

# AUDIO ELEMENTS BASED AUDITORY SCENE SEGMENTATION

Lie Lu<sup>1</sup>, Rui Cai<sup>2</sup>, and Alan Hanjalic<sup>3</sup>

<sup>1</sup>Microsoft Research Asia, Beijing, P.R. China

<sup>2</sup>Department of Computer Science and Technology, Tsinghua University, Beijing, P.R. China

<sup>3</sup>Department of Mediamatics, ICT Group, Delft University of Technology, Delft, The Netherlands

## ABSTRACT

Auditory scene segmentation is an important step in the process of high-level semantic inference from audio data streams, and in particular, a prerequisite for auditory scene categorization. In this paper, we analyze the limits of previous works on auditory scene segmentation, and then propose a novel method that, conceptually, is inspired by the ideas used in text and video scene segmentation, and is based on an analysis of *audio elements* and *key audio elements*, which can be seen as equivalents to the words and keywords in a text document, respectively. Experiments performed on 1.5 hours of audio data indicate that the proposed approach is promising.

## 1. INTRODUCTION

Nowadays, more and more composite digital audio data appear in various multimedia databases, either stand-alone (e.g. radio broadcasts) or combined with other media (e.g. visual and/or textual) into multimedia documents. As opposed to single-modal audio (e.g. pure music or speech), composite audio usually contains multiple audio modalities such as speech, music and various audio effects, which are either mixed together or follow each other in a sequence. Because most of the audio data streams appearing in multimedia applications are composite, building a system for content-based composite audio analysis is likely to facilitate the management of audio data and support various multimedia applications where this data plays a role.

A typical approach to content-based composite audio analysis can be represented by the flowchart shown in Fig. 1 [1]. The framework represents a generic process of audio content understanding, from low-level features, via mid-level content representation, to high-level semantics. In this flow, the input audio stream is first segmented into different *audio elements* such as speech, music, various audio effects and any combination of these. Then, the *key audio elements* are selected, being the audio elements that are most characteristic for the semantics of the analyzed audio data stream [1][2]. (Key) audio elements serve as mid-level representation of audio content. Introducing this middle level enables us to divide the semantics inference into two steps, each of which can be realized in a much more robust way compared to one-step inference (i.e. inferring the semantics directly from low-level features). Then, the *auditory scenes*, which are the temporal segments with coherent semantic content, are detected and classified based on the (key) audio elements they contain. For example, in [4], the audio elements such as *applause*, *ball-hit*, and *whistling*, are used to detect the highlights in sports videos; and in movie indexing [1][3] *humor* and *violence* scenes are categorized by detecting the key audio elements like *laughter*, *gun-shot*, and *explosion*. Further examples of previous works related to the components of the scheme in Fig.1

include speech/music classification [5][6], audio effects detection [7] and auditory scene classification [8].



Fig. 1. A typical approach to content-based composite audio analysis [1]

The least addressed of all components in Fig.1 is the module for auditory scene segmentation. In fact, most of previous works on audio classification either assume the auditory scenes are manually pre-segmented [3][8], or use simple techniques without utilizing all the potential of the available audio information [1][9][10]. In this paper, we analyze the drawbacks of the existing methods and propose a novel approach to auditory scene segmentation based on the obtained audio element sequence.

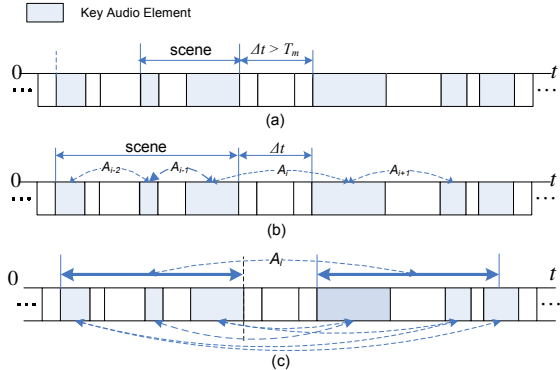
The rest of the paper is organized as follows. Section 2 analyzes the limits of the previous approaches and identifies the possibilities for improving the auditory scene segmentation. Section 3 presents the proposed approach, which is then evaluated experimentally in Section 4. Section 5 concludes the paper.

## 2. ANALYSIS OF RELATED WORK

In most previous works, auditory scenes were defined in the way to coincide with audio segments characterized by consistent low-level feature behavior. Such a definition served as a basis for numerous approaches for audio segmentation in speech, music and background noise [6][9][10]. For example, Venugopal [10] presented a work to segment an audio stream in terms of gender, speech, music and speaker, based on the features including tonality, bandwidth, excitation patterns, tonal duration and energy. Sundaram [9] presented a work on scene segmentation using low-level features, such as cepstral and cochlear decomposition, combined with listener model and various time scales. The definition of an auditory scene we apply in this paper, however, is much broader and is an analogy to the concept of a logical story unit [11]. That is, an auditory scene may consist of multiple, semantically related audio elements. An example of such an auditory scene is a *humor* scene consisting of several segments of *speech*, *laughter*, *cheer*, and possibly also some *music*.

In our previous work [1] we presented a simple scheme of auditory scene segmentation based on the time interval between consecutive key audio elements. Since key audio elements are most characteristic for the semantics of the analyzed audio data stream, only key audio elements are used for segmentation. As shown in Fig.2 (a), two adjacent key audio elements are assumed to be in the same auditory scene if the time interval between them is sufficiently short. Clearly, the algorithm is quite simple

and does not fully exploit the relationship between audio elements and auditory scenes.



**Fig.2** Illustration of previous approaches to auditory scene segmentation based on key audio elements: (a) [1], (b) [2], (c) possible scheme to investigate the relationship of key audio elements on a large scale

Then, in [2], we consider the *semantic affinity* between every two contiguous key audio elements, as well as the time interval between them, to locate the auditory scene boundaries. We introduced semantic affinity as a measure for the possibility that two key audio elements will appear together in the same auditory scene. As Fig.2 (b) shows, an auditory scene boundary is found between two audio element segments, if semantic affinity is low and the separating time interval is large. However, it is a little strict to base the detection of auditory scene boundaries on the comparison of two subsequent key audio elements only. A more intuitive approach would be to allow more flexibility in the ordering of key audio elements, as long as their mutual distance remains acceptable, similar to some classical video scene segmentation approaches [11][12]. As shown in Fig.2 (c), an approach in this direction would give the decision on the presence of scene boundary at the observed time stamp based on investigation of semantic affinity of (key) audio elements taken from a broader range and surrounding this time stamp.

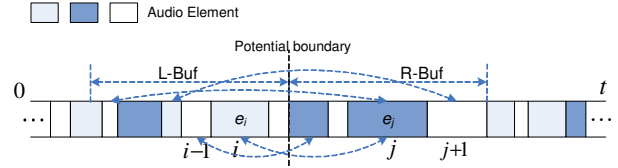
Moreover, the performance of the existing segmentation methods based on key audio elements strongly depends on the definition of a key audio element and the reliability of its detection. Crisply defining key audio elements and detecting them in a composite audio data stream may be rather difficult due to multiple superimposed audio modalities. Therefore, a more reliable solution would be to work with general audio elements instead, and for each element to compute a confidence that it can be considered a key audio element. Including this confidence value is likely to contribute to increasing the robustness of the segmentation scheme.

### 3. THE PROPOSED APPROACH

Based on our previous work [1][2] and the discussion from the previous section we propose in this section a novel approach to auditory scene segmentation, in which we draw analogies to text and video scene segmentation, and exploit the confidence values related to key audio elements. .

Fig. 3 shows an example audio element sequence obtained from an audio stream (the detailed algorithms are described in [1][2], both for the supervised and unsupervised approach), where each temporal segment (a block in the figure) belongs to an audio element and where different classes of audio elements are repre-

sented by different grayscales. Each time stamp separating two audio segments can be considered a potential auditory scene boundary. The confidence for having an auditory scene boundary at the observed time stamp is obtained by computing the semantic affinity between the audio segments surrounding the observed time stamp.



**Fig.3** Illustration of the proposed approaches to audio element based auditory scene segmentation

In the following section, we first define a new measure for semantic affinity between audio segments. Then, a new segmentation scheme is presented in which the proposed affinity measure is used to compute the confidence of having an auditory scene boundary at a given time stamp in a composite audio stream.

#### 3.1 Semantic Affinity Measure

The new definition of the semantic affinity is based on the following assumptions:

- i) there is a high affinity between two segments if the corresponding audio elements usually occur together;
- ii) the larger the time interval between two audio element segments, the lower their affinity;
- iii) the higher the confidence that an audio element is a key audio element, the more important role this element will play in the auditory scene segmentation.

In view of the above, the semantic affinity between the segments  $s_i$  and  $s_j$  can be computed as a function consisting of three components, each of which reflects one of the assumptions stated above:

$$A(s_i, s_j) = Co(e_i, e_j) e^{-T(s_i, s_j)/T_m} P_{e_i} P_{e_j} \quad (1)$$

Here,  $e_i$  and  $e_j$  are the audio element identities of  $s_i$  and  $s_j$ ; while  $P_{e_i}$  and  $P_{e_j}$  are the probabilities (confidences) that audio elements  $e_i$  and  $e_j$  are key audio elements.  $T(s_i, s_j)$  is the time interval between the audio segment  $s_i$  and  $s_j$ ; and  $T_m$  is a scaling factor, which is set to 16 seconds in our experiments, following the discussions on human memory limit [9] and our previous works [1][2]. Exponential expression in (1) is inspired by the content coherence computation formula introduced in [12].  $Co(e_i, e_j)$  measures the co-occurrence between two audio elements,  $e_i$  and  $e_j$ , in the entire audio stream. This value is computed in the following 3 steps:

- 1) First, we calculate  $D_{ij}$ , the average time interval between audio elements  $e_i$  and  $e_j$ . This value is obtained by investigating the co-occurrences of the observed audio elements in the input audio stream. For each audio segment belonging to audio element  $e_i$ , the nearest segment of  $e_j$  is located, and then  $D_{ij}$  is the average interval of each pair of  $e_i$  and  $e_j$ .
- 2) As an analogy to  $D_{ij}$ , we also calculate  $D_{ji}$ . It is clear that  $D_{ij}$  may be not equal to  $D_{ji}$ , in some cases.
- 3) We compute the co-occurrence value as

$$Co(e_i, e_j) = e^{-\frac{D_{ij} + D_{ji}}{2\mu_D}} \quad (2)$$

where  $\mu_D$  is the average of all  $D_{ij}$  and  $D_{ji}$ . The choice for an exponential formula in (2) was made to keep the influence of segment co-occurrence on the overall semantic affinity comparable with the influence of the time interval between the segments (1).

Having the affinity (1), we can now compute the confidence of having an auditory scene boundary at the time stamp  $t$  simply by averaging the affinity values computed for all pairs of segments  $s_i$  and  $s_j$  surrounding the  $t$ , that is,

$$C(t) = \frac{1}{W} \sum_{s_i \in L_t} \sum_{s_j \in R_t} A(s_i, s_j) \quad (3)$$

Here,  $L_t$  and  $R_t$  are the ranges of audio segments to the left and right from the time stamp  $t$ , respectively, which we set both to 16 seconds.

The issue related to confidence computation that requires careful consideration is the definition of the weighting coefficient  $W$  in (3). Weighting is necessary, since the number of audio segment pairs in (3) may be quite different for different observed time stamps. One possible approach would be to set  $W$  equal to the number of considered audio segment pairs. However, in this way, the importance of different audio elements is not taken into account, and the confidence will decrease too much when there are many short segments in the left and right segment range, even if the co-occurrence between the corresponding audio elements is high. Another possibility would be to define the weighting formula, which takes into account the importance of each audio segment, for instance, as

$$W = \sum_{i, s_i \in L_t} \sum_{j, s_j \in R_t} P_{e_i} P_{e_j} \quad (4)$$

However, our experiments showed that in the case that the audio elements around a potential boundary are all with low importance, the weight value (4) will be too small and the resulting confidence therefore too large. This will clearly result in missed true scene boundaries.

To deal with the weighting problem, we slightly revise the affinity measure (1) by using the pairs of *audio frames* instead of audio segments. An audio frame can be defined as an elementary part of an audio data stream having a fixed length of, in our case, 0.5 seconds. Reducing the analysis to frame pairs has as the consequence that the numbers of analysis units in both left and right range remain constant for all  $t$ . If we replace the segment pairs  $(s_i, s_j)$  by frame pairs  $(f_m, f_n)$  in (1), the new formula for semantic affinity becomes

$$A(f_m, f_n) = Co(e_m, e_n) e^{-T(f_m, f_n)/T_m} P_{e_m} P_{e_n} \quad (5)$$

which leads to the new formula for the boundary confidence:

$$C(t) = \frac{1}{N_l N_r} \sum_{m=1}^{N_l} \sum_{n=1}^{N_r} A(f_m, f_n) \quad (6)$$

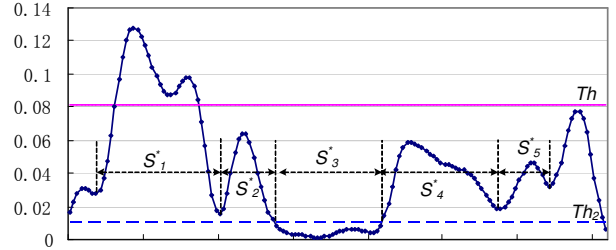
where  $N_l$  and  $N_r$  are the frame numbers in the left and right range of a potential boundary, respectively; and,  $e_m$  and  $e_n$  are the audio element identities of the frames  $f_m$  and  $f_n$ , respectively. In our experiments, each frame is set 0.5 seconds, since it is the basic unit in our audio element discovery [2]. Thus, we have 32 frames on each side.

### 3.2 Segmentation Scheme

Using (6), a confidence curve can be obtained, as illustrated in Fig. 4. Auditory scenes can be obtained simply by searching for local minima of the curve. In our approach, we first smooth the curve by using a median filter and then find the auditory scene boundaries with the following criterion.

$$C(t) < C(t+1); \quad C(t) < C(t-1); \quad C(t) < Th \quad (7)$$

where the first two conditions guarantee a local valley; while the last condition prevents high valleys from being detected. The threshold  $Th$  is set experimentally as  $\mu_a + \sigma_a$ , where  $\mu_a$  and  $\sigma_a$  are the mean and standard deviation of the curve, respectively.



**Fig.4** An example of the smoothed confidence curve and the auditory scene segmentation scheme

The obtained confidence curve is likely to contain long sequences of low confidence values, as shown by the segment  $S_3^*$  in Fig. 4. These sequences typically consist of the background (noise) audio elements which are weakly related to each other and also have low probabilities of being the key audio elements. As such a sequence clearly has different properties than an auditory scene that complies with the definition we introduced before, we choose to isolate these sequences by including all consecutive audio segments with low affinity values into a separate auditory scene. Detecting the boundaries of such scenes is analogous to detecting pauses in speech. Inspired by this, we set the corresponding threshold by using a similar approach to background noise level detection in speech analysis [13].

## 4. EVALUATION

In this section, the proposed approach is evaluated on the basis of the results we obtained by analyzing a 1.5-hour composite audio stream extracted from the video of ‘59th Annual Golden Globe Awards Ceremony’. The sound track contains an abundance of different audio elements, including speech, music and various audio effects like laughter, applause, and different combinations of these.

The audio stream is in 16 KHz, 16-bit and mono channel format, and is divided into frames of 25ms with 50% overlap for feature extraction. Following our previous work [2], a sliding window of one second with 0.5 seconds overlap is selected as the basic unit in audio element discovery, and spectral clustering with a self-tuning strategy is employed to decompose the audio stream into audio elements. Moreover, four heuristic importance indicators [2], including one occurrence frequency related and three duration related, are used to measure the probability that an audio element is a key audio element. The method results in 11 audio elements, which are listed in Table 1 together with a description of their content, and the confidence of being key audio elements.

It can be seen that some audio types are represented by several audio elements (like speech), but this is quite understandable in view of large variations one can expect in the properties across the segments containing the same audio type in general audio data streams.

**Table 1.** The list of obtained audio elements with description

No.	Description	Conf.	No.	Description	Conf.
1	speech1	0.380	6	music with speech	0.510
2	speech2	0.150	7	applause	0.959
3	speech3	0.001	8	applause with speech	0.553
4	Music	0.366	9	applause with dense-music	0.820
5	Noise	0.470	10	applause with light-music1	0.880
			11	applause with light-music2	0.544

In the process of creating the ground truth we observed that the award ceremony basically consists of a series of different events, like when the host announces the nominees and the winner, or when the winner approaches the stage while the audience is applauding. In view of this, and with help of a panel of unbiased persons, we selected the true auditory scene boundaries at the break points between different events. Moreover, we also annotated a number of ‘probable boundaries’ at places where the presence of a true boundary is unclear. For example, the turn between the played ‘nominated movies’ when a host announces the nominees, can be seen as a probable boundary. In total, we obtained 96 true boundaries and 60 probable boundaries.

In the experiments, a detected boundary is associated with annotated boundary if they are mutually to each other. Table 2 shows the evaluation results of the proposed approach, where different approach variants and different test configurations are compared. The table lists the recall of ‘ground-truth’ (R1) and ‘probable’ boundaries (R2) for different boundary shifts (where the notation “<*n*” and “+*n*” indicates that the boundaries are detected with the shift of less or more than *n* seconds, respectively, from the corresponding true boundary), and the corresponding false alarms (FA).

**Table 2.** The results of the auditory scene segmentation, comparing with different approach variants and different configurations

	R1(out of 96)	R2(out of 60)	<3	<6	<9	<12	+12	FA
Frame [Prob.]	72	36	62	20	8	7	11	12
Frame (0, 1)	60	23	46	20	8	5	4	11
Frame (1, 1)	72	29	25	31	21	10	14	25
Seg + W1	71	35	38	21	13	8	27	25
Seg + W2	70	37	43	22	18	11	13	43

The approach variants compared in Table 2 are (in the order of appearance) three variants of the frame-based approach (6), and two variants of the segment-based approach (3). The variants of the frame-based approach differ from each other in the way the key audio elements are defined. In the variant ‘Frame [Prob.]’, the entire value range of probabilities is used, while the variant ‘Frame (0,1)’ works with crisp definitions of key audio elements (audio elements 7-11 in Table 1 are selected as key audio elements based on [2], and their probabilities are set to 1 while others are set to 0). The third variant, ‘Frame (1,1)’, considers all obtained audio element as key elements (all probabilities set to 1). The first segment-based variant ‘Seg+W1’ works with the weight *W* set as the total number of audio segment pairs, while in the variant ‘Seg+W2’, *W* is set according to (4).

The results show that the approach variant Frame [Prob.] has the best recall and the smallest average boundary shift. It recalls 72 out of 96 ‘ground-truth’ boundaries and 36 out of 60 ‘probable’ boundaries, and the average boundary shift is about 4.6s. Although ‘Frame (1,1)’, ‘Seg+W1’ and ‘Seg+W2’ have similar recall values, their boundary shifts are much larger and they also result in more false alarms. Clearly, these detection schemes do not result in reliable confidence curves, which lead to many missed and misplaced scene boundaries.

## 5. CONCLUSION

Auditory scene segmentation is a non-trivial step towards high-level semantic inference of composite audio. In this paper, we analyze the limits of existing methods, and propose a novel approach to auditory scene segmentation based on audio elements contained therein. The approach not only investigates a broad range of audio elements surrounding an evaluated potential boundary, but also is flexible in dealing with audio elements with definitive or probabilistic decision. Evaluations performed on 1.5 hour of diverse test data indicate that the proposed approach is promising.

We see a number of possibilities to further improve the proposed approach. For example, pauses between different events might be further used to refine the auditory scene boundary.

## 6. REFERENCES

- [1] L. Lu, R. Cai, and A. Hanjalic. “Towards a unified framework for content-based audio analysis”. *Proc. ICASSP05*, vol. II, 1069-1072.
- [2] R. Cai, L. Lu, A. Hanjalic. “Unsupervised Content Discovery in Composite Audio”, *Proc. ACM Multimedia 2005*, 628-637, 2005.
- [3] R. Cai, L. Lu, and L.-H. Cai. “Unsupervised Auditory Scene Categorization via Key Audio Effects and Information-Theoretic Co-Clustering”, *Proc. ICASSP05*, Vol. II, 1073-1076, 2005
- [4] M. Xu, N. Maddage, C.-S. Xu, M. Kankanalli, and Q. Tian. “Creating audio keywords for event detection in soccer video”. *Proc. ICME03*, vol. 2, 281-284, 2003.
- [5] J. Saunders. “Real-time Discrimination of Broadcast Speech/ Music”. *Proc. ICASSP96*, Vol.II, pp.993-996, 1996
- [6] L. Lu, H.-J. Zhang, and H. Jiang. “Content analysis for audio classification and segmentation”. *IEEE Trans. Speech Audio Processing*, vol. 10, no. 7, pp. 504-516, Oct. 2002.
- [7] R. Cai, L. Lu, H.-J. Zhang, and L.-H. Cai, “Highlight sound effects detection in audio stream”. *Proc. ICME03*, vol. 3, 37-40, 2003.
- [8] W.-H. Cheng, W.-T. Chu, and J.-L. Wu. “Semantic context detection based on hierarchical audio models”. In *Proc. of the 5<sup>th</sup> ACM SIGMM-MIR*, 109-115, 2003.
- [9] H. Sundaram S.F. Chang “Audio Scene Segmentation Using Multiple Features, Models And Time Scales”, *Proc. ICASSP 2000*
- [10] S. Venugopal, K.R. Ramakrishnan, S.H. Srinivas, and N. Balakrishnan, “Audio scene analysis and scene change detection in the MPEG compressed domain,” *Proc. MMSP99*, 191-196, 1999.
- [11] A. Hanjalic, R. L. Legendijk, and J. Biemond. “Automated high-level movie segmentation for advanced video-retrieval systems”. *IEEE Trans. Circuits and Systems for Video Technology*, vol. 9, no. 4, 580-588, 1999.
- [12] J. R. Kender and B.-L. Yeo. “Video scene segmentation via continuous video coherence”. *Proc. CVPR98*, 367-373, 1998
- [13] D. Wang, L. Lu, H.-J. Zhang. “Speech Segmentation without Speech Segmentation”, *Proc. ICASSP03*, Vol. I, 468-471, 2003.