Using High-level Information to Detect Key Audio Events in a Tennis Game

Qiang Huang and Stephen Cox

School of Computing Sciences
University of East Anglia, UK
h.qiang@uea.ac.uk, s.j.cox@uea.ac.uk

Abstract

This paper describes how the detection of key audio events in a sports game (tennis) can be enhanced by the use of high-level information. High-level features are able to provide useful constraints on the detection procedure, and thus to improve detection performance. We define two types of event based information: event dependency and inter-event timing. These respectively characterize the identity of the next event and the time at which the next event will occur. Probabilistic models of high-level constraints are developed, and then integrated into our event detection framework. We test this approach on audio tracks extracted from two different tennis games. The results show that significant improvements in both accuracy and computational efficiency are obtained when applying high-level information.

Index Terms: Event detection, context information, audio information

1. Introduction

Our ambitious long-term goal is to understand multimodal interaction between humans, and we use a sports game, tennis, as a starting-point. In a sports game, the goals of the interaction are simple and clearly defined and the interaction is subject to clear rules. As such, it can be analysed in terms of a sequence of "events". Our work focuses on the retrieval of this sequence from audio information associated with the game: in later work, we will combine this with information obtained from analysing the video signal from the event.

Other work has shown that the context in which objects and events occur plays a crucial role in technologies such as video retrieval and image understanding. [1, 2, 3]. Objects and events never occur in isolation: they co-occur and co-vary with other objects and in particular environments, and this behaviour provides a rich source of contextual associations that provide clues to the identity of the object or events. In this paper, we utilize context information to improve the performance of automatically detecting and identifying the match events occurring within a tennis game.

Research on audio event detection has already been widely developed [4, 5, 6, 7, 8, 9]. Some of this work [6, 8, 7] focuses on the classification of isolated sound classes. It is mainly concerned with low-level perceptual features and the use of spectral clustering techniques. Similar work has also been done on event detection in sports games, such as tennis [12, 9, 13], football[15] and basketball[14]. These work has mainly relied on the use of visual information to find the boundaries of scenes and events, with audio information only adding ancillary information. Other work [15, 14] has even attempted to retrieve timings from a time stamp on the video.

By contrast, we focus on deriving the key events in a tennis game using purely audio information–later, we will combine this with information derived from visual analysis, and we expect that the two media streams will provide complementary and synergistic information. In previous work [11], we have demonstrated that it is possible to obtain a sequence of audio events with good accuracy. However, not all events convey the same amount of information: for instance, the event "crowd applause" may function as a useful contextual marker for the progress of the game, but gives little information on its state. We focus here on detecting three key audio events, which provide information about the state of the game and how it has changed since the last state. Interestingly, these events (the umpire's speech, the sound of the automatic detector that adjudges a serve as "let", and the call of a line judge in reporting ball out) are only available from the audio channel, and so make a convincing argument for combining video and audio information. These three events are difficult to detect, because there are often overlapping interfering noise with them and the last two are short and of low amplitude. We make use of two types of context information, event dependency and inter-event timing. These provide constraints about the identity of the most probable event to occur next, and where it is likely to be located in time.

This paper is organised as follows: the data and basic analysis techniques are introduced in Section 2. Section 3 describes what context information we use and how it is integrated into our framework. Section 4 describes the experimental procedure. Experiments and analysis are presented in Section 5, and we discuss and conclude in Section 6.

2. Data

The data used was mainly extracted from DVDs of Wimbledon singles tennis matches played in 2008. It consists of ten audio tracks, each lasting about 22 minutes (eight minutes for Track 8), taken from video recordings of two different tennis matches. Nine of the tracks (Track 1–9) are taken from the same tennis match (Murray vs. Gasquet) and Track 10 is from the other one (Federal vs. Nadal).

Audio analysis was standard: the audio sequence was windowed into 30ms-length frames with 20ms overlapping from which 39-D MFCC vectors were generated. Cepstral mean normalization was applied at the track level. Frame-based classification was done by using a Gaussian mixture model for each event [11], which is trained with the audio signals from Track 1, 2, and 3. Each audio track was manually segmented into a sequence of "audio events", which represent who or what makes the sound. The details are shown in Table 1.

In this paper, we specifically focus on the detection of the three audio events UMP, LJ and BP. Our previous work showed

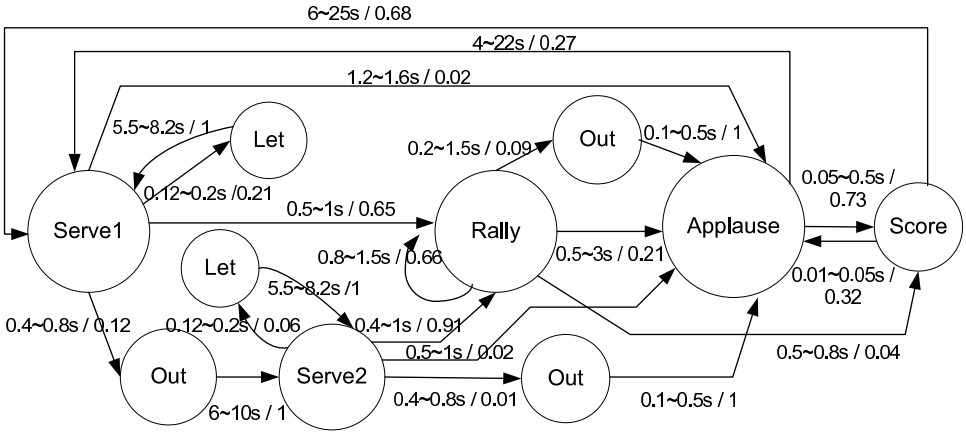

Figure 1: A single point of a tennis game represented as a probabilistic finite-state automaton.

Table 1: The Audio Events Classified

Audio Event	Name	Match Event
Chair umpire's speech	UMP	Report Score
Line judge's shout	LJ	Report serve out, fault etc.
Sound of ball hit	BS	Serve, Rally
Crowd noise	CN	
Beep	BP	Let
Commentators' speech	COM	
silence	SIL	-

that these were more difficult to detect than the other four events shown in Table 1, but they are highly informative about the match state and progress.

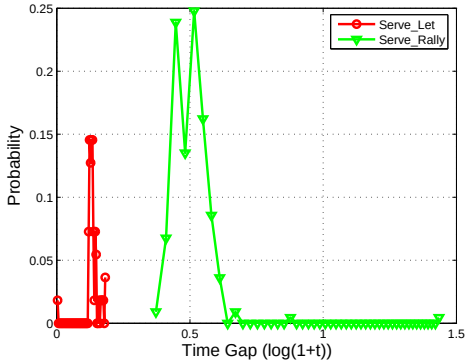

Figure 2: Examples of the distribution of time gap

3. Theoretical Framework

3.1. Definition of event level constraints

To improve the accuracy of the audio event detection, we utilize two kinds of event based information. Figure 1 represents a single point of a tennis game as a probabilistic finite-state automaton, with the nodes of the automaton representing match events. Events in a game do not immediately follow each other: there is always a period between two successive events where the previous event has finished and the next one is yet to begin, and such periods are labelled as "null" in our marked-up training-data. In this figure, the range of observed times from the end of one event to the beginning of another event is shown on the arc connecting the two events, as are the probabilities that an event succeeds another event to which it is connected. This data underlying this diagram enables us to estimate two useful pieces of contextual information:

1. the N-gram probability of an event given a history of the $N-1$ previous events, analogous to a language-model in speech recognition (in practice we confine ourselves to $N=2$ because of data sparsity);
2. the probability density functions of the time-gaps between events.

Figure 2 shows the probability density functions (PDFs) of the time gap between two event pairs: "Serve–Let" and "Serve–Rally" (Note that because the x-axis is logarithmic, the lower x-values are compressed: both PDFs do sum to one). The fact that these PDFs are well-separated is very useful: firstly, it will discriminate between the events "Let" and "Rally" given that the event "Serve" has been detected. It also helps to make searching more efficient: given that the event "Serve" we know that there is no need to search for the event "Let" at any time beyond the end of the left curve of Figure 2.

3.2. Modeling context for event detection

Although this contextual information is not on its own sufficient for event identification, it provides very useful constraints on the identity of an event. We hence integrate this contextual information into our event detection system.

If we denote a match event as ME and an observed audio signal (i.e. a sequence of frames) as O, for recognition purposes, we estimate

$$\Pr(ME|O) = \Pr(O|ME)\Pr(ME). \quad (1)$$

$\Pr(O|ME)$ and $\Pr(ME)$ can be viewed as an acoustic model and a event based language model, respectively. The term $\Pr(ME)$ depends on both the previous event and the time-gap since the previous event, and so can be approximated as:

$$\Pr(ME) \approx \prod_{i=2}^{Nevents} \Pr(T_{i-1,i}|ME_{i-1}, ME_i)\Pr(ME_i|ME_{i-1}) \quad (2)$$

where $\Pr(T_{i-1,i}|ME_{i-1}, ME_i)$ is the probability that we observe a time-gap of $T_{i-1,i}$ seconds between events ME_{i-1} and

ME_i. Experimental observation showed that the probability distribution of the time gap $\Pr(T_{i,i+1}|ME_{i-1}, ME_i)$ between two adjacent events could be well-approximated by a Gaussian distribution, whose mean and variance are estimated from the training data.

By combining equation 2 and equation 1, we obtain:

$$Pr(ME_i|O_i) = Pr(O_i|ME_i)Pr(T_{i-1,i}|ME_{i-1}, ME_i) \times Pr(ME_i|ME_{i-1})$$

Hence $Pr(T_{i-1,i}|ME_{i-1}, ME_i)$ can be viewed as a "weighting term" on $Pr(O_i|ME_i)$ that highlights regions where event ME_i is most likely to occur after time $T_{i-1,i}$ given event ME_{i-1}. There is a practical problem of balancing the contribution of this probability with the acoustic probability $\Pr(O_i|ME_i)$. We have found it useful to use an exponential function to smooth the weight of the time-gap contribution, and we therefore change the time gap weight from $\Pr(T_{i-1,i}|ME_{i-1}, ME_i)$ to $exp^{\lambda * \Pr(T_{i-1,i}|ME_{i-1}, ME_i)}$. λ is a parameter that is determined experimentally. In this paper, it is set to 5. In practice, performance varies little with λ.

4. Experimental Setup

We evaluated the performance of the model for both accuracy and computational efficiency. The baseline is obtained by using a set of Gaussian mixture models (GMMs), one for each audio event [11], to classify each audio frame. After classification, any short silence periods (less than 100ms in duration) are removed, and then the labels of frames that have the same labelling are merged to create a (noisy) sequence of events. The acoustic models and the models for event dependency and inter-event timing are trained on the manual transcriptions of Tracks 1, 2, and 3. Testing is performed on the remaining tracks.

To evaluate the detection performance, we use the F-score measure, defined as:

$$Fscore = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \quad (3)$$

$$Precison = \frac{\#Correctly\ \ Detected\ \ Events}{\#Detected\ \ \ Events} \quad (4)$$

$$Recall = \frac{\#Correctly\ \ Detected\ \ Events}{\#True\ \ \ Events} \quad (5)$$

A *Correctly Detected Event* is one that occurs in an approximately correct region. To determine these regions, we compute the mean position in time (MT) of each hand-labelled event using its start time (ST) and end time (ET):

$$MT_{event} = (ST_{event} + ET_{event})/2 \quad (6)$$

If MT_{event} is located within the time range of a detected event with the same labelling, then the detected event is viewed as a correct detection. A t-test is used to decide whether results are significantly different.

5. Experimental Results and Analysis

5.1. Effectiveness

Figures 3, 4, and 5 show the performance of detecting the umpire's speech, the line judge's shout, and the sound of beep on the test set, respectively. In the three figures, "Depend." means that the event dependency (the "language model" of events) was used, and "IET" means the inter-event timing was used.

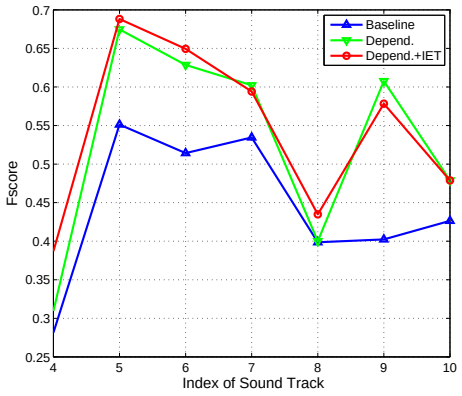

Figure 3: Performances of detecting umpire's speech

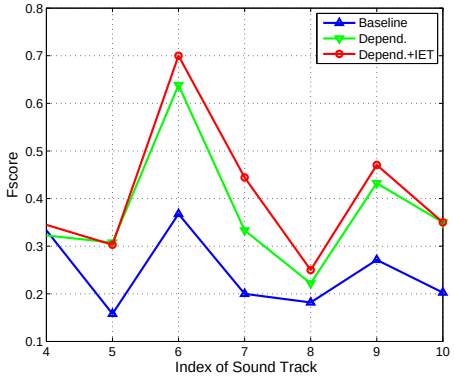

Figure 4: Performances of detecting a line judge's shout

To check whether combining event dependency with inter-event timing is significantly better than baseline, we run a significance test (t-test). By the central limit theorem, the set of F-scores obtained from the tests has an asymptotically normal distribution, and so we are justified in using a t-test to evaluate their significance.

Table 3: p-values of the t-test comparing performance

	Dependency *vs.* Baseline	Depend.+ IET *vs.* Baseline	Depend.+IET *vs.* Depend.
UMP	0.015	0.002	0.274
LJ	0.009	0.015	0.050
BP	0.002	0.006	0.356

Table 3 shows that the use of both event dependency and inter-event timing is significantly better ($p \leqslant 0.05$) than the baseline at detecting all three of the events under consideration here. Adding inter-event timing when event dependency is already being used only significantly increases performance for the line judge class. In general, adding the time-gap information adds only a small amount to performance and most of the gain comes form using the "syntax" of the game i.e. the event dependency. However, the time-gap information is useful in improving the efficiency of the computation (see next section).

Table 2: Computational efficiency of detecting audio events and its effect on F-score on the test set

Track Number	4	5	6	7	8	9	10
Reduction in computation (%)	8.01	9.49	11.15	15.73	16.71	18.61	15.62
Depend. (F-score)	0.6571	0.7504	0.8036	0.7783	0.7547	0.7117	0.7601
Depend.+IET (F-score)	0.6646	0.7577	0.8123	0.7791	0.7561	0.7078	0.7614
Depend.+IET+Eff. (F-score)	0.6656	0.7513	0.8108	0.7611	0.7607	0.6894	0.7554

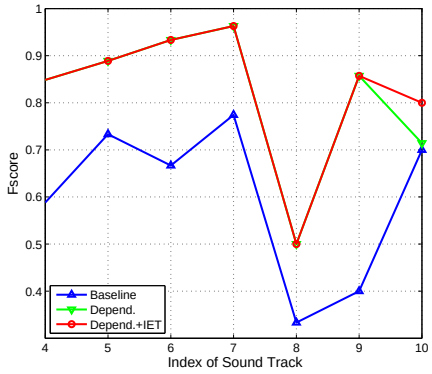

Figure 5: Performances of detecting the sound of the beep

5.2. Efficiency

In our experiments, the computation on each audio frame with GMM for seven sound classes is a very time-consuming task. To reduce the computation, we make use of the event dependency and inter-event timing shown in figure 1. Two methods for computation reduction are used:

1. We can avoid performing this computation for every frame by skipping some regions where no match events would occur except silence (marked as "null" in our data), such as the time region before a "Serve" event.
2. Given a certain hypothesized event, we only need to compute the occurrence probability of a subset of next events in regions.

For the first case, we can skip several hundred frames without any computation work, while, for the latter one, we also do not need to compute the acoustic probability for all sound classes.

Row 2 of Table 2 shows the efficiency gain on the seven sound tracks, which has a mean value of 13.11%. Row 3 shows the F-score using event dependency, and Row 4 the F-score using event dependency and inter-event timing without the use of any computational reduction. The bottom row shows the F-score when the efficiency techniques described above are used: we see that the performance is substantially the same, for a useful gain in computational efficiency.

6. Conclusions and Future Work

In this paper, a new framework was developed to improve detection performance of key audio events in a tennis game. The technique takes account of two contextual features, namely event dependency and inter-event timing. The results obtained show significant improvements in comparison with the baseline in both performance (as measured by F-score), and a small gain in computational efficiency.

Our immediate future work is to look in more detail at the issue of how to balance the probabilities from the different models used here, and how to apply them to more complex event detection tasks. We will then begin to incorporate and integrate information derived from computer vision techniques. Our long-term goal is to produce a system that can understand a tennis game completely, and that is capable of being re-engineered for a similar game with minimal human intervention.

ACKNOWLEDGMENT: This work was supported under a UK Engineering and Physical Sciences Research Council Grant number EP/F069626/1.

7. References

[1] Divvala, K. S. and Hoiem, D. and Hays, H. J. and Efros A. A. and Hebert, M., "An Empirical Study of Context in Object Detection", In Proceedings of CVPR, Florida, June 2009.

[2] Oliva, A. and Torralba, A., "The role of context in object recognition", Trends in Cognitive Sciences, vol 11(12):523—527, 2007.

[3] Chu, W. and Cheng W. and Wu, J., "Semantic Context Detection Using Audio Event Fusion", Journal of Applied Signal Processing, 1–12, 2006.

[4] Cai, R. and Lu, L. and Hanjalic, A., "Unsupervised Content Discovery in Composite Audio", In Proceeding of Int. Conf. of MM, Singapore, November 2005.

[5] Cai, R. and Lu, L., Zhang, H.-J., and Cai, L.-H., "Highlight sound effects detection in audio stream", In Proceedings of ICME, 2003.

[6] Zhuang, X. and Zhou, X. and Huang, T and Hasegawa-Johnson, M., "Feature analysis and selection for acoustic event detection", In Proceedings of ICASSP, 2008.

[7] Lu, L.,"Content analysis for audio classification and segmentation", IEEE Trans. Speech and Audio Processing, vol 10:504–516, 2002.

[8] Atrey, P. and Maddage, N. and Kankanhalli, M., "Audio Based Event Detection for Multimedia Surveillance", In Proceedings of ICASSP, 2006.

[9] Kijak, E. and Gravier, G. and Oisel, L. and Gros, P., "Audiovisual Integration for Tennis Broadcast Structure", Source Multimedia Tools and Applications archive, vol 30(3):289–311, September 2006.

[10] Lu, L. and Cai, R. and Hanjalic, A., "Towards a Unified Framewrok for Content-based Audio Analysis", In Proceedings of ICASSP, 2005.

[11] Huang, Q. and Cox, S., "Hierarchical Language Modeling for Audio Events Detection in a Sports Game", In Proceedings of ICASSP, Dallas, USA, 2010.

[12] Dahyot, R. and Kokaram, A. and Rea, N. and Denman, H., "Joint Audio Visual Retrieval for Tennis Broadcast", In Proceedings of ICASSP, 2003.

[13] Rea, N. and Dahyot, R. and Kokaram, A., "Classification and representation of Semantic Content in Broadcast Tennis Videos", In Proceedings of ICIP, 2005.

[14] Zhang, Y.-F. and Xu, C.-S. and Rui, Y. and Wang, J.-Q. and Lu, H.-Q., "Semantic Event Extraction from Basketball Games Using Multi-Modal Analysis", In Proceedings of ICME, 2007.

[15] Xu, C.-S. and Wang, J.-J. and Wan, K.-G. and Li, Y.-G. and Duan, L.-Y., "Live Sports Event Detection Based on Broadcast Video and Web-casting Text", In Proceedings of ACMMM, 2006.