euronoise

The Use of Mel Cepstral Coefficients and Markov Models for the Automatic Identification, Classification and Sequence Modelling of Salient Sound Events Occurring During Tennis Matches

G. J A Hunter, K. Zienowicz and A. I Shihab

Kingston University, Faculty of Computing, Information Systems and Mathematics, Penrhyn Road, KT1 2EE Kingston-upon-Thames, UK
g.hunter@kingston.ac.uk

Some significant events in sports matches occur too quickly to be detected by conventional video. Audio signals, normally sampled at a much higher rate, provide a way to detect such short events.

Here, we employ approaches inspired by methods used in automatic speech recognition – use of templates of Mel Frequency Cepstral Coefficients (MFCCs) compiled over several adjacent time windows, together with Principal Components Analysis (PCA) – to classify sound events, including different tennis strokes, bounces of the ball, echos, speech and audience applause, occurring in the relatively controlled situation of major championship tennis matches. Good success rates were obtained for classification of the 1504 sound events in the available recordings. We go on to use Markov models to predict sequences of strokes (i.e. produce "synthetic rallies") and combine the predictions of the acoustic classifier and the Markov model, using a Bayesian approach to produce a hybrid classifier. These approaches could yield valuable information, of benefit to spectators, match officials and coaches in tennis and other sports (including cricket, baseball and golf) , for making video games (such as the Nintendo *Wii*) more realistic and also help identify "unusual" or "unexpected" salient sounds.

1 Introduction

Modern T.V. coverage of major sports events offers both home viewers and spectators watching "live" a wide range of views – from various camera angles and slow-motion replays – and additional information such as match statistics and the speed of, or distance travelled by, the ball. However, in some ball games – including tennis, squash, golf, cricket and baseball – the ball is typically in contact with the bat, racket or club for a few milliseconds [1] : a period rather shorter than the interval between successive video frames. This makes precise detection and analysis of the impact impossible from use of the video footage alone. Some systems – such as *Hawkeye* [2], now used in major tennis tournaments to decide whether the ball landed in or out of the court – make use of several video cameras working at a very high frame rate. However, such technology is expensive and highly complex.

Detailed analysis of strokes, including mis-hits – whether in tennis, squash, golf, cricket or baseball - could be highly valuable as a coaching aid. However, in order for this to be accessible to "club level" players and coaches, the technology required would have to be inexpensive and easy to use.

An alternative to relying on video footage alone is to employ the acoustic signal(s) recorded during the match. Although there are several complications – including background noise, latencies due to the finite speed of sound and the effects of echos – all of which have to be allowed-for, the acoustic signal is sampled at a much higher temporal rate than is normal for video frames (several thousand times a second as opposed to 25 or 30 times a second). This means that the acoustic signals offer the opportunity to detect and analyse events which take place over very short timescales – including racket, club or bat on ball impacts and bounces of the ball – and would thus be difficult or impossible to detect using conventional video footage.

In this paper, we discuss our recent work on the analysis of the acoustic signals recorded with T.V. footage during a major tennis championship (Wimbledon 2005). We have used the video to "mark-up" notable "sound events" – including various types of tennis strokes, echos, bounces of the ball, audience applause, footsteps, speech and other vocalizations – audible on the sound signal and/or visible in its spectrogram. We have then made use of some methods inspired by techniques used in automatic speech recognition (matching of templates of Mel Frequency Cepstral Coefficients (MFCCs) [3]) and pattern recognition - Principal Components Analysis (PCA) [4], Markov models and Bayesian decision-making [5] in an attempt to classify examples of these various types of "salient sound events" correctly. Our initial results are promising, suggesting that, at least in the relatively controlled environment of championship tennis matches, these techniques can be reliably used for the detection and classification of notable sounds. Although the problems are likely to be more challenging in less controlled situations, the methods we have employed may prove useful in wider applications, such as monitoring the well-being of sick or elderly people, or security surveillance (see, e.g.. [6, 7]).

2 Audio-Visual Tennis Dataset

2.1 Wimbledon 2005 Data

Our data comes from TV broadcasts of four matches from the Wimbledon Lawn Tennis Championships of 2005, totalling over 95 minutes. We used the video footage to manually "mark-up" all notable "sound events" – over 1400 in all, of which around 800 were tennis strokes. These were put into 14 classes : 9 different types of tennis strokes and 5 categories of other sounds (see Table 1).

The audio accompanying the video footage was sampled at 48kHz, which allowed spectrographic analysis in the frequency range 0 – 24kHz .

2.2 Play and Non-Play Sounds

The "sound events" occurring during a tennis match can be broadly categorized into three types : sounds arising due to the play of the game itself (including the sound of the racket on ball impacts from the various tennis strokes, bounce of the ball on the surface of the court,) on court noise – speech from the players and match officials, grunts and other vocalizations from the players, echos and (from the 2006 and earlier championships) the "bleep" from the *Cyclops* system used (at that time) to detect when the ball landed just out of court, or clipped the net) – and background noise : speech, other vocalizations (including cheering) and applause from the spectators, announcements and speech from the TV commentators, and "background noise" from outside traffic, aircraft flying overhead and similar sources. Fortunately, the latter category did not prove to be significant in our dataset.

Code	Meaning
S1	First Serve
S2	Second Serve
FD	Forehand Drive
BD	Backhand Drive
BD2H	Backhand Drive Two Handed
SM	Smash
VO	Volley
SS	Stop shot
LO	Lob
F	Failure of Stroke
W	Stoke Wins the Point
E	End of rally
BC	Bounce of Ball
EC	Echo
AP	Applause
SP	Speech and other vocalisations
SL	Silence

Table 1. The codes we use in this paper for tennis stokes and other notable events : 9 distinct types of tennis strokes, 5 other "sound events", plus markers for a stroke failing, a stroke winning the point, and the end of the current rally.

3 Previous Work in the Area

3.1 Other Studies

Previous studies on the automated identification of "salient sound events" have followed diverse approaches, with varying degrees of success. Although some have attempted methods based on template matching, these have not always achieved good results. For example, [13] found that an approach based on template matching did not work well in the context of detecting the sound of "dribbling" in basketball. This led to other authors, including [14], rejecting template matching as a possible method for detecting the sounds of tennis strokes, despite the acoustic environments of tennis and baseball matches being very different – in tennis, many of the salient sound events are of short duration (of the order of 10-20 ms) but separated by intervals of order 1 second, whereas in basketball the ball bounces can occur several times a second. Furthermore, audiences in basketball matches tend to be quite noisy for much of the time. In major tennis championships, the spectators are normally quiet most of the time – except in situations of highly partisan support (e.g. "Henman Hill" at Wimbledon in recent years) ! Although some authors (e.g. [15] studying golf strokes and [16] detecting bat on ball impacts in baseball) have used an approach which has some aspects in common with our "template matching" approach, they have only used the signal power in a very small number of frequency bands [16] or a very small number of MFCC coefficients in conjunction with a neural network [15]. Some authors have only tried to *detect* impulsive sounds [14, 17], not classify them. Several studies have used peaks in Short Term Energy (STE) - the signal power averaged over a very short time window [17, 18].

Amongst the relatively small number of studies which attempt to classify different types of sounds [19, 20, 21, 22] (rather than just detect salient impulsive sounds), some only attempt to distinguish between radically different sounds [20] (periodic, impulsive, close to monotone or of very limited spectral range) and/or sounds of a small number of distinct classes (6 classes in the case of [19]). Zhang & Kuo [21, 22] used a "hierarchical" approach, using three levels of "coarse", "intermediate" and "fine" discrimination between sound types. At the "coarse" level, they used the power, zero crossing rate and fundamental frequency of the signal to distinguish between speech, music, environmental sounds and silence. The second and third levels employed Hidden Markov Models (HMMs) to sub-classify each category (e.g. for speech, whether the speaker is an adult or child, male or female; for music which genre – classical, jazz, rock, etc. - the sample belongs to), followed by a "querying/retrieval" approach to identify the most similar "previously heard" sounds to the current example.

3.2 Our Previous Work

In our previous work on this topic [8, 9, 10], we studied the dependence of the time interval between successive strokes and the acoustic energy from the first of those two strokes. It was found that, as predicted by Newtonian dynamics (assuming that the acoustic energy was proportional to the kinetic energy imparted to the ball in the stroke which produced that sound), the logarithm of the energy was, to a good approximation, linearly dependent on the logarithm of the time until the next stroke [8]. This implies that the energy in the acoustic signal can be a reasonable predictor of the time when the next stroke occurs.

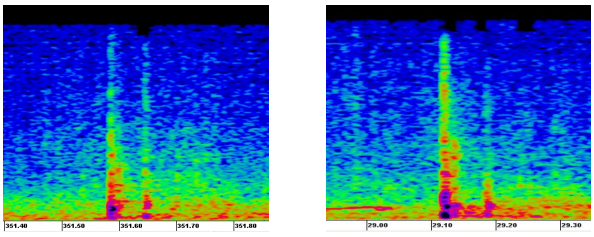
We also considered the statistical distributions of both the acoustic energy and the time interval before the next stroke for various types of stroke [8]. Not surprisingly, some strokes (such as first serves and forehand drives) were found to be consistently of high energy and a short time until next stroke, whilst others (such as lobs) were of lower, but more widely varying, energy and longer and widely varying time intervals before next stroke. We went on to study Markov models of stroke sequences [10] and the feasibility of using templates of MFCCs [3, 9] obtained from the acoustic signal to classify the "salient sounds" in tennis matches. It was found that only the "observable state" Markov model (and not HMMs) generated realistic length rallies [10], but on a limited dataset, the MFCC template matching approach did show encouraging success rates for classifying the sound events [9].

In the present paper, we extend this work by using a more comprehensive approach with partitioned datasets than in our previous studies, and investigate the feasibility of using a "hybrid" classification strategy, combining "prediction information" from the Markov model sboth detecting and correctly classifying tennis strokes and other "salient sounds" automatically, solely on the basis of evidence present in the acoustic signal.

4 Methodology

4.1 Acoustic Analysis

Many of the sound events of interest here are "impulsive" and some of the types considered have rather similar power and spectrographic profiles. This would make distinguishing between them, based on visual inspection of spectrograms alone, rather difficult. Two examples of such spectrograms are shown in Figure 1 . Further examples can be found in [9].



(a) (b)

Figure 1 : Example Spectrograms for "similar" sound events : (a) Forehand drive, (b) Two-handed backhand drive. These appear very similar to the human eye.

In our attempt to classify the "salient sounds" occurring during the tennis matches, we have created "spectrographic templates". Following a standard procedure used in speech processing, we have used 13 Mel Frequency Cepstral Coefficients (MFCCs) [3] – 12 for frequency bands covering the range 0 – 15 kHz and one representing the power in the audio signal. For each MFCC calculation, we use a window of duration 10 ms. We then formed a "template" of 20 such MFCC values, calculated over 20 windows, with any given MFCC window covering a period which overlapped by 5 ms with its predecessor and by 5ms with its successor in the template). Each template thus encodes sound events occurring within a period lasting 100ms in total. This value was chosen since it is relatively long compared with the duration of "impulsive" sound events (such as racket on ball impacts or bounces of the ball) occurring during the tennis matches, but short compared with the typical intervals (of the order of 1 second) between such events.

4.2 Principal Component Analysis

Our method of "encoding" the sound events in templates of MFCC values results in feature vectors of rather large dimensionality (20 x 13 = 260 in this case) for each example. In order to perform classification of examples into the 14 categories of sound events, it is highly desirable to work with a smaller number of "most useful" features. Principal Components Analysis (PCA) [4] is one method for achieving this, by choosing the "best" linear combinations of the original features. PCA proceeds by computing the eigenvalues and corresponding eigenvectors of the covariance matrix of all the examples for each category or class in the dataset of interest. In general, for data in N dimensions, there will be N eigenvalue-eigenvector pairs (not necessarily all distinct, but all the eigenvalues are non-negative). In PCA, the largest M eigenvalues (for M < N) are retained along with their corresponding eigenvectors. These will form a basis for a "reduced dimensional" space which will be the "most useful M dimensions" for identifying data of that category. The choice of M can be empirical in order to obtain satisfactory classification performance. In application, data points (spectrographic templates in this case) in the original space are projected onto the reduced (M dimensional) eigenspace for each category in turn. The example is put into whichever category gives the smallest distance (using an appropriate metric) between the original data point and its projection.

Here, each template is regarded as a "feature vector" with N = 260 components. Having used the video footage to identify the appropriate category (9 different types of tennis strokes and 5 "other events" – echo, ball bounce, applause, speech or silence) for each "sound event", the covariance matrix over all examples of that type was found. All the eigenvalue-eigenvector pairs for each category/matrix were then computed and the appropriate reduced eigenbasis for each class found by PCA. For various values of M << 260, we studied the effects on the successful classification rates of retaining only those dimensions in the eigenbasis corresponding to the largest M eigenvalues. In application, we used a Euclidean distance metric for determining the "closest" category (in the feature space) to any given example sound event, and that example was then classified as belonging to that category.

4.3 Markov Models

In our previous studies [8, 10], we found that, to a reasonable approximation, the distribution of rallies by number of strokes could be modelled by a relatively simple probabilistic, Markov type, model. Models where the states directly corresponded to actual classes of tennis strokes approximated the data from the videos of real matches rather better than Hidden Markov Models (HMMs), where the states could generate strokes (without a particular state necessarily being in 1-1 correspondence with any given type of tennis stroke) [10]. In this present paper, we extend this by (a) performing our experiments on partitioned datasets, in a "cross-validation" approach [11] and (b) looking at the distributions of short rallies (of up to 4 strokes), where meaningful statistics can be calculated for explicit distinct sequences of strokes (e.g. Serve followed by Forehand Drive, followed by Backhand Drive followed by Two-Handed Backhand Drive), at least for the more common such sequences. Using one subset of the available data to calculate a "transition matrix" of probabilities of one stroke type following another, we then generate "synthetic rallies" using this probabilistic model. The distribution of the set of sequences so generated are then compared with the distributions of the corresponding sequence of strokes (if it is present) in the dataset of "real rallies" from the TV footage, using a χ^2 "goodness of fit" test (with sequences where the simulations gave a number small number of occurrences being group together). Some examples of "real" rallies found in our dataset are given in table 2.

S1,F
S1,BD2H,FD,F
S1,FD,FD,BD,VO,W
S2,BD2H,FD,FD,FD,BD2H,SS,W
S1,FD,SS,LO,SM,BD2H,BD,FD,FD,BD2H,BD,FD,BD,W

Table 2 Examples of rallies in our dataset. The codes used for the tennis strokes are as specified in Table 1.

4.4 Hybrid Classifiers

Inspired by the "Acoustic Model & Language Model" approach widely used in Automatic Speech Recognition [12], we have attempted to improve the success of our classifier by combining information from the acoustic signal with predictions about which stroke types are likely or unlikely to occur, based on the statistical distribution of individual strokes in the dataset and/or on statistics relating to pairs of successive strokes (from the Markov model), using a "Naïve Bayesian" approach [5] to make decisions based on more than one source of information.

5 Results

5.1 Acoustic Template Matching

We have performed an N-fold cross-validation [11], after partitioning the data N ways and using (N - 1) parts for training and 1 part for testing in each case. (This contrasts with our previous experiments [9, 10] which were on a more limited dataset which was not partitioned.) Reasonably consistent results were obtained across the different partitions for both N = 2 and for N = 5. (The dataset contained too few examples of some strokes, such as smashes and lobs, to give meaningful results if larger values of N were used.) Serves were generally classified correctly as serves (83%), although second serves were sometimes mis-classified as first serves, and forehand drives were also normally classified correctly (73%). However, the other powerful "play" strokes : one- and two-handed backhand drives, were quite often mis-classified as forehand drives (43%). This is not so surprising since these strokes are spectrographically rather similar. The results on lobs, volleys, smashes and stop shots were also poor, although the data available for these was very limited. For the other sounds, 70% of echos, 75% of ball bounces, 73% of examples of speech and all examples of applause and "silence" were correctly classified. Around 15% of echos were mis-classified as forehand drives (of which they were probably the reflections) and around 13% of bounces of the ball were classified as "silence", probably because the signal power was then very low. No more than 10 PCA dimensions were retained for any one category in any case.

5.2 Markov Model "Simulated Rallies"

We generated a large number of simulated rallies using our "observable" Markov model and compared the first N strokes of these with the first N strokes of the real rallies in our dataset from the TV coverage. Meaningful results were only obtainable for $N \leq 4$, since the real data was two sparse for longer rallies. Scaling the simulated data to give an equivalent total number of rallies to the number in the real data (i.e. 217 rallies), and grouping together all distinct sequences for which the simulated data predicted fewer than 5 examples, we performed a χ^2 "goodness of fit" test to investigate whether our Markov model generated realistic proportions of each distinct rally sequence. For the "training" portion of the real data., an excellent fit (χ^2 = 0.4263, DF = 5, p = 0.995) was obtained, whilst with the "test" portion of the real data (χ^2 = 3.0919, DF = 5, p = 0.686) a satisfactory fit was obtained, suggesting that the Markov model did provide a reasonable explanation of the observed rally data. This complements our previous work which looked only at the lengths of rallies generated by the Markov model in comparison with the real data [10].

5.3 Hybrid Classifiers

We have compared the "successful classification rates" for tennis strokes, when applied to previously unseen data (i.e. data not used in training), of methods using the acoustic templates (with PCA) alone, the statistical distribution of individual strokes across the dataset (which can be considered equivalent to a "unigram" model in Statistical Language Modelling), and the predictions of the Markov model alone (equivalent to a "bigram" language model). Initially we employed an "Argmax" criterion, so that the category with the highest probability was selected. Under this approach, the acoustic model was most successful (overall 55% classified correctly), and the "unigram" model had the second best success rate (48%). However, due to the "Argmax" method, only forehand drives (the most common type of stroke in the data) were correctly identified, highlighting the limitations of this approach ! Under these conditions, the Markov (bigram) model gave 45% successful classifications. We then used a "Naïve Bayesian" scheme [5] to combine information from more than one source into our classification process. Still using the "Argmax" approach, hybrid Acoustic-Unigram, Acoustic-Markov, Unigram-Markov and Acoustic-Unigram-Markov all gave similar success rates to those of Unigram or Markov alone (45 to 48%).

In order to address the issue of "Unigram with Argmax" only being able to correctly classify the most common type of stroke, we investigated an alternative strategy where, if a sound event was given a probability $\boldsymbol{p_i}$ of being of class $\boldsymbol{i}$, $\boldsymbol{p_j}$ of being of class $\boldsymbol{j}$, etc., then it would randomly be assigned to class $\boldsymbol{i}$ with probability $\boldsymbol{p_i}$, etc. This approach did not yield good results for "successful classification rates" when applied to the "acoustic plus PCA" method on its own, and made little overall difference to the success rates for "unigram only" or "Markov only" classifiers. However, successful classification were made across all the different stroke types, in contrast to the results for the "Argmax" approach. The successful classification rates for the "hybrid" classifiers were also better with this "random" approach than using "Argmax", but, on the data studied to date, have never exceeded the success rates obtained using the "Acoustic model with PCA" under the Argmax scheme. This is currently under further investigation.

6 Conclusions and Further Work

We have shown that acoustic template matching using MFCCs and PCA can lead to reasonably successful classification of "salient sound events" in the controlled environment of championship tennis matches, and that an observable state Markov model can generate realistic rallies which provide a "satisfactory fit" (according to χ^2 tests) to the distribution of specific stroke sequences found in real data. Our basic acoustic classifier has been more successful in classifying some types of events than others. We have attempted to combine it with "predictions" based on the statistical distributions of single strokes ("unigram" model) and the Markov model ("bigram" model). To date these have had limited success, but we are investigating them further, along with statistics on sequences of 3 successive strokes ("trigram" models), which have been employed very successfully in statistical language modeling [12].

We also intend to generalise this work to audio-visual data obtained from "club level" tennis matches (aiming to incorporate our methods into coaching aids), to other sports such as golf, cricket and baseball, and to a wider range of situations, such as detecting "salient sounds" in security surveillance and in the care of the sick or elderly. Such environments are less controlled and predictable than championship tennis matches, with greater challenges.

Acknowledgments

This work has been funded by the EPSRC of the U.K. as part of the CAPS project (Grant GR/S78841/01). Krzysztof Zienowicz is grateful to the EPSRC for financial support.

References

[1] H. Brody, R. Cross & C. Lindsey, "*The Physics & Technology of Tennis"*, Racquet Tech (USRSA) Publishing, Solana Beach, U.S.A. (2002)

[2] Hawk-Eye Innovations, "Instant Replay Comes to Tennis" http://www.hawkeyeinnovations.co.uk (2006)

[3] J. Holmes & W. Holmes "*Speech Synthesis and Recognition"*, Taylor & Francis, London, U.K. (2001)

[4] I.T. Jolliffe, "*Principal Component Analysis"*, 2nd Edition, Springer, New York, USA (2002)

[5] R.O. Duda, P.E. Hart & D.G. Stork, *Pattern Classification*, Wiley, New York, pp 20-63 (2001)

[6] S. Bahadori A.Cesta, L. Iocchi, G.R. Leone, D. Nardi, F. Pecora, R. Rasconi & L. Scozzafava, "Towards Ambient Intelligence for the Domestic Care of the Elderly", in *Ambient Intelligence : A Novel Paradigm* (Editors : P. Remagnino, G.L. Foresti & T. Ellis), Springer, Chapter 2, pp 15-38 (2005)

[7] F. Rivera-Illingworth, V. Callaghan & H. Hagras, "A Neural Network Agent Based Approach to Activity Detection in AmI Environments", *Proceedings of the IEE International Workshop on Intelligent Environments (IE'05)*, Colchester, U.K., pp 92 – 100 (2005)

[8] G. Hunter, A. Shihab & K. Zienowicz, "Modelling Tennis Rallies Using Information from both Audio and Video Signals", *Proceedings of the I.M.A. International Conference on Mathematics in Sport*, pp 103-108, Salford, Manchester, U.K., June (2007)

[9] K. Zienowicz, G. Hunter & A. Shihab, "The Use of Spectrographic Template Matching to Identify and Classify Salient Sound Events in Tennis Matches", *Proceedings of the Institute of Acoustics, U.K.*, Vol. 30, Part 2, pp 171-179 (2008)

[10] K. Zienowicz, A.I. Shihab & G.J.A.Hunter, "Detecting, Classifying and Predicting Salient Events Using Acoustic Signals and Markov Models", *Proceedings of the 4th IET/IEEE/AAAI International Conference on Intelligent Environments (IE'08)*, Seattle, U.S.A. (2008)

[11] T.M. Mitchell, "Machine Learning", McGraw-Hill International, New York, pp 111 – 112 (1997)

[12] S.J. Young, "Large Vocabulary Continuous Speech Recognition : A Review", *IEEE Signal Processing Magazine*, Vol. 13(5), pp 45-57 (1996)

[13] D. Zhang & D. Ellis, "Detecting Sound Events in Basketball Video Archive", *Technical Report, Electrical Engineering Department,* Columbia University, U.S.A. (2001)

[14] R. Dahyot, A. Kokaram., N. Rea & H. Denman, Joint Audio-Visual Retrieval for Tennis Broadcasts", *Proceedings of the International Conference on Acoustics, Speech and Signal Processing* (2003)

[15] W. H-M. Hsu, "Golf Impact Detection with Audio Clues", http://www.ee.columbia.edu/~winston/courses/speechaudio/project/E6820PrjRpt.pdf (2002)

[16] Y. Rui, A. Gupta & A. Acero, "Automatically Extracting Highlights for TV Baseball Programs", *Proceedings of ACM Multimedia 2000*, Los Angeles, CA, pp 105-115 (2000)

[17] B. Zhang, W. Dou & L. Chen, "Ball Hit Detection in Table Tennis Games Based on Audio Analysis", *18th International Conference on Pattern Recognition (ICPR'06)*, pp 220-223 (2006)

[18] W. Lao, J. Han, & P.H.N. de With, "Automatic Sports Video Analysis using Audio Clues and Context Knowledge", *Proceedings of EuroIMSA* 2006, pp 198-202 (2006)

[19] A. Dufaux, L. Besacier, M. Ansorge, F. Pellandini, "Automatic Sound Detection and Recognition for Noisy Environment", *Proceedings of the X European Signal Processing Conference* (2000)

[20] D. Hoiem, Y. Ke, and R. Sukthankar, "SOLAR: Sound Object Localization and Retrieval in Complex Audio Environments ", *Proceedings of ICASSP* (2005)

[21] T. Zhang & C. Kuo, "Content-Based Classification and Retrieval of Audio", *SPIE Conference on Advanced Signal Processing Algorithms, Architecture and Implementations VIII* (1998)

[22] T. Zhang & C. Kuo, "Hierarchical System for Content-Based Audio Classification and Retrieval" *Proceedings of the ICASSP* (1999)