IMPROVED AUDIO EVENT DETECTION BY USE OF CONTEXTUAL NOISE

Qiang Huang and Stephen Cox

School of Computing Sciences
University of East Anglia, Norwich, UK.

ABSTRACT

This paper presents new approaches to improve the detection of two key audio events in a sport game (tennis) using contextual information. When analysing a tennis match using only audio information, the sound of the ball being struck and the occurrence of a line judge's shout can be obscured by players' grunts or shouts. Furthermore, if models of these two important events are trained from labelled training-data, there is often considerable audio mis-match with the test-data, which means that detection performance can be very poor. To handle this problem, we regard the players' grunts as useful contextual information that indicates the position of the events of interest. We show how to use an unsupervised learning method to build an improved model of the ball-hit event using grunt information. We can then use high-level information to distinguish grunts from line-judge shouts. This approach gives simultaneous improvements in detection of both ball-hits and line judge shouts, and is portable between different matches, unlike approaches based on the use of manually labelled training-data.

Index Terms— Audio event, unsupervised detection

1. INTRODUCTION

Our long-term goal is to develop machines with the ability to analyse and participate in complex human activities by information acquisition and learning. We begin with the analysis of tennis games. which contain rich and tightly linked audio and visual information. Our current focus is the detection of key audio information, such as the event of the racquet hitting the ball, or the call from a line judge ball to indicate a fault or a ball that is "out".

However, when processing some tennis matches, we have found that ball hit detection performance is often very poor for two reasons:

1. acoustic mismatch between the training and the test data;

2. interfering noise in the form of players' grunts, foot steps, and general background noise.

Most previous work [1, 3] in the analysis of tennis games ignores these issues, which means that portability of techniques across games is impossible (because of mismatch), and the practical difficulty of dealing with interference is ignored. In our own previous work, the dependencies between audio events were used to enhance the robustness of audio event detection [5], and later, the problem of acoustic mismatch in detecting umpire calls was considered [7]. This work develops and combines aspects of these two papers.

In the modern tennis game, players are increasingly likely to vocalize when they strike the ball—colloquially known as a "grunt". This is unfortunate for the key task of detecting ball hits, because the grunt can mask the hit sound, although more often, they are separate events. Here, we make a virtue out of necessity, and rather than regarding the players' grunts as interfering noise, we use them as contextual information during rallies. In fact, if they occur, they are near perfect indicators of the nearby occurrence of a ball hit. Some players grunt every time the hit the ball, others less frequently or not at all, but every single match we have processed contains grunts. Hence in this work, we use only ball-hits that are indicated by grunts, and we find that this give us sufficient training-data to build a reliable model of ball-hits.

However, there are other vocalisations on the soundtrack, which are the umpire's speech, the line judge's shouts and the commentary. We show in this paper how these are discriminated from grunts and thus from ball-hits. Our approach is in four steps:

1. Location of possible players' grunts or line judge calls by pitch information;

2. Location of candidate ball hit locations using the above information;

3. Initialisation and refinement of the acoustic model of the sound class of the ball hit;

4. Discrimination of players' grunts from other vocalisations.

2. DATA

In this paper, we use soundtrack data from a total of five tennis matches, one for training and the other four for test. Table

1 gives basic information about these matches. The training

	Game	Type	Dur. (mins.)	# ball hit	# line judge call
Train.	Wim-08	singles	180	1528	128
Test (1)	AUS-09	singles	136	1259	92
Test (2)	US-06	doubles	48	510	76
Test (3)	AUS-09	doubles	98	796	79
Test (4)	Wim-09	singles	131	838	89

Table 1: Data for training and test

data is extracted from a men's single match of the Wimbledon Open (2008), while the test matches are from the Australian Open (Test 1 and 3), the US Open (Test 2), and the Wimbledon Open (Test 4). The soundtracks of the five matches are segmented into 30 ms frames using a sliding window with a 20-ms overlap. Each audio frame is converted into a vector of 39-D MFCCs (13 static components, plus velocity and acceleration). As in our previous work [2], we define seven classes of audio events, each of them being modelled with a Gaussian mixture model (GMM) built from frames labelled as belonging to that audio class. These are: Chair umpire's speech, Line judge's shout, Sound of ball hit, Crowd noise, Electronic beep, Commentators' speech and Silence. The number of mixture components in the GMMs ranges from three to seven.

3. EVENT DETECTION

3.1. Location of players' grunts

As we have indicated in the Introduction, the ball-hit model is much affected by acoustic mismatch and interfering noise. Hence in this work, we replace the GMM of the ball-hit model formed from the training data with a new model, trained in an unsupervised fashion from the test data, that relies on grunt detection to locate the position of a ball-hit.

For detection of players' grunts, we do not build a specific model because the acoustic characteristics of these vocalisations vary hugely depending on the player, the location of the tournament, the microphone positions etc. Instead, we locate them using pitch information. To extract pitch information from a segment of audio, we compute the "subharmonic-to-harmonic ratio": a detailed description of this technique can be found in [4]. Figure 1 shows the distribution of the fundamental frequency (F0) from different speakers in a tennis match over the training data. For the time being, we regard line judges' shouts and players' grunts as being in the same class. The figure shows that "proper" speech (even the few words spoken by the umpire) has a very different F0 distribution from the short vocalisations produced by the players and the line judges. The pitches of commentators' and chair umpire's speech lie mainly within the range of 100-200 Hz, while much of the pitch extracted from players' grunts and

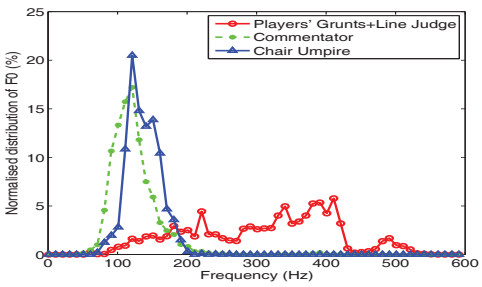

Fig. 1: Normalised distributions of F0 of the voices from different speakers, including commentators, chair umpire, line judges and players.

line judge calls is higher than 250 Hz. This difference enables us to coarsely locate the position of grunts/calls in the sound track.

3.2. Finding candidate ball-hit locations

The construction of a GMM for the sound of ball hit from the soundtrack of a previously unheard match relies on the detection of the locations of players' grunts. Figure 2 illustrates

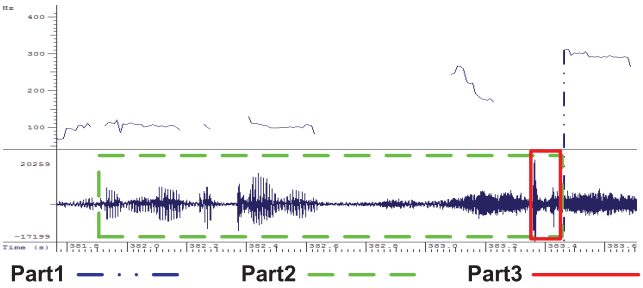

Fig. 2: An illustration of how we approximately locate the sound of ball hits using pitch information extracted from the players' grunts. The upper part of this figure shows the pitch information (F0) of the audio signals plotted at the bottom. X-axis indicates timing information, and y-axes represent frequency for F0 and amplitude for audio signals, respectively.

how we locate possible ball-hit locations (Step 2). The upper part of the figure shows the estimated F0 contour of the signal, and we see that at the end of the signal, there is a section which is above 300 Hz for about 200 ms, which may indicate a grunt/line-judge shout event. A 1.5s window before this point indicates the region in which a ball-hit is likely to be present: 1.5s of signal *before* the grunt is used because analysis of the soundtrack revelas that ball-hits almost always occur before grunts, and that the timing gap τ between any two adjacent ball hits lies in the interval $0.5 < \tau < 1.5$ [6]. We then locate the frame whose sound intensity is the largest within this window, and we assume that this frame is the beginning of the ball-hit. The following 100ms of signal is labelled as

"ball-hit". The figure of 100 ms is derived from work done in [2], which analysed the duration distribution of the ball hits. The accuracy of ball-hit candidate selection on our test games is actually 73.68%, which is sufficient to build a better model of the ball-hit class than that obtained from the training game.

3.3. Model refinement and confidence measure

The GMMs of all the audio event classes *except* the ball-hit class are used for recognition of test material. The ball-hit class is rebuilt for each new test data set using the procedure described above to identify ball-hit events. However, because this procedure is subject to error, it is necessary to exclude ball-hit segments that we have low confidence in. We define a confidence measure for any frame whose highest likelihood is produced by the class "ball-hit" as follows:

$$CM(f_i) = LL_{max}(f_i|C_B) - LL_{max2}(f_i|C_k) \qquad (1)$$

where LL_{max} is the highest log likelihood from class C_B = "ball-hit", and LL_{max2} the next highest likelihood from a different class C_k. Any frames whose value of $CM(f_i)$ is above some threshold will be used for training a new ball-hit GMM: frames whose value of $CM(f_i)$ is below threshold are discarded. The threshold is a parameter that is varied in our experiments.

3.4. Detection of Line Judge's Shout

To distinguish the line judge's shouts from the players' grunts, we take into account both their differences in acoustic characteristics and occurrence positions in a match. Players' grunts occur almost exclusively between two adjacent ball hits, and the timing interval ($T_{interval}$) between two "grunt" events is much shorter than the time between a grunt/line-judge shout event and the next line judge shout, since line judge shout events events occur only at the end of a game point or after a serve. Using an appropriate timing interval threshold enables us to distinguish grunts from line judges' shouts.

4. EXPERIMENTAL SET-UP

In our experiments, we compare the detection performances of the sound of ball hits with and without using players' grunts. The model without using players' grunts uses the ball-hit GMM built from the labelled training-data. Because this can be applied to any test match, we call this the "match-independent" model (**MIM**). The model that makes use of players' grunts is specific to a given test match, and hence we term this a "match-dependent" model (**MDM**). In addition, we test the effect of the CM threshold on ball hit detection performance.

To evaluate the detection performance for the event "ball hit" or the event "line judge shouts", the *F-score* is used, defined below as:

$$P = \frac{\text{\# correctly detected events}}{\text{\# detected events}} \qquad (2)$$

$$R = \frac{\text{\# correctly detected events}}{\text{\# audio events in ground truth}} \qquad (3)$$

$$F = \frac{2PR}{P+R}.$$

In equation 3, an "event" is either a ball-hit or a line-judge shout. An event is considered to be correctly detected when the maximum likelihood value of the detected event is located within the manually annotated range of an event with the same label. Detected events that are not within an audio event that has the same label are regarded as false positives, and undetected events are false negatives.

5. RESULTS AND ANALYSIS

Table 2 shows the detection performance for the sound of ball hits in different matches using the match-independent (MIM) and match-dependent model (MDM), respectively. The CM value was manually selected to give the highest obtainable F-score, and is also shown. Using the MIM model, the acoustic mismatch between the training and the test data severely impacts detection performance on the four test matches, especially on Tests 1 and 3. Using the MDM significantly outperforms the MIM on Tests 1, 2 and 3, although there is a relative smaller improvement (11%) on Test 4. This is almost certainly because Test 4 and the training match are from the same tournament, the Wimbledon Open, so the acoustic mismatch is not too great. Table 2 also shows that the optimum CM thresholds vary considerably from match to match.

Type	Train.	Test 1	2	3	4
MIM	85.3	31.07	51.57	16.99	73.01
CM-thresh	10	1	6	1	4
MDM	85.3	79.47	76.36	76.70	81.41
CM-thresh	10	15	18	18	11
Improv.	0	**+156%**	**+48%**	**+351%**	**+11%**

Table 2: Ball-hit detection performance (F-score) using the MIM and MDM with manually selected CM threshold

Figures 3- 6 plot the variation of the F-score with the confidence measures used in the MIM (blue) and MDM (red) models. Also plotted (green) is the performance obtained when the CM threshold is set CM_{mean}, defined as:

$$CM_{mean} = \frac{1}{N}\sum_{i=1}^{N} CM(f_i) \qquad (4)$$

These figures confirm that the choice of CM value has a huge effect on detection performance. However, it is very encouraging that by using CM_{mean} as the threshold, the performance in most cases is near optimum for the proposed

(MDM) technique. This means that a threshold can be automatically set for determining which signal segments are to be used in building the ball-hit GMM without compromising performance.

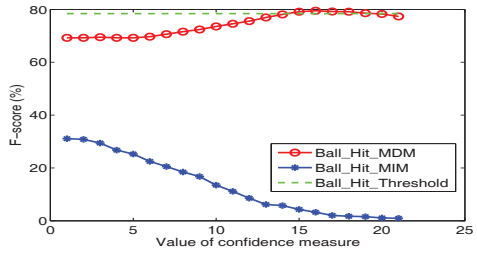

Fig. 3: Detection performances of ball hit on Test 1

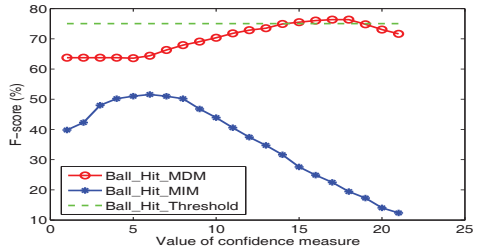

Fig. 4: Detection performances ball hit on Test 2

Table 3 shows the detection performance for line judges' shouts. The improvement is less dramatic than for ball-hit detection, but is still useful. However, correct detection of ball-hits is more important to us in this task than detection of line-judges' shouts.

Type	Training	Test 1	2	3	4
MIM	42.99	38.10	40.76	36.79	42.71
MDM	42.99	43.49	44.71	44.31	47.19
Improv.	0	**+14.3**	**+9.7**	**+20.7**	**+10.5**

Table 3: Detection performance (*F-score*(%)) of the line judges' shouts using the timing intervals after employing the MDM to improve ball hit detection

6. CONCLUSION AND FUTURE WORK

In this paper, we regarded players' grunts not as interfering noise, but rather as useful contextual information to aid us in locating ball-hit positions. We made use of pitch information to identify grunts and line judges' shouts, which are acoustically similar, and then used high-level information (inter ball-hit timing) to separate these two audio events. We showed how to use an unsupervised learning method to selects segments of signal to train improved models of ball-hits, and that a confidence measure could be successfully automatically determined for this selection, which is very important for portability to different matches.

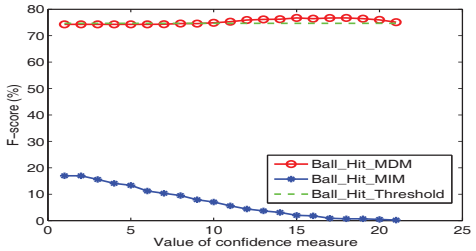

Fig. 5: Detection performances ball hit on Test 3

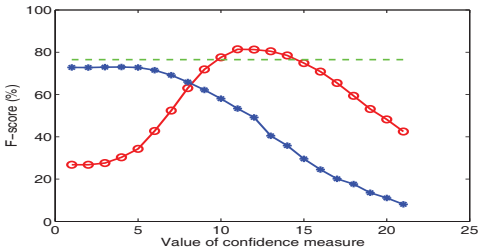

Fig. 6: Detection performances ball hit on Test 4

Our future work will firstly be to improve our ability to identify more types of audio signals in different environments. We will also begin to incorporate and integrate information derived from computer vision techniques to build more accurate high-level game structures for understanding human actions by learning the multimodal information in games.

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

7. REFERENCES

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

[2] Huang, Q. and Cox, S., "Hierarchical Language Modeling for Audio Events Detection in a Sports Game", In Proceedings of ICASSP, pp.2286–2289, Dallas, USA, 2010.

[3] Rea, N. and Dahyot, R. and Kokaram, A., "Classification and representation of Semantic Content in Broadcast Tennis Videos", In Proceedings of ICIP, pp.1204–1207, 2005.

[4] Sun, X., "Pitch Determination And Voice Quality Analysis Using Subharmonic-To-Harmonic Ratio", In Proceedings of ICASSP, pp.200–203, 2002.

[5] Huang, Q. and Cox, S., "Using High-level Information to Detect Key Audio Events in a Tennis Game", In Proceedings of InterSpeech, pp.1409–1412, 2010.

[6] Huang, Q. and Cox, S., "Improved Detection of Ball Hit Events in a Tennis Game Using Multimodal Information", In Proceedings of AVSP, pp.127–130, Voterral, Italy, 2011.

[7] Huang, Q. and Cox, S., "Iterative Improvement Of Speaker Segmentation Using High-Level Knowledge", In Proceedings of Interspeech, 2011