Dipartimento di Informatica
Università del Piemonte Orientale "A. Avogadro"
Via Bellini 25/G, 15100 Alessandria
http://www.di.unipmn.it

An Audio-Video Summarization Scheme Based on Audio and Video Analysis

Marco Furini and Vittorio Ghini
(marco.furini@mfn.unipmn.it, ghini@cs.unibo.it)

TECHNICAL REPORT TR-INF-2005-10-04-UNIPMN
(October 2005)

The University of Piemonte Orientale Department of Computer Science Research Technical Reports are available via WWW at URL `http://www.di.mfn.unipmn.it/`. Plain-text abstracts organized by year are available in the directory

Recent Titles from the TR-INF-UNIPMN Technical Report Series

2005-03 *Achieving Self-Healing in Autonomic Software Systems: a case-based reasoning approach*, Anglano, C., Montani, S., October 2005.

2005-02 *DBNet, a tool to convert Dynamic Fault Trees to Dynamic Bayesian Networks*, Montani, S., Portinale, L., Bobbio, A., Varesio, M., Codetta-Raiteri, D., August 2005.

2005-01 *Bayesan Networks in Reliability*, Langseth, H., Portinale, L., April 2005.

2004-08 *Modelling a Secure Agent with Team Automata*, Egidi, L., Petrocchi, M., July 2004.

2004-07 *Making CORBA fault-tolerant*, Codetta Raiteri D., April 2004.

2004-06 *Orthogonal operators for user-defined symbolic periodicities*, Egidi, L., Terenziani, P., April 2004.

2004-05 *RHENE: A Case Retrieval System for Hemodialysis Cases with Dynamically Monitored Parameters*, Montani, S., Portinale, L., Bellazzi, R., Leonardi, G., March 2004.

2004-04 *Dynamic Bayesian Networks for Modeling Advanced Fault Tree Features in Dependability Analysis*, Montani, S., Portinale, L., Bobbio, A., March 2004.

2004-03 *Two space saving tricks for linear time LCP computation*, Manzini, G., February 2004.

2004-01 *Grid Scheduling and Economic Models*, Canonico, M., January 2004.

2003-08 *Multi-modal Diagnosis Combining Case-Based and Model Based Reasoning: a Formal and Experimental Analisys*, Portinale, L., Torasso, P., Magro, D., December 2003.

2003-07 *Fault Tolerance in Grid Environment*, Canonico, M., December 2003.

2003-06 *Development of a Dynamic Fault Tree Solver based on Coloured Petri Nets and graphically interfaced with DrawNET*, Codetta Raiteri, D., October 2003.

2003-05 *Interactive Video Streaming Applications over IP Networks: An Adaptive Approach*, Furini, M., Roccetti, M., July 2003.

2003-04 *Audio-Text Synchronization inside mp3 file: A new approach and its implementation*, Furini, M., Alboresi, L., July 2003.

2003-03 *A simple and fast DNA compressor*, Manzini, G., Rastero, M., April 2003.

An Audio-Video Summarization Scheme Based on Audio and Video Analysis

Marco Furini
Computer Science Department
University of Piemonte Orientale
15100 Alessandria, Italy
Email: furini@mfn.unipmn.it

Vittorio Ghini
Computer Science Department
University of Bologna
40127 Bologna, Italy
Email: ghini@cs.unibo.it

TECHNICAL REPORT TR-INF-2005-10-04-UNIPMN

Abstract

The availability of video files in the Internet is growing at an exceptional speed and in the near future video browsing will be a common activity. To facilitate such activity it will be necessary to have a small clip for any given video. Currently, video skimming and video summarization techniques can reduce the temporal representation of a given video. However, most of these techniques do not include audio in the produced summaries. Here, we propose a mechanism that, using audio and video analysis, produces video summaries coupled with intelligible audio. Experimental results show that the summaries are largely reduced (up to 50%) and that the perceived video quality may be comparable to the one of the original video (in term of jerkiness). Consumers satisfaction has been investigated through MOS and results show that our summaries can be considered as an alternative to the original videos.

1 Introduction

Thanks to the advances in networking and multimedia technologies, video information are massively entering our life. Nowadays, a variety of devices (palms, cellulars, laptops) has Internet connection (either Fiber, DSL, Cable, dial-up, UMTS, GPRS) and can easily play out video files.

The popularity of video files is highlighted by the increasing number of TV-show downloads that P2P systems are experiencing and by the video searching

mechanisms recently introduced by Google and Yahoo. This popularity is expected to increase in the near future thanks to the development of portable devices with LCD color screen, high storage capacity and broadband Internet connection. Moreover, the availability of video contents will increase as the TV industry is ready to provide television programs available for a fee, on legitimate websites, in order to contrast piracy [1].

The growing availability of videos will expose consumers to video library with ten of thousands videos. Hence, video browsing will be a common activity aimed at finding the right video. To facilitate such activity, each available video should be represented with a temporal reduced version so that the browsing may be performed on the reduced version and the consumer can decide which video he/she wishes to buy. Since it would be too expensive to manually produce such reduced versions, it is necessary to develop mechanisms that automatically produce such versions. Needless to say, the video summary should give a global picture of the video and should be as short as possible.

Currently, video skimming or video summarization techniques can be used to produce a small clip of a given video. These techniques include two main approaches: one uses only video analysis to produce summaries, while the other uses both audio and video analysis. In both approaches, the produced summaries are usually not provided with audio features. Among the ones that consider only video analysis, some selects a set of key video frames and re-arranges them in order to form a new video (see for example, [2, 3, 4, 5]); others propose to perform fast-forward playback or to skip video frames at fixed intervals, while different others use video information (color, shape, motion, scene change, luminance, etc.) to extract significant video segments from the original video. Audio and video analysis are sometimes applied to particular videos. For instance, *He* et. al. [6] summarize videos of talks that are accompanied by PowerPoint slides and use the slides change to determine a video segment; *Tjondronegro* et al. [7] focus on sports video summarizations and use the audio features of the video stream to produce highlights. An interesting survey of these techniques is given in [8].

Although proposed for particular types of video, the approaches that use both audio and video analysis show the benefits of using audio information in producing video summaries. If we consider that audio analysis is computationally cheaper than visual analysis, it follows that audio analysis should definitely take part of the video summaries production. Further, by analyzing the audio, it would be possible to select significant part of it so that the produced summaries can be provided with audio features.

The contribution of this paper is the proposal of a mechanism that automatically produces video summaries by performing both audio and video analysis of a given video stream. In contrast to the works discussed above, our paper introduces a more general framework which is not limited to a certain type of video, but can be applied to many different classes of videos. By using both audio and video analysis, our mechanism is designed to produce summaries with completely intelligible audio. In this way the produced summaries are complete audio/video products and can therefore be considered as an alternative to the

original videos.

Video summaries are produced with three different heuristic algorithms: one is oriented to the production of summaries for browsing purpose; another is oriented to the production of summaries that can be considered as an alternative to the original videos and the third algorithm produces summaries that are a trade-off between the summaries produced by the previous algorithms.

The evaluation of our mechanism has been done by considering different types of video (TV-show, TV-News, movies, talk-show, sport, distance learning). Results show that the produced summaries may be reduced up to 50%. The perceived play out quality is also investigated by meaning of jerkiness (i.e., whether a video play out is perceived as a continuous motion or as a sequence of distinct snapshots). In fact, if the produced summaries have a high degree of jerkiness then the summaries may be useless as people tend to avoid such annoying videos. Our investigation show that jerkiness is kept under control by our heuristic algorithms.

Further, since our summaries are complete audio/video products, we analyze the consumer satisfaction by meaning of Mean Opinion Score (MOS). Results show that our summaries can be considered as an alternative to the original videos.

In summary, our mechanism may be useful to those systems that offer video browsing facilities and can also be useful to video content providers. For instance, categories of people like commuters and young consumers can benefit from having different length versions of the same video. Thanks to our mechanism, the same TV-News report or the same TV-show can be watched in a shorter time.

The remainder of this paper is organized as follows. In Section 2 we present details of our proposal; in Section 3 we present the experimental results and the analysis of the perceived quality; in Section 4 we present the MOS results; Conclusions are drawn in Section 5.

2 Audio Video Summarization Scheme

In this section we present details of our proposal, a mechanism to automatically produce audio/video summaries. The mechanism is named AVSS (Audio Video Summarization Scheme) and is based on audio and video analysis. It aims at producing summaries with audio and video features by achieving the following goals:

1. Temporal Reduction: Since a summary is a shorter version of a given video, the play out length must be reduced with respect to the one of the original video;
2. Audio Continuity: Audio is provided with the summary and hence it must be completely intelligible;
3. Video Continuity: The video must be as smooth as possible (video jerkiness may be annoying to users).

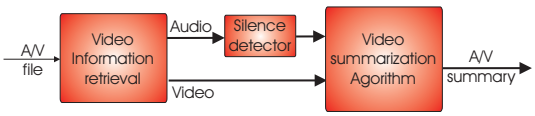

Figure 1: Audio Video Summarization Scheme.

To achieve the above goals, our mechanism is composed of three different steps as depicted in Fig. 1. The idea is to perform audio analysis in order to identify all the silent video portions (parts of the video with no sound) and to act on these in order to produce summaries. The idea is based on the fact that the percentage of silent parts is usually very high and, by acting on these it is possible to meet goals 1 and 2. In fact, by removing the silent portions, the overall play out time is reduced and audio does not suffer of jerkiness. Goal 3 is against goal 1, as by removing several portions of the video, the video jerkiness increases. On the other hand, if we don't remove several video portions then goal 1 is not met. Hence, the task of the video analysis is to find a reasonable trade-off between goal 1 and goal 3. For this reason, our mechanism is provided with three different heuristic algorithms that aim at finding a good trade-off between goal 1 and goal 3.

It is to note that, by acting on the silent video portions, our mechanism can be applied to a large variety of videos (with the exception of music video clip). In fact, as we better show in the experimental section, the percentage of silence is very high not only in speech-based videos (TV-News and Talk-shows), but also in movies and TV-shows.

By meeting the 3 goals above, the summaries produced by our mechanism can be useful to those systems that provide browsing features (the produced summary is shorter that the original video), and, being a complete multimedia product, can also be useful to video providers that want to offer different length versions (hopefully with different fees) of a given video to clients in order to reach a large number of customers.

In the following we present details of the three steps involved in our mechanism.

2.1 Video Information Retrieval

The first step of our mechanism is the retrieval of audio/video information like audio/video encoding and the fps (frames per second). All these information are provided in the header file and hence can be easily retrieved. These information will be used in the audio/video analysis.

2.2 Audio Analysis

The audio stream is analyzed in order to identify the portions of the audio stream that correspond to silence. The analysis is performed by a silence detector algorithm. In its simplest form the silence detection can be a magnitude based

decision: the silence detector algorithm compares the magnitude of the signal against a preset threshold and if a percentage of the data is smaller than the threshold, silence is declared [9]. Although the magnitude based algorithm has fairly mediocre performance in the presence of any background noise, it does not require much complexity. The Robust Audio Tool (RAT) uses a similar approach, where the threshold is automatically adjusted according to audio characteristics [10]. Although more sophisticated approach may be used, in this paper we consider the RAT approach and results were satisfactory.

Before performing a silence analysis, the audio stream is decoded to PCM, in order to apply our silence detector algorithm, and is divided into consecutive audio blocks. Each audio block has a temporal length equal to the one of a video frame (hence the number of audio blocks per second and the number of video frames per second is the same). This is done to have a perfect synchronization between an audio block and a video frame (and is done to facilitate the video analysis).

The silence detection is then applied to each audio block in order to identify all the audio blocks that contain no audio. In this way, a set of silent audio blocks is created and is given to the following step (video analysis).

2.3 Video Analysis

Video analysis is done to select the portions of the video that have to be removed. It uses the set of silent audio blocks to identify the correspondent set of video frames (silent video frames). In particular, the video is considered composed of video segments, each of them is associated either with sound (non silent video segment) or with silence. In this way, the video analysis can select the video frames to drop among the silent video segments (this meets goal 2).

Since the video may be encoded with inter-frames techniques (e.g., MPEG), the discard of a video frame may cause problems in decoding its neighbor frames. For this reason, the video file is temporarily decoded so that each frame can be independently decoded and hence any video frame can be dropped without any problem.

We recall here that the produced summary should meet three goals: temporal length reduction, audio and video continuity. We already mentioned that video continuity moves against temporal reduction and hence we propose three different heuristic algorithms that produce summaries with different trade-off between temporal reduction and video continuity. In particular, we propose the following heuristics:

ALL: it removes all the silent video segments from the original video. The goal here is to highly reduce the temporal length of the video. The drawback is that jerkiness may be high as complete portions of the video are removed.

2x: it removes a video frame every two in a silent video segment. The goal here is to maintain the video as smooth as possible. Temporal reduction is not effective as with the ALL heuristic, but the visual information of a silent video segment is preserved, although, due to the skipping frames, the user will

perceive the play out of the silent video segment at a speed factor of 2x.

3x: it is a variation of the previous one; The goal here is to reduce the temporal length while preserving smoothness. It can be seen as a trade-off between the ALL and 2x heuristics. In this case, the algorithm skips two frames every three, causing the user to perceive the play out of silent video segments at a speed factor of 3x.

Once the video frames (and audio blocks) are removed, the audio and video are re-encoded and the summary is released.

Before presenting the experimental results, it is worth spending some words on the computational cost of our mechanism. In fact, our mechanism performs an audio PCM transformation, a video decoding and a final audio/video encoding. Although the transformation overhead can be handled in real-time by general purpose processors (for instance, over a 512MB laptop with 1.5Ghz Centrino), if several encoding/decoding requests happen at the same time (for instance, when many consumers want the same summary of the same video) some computational problems may arise. For this reason, the summaries production should be done off-line, so that it is done only once and the system is not overloaded with encoding/decoding processes.

3 Experimental Results

In this section we present results obtained from analyzing our mechanism with 25 fps encoded videos (AVI format, DiXV encoded). Various video types are analyzed (TV-news, TV-shows, talk-shows, movies, sport events and cartoons) in order to test our proposal in speech based videos (TV-news and talk shows), high action videos (TV-shows, cartoons and movies) and simple motion videos (TV-shows, movies and sport events). Each video has been tested in its original length.

3.1 Temporal Length

Table 1 summarizes the results obtained while analyzing some TV-shows. For each TV-show and for each proposed heuristic algorithms, it is reported the original length, the summary length and the percentage of reduction. It is not surprisingly to notice that the algorithm ALL produces summaries much shorter than 2x and 3x, and that 3x produces shorter summaries than 2x. In fact, by acting on the same silent video segments, the ALL algorithm entirely removes any silent video segment, while the others simply skip video frames to produce the speed up effect of 2x and 3x. However, it is interesting to note that the 2x algorithm allows reducing the original video of a percentage that ranges between 15% (Friends) and 36% (Smallville). The result is remarkable if we consider that all the audio information have been preserved and that the user does not miss any video scene (although this is an eye illusion, since some parts are actually missing; the user only perceives the play out of some video segments faster and hence he/she has the complete video information).

TV-show	Original	2x	3x	ALL
Lost	36.00	26.38 (26%)	23.31 (35%)	18.09 (50%)
Ally McBeal	39.19	30.20 (23%)	27.28 (31%)	22.31 (43%)
Smallville	37.34	24.08 (36%)	19.41 (48%)	11.51 (69%)
Sex & the city	26.48	20.43 (23%)	18.40 (31%)	15.33 (42%)
Desperate Housewives	40.17	31.31 (22%)	28.36 (29%)	23.55 (41%)
Friends	19.34	16.46 (15%)	15.51 (19%)	14.26 (27%)
L-World	46.38	31.29 (33%)	26.26 (43%)	17.53 (62%)

Table 1: Video Length (in min.) of the produced TV-show summaries.

Table 2 reports the results obtained from analyzing other types of video. Also in this case, the three proposed algorithms highly reduce the temporal length of the original video. However, a surprisingly result is the one related to talk-show. Since they are speech-based, high reduction was expected. However, by watching carefully at these videos, it is possible to note that the percentage of silence is not very high, as people usually talk very fast and try to talk each other down.

Figure 2 summarizes the obtained temporal length reduction (in percentage) for some of the analyzed categories. From these results the ALL algorithm should be preferred, but jerkiness perception has not been analyzed yet.

3.2 Jerkiness Perception

The temporal length reduction is only one goal of our mechanism. The other two goals involve audio and video continuity and hence it is necessary to investigate how the proposed heuristics affect the jerkiness perception (i.e., is the play out a continuous motion or a sequence of distinct snapshots?). However, since audio continuity is not affected as the heuristic algorithms act on silent video segments, here we investigate the video continuity.

We recall that a user perceives a video play out discontinuity when there is a video cut. A cut happens when the transition between the display of a frame and its successive is not smooth. The number of video cuts is established by the director as he/she defines the length of any video shot (the set of frames between two consecutive cuts).

In video summaries, the number of cuts is usually increased as portions of the video are removed. By comparing the number of cuts it is possible to have

Video Type	Original	2x	3x	ALL
TV-News Sky Report	30.27	26.58 (12%)	25.45 (16%)	24.36 (20%)
Talkshow Sky	21.24	18.52 (12%)	17.58 (17%)	17.03 (20%)
Talkshow MTV TrueLine	15.05	13.50 (9%)	13.23 (12%)	13.01 (14%)
Distance Learning Math Lesson	30.00	20.42 (31%)	17.30 (42%)	12.30 (59%)
Sport Soccer	92.04	70.44 (23%)	63.38 (31%)	53.30 (42%)
Cartoon Futurama	19.00	15.56 (16%)	14.55 (22%)	13.25 (30%)
Cartoon The Simpsons	21.50	18.03 (18%)	16.46 (23%)	14.50 (32%)
Runaway Jury Movie	125.58	83.57 (40%)	69.49 (45%)	46.41 (63%)
The 6th Sense Movie	102.58	71.01 (31%)	60.22 (42%)	41.16 (60%)

Table 2: Video Length (in min.) of the produced video summaries.

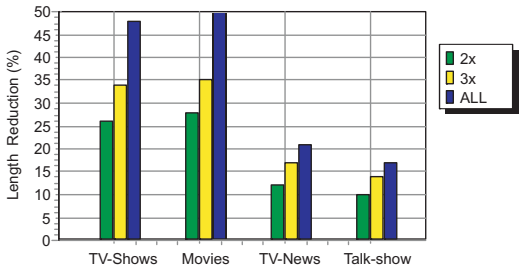

Figure 2: Average length reduction (%) of different categories.

an indicator of the introduced jerkiness. The more cuts are inserted, the more the jerkiness perception is affected. However, since we are comparing videos with different length (and hence with a different number of frame), the number of cuts may be misleading. For this reason we investigate the average length of a video shot. Note that the different video lengths does not allow to investigate the video quality with classic techniques like PSNR that compares each frame of the original video with the correspondent frame of the compressed video.

Cuts can be detected by analyzing and comparing any single frame. Different techniques are possible: histogram changes, edges extraction, chromatic scaling, DCT distribution change. In this paper we consider a variation of the histogram analysis: For each video frame we compute the average value of the three components (YUV) and we combine them considering that human vision is more sensitive to brightness than to color. Hence, the following combination is used:

Definition 1 *A perceptual representation of a video frame i is given by a combination of the luminance (Y) and the two chrominance components (U and V) and is given by:*

$$YUV(i) = 0.5Y(i) + 0.25U(i) + 0.25V(i) \tag{1}$$

The perceptual difference between two consecutive frames is computed as in the following definition.

Definition 2 *The perceptual difference between a frame i and a frame $i-1$ is defined as:*

$$PD(i) = YUV(i) - YUV(i-1) \tag{2}$$

The perceptual difference allows identifying all the cuts and hence can be used to measure the video continuity. Note that a cut happens when the perceptual difference is above a pre-defined threshold (here equal to 10, a number obtained from analyzing several videos).

Definition 3 *The video continuity, VC, of a given video is defined as the number of cuts perceived by the user and is equal to :*

$$VC = \sum_{i=1}^{N} D(i) \tag{3}$$

where $D(i) = 0$ if $PD(i) < 10$ otherwise $D(i) = 1$, and N is the number of video frames that composes the video.

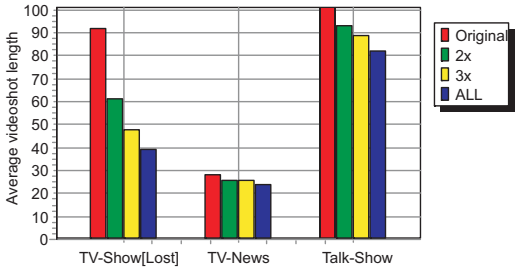

Figure 3: Average length (in number of frame) of a video shot.

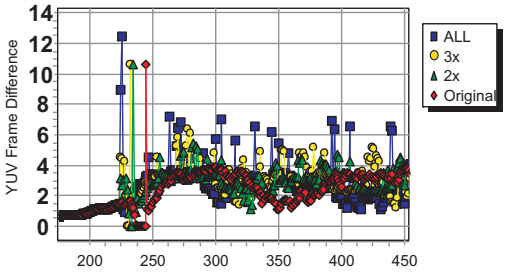

Figure 4: TV-Show: YUV Difference between two consecutive video frames.

The average length of a video shot is used to measure the jerkiness of a video. The more this number is similar to the one of the original video, the more the jerkiness has not been affected by the summarization technique.

Figure 3 shows the average length (in frame) of a single video shot for three different video (TV-Show Lost, TV-News and Talk-show). With respect to the TV-show original video, a video shot has an average length of 92 video frames, while the summaries have values of 60, 48 and 39 for the 2x, 3x and ALL algorithms, respectively. In this case, ALL performs much worse that the other, and 2x should be preferred. For TV-news and Talk-show the difference is not so high and hence the three heuristics perform similarly.

A deeper investigation is presented in Figure 4, which depicts a close-up of the perceptual difference measured for each video frame. It is interesting to note the smoothness of any single curve. While the original video has a smooth curve, the others have a higher variability. Although most of this variability is not perceived as a cut, and hence is not accounted in the previous analysis, the user perceives little jerkiness. This is a further confirmation that the algorithm 2x should be preferred as it has less variability than 3x and ALL. Also in this case, ALL performs worse than 2x and 3x.

Video Type	2x	3x	ALL
Movies	3.2	3.0	2.1
TV-Show	4.0	3.5	2.4
TV-News	4.8	4.4	4.4
Talk-show	4.2	3.8	3.8
Soccer	3.2	3.0	2.7

Table 3: Average MOS for the overall video quality.

4 Consumer Satisfation

The attractiveness of the produced summaries is analyzed with Mean Opinion Score that measures the consumer satisfaction (on a scale of 1 to 5). Using a collection of different videos, a group of 30 people has evaluated the quality of the produced summaries.

Table 3 show results obtained from asking to rate the overall video quality. Results show that TV-News summaries can be considered as an alternative to the original videos. TV-Shows and talk-show obtained very good scores for the 2x summaries and acceptable scores for the 3x summaries. Movies obtain good scores only for 2x summaries.

We also asked whether the summaries can be used for browsing purposes and all the audience agree on that. Hence, the ALL algorithm should be preferred in this case.

5 Conclusions

In this paper we proposed AVSS, a mechanism that uses audio/video analysis to automatically produce audio/video summaries of a given video. The produced summaries have intelligible audio information. Experimental results showed that original videos can be largely reduced by acting on the silent video parts. A video evaluation showed that the three proposed heuristic algorithms can keep the introduced jerkiness under control. Customer satisfaction has been investigated and MOS results showed that many produced summaries can be considered as an alternative to the original video.

We are currently working on a deeper analysis of both video and audio (for instance by introducing a speech detector).

Acknowledgment

This work has been partially supported by the WEBMINDS FIRB project of the Italian MIUR.

References

[1] Fox Italia Mobile Service, [on-line] http://www.foxtv.it

[2] D.DeMenthon, D.S.Doermann and V.Kobla, *Video Summarization by Curve Simplification*, ACM Multimedia. 1998, pp. 211-218.

[3] S. Ju, M. Black, S. Minneman, and D. Kimber. *Summarization of Video-taped Presentations: Automatic Analysis of Motion and Gesture*, IEEE Trans. on Circuits and Systems for Video Technology, Vol. 8(5), 1998, pp. 686-696.

[4] W. Zhou, A. Vellaikal, and C.C. Jay Kuo, *Rule-Based Video Classification System for Basketball Video Indexing*, Proc. ACM Multimedia Workshop, 2000, pp. 213-216.

[5] Y.P. Ma, L. Lu, H.J. Zhang, and M. Li, *A User Attention Model for Video Summarization*, Proc. ACM Multimedia, 2002.

[6] L.He, E.Sanocki, A.Gupta and J.Grudin, *Audio-Summarization of Audio-Video Presentations*, ACM Multimedia, 1999, pp.489-498.

[7] D. Tjondronegro, Y.P.P. Chen, B.Pham, *Sports Video Summarization using Highlights and Play-Breaks*, Proceedings of ACM Multimedia Information Retrieval 2003, pp. 201-208, Berkeley, CA, USA.

[8] M.Mentzelopoulos, A.Psarrou, *Key-Frame Extraction Algorithm using Entropy Difference*, Proceedings of Multimedia Information Retrieval 04, New York, USA, October 2004, pp. 39-45.

[9] Rabiner L.R., Schafer R.W., *Digital processing of Speech Signals*, Prentice Hall, 1978.

[10] M. Roccetti et al.,*Design and Experimental Evaluation of an Adaptive Play-out Delay Control Mechanism for Packetized Audio for Use over the Internet*, Multimedia Tools and Applications, 14(1), 2001, pp.23-53.