TOWARDS A UNIFIED FRAMEWORK FOR CONTENT-BASED AUDIO ANALYSIS

Lie Lu[1], Rui Cai[2], and Alan Hanjalic[3]

[1]Microsoft Research Asia, Beijing, P.R. China
[2]Department of Computer Science and Technology, Tsinghua University, Beijing, P.R. China
[3]Department of Mediamatics, ICT Group, Delft University of Technology, Delft, The Netherlands

ABSTRACT

Audio content analysis is helpful in many multimedia applications. In this paper, we present a unified framework for content analysis of composite audio. The framework is designed to extract relevant information from different available audio modalities and to discover high-level semantics conveyed by the data. We also demonstrate an implementation of the proposed framework for scenes and events detection in various TV shows and movies, in which key audio effects are first extracted as mid-level representation and then Bayesian Network is used for high-level semantics inference. Experiments on 12-hour audio data indicate that the proposed framework has a satisfying performance.

1. INTRODUCTION

Nowadays, more and more digital audio data appear in various multimedia databases, either stand-alone (e.g. music, radio broadcasts) or combined with other media (e.g. visual and/or textual) into multimedia documents. However, most of the audio data are not indexed, which makes the contained information difficult or inconvenient to reuse. Building a system for audio content indexing is likely to facilitate the management of audio data and support various multimedia applications where this data plays a role.

To be able to index audio data, semantic information needs to be extracted from data. Most existing audio indexing systems focus on speech signals. This is, first, because reliable speech recognition tools are available, using which audio signals can be transcribed into the text domain; and second, because advanced algorithms for text information analysis and retrieval can then be applied to reveal the topic structure and items of a spoken audio document. Fewer solutions have been proposed so far addressing non-speech and composite audio signals. Saunders [1] presents a speech/music classifier based on simple features such as zero-crossing rate and short-time energy for radio broadcast. Lu et al. [2][3] present schemes to classify audio signals into four classes, including speech, music, noise, and silence, by using heuristic rules or SVM. Wold et al. [4] propose an approach to detecting more specific audio categories, such as animal sounds, bell, crowds, laughter, machine and musical instruments. The scope of existing non-speech indexing approaches also includes methods for extracting audio keywords [5] or highlight sound effects [6] from audio signals.

To be able to extract more semantic information from a multimodal data stream, research has been reported on audio scene classification and highlight detection to support video content analysis. For example, Liu et al. [7] study the problem of classifying TV broadcast into five different categories: news, commercial, weather forecast, basketball game, and football game by using a 3-layer feed forward neural network classifier. In sports video analysis [8], highlight events are detected based on special audio effects like cheering, ball-hit, and whistling; while in film indexing [9] sounds like car-braking, siren, gun-shot, and explosion are used to identify violent scenes in action movies.

Compared to previous approaches on analyzing composite audio, which are usually heuristic or constrained on a certain audio type, we develop in this paper a unified framework for composite audio content analysis applicable to a large variety of multimedia data streams. The framework is designed to integrate information from different audio modality and to discover high-level semantic concepts from mid-level representations instead of from low-level features directly. We also demonstrate the effectiveness of the proposed approach on a broad set of test videos including entertainment TV shows and action/war movies.

The rest of the paper is organized as follows. Section 2 presents the unified framework for composite audio analysis. Section 3 demonstrates an implementation of the proposed framework and the corresponding modules. Evaluation and discussion are given in Section 4. Section 5 concludes the paper.

2. THE UNIFIED FRAMEWORK

Our proposed unified framework for content-based analysis of composite audio is illustrated in Fig. 1. Here, the attribute "unified" refers to the ability to deal with an arbitrary audio source, and thus makes the framework suitable for various applications.

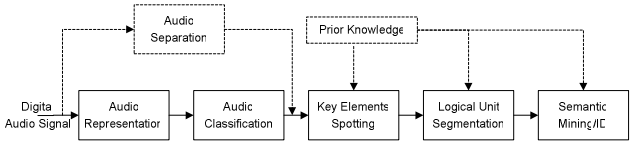

Fig. 1. A unified framework for content-based audio analysis

The framework actually represents the generic process of audio content understanding, from low-level features, via mid-level content representation, to high-level semantics. It consists of five main modules: audio representation, audio classification, key-elements spotting, logical unit segmentation, and semantic mining. Moreover, prior knowledge can also be integrated in order to improve the performance of key-elements spotting, logical unit segmentation and semantic mining.

The basic processing flow of the framework is as follows. Firstly, in the Audio Representation module, the input audio

signal is represented by low-level features (including temporal, spectral or structural features), which should have enough discrimination capability regarding different audio types. Then, in Audio Classification module, the audio signal is classified and temporally segmented into a number of general audio modalities such as speech, music, and background sounds. Subsequently, key components, such as a keyword in speech and a key audio effect of a background sound, are detected from audio signals in the Key-Elements Spotting module. A key component here is a mid-level representation of an audio signal and serves to enable direct links to semantics. As such, it provides the basis for further semantics discovery. Based on the obtained low-level and mid-level representations, the audio signal is further segmented into homogeneous logical units in the module called Logical Unit Segmentation. A logical unit is a segment which has a coherent semantic content, such as an event, a scene, a topic or an episode. Finally, in the Semantic Mining module, the semantic concept of each logical unit is discovered.

The audio separation module in the dashed block can be used to spectrally separate different audio modalities from each other. Thus, together with Audio Classification module where temporal signal segmentation is performed, a composite audio signal can be spectrally and temporally separated. As spectral audio separation can still not be performed reliably on general acoustic data, we consider in our implementation the temporally composite audio only. Once reliable algorithms for blind-source signal separation are available, they can easily be incorporated into our framework.

3. IMPLEMENTATION

Using the proposed framework we may extract different features, key elements and target semantics for different applications. In this section, we demonstrate an implementation of the framework for detecting scenes or events from entertainment TV shows and action/war movies. In this particular application we define five semantic categories that are to be extracted, namely excitement, humor, pursuit, fight, and air-attack. Each of these categories can be characterized by a number of representative audio effects. For example, the category "Excitement" could be detected based on cheering and applause, while laughter is an important indication for the "Humor" category. "Pursuit" can usually be found in action movies and could be detected through the presence of car-crash and car-braking sounds, sirens, helicopter sounds, gun-shots and explosions. The categories "Fight" and "air-attack" often occur in action and war movies. Scenes of fight mostly contain gun-shots and explosions, while explosions and sound of planes could be the indications of the "air-attack" category. Based on the above, ten key audio effects can be selected to support the detection of the defined high-level semantic categories: *applause*, *laughter*, *cheer*, *car-braking*, *car-crash*, *explosion*, *gun-shot*, *helicopter*, *plane*, and *siren*. Detailed implementations of each module are explained in the following.

3.1 Audio Representation

Many audio features have been proposed so far in the context of content-based audio analysis and have been proved to be effective in providing the base for interpreting audio signals at the semantic level [3][4][7]. Building on these previous works, we choose to extract a number of "traditional" temporal and spectral low-level features for the application context defined above.

In the temporal domain, we extract short-time energy (STE) and zero-crossing rate (ZCR). STE provides a good representation of the amplitude or the loudness, while ZCR gives a rough estimation of the frequency content in an audio signal. Our selected set of spectral features consists of band energy ratios (BER), brightness, bandwidth and Mel-frequency Cepstral Coefficients (MFCCs). BER describes the characteristics of spectral energy distribution. In our experiments, the spectral domain is equally divided into 8 sub-bands and the energy in each sub-band is then normalized by the whole spectrum energy. Brightness and Bandwidth are the first-order and second-order statistics of the spectrogram respectively. They roughly measure the timbre quality of a sound. MFCC is a sub-band energy feature in mel-scale, which gives a more accurate simulation of human auditory system. As suggested in [2][3], 8-order MFCC is used in the experiments.

Besides the above features, two new spectral features, namely *sub-band spectral flux* and *harmonicity prominence*, are also used, based on our previous works on audio representation [10]. *Sub-band spectral flux* is used to measure whether there are salient frequency components in each sub-band, while *harmonicity prominence* estimates the harmonic degree of a sound.

3.2 Audio Classification & Key Element Detection

In this step a composite audio is temporally segmented into (close-to) mono-modal segments including speech, music or background noise, and then the key elements are extracted from the obtained segments. As explained above, the key elements are the effects serving as indications for the presence of a high-level semantic concept. Since "speech", "music" or "noise" can also be considered a key element in semantics discovery, in our implementation, audio classification and key effect detection are implemented together. Besides the "speech", "music" and "background sound" label, ten special key audio effects are also extracted: applause, laughter, cheer, car-braking, car-crash, explosion, gun-shot, helicopter, plane, and siren.

We employ Hidden Markov Model (HMM) for key audio effect modeling since HMM provides a natural way for modeling time varying processes [10], and has been proven to be effective for audio effect modeling in many approaches proposed before [6]. Unsupervised k-mean clustering with Bayesian Information Criterion (BIC) [15] is performed on the training sets to estimate the HMM states of each key audio effect model,. The results are listed in Table 1. Here, more states are used than in our previous work [6], in order to catch a large variety of training samples. Also, due to a large variety of general audio classes, such as speech and music, 128 Gaussian mixtures are used for them in our approach.

Table 1 HMM States for Key Audio Effects Modeling

Key Audio Effects	States	Key Audio Effects	States
applause	8	*gun-shot*	10
car-braking	9	*Helicopter*	7
car-crash	9	*Laughter*	8
cheer	5	*Plane*	10
explosion	9	*Siren*	11

As for the topology of HMM, the most popular topologies are left-to-right and fully connected. The left-to-right structure only permits transitions between adjacent states; while fully connected structure allows transitions between any two states in the model. In our approach, we use both structures to model key effects with different properties:

- For key effects with obvious characteristics in their progress phases, such as car crash and explosion, left-to-right structure is adopted.
- For key effects without distinct evolution phases such as applause and cheer, or general audio classes, fully connected structure is applied.

3.3 Logical Unit Segmentation

Having extracted the key effects from an audio signal, we are now searching for potential temporal segments which are most likely characterized by a coherent semantic content. We model this coherence using auditory contexts, as illustrated in Fig 2.

An auditory context is defined by a number of subsequent key effects. Two adjacent key effects are assumed to be in the same context if the time interval between them is sufficiently short. In the same way, a new context is started if the time interval between two key effects is larger than a pre-defined threshold T_m. The threshold T_m can be determined based on the estimate of the upper limit for the length of human memory window to perceive a consecutive scene. We set this threshold to 16 seconds, following the discussion in [12]. It is noted that, with this scheme, two neighboring logical units may have the same semantic meaning.

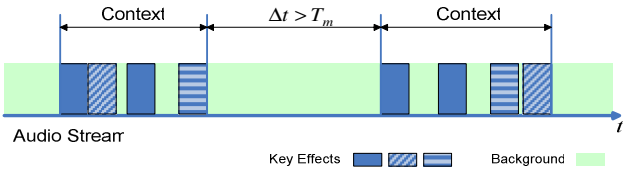

Fig2. Examples of logical unit (auditory context) in audio stream

3.4 Semantic Concept Detection

Each of the detected logical units is now tested for the presence of the defined five target semantic categories: excitement, humor, pursuit, fight, and air-attack. To infer high-level semantics from obtained key effects, most of previous works utilize rule-based approaches [5] or statistical classification [9]. Although heuristic inference is straightforward and easy to be applied in practice, it is laborious to find a proper rule set if the situation is complex. Also, some rules may be in conflict with others, and some cases may not be well-covered. For a classification-based approach, the inference performance highly relies on the completeness and the size of the training samples. Without a sufficient training database, positive instance not included in the training set will usually be misclassified. To solve these disadvantages, Bayesian Network is used for semantics inference in our framework.

A Bayesian Network [13] is a directed acyclic graphical model that encodes probabilistic relationships among nodes which denote random variables related to semantic concepts. Bayesian Network can handle situations where some data entries are missing, as well as avoid the over-fitting of training data [13]. Thus, it weakens the influence from the unbalanced training samples. Furthermore, a Bayesian Network can also integrate prior knowledge by specifying its graphic structure.

Fig 3 illustrates the graphic topology of a two-layer Bayesian Network. There are 13 nodes in the bottom layer to represent ten key audio effects and three general audio classes in each logical unit. In this way, the information from different audio modality is roughly integrated. Further, there are 5 nodes in the top layer, denoting the five predefined semantic categories. The structure can be manually specified according to the causal relationships between semantic categories and audio effects. That is, the children of a category node only include those key effect nodes which have relationships with this semantic category. Moreover, the nodes in the top layers are assumed to be discrete binaries denoting the presence or absence of the corresponding category; and the nodes in the bottom layer are continuous-valued with Gaussian distribution. The model parameters are uniformly initialized, and then the model training and semantic inference are implemented with the Bayes Net Toolbox [14].

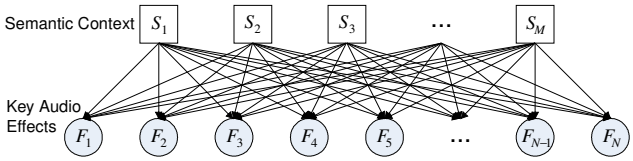

Fig. 3 A Bayesian Network for auditory context inference

4. EVALUATION

We evaluated the proposed system on 12 hours of audio extracted from various video genres, including action/war movies and entertainment TV shows. Detailed information on the audio tracks used is listed in Table 2. All audio streams are in the 16 KHz, 16-bit and mono channel format, and are divided into frames of 30 ms with 50% overlap for feature extraction. In total, around 100 training samples are annotated for each key audio effect, and 5 hours for general audio classes: music, speech and noise.

Table 2. Description of the test audio data set

Movie/TV Title	Genre	Duration
Saving Private Ryan	war	2:41:41
Enemy at the Gates	war	2:11:08
Swordfish	action	1:39:17
The Rock	action	2:16:35
3rd Rock from the Sun	situation comedy	0:30:05
Hollywood Squares	TV shows	0:25:43
59th Annual Golden Globe Awards	TV shows	2:14:50

We firstly evaluate the performance of key effect detection. Table 3 lists the precision, recall and F1-measure obtained for each target key audio effect. It can be seen that most effects are detected with high recall and precision.

Table 3. Results on audio effect detection

	Recall	Precision	F1
applause	97.66%	91.40%	94.43%
cheer	97.17%	65.84%	78.50%
laughter	99.09%	71.09%	82.79%
car-brake	86.54%	85.64%	86.08%
car-crash	98.17%	69.36%	81.29%
explosion	80.00%	76.04%	77.97%
gun-shot	80.50%	76.83%	78.62%

helicopter	98.66%	85.01%	91.33%
plane	92.79%	89.05%	90.88%
siren	99.10%	93.05%	95.98%

Then we evaluate the performance of semantics inference. To obtain a better picture about the performance, three inference approaches are compared: rule-based approach, SVM-based approach, and our Bayesian Network-based approach. In our approach, the heuristic rules are mainly based on the description in Section 3. The inputs for SVM and Bayesian network are the duration ratio and confidence of each key audio effect. The comparison results are listed in Table 4.

Table 4 Comparison among different Inference Engines regarding their ability to detect the defined semantic concepts

Inference Engine	Recall	Precision	F1
Heuristic Rule-based	84.87%	71.47%	77.60%
SVM-based	77.21%	83.33%	80.15%
Bayesian Network	***89.34%***	***82.37%***	***85.71%***

From Table 4, it can be seen that the heuristic rule-based method usually results in a high recall but low precision. This is because the rules are usually set for the target phenomena, so that the negative samples are often misclassified. Opposed to this, a SVM-based approach obtains a high precision but low recall, since it typically cannot detect the instances not included in the training set. Comparatively, a Bayesian Network can handle situations where some data entries are missing so that it has both a high recall and a similarly high precision and shows a better overall performance than the other two approaches. As shown in Table 4, the Bayesian Network offers similar precision to SVM while improve recall by about 12%, and obtains similar recall to the rule-based approach while improve the precision by about 11%. A confusion matrix of high-level semantic inference based on Bayesian Network is listed in Table 5

Table 5 Confusion matrix of semantic inference based on BN

	excitement	humor	pursuit	fight	air-attack
excitement	29	1	0	0	0
humor	2	36	0	1	0
pursuit	0	0	81	10	0
fight	0	0	5	72	0
air attack	0	0	1	1	25

Finally, we also compared the performance of semantic inference based on low-level features directly and based on mid-level representations (key audio effects). The mapping between low-level features and semantics are modeled by GMM with 128-mixtures in our approach. From Table 6 it can be seen the semantic inference based on key audio effects results in a higher overall performance.

Table 6. Performance comparison of high-level semantic inference from key audio effects and from low-level features

Category	From key audio effects		From low-level features	
	Recall	Precision	Recall	Precision
excitement	93.55%	93.55%	*78.13%*	*92.59%*
humor	94.74%	92.31%	*94.74%*	*83.72%*
pursuit	91.01%	82.65%	*85.23%*	*77.32%*
fight	80.90%	72.73%	*82.76%*	*80.90%*
air-attack	100%	89.29%	*55.56%*	*93.75%*
Average	89.34%	82.37%	82.07%	82.27%

5. CONCLUSION

In this paper, we presented a unified framework for content analysis of composite audio and demonstrated its implementation for a specific application. The framework utilizes the mid-level representation to discover high-level semantic concepts, with a Bayesian Network based approach. It can also use the information of multiple audio modalities. Experiments indicated that the framework has a good flexibility and a satisfying performance.

We see the possibilities to further improve the proposed framework mainly by finding better ways of integrating different audio modalities, and by increasing the robustness of logical unit segmentation.

6. REFERENCES

[1] J. Saunders. "Real-time Discrimination of Broadcast Speech/ Music". *Proc.ICASSP96*, Vol.II, pp.993-996, 1996

[2] L. Lu, H.-J. Zhang, H. Jiang, "Content Analysis for Audio Classification and Segmentation", *IEEE Trans. on Speech and Audio Processing*, Vol.10, No.7, pp.504-516, 2002.

[3] L. Lu, H.-J. Zhang, S. Li, "Content-based Audio Classification and Segmentation by Using Vector Machines", *ACM Multimedia Systems Journal* 8 (6), pp. 482-492, March, 2003.

[4] E. Wold, T. Blum, and J. Wheaton. "Content-based Classification, Search and Retrieval of Audio". *IEEE Multimedia*, 3(3), pp.27-36, 1996

[5] M. Xu, N. Maddage, C.-S. Xu, M. Kankanhalli, and Q. Tian, "Creating Audio Keywords for Event Detection in Soccer Video," *Proc. of ICME* Vol.2, pp.281-284, 2003.

[6] R. Cai, L. Lu, H.-J. Zhang, and L.-H. Cai, "Highlight Sound Effects Detection in Audio Stream," *Proc. of ICME* Vol.3, pp.37-40, 2003.

[7] Z. Liu, Y. Wang, and T. Chen, "Audio Feature Extraction and Analysis for Scene Segmentation and Classification," *J. VLSI Signal Processing Sys. Signal, Image, Video Technology*, Vol. 20, pp. 61-79, Oct.1998.

[8] Y. Rui, A. Gupta, and A. Acero, "Automatically Extracting Highlights for TV Baseball Programs", *Proc. of 8th ACM Multimedia*, pp.105-115, 2000.

[9] S. Moncrieff, C. Dorai, and S. Venkatesh, "Detecting Indexical Signs in Film Audio for Scene Interpretation", *Proc. of ICME*, pp. 1192-1195, 2001.

[10] R. Cai, L. Lu, H.-J. Zhang, L.-H. Cai, "Improve Audio Representation by Using Feature Structure Patterns", *Proc. of ICASSP 2004*.

[11] L.R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257-286, Feb. 1989

[12] H. Sundaram and S.-F. Chang, "Audio scene segmentation using multiple features, models and time scales," *Proc. of ICASSP'00*, vol. 4, pp. 2441-2444, 2000.

[13] D. Heckerman, "A tutorial on learning with Bayesian networks," *Microsoft Research, Tech. Rep*. MSR-TR-95-06.

[14] K. Murphy, "The Bayes net toolbox for Matlab", *Computing Science and Statistics*, vol. 33, 2001.

[15] D. Pelleg and A.W. Moore, "X-means: Extending K-means with Efficient Estimation of the Number of Clusters," *Proc. of ICML*, pp.727-734, 2000.