New Process to Identify Audio Concepts Based on Binary Classifiers Framework

Issam Feki, Anis Ben Ammar, and Adel M. Alimi

***Abstract*—Major researches in the field of audio classification have neglected the importance of preprocessing step of environmental sound recognition, or using usually classic sound classifiers. The originality of this paper is to construct a complete three modules process, acting dependently, with well defined functions. New method of acoustic sources separation are offered by our process, as well as, a sophisticated framework of binary classifiers is used to promote environmental sound classification. The results show that the proposed system has achieved accuracy higher than 90% for audio concept identification.**

***Index Terms*—Audio, concept, classification.**

I. Introduction

Audio, which includes voice, music, and various kinds of environment sounds, is an important type of media, and also a significant part of audiovisual data. As there are more and more digital audio databases in place at present, researchers start to realize the importance of audio database management relying on audio content analysis. Content-based audio classification and indexing have a wide array of application: audiovisual archiving management, supervising, guides system for deaf disability etc. For example, it will be very helpful to be able to identify sound effects automatically from a very large audio database of any kind of movies, which contains sounds of many concepts: explosion, animals, etc. Content-based audio could be an ideal approach for sound indexing and search. Furthermore, content analysis on audio is useful in audio-assisted video analysis. Possible applications include video scene classification, automatic segmentation and indexing of raw audiovisual recordings and audiovisual database browsing [1]. During the recent years, there have been many studies on automatic audio classification and segmentation to use several features and techniques. The most common problem with audio classification is speech/music classification, in which the highest accuracy has been achieved, especially when the segmentational information is known beforehand. In [2], wavelets are first applied to extract acoustical features such as subband power and pitch information. The method uses a bottom-up SVM over these acoustic features and additional parameters, such as frequency cepstral coefficients, to accomplish audio classification and categorization. In the same context, wavelet networks are used to recognize different phonemes in [3]. An audio feature extraction and a multigroup classification scheme that focuses on identifying discriminatory time–frequency subspaces using the local discriminate bases (LDB) technique is described in [4]. A new approach towards high performance speech/music discrimination on realistic tasks related to the automatic transcription of broadcast news is described in [5], in which an artificial neural network (ANN) and hidden Markov model (HMM) are used. The method proposed in [6], investigates the feasibility of an audio-based context recognition system where simplistic low-dimensional feature vectors are evaluated against more standard spectral features. Using discriminative training, competitive recognition accuracy is achieved with very low-order hidden Markov models [5]. The classification of continuous general audio data for content based retrieval was addressed in [7]; the audio segments are classified based on MFCC and LPC. The authors also showed that cepstral-based features gave better classification accuracy. The audio signals were decomposed in [8], using an adaptive time frequency decomposition algorithm, and the signal decomposition parameter based on octave (scaling) was used to generate a set of 42 features over three frequency bands within the auditory range. These features were analyzed using linear discriminated functions and classified into six music groups. An approach given in [9] uses support vector machine (SVM) for audio scene classification, it classifies audio clips into one of five classes: pure speech, non-pure speech, music, environment sound, and silence. In [10] a speech/music discrimination system was proposed based on Mel-Frequency Cepstral Coefficient (MFCC) and GMM classifier. This system can be used to select the optimum coding scheme for the current frame of an input signal without knowing a priori whether it contains speech-like or music-like characteristics. Through [11] an overview of audio-based multimedia indexing and retrieval scheme within MUVIS framework was presented .In that meaning, [12] experimentation tested such framework and if it would significantly increase the efficiency and the accuracy of audio-based retrieval especially in large multimedia databases. Also, in the context of content based multimedia indexing and retrieval, the concept is a cue. Therefore, automatic concept detection can be a tool in a content-based multimedia indexing system. This paper describes a novel approach for audio concept identification in content-based indexing. Several acoustic conditions exist in audio-visual data: compressed signal, noisy signal, and signal over background music. Thus, a concept identification system should be able to process this variety of acoustic signal conditions with acceptable performance. That's why; we have focused on environmental sound extraction and identification, to fill the gap in the acoustic research classification. We have used a number of features such as MFCC to characterize the audio content, in a first time. In order to classify audio concepts, we have used a binary

Manuscript received June 15, 2012; revised July 19, 2012.

The authors are with REGIM: Research Group on Intelligent Machines University of Sfax, ENIS, BP1173 - 3038, Sfax, Tunisia (e-mail: feki_issam@ieee.org, anis.benammar@ieee.org, adel.alimi@ieee.org)

classifiers framework, in the second time. This paper is organized as follows: The framework of our identification concepts system is described in Section 2. Experimental results using SVM and HMM are reported in Section 3. Finally, conclusion and future works are shown in Section 4

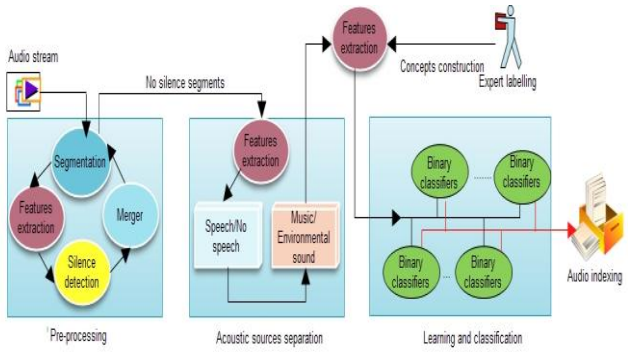

Fig 1. Framework of the proposed audio concepts identification system

II. Audio Identification Concepts System Framework

As per Fig1, we have tried to construct a complete process of ACIS (Audio Concepts Identification System) compounded by three modules acting dependently. The first one has as function segmentation and extraction of silence and no silence segments which are automatically sent to the second one which has two major functions: Discriminating between speech and non speech signals, in one hand, and separating the latter to music and environmental sounds, in the other hand. Finally, the third module would learn the environmental sound and classify them into concepts desired by an expert labeling.

A. Pre-processing Module

The segmentation modules in our system aim to separate heterogeneous input audio into speech, music and environmental sound regions. Initial segmentation is achieved in the time domain based on silence detection. The audio signal is sampled at 22050 Hz and 16 bits is ample. The audio stream is segmented into clips that are 3 seconds long with 1 second overlapping with the previous ones. Each clip is then divided into frames that are 512 samples long and are shifted by 256 samples from the previous frames. For each frame we extract STE (Short Time Energy) value [4], set that will be used to determine a silence segments. These segments are automatically eliminated because they are not part of our indexing. A merge module of no silence segments remaining runs to the preparation to a new segmentation. This segmentation is well oriented to the detection of speech and music classes of the audio stream obtained.

B. Acoustic Sources Separation Module

A two-step scheme is proposed to classify audio clips into one of three audio classes: speech, music, and environment sound. First, the input audio segments are separated into speech and non-speech segments by two features: Low short time energy ratio (LSTER) and Spectrum flux (SF). LSTER is defined as the ratio of the number of frames who's STE is less than 0.5 times of average short time energy in a one second window. LSTER is an effective feature, especially for discriminate speech and other no speech signals [13]. In general, there are more silence frames in speech, so the LSTER measure will be much higher for speech than that for no speech segments. Spectrum Flux (SF) is defined as the average variation value of spectrum between two adjacent frames in one second window [13]. In our experiments, we found that, in general, the SF values of speech are higher than those of all other no speech segments. Second, no speech segments are further classified into music and environmental sound, by a Band Periodicity feature (BP). Band periodicity is defined as the periodicity of each sub-band. It can be derived from sub-band correlation analysis. It is observed that the music band periodicities are generally much higher than those of environment sound [14]. This two-step scheme is suitable for different applications, and it can achieve high classification accuracy. Then, the segmentation of an audio stream can be carried out, by using these classification results. In extracting audio features in our classification scheme, whatever the sample rate of input signal could be, we all down sample it into 16 KHz sample rate and then segment it into sub segments by one-second window. This one-second audio clip is taken as the basic classification unit in method. It is further divided into forty 20 ms frames. Each feature is extracted based on these frames in one-second audio clip. In the end of this module, speech and music classes are considered in final audio data indexing. Thereafter, environment sound segments are the input of recognition concepts module of our system.

C. Learning and Classification Module

Labeling user sets 18 audio concepts for identification. His choice is based on an expanded study of the sound taxonomy which can give semantic meaning to the final user of a video search system. So the audio samples of each concept are introduced by a cepstral description. To obtain a good description for environmental sound segments, we extract MFCC feature. MFCCs are a set of perceptual parameters calculated from the STFT [4]. They provide a compact representation of the spectral envelope, such that most of the signal energy is concentrated in the first coefficient. After introducing training signal audio concepts, the feature extraction module gets in output a vector for each sound frames.

In this work, we designed a new framework of binary classifiers. This framework is continuously developing with the current user session; thereafter we get a most advantageous pattern of audio classifiers. Each binary classifier presented a single semantic audio concept. Each binary classifier processes the vectors outcome from a single descriptor. Therefore, whenever a new signal description is extracted, a binary classifier will be created. We preserved the past configuration as they are. Whenever an existing description is removed, the binary classifier will match simply deleted the system. Finally another binary classifier located in the output layer merges the binary outputs of all input layers. A simple binary output indicates the weight of each signal to its concept. This made the system scalable to any number of concepts since each time a new sound concept is defined by the expert; the system can simply create new binary classifier for the concept. So the classifier framework dynamically adapts to the demand for expert to the definition of a sound concept from the database. Each binary classifier is charged to learn the meaning of an individual part feature

vector corresponding to the discrimination of its concept. The main idea of this approach is to use the required number of classifiers in order to divide an enormous problem for learning in many units of classifiers and prevent the need for complex classifiers.

III. Experimental Corpus and Results

A. Corpus

The evaluation of the proposed concepts identification system has been performed by using a generic audio database totaling of 120 movie clips, 70 sports clips and 50 news clips. Audio samples are of different lengths, ranging from 1 s to about 10 s, with a sampling rate of 8 kHz and 16-bits per sample. The training data should be sufficient to be statistically significant. The training data is segmented into fixed-length and overlapping frames (in our experiments we used 30 ms frames). When neighboring frames are overlapped, the temporal characteristics of the audio content can be taken into account in the training process. Due to radiation effects of the sound from lips, high-frequency components have relatively low amplitude, which will influence the capture of the features at the high end of the spectrum.

B. Experimental Results

In our system, every training or testing vector is 26 dimensional, including 12-order MFCC and one log energy and their first time derivatives. The structural design for each MLP is defined by the layer range for the minimum and maximum number of layers. One for minimum and the other for maximum number of neurons allowed for each layer the size of both arrays is naturally max+1 where corresponding entries define the range of the l^{th} hidden layer for all those MLPs having an l^{th} hidden layer. We use *hyperbolic tangent* as the activation MLPs function:

$$\tanh(x) = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}} \quad (1)$$

The size of input $\{N_I\}$ and output layers $\{N_O\}$, is fixed (N_O =2 for all binary classifiers) and same for all configurations in an structural design space within which any *l*-layer MLP can be defined providing that $Min \le I \le Max$. For results evaluation, we use accuracy to estimate the audio identification method. The accuracy equation is presented as follows:

$$Detection\ accuracy = \frac{\#(reievant\ \ items)}{\#(All\ \ relevant\ \ items)} \quad (2)$$

For contrast, we also perform experiments on other classification systems. The first one is SVM-based classification with the following parameter settings: The penalty factor C is set to 1 and the kernel function is Gaussian kernel with α =10. [12]

$$K(x, y) = \exp\left(\frac{-\|x - y\|^2}{2, \alpha^2}\right) \quad (3)$$

The second classification system is a hidden Markov classification model (HMM). The third one is our classification system in which the framework of binary classifiers are set. The experimental results of the proposed sound classification systems are listed in Table I.

C. Discussion

Compared to HMM and SVM approaches, our system shows the best results for all concepts (accuracy about 85%). It is also seen that speech and music have higher classification accuracy than different environment sound. This is not surprising, since both speech and music are pure types and they are easily classified. Environment sound can include sound of animal (e.g. dog barking), some man-made sound (e.g. police alarm or breeze glasses), and sound of nature (e.g. wind), etc. Therefore, the characteristics displayed by different environment sounds may be quite different, which results in lower accuracy compared with speech and music. However, in all cases, features obtained from our framework performs better than other kinds of clip-level features, this proves that MFCC feature integration using framework of binary classifiers is effective. However, there is one interesting observation in the HMM classifier experiments. We found 28 of 35 helicopter's sound files were misclassified as explosion. This phenomenon can be explained by Figs. 2 and 3. In the view of spectrogram, some helicopter's sound files to be classified are very similar to the explosion's sound files in the database. However, this drawback was lessened drastically by using the proposed framework of binary classifiers set. The misclassified helicopter's sound files were reduced from 28 to six. This experiment shows efficiency of the proposed system.

TABLE I. Classification Accuracy for Different Systems

Audio Concepts	SVM	HMM	Our System
Speech	96,01%	98,33%	99,20%
Music	92,20%	95,87%	96,43%
Ringtones	85,08%	88,07%	91,27%
Train	77,13%	80,12%	85,62%
Motorcycle	72,89%	74,17%	76,91%
Explosion	85,50%	17,12%	91,00%
Helicopter	84,80%	94,89%	89,93%
Slamming door	79,06%	82,19%	84,48%
Dog barking	83,07%	84,66%	87,91%
Bird	82,19%	82,29%	86,67%
Breeze glasses	89,01%	90,09%	94,00%
Applause	87,88%	87,72%	92,07%
Horse	88,04%	89,00%	93,12%
Cat	89,00%	89,11%	94,13%
Care	76,29%	79,32%	81,92%
Slot machine	73,87%	71,56%	74,11%
Wind	72,37%	73,09%	73,33%
Plane	79,97%	81,64%	84,99%
Laugh	82,59%	84,56%	87,03%
Police alarm	91,02%	94,11%	96,44%

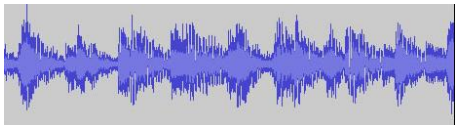

Fig. 2. Spectrogram of explosion's sound file

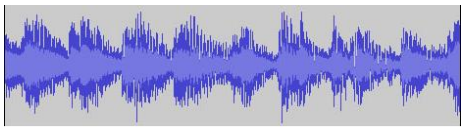

Fig. 3. Spectrogram of helicopter's sound file

IV. Conclusion

This paper presented an environmental sound extraction and incremental learning approach. It had involved a new classification scheme which has shown its flexibility and efficiency in signal audio identification. Better results should be obtainable after combining both audio and visual information (image or text) and setting prior known constraints to the model parameters. This work is currently underway within our RegimVid (REsearch Group on Intelligent Machines, Video) team dedicated to video processing based on various modalities. Indeed, integration with a multimodal indexing system may give access to more semantic content than that stated above (extract the meaning of the audio signal). The type of information we can extract could be: keyword detection, speaker recognition, recognition of jingles / music, information on the acoustic environment, information on mounting the soundtrack

Acknowledgment

The authors would like to acknowledge the financial support of this work by grants from the General Direction of Scientific Research and Technological Renovation (DGRSRT), Tunisia, under the ARUB program 01/UR/11/02. A special thanks to Pr Moncef Gabbouj (Professor, Senior Scientist; Tampere University of Technology) and all his partners especially Pr Serkan Kiranyaz (Professor; Tampere University of Technology) who generously shared their expertise during my probabation period in their research unity (http://www.cs.tut.fi).

References

[1] H. Karray, A. Wali, N. Elleuch, A. Ben Ammar, M. Ellouze, I. Feki, and Adel M.Alimi, "REGIM at TRECVID2008: High-level Features Extraction and Video search," in *Proc. of the International Conference TREC*, 2008.

[2] C. Lin, S. Chen, T. Truong, and Y. Chang, "Audio classification and categorization based on wavelets and support vector machine," *IEEE Transactions on Speech and Audio Processing*, vol. 13, pp. 644–651, 2005.

[3] R. Ejbali, Y. Ben Ayed, M. Zaied, and Adel M. Alimi, "Wavelet Networks for phonemes Recognition," *presented at the International Conference on Systems and Processing Information*, Algeria, 2009.

[4] K. Umapathy, S. Krishnan and R. Rao, "Audio signal feature extraction and classification using local discriminate bases," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, pp. 1236–1246, 2007.

[5] J. Ajmera, I. McCowan, and H. Bourlard, "Speech/music segmentation using entropy and dynamism features in a HMM classification framework," *Speech Communication*, vol .40, pp. 351–363, 2003.

[6] A. Eronen , T. Peltonen V., T. Tuomi J, P. Klapuri A., and S. Fagerlund, "Audio-based context recognition," *IEEE Transactions on Audio, Speech and Language Processing*, vol .14, pp. 321–329, 2006.

[7] Li. Sethi , N. Dimitrova, and T. McGee, "Classification of general audio data for content-based retrieval," *Pattern Recognition Letters*, vol. 22, pp. 533–544, 2001.

[8] K. Umapathy, S. Krishnan, and S. Jimaa, "Multigroup classification of audio signals using time–frequency parameters," *IEEE Transactions on Multimedia*, vol. 7, pp. 308–315, 2005.

[9] H. Jiang, J. Bai, S. Zhang, and B. Xu, "SVM-based audio scene classification," in *proc. of IEEE*, pp. 131–136, 2005.

[10] M. Mubarak, E. Ambikairajah, and J. Epps, "Analysis of an MFCC-based audio indexing system for efficient coding of multimedia sources," in *Proc. IEEE international conference on acoustics, speech and signal processing*, pp. 619–622, 2005.

[11] M. Gabbouj, S. Kiranyaz, K. Caglar, E. Guldogan, and O. Farooq A, "Audio-based Multimedia Indexing and retrieval scheme in MUVIS framework," in *Proc. of IEEE International Symposium on Intelligent Signal Processing and Communication Systems*, 2003.

[12] S. Kiranyaz, F. Qureshi A., and M. Gabbouj, "A Generic Audio Classification and Segmentation Approach for Multimedia Indexing and Retrieval," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 13, pp. 1062-1081, 2006.

[13] E. Scheirer and M. Slaney, "Construction and Evaluation of a Robust Multifeature Music/Speech Discriminator," in *Proc. of the International Conference Acoustic Speech and Signal Processing*, pp. 1331-1334, 2007.

[14] K. El-Maleh, M. Klein, G. Petrucci, and P. Kabal, "Speech/music discrimination for multimedia application," in *Proc. of the International Conference Acoustic Speech and Signal Processing*, 2000.

Issam Feki was born in 1977 and grew up in Sfax (Tunisia). He received his M.Sc. degree in computer science in 2005 from the National School of Engineers, University of Sfax, Tunisia. He is an assistant with the Department of Computer Science, Higher Institute for Computer Sciences and Management -University of Gabes, Tunisia.

He is currently pursuing the Ph.D. degree at the University of Sfax, and is a member of the Research Group on Intelligent Machines (REGIM) at university of Sfax, supervised by Professor Adel M. Alimi and Dr. Anis Ben Ammar (from university of Sfax). His dissertation focuses on multimedia treatment and audio signal processing. His research interests include audiovisual analyzing systems, audio systems, indexing based on audio concepts.

Mr. Issam is IEEE member since 2007.

Anis Ben Ammar was born in Sfax (Tunisia) in 1975. He graduated in Computer Sciences 1998, obtained a Master degree 1999 and Ph.D. 2003. He is now an assistant professor in Computer Sciences at the University of Sfax.

His research interest includes information retrieval systems. He focuses especially on Multimedia retrieval in large scales collection. He is interested in several fields within information retrieval framework. The audiovisual data indexing is the main research axis including applications of intelligent methods (neural networks and fuzzy logic).

Dr, Anis is the leader of the REGIM (Research Laboratory Attachment) participation in Benchmarks evaluation as TRECVID and Image CLEF.

Adel M. Alimi was born in Sfax (Tunisia) in 1966. He graduated in Electrical Engineering 1990, obtained a Ph.D. and then an HDR both in Electrical & Computer Engineering in 1995 and 2000 respectively. He is now professor in Electrical & Computer Engineering at the University of Sfax.

His research interest includes applications of intelligent methods (neural networks, fuzzy logic, evolutionary algorithms) to pattern recognition, robotic systems, vision systems, and industrial processes. He focuses his research on intelligent pattern recognition, learning, analysis and intelligent control of large scale complex systems.

He is associate editor and member of the editorial board of many international scientific journals (e.g. "Pattern Recognition Letters", "Neuro Computing", "Neural Processing Letters", "International Journal of Image and Graphics", "Neural Computing and Applications", "International Journal of Robotics and Automation", "International Journal of Systems Science", etc.).

Prof. Alimi was guest editor of several special issues of international journals (e.g. Fuzzy Sets & Systems, Soft Computing, Journal of Decision Systems, Integrated Computer Aided Engineering, Systems Analysis Modeling and Simulations).He was the general chairman of the International Conference on Machine Intelligence ACIDCA-ICMI'2005 & 2000. He is an IEEE senior member and member of IAPR, INNS and PRS. He is the 2009-2010 IEEE Tunisia Section Treasurer, the 2009-2010 IEEE Computational Intelligence Society Tunisia Chapter Chair, the 2011 IEEE Sfax Subsection, the 2010-2011 IEEE Computer Society Tunisia Chair, the 2011 IEEE Systems, Man, and Cybernetics Tunisia Chapter, the SMCS corresponding member of the IEEE Committee on Earth Observation, and the IEEE Counselor of the ENIS Student Branch.