**SEÁN MARLOW**
**DAVID A. SADLIER**
**NOEL O'CONNOR**
**NOEL MURPHY**
**Dublin City University, Ireland**
**marlows@eeng.dcu.ie**

# *Voice Processing for Automatic TV Sports Program Highlights Detection*

## I INTRODUCTION

The imminent rapid expansion in the number of TV channels is driving the need for efficient digital video indexing, browsing and playback systems. For the past four years, the Centre for Digital Video Processing in DCU has been working towards the provision of such a system. The current stage of development is demonstrated on our Web-based digital video system called *Físchlár* [1,2], which is now in hourly use by over 1000 registered users. At present a user can pre-set the recording of TV broadcast programmes and can choose from a set of different browser interfaces which allow navigation through the recorded programmes. As our research has developed we have been adding increased functionality such as personalisation and programme recommendation, automatic recording, SMS/WAP/PDA alerting, searching, summarising and so on.

There have been many approaches to automatic highlight detection and summarization of sports programmes [3-6]. However, most methods are oriented to particular sports and/or specific edit effects or need much computational effort. Therefore we decided to take a more basic and generic approach based on the observation that, in most common sports coverage with a commentator and spectators, exciting events are usually characterized by an increase in the audio amplitude, particularly in the speech band. Measurement of the audio amplitude is described in Section II and a case study is presented in Section III. Section IV summarises results while conclusions and areas of continuing and future work are outlined in Section V.

## II AUDIO PROFILE GENERATION

### *a) MPEG Bitstream Processing*

The *Físchlár* system captures television broadcasts and encodes the programmes according to the MPEG-1 digital video standard with the audio signal coded in line with the Layer-II profile [8]. Unlike many other audio compression algorithms, which make assumptions about the nature of the audio source, MPEG-1 Audio exploits the perceptual restrictions of the human auditory system, via psychoacoustic weighting of the bit allocation for each frequency subband, to attain its compression [7].

The MPEG-1 Layer-II compression algorithm encodes audio signals by dividing the frequency spectrum of the audio signal, bandlimited to 20kHz, into 32 subbands. The subbands are assigned individual bit-allocations according to the audibility of quantisation noise within each subband. A pyschoacoustic model of the ear analyses the audio signal and provides this information to the quantiser. Layer-II 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 (given in a look-up table) to the maximum of the absolute values of the 12 samples. Thus it provides an indication of the maximum power exhibited by any one of the 12 samples within the group.

### *b) Amplitude of the Speech Band*

Most of the energy in a speech signal lies between 0.1kHz – 4kHz. According to the MPEG-1 Layer-II audio standard, the maximum allowable frequency component in the audio signal is at 20kHz. At the encoder, the frequency spectrum (0 – 20kHz) is divided uniformly into 32 subbands, each having bandwidth of 0.625kHz [7]. Thus, subbands 2 through 7 represent the frequency range from 0.625kHz – 4.375kHz. See Figure 1.

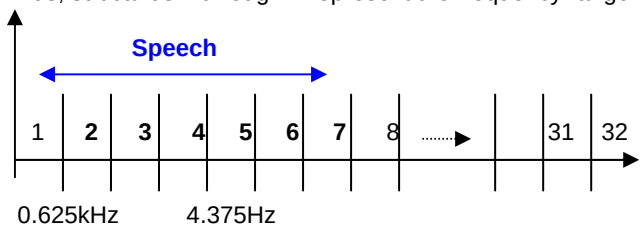


**Figure 1:** MPEG-1 Layer-II Frequency Subbands

By limiting the audio examination to these subbands, which approximate the range of the speech band, we further concentrate the audio investigation on commentator vocals. Therefore, the influence of the commentator on the

generation of the audio amplitude profile is increased. It was expected that the examination of subbands 2 though 7 would provide for a reasonable trade-off between rejection of low-frequency background noise (typically present in sports programmes which would naturally upset results) and the capture of excited speech.

*c) Boundary Detection*

One of the problems with the audio amplitudes technique is caused by the inclusion of supplementary content which typically accompanies the main event in a sports programme. Features such as player profiles, highlights of recent events etc. tend to contain attributes such as commentator dialogue and spectator noise, similar to that of the main event. The problem is that these features generally have audio amplitudes comparable to that of the event of interest. To combat this problem, 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 audio track for extended periods of sustained volume.

Segments such as interviews, studio discussions, archive video clips, etc. which make up the peripheral content, are flagged by the intermittent occurrence of brief moments of silence. For example, short silences exist in between sentences spoken by an anchorperson, when switching from anchorperson to video clips, or between advertisements. In contrast, the main event in a sports broadcast features relatively long periods of sustained volume due to the continuous presence of 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.

## III CASE STUDY

*a) Task*

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.

*b) Amplitude Profiles*

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

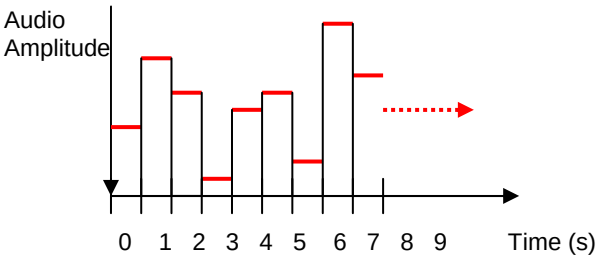


**Figure 2:** Per-Second Audio Amplitude Profile

A frame-by-frame audio amplitude profile was established by a superposition of all the scalefactors from subbands 2-7 over a window of length corresponding to one video frame (˜1/25s). See Figure 3.

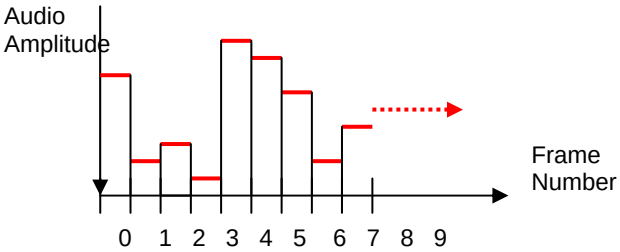


**Figure 3:** Per-Frame Audio Amplitude Profile

*c) Boundary Detection*

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

Ads ~3 mins

| | | |
|---|---|---|
| | Interviews | ~14mins |
| * | 1st half | ~51mins |
| | Studio analysis | ~14mins |
| * | 2nd half | ~49mins |
| | Studio analysis | ~4 mins |
| * | Extra time | ~26mins |
| | Studio analysis | ~6 mins |

Boundary Detection using the per-frame audio amplitude profile with a silence threshold of

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

eliminated all but the 1st half, 2nd half and extra time from subsequent processing.

*d) Summary Generation*

The per-second audio amplitude profiles of segments 1-3 (above) were examined. A loudness threshold, $\mathbf{L_{th}}$, was defined and initalised to the value corresponding to the largest peak found. See Figure 4.

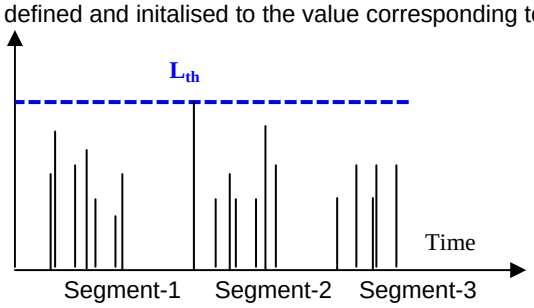


**Figure 4:** Examination of Segments 1-3

An audio amplitude peak is defined as *loud* if it exceeds $\mathbf{L_{th}}$. Ignoring isolated peaks, $\mathbf{L_{th}}$ was gradually reduced until it began to pick out loud periods of at least 3-seconds in duration (*audio surges*). See Figure 5.

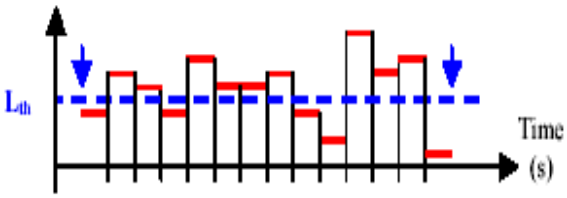


**Figure 5:** Decreasing $\mathbf{L_{th}}$ and detecting *audio surges*

Figure 5 shows three sections which extend beyond the current value of $\mathbf{L_{th}}$. The second and third have time spans of 4 seconds and 3 seconds respectively. Thus both are recognised as *audio surges*. The first section is ignored since with a length of 2 seconds, it does not meet the minimum *surge* threshold of 3-seconds. $\mathbf{L_{th}}$ 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 video clips within the combined audio/video track which temporally align with the *audio surges.* Then, a pre-clip buffer of 1 shot and a post clip buffer of 2 shots was appended (to make viewing the amalgamation less visually disturbing). Finally these clips were extracted from the audio/video stream and (chronologically) concatenated to generate the highlights summary. See Figure 6.

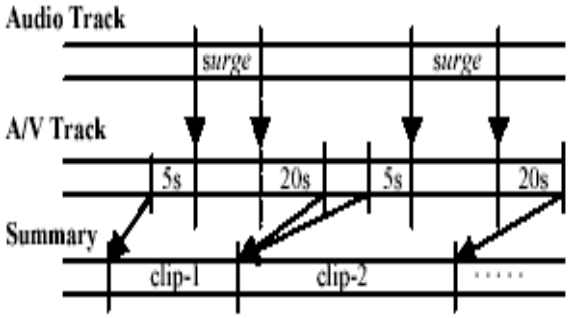


**Figure 6:** Summary generation

## IV RESULTS

The analysis returned 18 individual clips corresponding to the following descriptions, comprising a summary length of just over 10-minutes:

1. Teams come out #
2. Goal ?
3. Goal ?
4. Penalty Offence ?
5. Goal ?
6. Irrelevant content ?
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, the twenty clips returned were examined and classified into four categories according to significance. 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 *semi-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 length of all *definite* and *semi-highlight* clips fails to satisfy the desired length of the summary.

A slight error in the boundary detection results 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. This result was labeled *error* (**?**), and inclusion of clips of this type in the summary is undesired under any circumstances.

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 noted that ten corresponded to largely important moments of the feature, a further 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 a full qualitative analysis of the video content has yet 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 being evaluated 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 significant material and viewed as a whole, the amalgamation gives a very coherent synopsis of the dramatics of the feature.

## V SUMMARY OF RESULTS

The work reported here is a preliminary investigation into the usefulness of pure audio analysis for summarisation of (limited types of) sports programmes. A further eight 10-minute summaries were generated from various other broadcast sports programmes.. The content of returned clips, which make up the final summary, is listed in Table 1. Again, a full qualitative analysis of the subject content has not yet been performed, so attention should be paid primarily to the number of true-positive returns (*definite* & *semi-highlights*).

To the best of our knowledge, examination of scalefactor weights represents the most efficient method of constructing the audio amplitude profiles from encoded MPEG audio. The example in Section III was performed on a P3-700MHz machine running the Mandrake Linux operating system, and took 2-minutes computing time to retrieve the 18 clip locations directly from the audio bitstream information.

| | | **Clip Classification** | | | |
|---|---|---|---|---|---|
| Sports Broadcast | Total Clips Returned | Definite **(?)** | Semi **(#)** | Low **(*)** | Error **(?)** |
| 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 1**: Classification of clips included in automatic summaries

## VI CONCLUSIONS AND FUTURE WORK

From the preliminary results presented above, we can see that the audio analysis we perform makes a significant contribution to summarisation of sports TV broadcasts. We are working to improve the usefulness of the sports summary by analyzing the prosody (pitch, duration of phonemes and amplitude) of the commentator's voice, detecting slow-motion replays, tracking individual players using word-spotting, removing ad breaks and detecting the goal area in football matches.

In a real scenario, automatic summarisation of such broadcasts would depend on some combination of an analysis of the closed captions (teletext), analysis at the visual level, as well as the analysis we have reported here. In separate work we are working on analysing TV broadcasts in those other domains.

## REFERENCES


[1] Lee H., Smeaton A., O'Toole C., Murphy N., Marlow S. & O'Connor N., *The Físchlár Digital Video Recording, Analysis, and Browsing System*, Proc. Content based Multimedia Information Access (RIAO 2000), Vol. 2, pp. 1390-1399, Paris, France, 12-14 April 2000.

[2] Center of Digital Video Processing/ Físchlár web site: http://www.cdvp.dcu.ie.

[3] Intille S. *Tracking using a local closed-world assumption: Tracking in the football domain,* Proc. SPIE Storage and Retrieval for Image and Video Databases, pp. 216-227, 1997.

[4] Kawashima T. et al, *Indexing of baseball telecast for content-based video retrieval,* Proc. IEEE Int. Conf. Image Processing, pp. 871-875, 1998.

[5] Saur D. et al, *Automated analysis and annotation of basketball video,* Proc. SPIE Storage and Retrieval for Image and Video Databases, pp. 176-187, 1997.

[6] Yow D. et al *Analysis and presentation of soccer highlights from digital video,* Proc. Asian Conf. Computer Vision, 1995

[7] Pan, D., *A Tutorial on MPEG/Audio Compression*, IEEE Multimedia Journal, pp. 60-74, 1995.