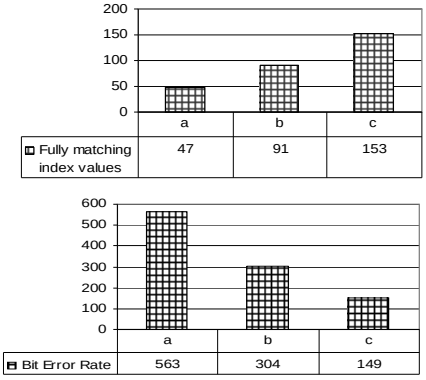

Figure 4. Amount of fully matching index values and respecting bit error rates for a)2048 samples of frame size – 128 samples of frame movement b) 4096 samples of frame size – 256 samples of frame movement c) 8192 samples of frame size – 512 samples of frame movement

When there are 4 Bytes of hash width respecting an audio track Px as shown in figure 5, normally use of 4 Bytes index pointers is necessary. During the training phase hash data is addressed in front of a cluster together with their position P(n,m). Here P(n,m) represents m_{th} 32Bits data in Audio record number n. Classification of all possible 32bits data variation in order to represent any value of P(n,m) requires 2^{32} Bytes=16GByte memory only for index pointers. This is a fixed value besides the additional memory requirement with respect to amount of recorded audio fingerprint. Computers having 32Gbytes or greater RAM are not cost effective since they are still not common for regular PCs (Commercially available mid-level personal computer main-boards were supporting upto 4Gbytes RAM and some special ones were able to use 16Gbytes in 2005). Using high capacity RAM disk banks for Content monitoring operation of 1000(s) channels could be a solution but the cost would be in million USD levels. We proposed a modification in the algorithm for solution of this problem. When we investigated the probabilistic data distribution in the frequency spectrum up to 3000Hz we proposed another solution. Highest order byte of each hash data (B3 for 32Bits) represents higher frequency components in the spectral distribution. Since the center of distribution is in the range of lower bands, less amount of changes occur in B3. We reduced 4Bytes hash data into 3 Bytes by applying logical "Exclusive Or" operation. Hence our Index pointer memory requirement reduced 256 times less to 64Mbytes and total memory requirement become quite feasible. On the other hand variations in B3 can still cause bit errors in recognition phase within an acceptable probability.

Indexing operation is represented with the block "I" in hash table construction and recognition schemes shown in figure 5 and figure 6. Since the same byte reduction method is applied in recognition phase, B3 has still positive addition in bit error counting (figure 6) that decreases false positive rate. Hash data I(B3, B2, B1, B0) indicates an index pointer. Matching memberships (m,n) are compared to the active track by adding the continuing components in time as (m,n-1) and (m,m+1). Each matching subfingerprint is byte by byte Exclusive-Ored in order to find bit error rate.

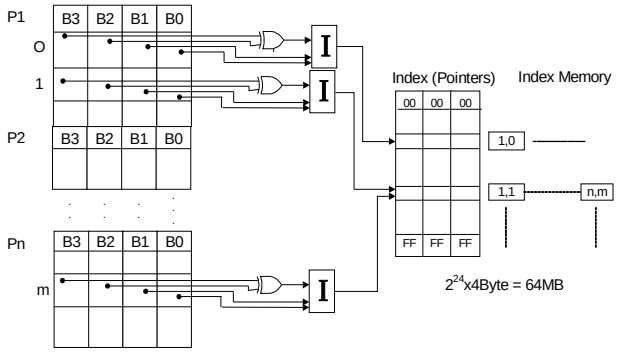

Figure 5. Indexing of hash files in memory

Bit error rate is count of logical "1"s since Ex-Or operand yields 1 only for differentiating bits as (0,1) or (1,0). It is inverse proportional to the likelihood of active fingerprint to be recognized and the ones having common index members within the recorded hash file. Likelihood level is entered as inverse of bit error threshold in recognition phase shown in figure 6.

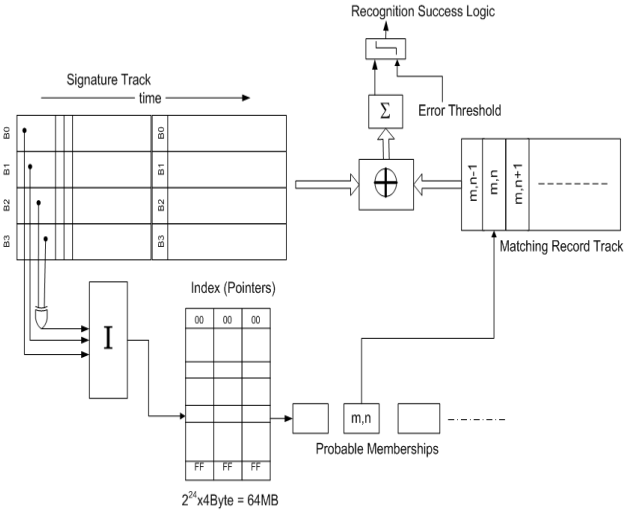

Figure 6.Modified recognition process over hash database

III. PERFORMANCE EVALUATIONS

Memory consumption of a computer in media content monitoring system is shown in figure 7. Index memory size is twice of hash record since track ID number (m) is stored together with the order number (n). Hash file is 2.5kbytes (11.025kHz sampling, 4096 samples-frame size, 256 samples frame movement rate) per minute for 1 single channel. A Pentium IV PC in the grid system performs 8 channels and meta-data is coupled through one operator in 8hours working for 24 hours record time. Variable part of the memory is totally equal to 7.5kByte/total record time (minute).

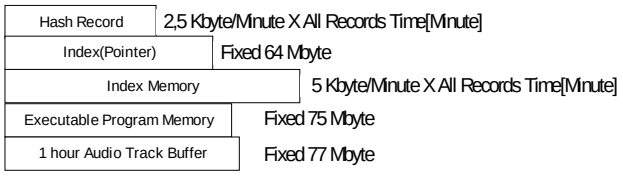

Figure 7.Memory budget with respect to total record duration

Distribution of 1 year content shown in figure 8 was extracted for a limited set that covers only 34 national radio channels in Turkey (9.8 Million plays of 9981 new advertisement, 1838 new song and 600 new generic program ID/year). It is different from distribution in terms of play time. When we consider that length of songs is centered around 200seconds with a normal distribution and advertisements are centered around 30 seconds then songs would appear as the highest portion if we had drawn the chart in terms of play (air) time. Total air time of the different content types is a function of repeat rate too. This figure reflects the need of operator work-time in coupling of the related meta-data for the first plays besides the automated recognition. We determined human support limit as 10..12channels/person but less ratio is used for long term efficiency and work comfort.

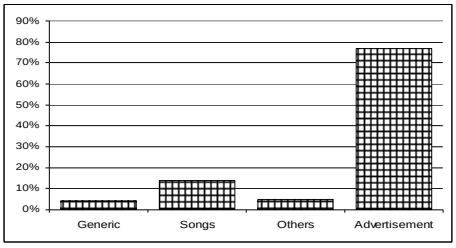

Figure 8. Distribution of new audio tracks in a year time,

Memory distribution after analysis of 9.8 million plays of more than 13.000 new media content is shown in figure 9.

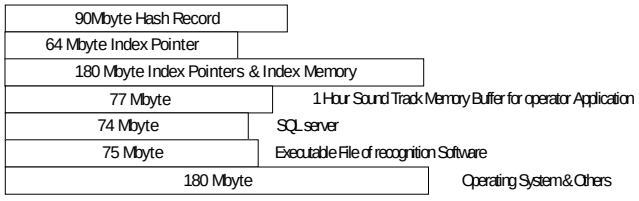

Figure 9. Optimized memory distribution in one PC unit at the end of one year record time.

IV.CONLUSIONS

Both the audio finger printing algorithm design and its parameter adjustments dramatically affects the hardware requirement for the specified recognition performance. We tested all variants of framing between 1024…8192sample width and 64…512 shifting rates on real data. Automated content recognition of 1hour radio record takes 35seconds with tuned parameters for 4096samples frame size/256bits shift rate (on a Pentium IV PC as explained in section 2). Reduction of 4 Bytes indexing of hash data down to 3 Bytes by using Ex-or operations allow us high amount of cost reduction although this does not cause significant change on recognition performance. Over all system error that includes human failures in meta-data connection was less than %0.05 by the end of 2005. These temporary errors are corrected manually in a secondary process. Highest CPU load comes from the FFT calculations in hash indexing phase. Cost effectiveness is important as the low failure rate especially for local radio stations because their income is quite low contrary to their population. Monitoring of the songs require rating together with the content reporting if it is intended to be used as royalty management mechanism. For this purpose, we also successfully tested and patented a new rating measurement method based on comparison of time stamped audio signature records [10].

REFERENCES

1. J. Haitsma and T. Kalker, “A highly robust audio fingerprinting system,” in Proc. of the 3rd Int. Symposium on Music Information Retrieval,, October 2002, pp. 144–148.
2. P. Cano, E. Batlle, T. Kalker, and J. Haitsma, “A review of algorithms for audio fingerprinting,” in International Workshop on Multimedia Signal Processing, December 2002.
3. F. Mapelli, R. Pezzano and R. Lancini : Robust Audıo Fıngerprıntıng For Song Identıfıcatıon: XII. European Signal Processing Conference September 6-10, 2004
4. Burges C.J.C., Platt J.C., Jana S.: Extracting Noise-Robust Features From Audio Data ICASSP, 2002.
5. J.Herre, O.Hellmuth and M.Cremer, “Scalable Robust Audio Fingerprinting Using MPEG-7 Content Description” IEEE International workshop on MMSP, 2002.
6. Jaap Haitsma and Ton Kalker**:** Speed-change resistant audio fingerprinting using auto-correlation:Acoustics : Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03). 2003 IEEE International Conference on Section IV- 728,731
7. Bruno Oliveira, Alexandre Crivellaro, Roberto M. César Jr.: Audio-Based Radio and TV Broadcast Monitoring Conference’05, Month 1–2, 2004, City, State, Country.
8. Cheung, T. Nguyen and Mining arbitrary-length repeated patterns in television broadcast, to appear at ICIP, Sept. 2005.
9. Arunan Ramalingam and Sridhar Krishnan : Gaussıan Mixture Modelig Using Short Time Fourier Transform Features For Audio Fingerprinting: Multimedia and Expo, 2005. ICME 2005. IEEE International Conference on 06-06 July 2005 Page(s):1146 - 1149
10. Ustundag, B., Ergür, C., "Content based rating system for audio and video broadcast", patent application, (2007), Turkish Patent Institute.