MPEG AUDIO BITSTREAM PROCESSING TOWARDS THE AUTOMATIC GENERATION OF SPORTS PROGRAMME SUMMARIES

David A. Sadlier, Sean Marlow, Noel O'Connor & Noel Murphy*

Centre for Digital Video Processing
Dublin City University, Ireland
sadlierd@eeng.dcu.ie

ABSTRACT

The frequency subband scalefactors are fundamental components of MPEG-1 audio encoded bitstreams. Examination of scalefactor weights is sufficient for the establishment of an audio amplitude profile of an audio track. If, for sports programme TV broadcasts, the audio amplitude is assumed to primarily reflect the noise level exhibited by the commentator (and/or attending spectators), then, this vocal reaction to the significance of unfolding events may be used as a basis for summarisation. i.e. by relying on the exhilaration, or otherwise, expressed by the commentator/spectators, individual clips of the programme (e.g. camera shots), may be ranked according to their relative significance. A summary may then be produced by amalgamating (chronologically) any number of these clips corresponding to selected audio peaks.

1. BACKGROUND

The Centre for Digital Video Processing at Dublin City University conducts concentrated research in developing innovative technologies fundamental to the realisation of efficient video content management. The current stage of development is demonstrated in the web-based digital video system, *Físchlár* [1]. *Físchlár* captures TV broadcast programmes and encodes them according to the MPEG-1 (Layer-II) video standard. *Físchlár* then provides for efficient analysing, browsing, and viewing of the recorded content. At present, a user can pre-set the recording of programmes selected from an online TV broadcast schedule, and then choose from a set of different browser interfaces which allow navigation through the recorded material. As the research develops, increased options such as personalisation and programme recommendation, automatic recording, SMS/PDA/WAP alerting, searching, etc. are being plugged in and utilised.

2. INTRODUCTION

Recent developments in video compression technologies have paved the way for substantial allowances in extensive archiving of video content. The limited bandwidth availability for an online video streaming application, rooted on such archives, suggests an increasingly crucial role for highlighting of video content. This point is accentuated when considering wireless video downloads to hand-held mobile devices.

In all observed previous works within the sports summarisation research field, the analyses have been specifically tailored towards the individual sports types. Rui *et al.* [2] & Zhang *et al.* [3], using purely audio based methods, both report moderate success in event detection within TV baseball and basketball programmes, respectively. Whereas, Yow *et al.* [4] & Ekin *et al.* [5] describe some purely visual based methods for highlight extraction from soccer content.

It is the authors' vision to develop a global sport broadcast analysis application which initially detects the particular sport type, and then using type-tailored audio/visual/textual tools, attempts to automatically summarise the content.

However, the first phase of the research was to quantify the effectiveness of unaccompanied, generic audio amplitude analysis on sports event broadcasts, within an limited type domain: TV events such as soccer, gaelic, rugby and hockey matches etc. are typically audio-busy broadcasts, prominently characterised by spectator cheering and enthused announcer commentary. Assuming a direct correlation between commentator/spectator zest and momentary significance, an amplitude profile, presenting a good indication of the variance of these features, may provide for effective highlight extraction. This paper treats of the procedures involved in the investigation of these ideas.

3. AUDIO AMPLITUDE PROFILE GENERATION

3.1 Frequency Band Scalefactors – MPEG-1 Audio

The MPEG-1 Layer-II (MP2) compression algorithm encodes audio signals as follows: the frequency spectrum of the audio signal, bandlimited to 20kHz, is uniformly divided into 32 subbands which approximate the ear's critical bands. The subbands are assigned individual bit-allocations according to the audibility of quantisation noise within each subband.

Layer-II audio frames consist of 1152 samples; 3 groups of 12 samples from each of 32 subbands. A group of 12 samples gets a bit-allocation and, if this is non-zero, a scalefactor. Scalefactors are weights that scale groups of 12 samples such that they fully use the range of the quantiser. The scalefactor for such a group is determined by the next largest value (in a look

0-7803-7304-9/02/$17.00 C2002 IEEE

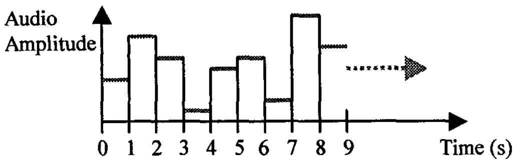

Figure-1: Per-Second Audio Amplitude Profile

up table) to the maximum of the absolute values of the samples. Therefore, scalefactors provide an indication of the maximum power exhibited by any of the 12 samples within the group.

3.2 Spectral Focus

In the MP2 audio compression standard, which *Físchlár* adheres to, the maximum allowable frequency component is at 20kHz. At the encoder, the frequency spectrum (0-20kHz) is divided uniformly into 32 subbands. Thus subbands 2 through 7 represent the frequency range from 0.625kHz – 4.375Khz. Hence, the span of these subbands approximates that of relaxed speech which typically ranges between 0.1kHz and 4kHz.

For sports programme audio tracks, by strictly focusing the audio analysis on the content within these subbands, which approximate the range of the speech band, we concentrate on the audio components containing commentator vocals. Therefore the influence of the commentator on the generation of the audio amplitude profile is increased. This is clearly desirable since it is assumed that commentator vocals represent the most reliably accurate noise-level indicator of event significance.

It was expected that the examination of subbands 2-7 would provide for a reasonable trade-off between rejection of low/high (and possibly destructive) background noise and the capture of the principal frequency content of excited commentator speech.

3.3 Preamble: Boundary Detection

One of the problems envisaged with the audio amplitudes technique for automatic sports summarisation is generated by the inclusion of supplementary content which typically accompanies the main event in a sports programme broadcast. Features such as athlete profiles, highlights of recent events etc. tend to contain attributes such as commentator dialogue and spectator noise which are comparable, in amplitude, to that of the event of interest. To combat this setback, the system must be able to detect the temporal boundaries of the main feature within the overall sports programme. This is done by searching through the entire audio track for extended periods of sustained volume. Peripheral programme segments such as interviews, studio discussions, archive video clips, etc. are flagged by the intermittent occurrence of very brief moments of silence. For example, infinitesimal silences exist in between sentences spoken by an anchorperson, or when switching from studio scenes to archive video clips, or between advertisements. In contrast the main event of a sports broadcast (of type within our limited domain) features relatively long periods of sustained volume due to the continuous presence background noise. On this basis it may be automatically distinguished from the supplementary content. i.e. the temporal boundaries of the main event within the overall programme may be detected. For the summary generation, the probing domain is restricted to lie within these boundaries.

4. CASE STUDY

4.1 Objective

The following is an illustration of the automatic generation of a 10-minute summary of a terrestrial TV broadcast of a sports event via the discussed technique. The experimental subject is the *UEFA Cup Final 2001* featuring *Liverpool FC* Vs *Alaves FC*. This was a near 3-hour soccer match broadcast, resulting in a 5-4 victory for *Liverpool FC*. The programme featured the main event plus studio discussions and analysis, player profiles highlights of related events and advertisement breaks.

4.2 Amplitude Profiles

A second-by-second audio amplitude profile was established by a superposition of all scalefactors from subbands 2-7 repeated over a window length of one second. See Figure-1.

Separately, at a higher temporal resolution, a frame-by-frame audio amplitude profile was established by a superposition of all scalefactors from subbands 2-7 repeated over a window of length corresponding to one video frame (≈1/25s).

4.3 Preamble: Boundary Detection

The overall structure of the near 3-hour subject (as captured by *Físchlár*) is described below. In terms of summary generation, segments of interest are identified by an asterisk.

Advertisements	≈ 3 mins
Prog. start: studio discussions, interviews	≈ 14 mins
First half *	≈ 51 mins
Studio analysis, discussions and archive video clips	≈ 14 mins
Second half *	≈ 49 mins
Studio analysis and discussion	≈ 4 mins
Extra time *	≈ 26 mins
Studio analysis (to prog. end)	≈ 6 mins
	167 mins

A silence threshold was empirically defined as

$$S_{th} = 0.033 * \text{overall mean audio amplitude}$$

Using the per-frame audio amplitude profile and S_{th}, the entire audio track of the subject was examined for periods of continuous volume lasting for (e.g.) at least 1-minute. It was found that sustained volumes exceeding S_{th} occur during the following video frames:

- 309 – 3504 .... ≈ 2 mins
- 3876 – 6608 .... ≈ 1 mins
- 6984 – 13577 .... ≈ 4 mins
- 15037 – 19467 .... ≈ 3 mins
- 19553 – 23405 .... ≈ 2 mins
- 26248 – 102751* .... ≈ 51 mins
- 106225 – 109935 .... ≈ 2 mins

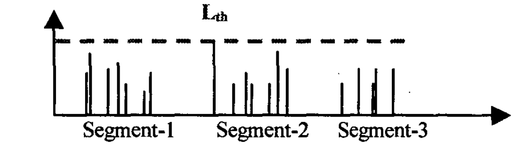

Figure-2: Exclusive Examination of Segments 1-3

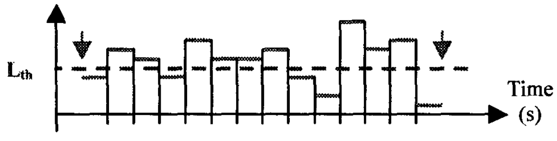

Figure-3: Decreasing L_{th} and detecting *audio surges*

- 112596 – 115354.. ≈ 1 mins
- 123629 – 198950*...................................... ≈ 50 mins
- 199586 – 201168.. ≈ 1 mins
- 201252 – 244086*..................................... ≈ 28 mins
- 245527 – 247690.. ≈ 1 mins

Further thresholding at (e.g.) a length of 10 minutes rejects all segments except for three (identified above by an asterisk), which correspond almost precisely to the segments of interest mentioned previously (i.e. the temporal boundaries of the match play segments were accurately detected). Changing units to seconds these are:

- Segment-1: 1050s – 4110s
- Segment-2 : 4945s – 7958s
- Segment-3: 8050s – 9763s

Only the content which resides within these boundaries is eligible for inclusion in the summary. Hence, further audio examination is restricted accordingly.

The boundary detection preamble is not a crucial component of the summarisation procedure i.e. in the event of failure, the main audio analysis procedure would still be expected to produce a moderately successful summary. However, it is beneficial tool which prevents the consideration of irrelevant material and in doing so, lightens the workload of subsequent procedures.

4.4 Summary Generation

The per-second amplitude profiles of segments 1-3 (above) were examined. A loudness threshold, L_{th}, was initialised to the value corresponding to the largest peak found. See Figure-2. An amplitude peak is defined *loud* if it exceeds L_{th}. Ignoring isolated peaks, L_{th} was gradually reduced until it began to pick out *loud* periods of at least 3-seconds in duration (*audio surges*).

Figure-3 shows three sections which extend beyond the current value of L_{th}. The second and third have time spans of 4 *loud* seconds and 3 *loud* seconds respectively. Thus both are recognised as *audio surges*. The first section is ignored since with a length of just 2 *loud* seconds, it does not (yet?) meet the minimum *surge* length threshold of 3-seconds.

L_{th} was further reduced until the amount of detected surges was sufficient such that a 10-minute video summary could be produced. The summary was then generated by first matching up the moments within the combined A/V track which temporally align with the *audio surges*. Then, appending a pre-clip buffer of (e.g.) 5-seconds and a post clip buffer of (e.g.) 20-seconds (to make viewing the amalgamation less visually disturbing). Then finally extracting and combining (chronologically) these clips from the audio/video stream. See Figure-4.

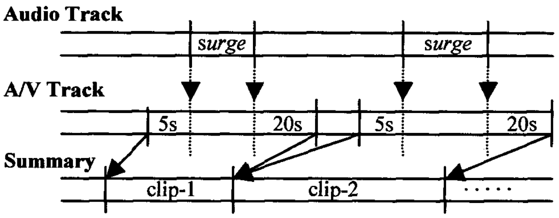

Figure-4: Summary generation

4.5 Results & Classification

The analysis returned 18 individual clips corresponding to the following descriptions, yielding a combined length of ≈10-mins:

1. Teams come out #
2. Goal ♥
3. Goal ♥
4. Penalty Offence ♥
5. Goal ♥
6. Irrelevant content Δ
7. Goal ♥
8. Goal ♥
9. Yellow Card Offence ♥
10. General Play *
11. General Play *
12. Substitution #
13. Controversial Foul #
14. Goal ♥
15. General Play *
16. Goal ♥
17. Red Card Offence ♥
18. General Play *

For the purposes of evaluation, several colleagues were requested to examine and pigeonhole each of the eighteen clips into one of four categories according to significance. The opinions were pooled and are now summarised: Ten clips (♥) seemed to depict very significant moments of the feature and hence were described as *definite highlights*. The inclusion of *definite highlights* in the summary is always preferred. Three of the clips returned seemed to represent moments of arguably lesser significance (#). These were described by the term *quasi-highlights*, and their inclusion in the summary is desired once all *definite highlights* already have been. The system returned four further clips containing content of considerably less significance (*), labeled *lowlights*. Inclusion of *lowlights* would typically not be tolerated except when the combined duration of all *definite* and *quasi-highlight* clips fails to satisfy the desired length of the summary.

A slight miscalculation in the boundary detection allowed for a minor amount of peripheral content to be included within the summary probing domain. Consequently, the summary contained a clip containing irrelevant content (Δ). This result was labeled *error*, and inclusion of clips of this type in the summary is undesired under any circumstances.

Sports Broadcast (program length)	Segments of interest	Results of Boundary Detection
1.Soccer (2.8hrs)	51min (1st half).... 49min (2nd half)... 26min (extra time).	51min segment detected 50min segment detected 28min segment detected
2.Gaelic Football (2hrs)	37min (1st half).... 38min (2nd half)...	37min segment detected 38min segment detected
3.Ice Hockey (2hrs)	21min (1st third)... 22min (2nd third)... 21min (3rd third)...	21min segment detected 22min segment detected 21min segment detected
4.Gaelic Football (2hrs)	39min (1st half).... 36min (2nd half)...	39min segment detected 36min segment detected
5.Rugby (2hrs)	42min (1st half).... 44min (2nd half)...	42min segment detected 44min segment detected
6.Soccer (2hrs)	47min (1st half).... 49min (2nd half)...	52min segment detected 49min segment detected
7.Rugby (2hrs)	43min (1st half).... 43min (2nd half)...	43min segment detected 43min segment detected
8.Soccer (2hrs)	46min (1st half).... 48min (2nd half)...	46min segment detected 48min segment detected
9.Field Hockey (2hrs)	37min (1st half).... 37min (2nd half)...	37min segment detected 37min segment detected

Table-1 Results of Boundary Detection Preamble

4.6 Evaluation

The objective was to automatically generate a 10-minute summary of the sports broadcast *UEFA Cup Final 2001.* Pure audio analysis yielded a 10.3 minute long amalgamation of 18 individual clips from the programme. From these, seventeen clips related to the main feature; it was of the opinion of several objective colleagues that ten of these corresponded to largely important moments of the feature, three exhibited content of a more debatable significance, and the remaining four presented content of a more inconsequential nature.

It is important to note that no qualitative analysis of the video content has been performed, therefore it currently remains unknown whether or not the programme contains any further, undetected, content which would have deserved the *definite-* or *quasi-highlight* label. If so, then a number of the included *lowlight* clips would represent false positives. This number is currently unevaluated and thus the inclusion of all the *lowlight* clips is not yet indisputably justified. However, concentrating on the amount of true-positives returned, 72% of the summary is comprised of at least *quasi*-significant material and viewed as a whole, the amalgamation provides a coherent synopsis of the dramatics of the feature.

5. EXPERIMENTAL RESULTS & CONCLUSION

In a similar manner, a further eight 10-minute summaries were generated from various other (appropriate) broadcast sports programmes. Again, no qualitative analysis of the subject matter has been performed, so attention should be paid primarily to the number of true-positive returns (*definite & quasi-highlights*).

Table-1 presents the results of boundary detection preambles.

		Clip Class Breakdown			
Sports Broadcast	**Total Clips Returned**	**Definite (♥)**	**Quasi- (#)**	**Low (*)**	**Error (Δ)**
1.Soccer	18	10	3	4	1
2.Gaelic Football	21	9	11	1	0
3.Ice Hockey	20	9	6	5	0
4.Gaelic Football	24	12	11	1	0
5.Rugby	18	7	8	3	0
6.Soccer	17	9	7	0	1
7.Rugby	20	8	6	6	0
8.Soccer	14	3	9	2	0
9.Field Hockey	22	8	9	5	0

Table-2 Clip Classification Breakdown of Summaries

Table-2 presents a summarisation of the opinions of several objective colleagues on the clip classification breakdown of all generated summaries.

From Table-1 we can clearly see that, in general, the boundary detection preamble provides for accurate pinpointing of the segments of interest within a sports programme broadcast.

The opinions given in Table-2 suggest that the audio analysis we perform makes a more than useful contribution to the summarisation task. Averaging the results of all nine experiments, over 43% of returned content was objectively said to relate to events of a *definite* significance. This was coupled with a further 40% corresponding to *quasi*-important moments.

The work reported here is a preliminary investigation into the usefulness of pure audio analysis for summarisation of (limited types of) sports programmes. Alternative audio-only techniques are described via [2] & [3], however they differ in the fact that they both attempt to obtain an immediate final solution for just a single sports type. These works represent the pursuit of a unique objective implemented via the tracking of specifically tailored audio metrics, which cannot be applied to other sports types. We have shown that our work makes a useful, more generic contribution to event detection. However, it is not solely expected to retrieve an ultimate solution. Rather, it represents the establishment of a generic audio-based retrieval foundation, upon which more sophisticated, type orientated, audio/visual techniques such as camera shot classification, voice prosody analysis, feature tracking etc. may be suitably developed.

6. REFERENCES

[1] McDonald, K. *et al., Use of the Físchlár Video Library System*, Int. Conf. User Modeling, Sonthofen, Germany 2001.

[2] Rui, Y., Gupta, A., Acero, A., *Automatically Extracting Highlights for TV Baseball Programs*, Proc. ACM Multimedia, Los Angeles USA, pp105 -115, 2000.

[3] Zhang, D., Ellis, D., *Detecting Sound Events In Basketball Video Archive*, http://www.ctr.columbia.edu/~dpwe/courses/e6820-2001-01/projects/dqzhang.pdf

[4] Yow, D. *et al., Analysis and Presentation of Soccer Highlights from Digital Video*, Proc ACCV '95, Singapore.

[5] Ekin, A., Tekalp, A.M., *A Framework for Tracking and Analysis of Soccer*, Proc. VCIP 2002, San Jose, CA.