MPEG-1 Bitstreams Processing for Audio Content Analysis

Roman Jarina*, Orla Duffner▽, Seán Marlow, Noel O'Connor, and Noel Murphy

Visual Media Processing Group
Dublin City University
Glasnevin, Dublin 9, Ireland
E-mail: *jarinar@eeng.dcu.ie ▽duffnero@eeng.dcu.ie

Abstract -- **In this paper, we present the MPEG-1 Audio bitstreams processing work which our research group is involved in. This work is primarily based on the processing of the encoded bitstream, and the extraction of useful audio features for the purposes of analysis and browsing. In order to prepare for the discussion of these features, the MPEG-1 audio bitstream format is first described. The Application Interface Protocol (API) which we have been developing in C++ is then introduced, before completing the paper with a discussion on audio feature extraction.**

Keywords – **audio, MPEG-1, scalefactors, speech, music**

I INTRODUCTION

Audio content classification is a very important task for the browsing, indexing and retrieval of audio/video databases. It includes speech-music discrimination, speaker segmentation, automatic keyword spotting etc. The development of standards for high-quality audio and video compression such as the MPEG family [1], coupled with the increased performance of computing enables easy recording, storage and manipulation of multimedia content.

Unlike MPEG video analysis, little work has been done in the study of the compressed MPEG audio bitstream. In [2] [3], an audio classification of the MPEG-1 compressed domain has been proposed. Both used short–term energy based features, which were computed from subband encoded audio samples. Another MPEG audio bitstream manipulation, which was published recently, includes audio bit rate scaling [4] and audio signal mixing for Internet applications [5].

This paper presents the MPEG-1 audio bitstreams processing work our research group is involved in. Audio features that can be extracted directly from the encoded bitstreams are reviewed. Dealing with the encoded bitstream has the following advantages:

- It can deal with long audio and video sequences (several hours of recordings).
- It has much smaller storage and computational requirements than the uncompressed signal processing, because the computationally difficult decoding process is not required.
- The audio signal analysis carried out during the encoding process can be utilized (e.g. subband filtering, volume estimation).

The work described in this paper is under the auspices of the Centre for Digital Video Processing (CDVP), in cooperation with the Research Institute for Networks and Communications Engineering (RINCE) at Dublin City University. The CDVP research work concentrates on digital video analysis and multimedia content management. The current results are demonstrated on the web-based digital video system called Físchlár [6],[7].

II MPEG-1 AUDIO BITSTREAM DESCRIPTION

The Moving Pictures Expert Group completed its first standard for video and audio, called MPEG-1, in November 1992. The MPEG-1 audio coding system, specified in ISO/IEC 11172-3 operates in single-channel or two-channel stereo modes at sampling frequencies of 32, 41.1, and 48 kHz, and is well documented [1],[8].

The MPEG-1 Audio uses a high performance perceptual coding scheme to compress the audio signal. This means it achieves higher compression by removing acoustically irrelevant parts of the audio signal (those not perceived by the human ear). The standard defines three layers with increasing

* also with the University of Žilina, Slovakia

computational complexity and better compression and performance in the higher layers.
The *Físchlár* system, which we have been developing as a partner with CDVP, stores video and associated audio in the MPEG-1 format. For audio, it uses the Layer-II compression scheme (MP2) with a sampling rate of 44.1 kHz. The following subsection gives a brief description of the Layer II encoded bitstream.

a) MPEG-1 Layer II

The MPEG-1 Layer-II compression scheme provides very high quality at a data rate of 128 kbit/s per channel [8]. A block scheme of the encoder is depicted in Figure 1. The input signal (PCM samples) is first decomposed into 32 critically sub-sampled subbands using a polyphase realization of a psuedo QMF filterbank. These 511th-order filters are equally spaced such that a 44.1 kHz input signal is split into 690 Hz subbands, with the subbands decimated at 32:1. For the purposes of psychoacoustic analysis and determination of masking thresholds, 1024 point FFT is computed in parallel with the subband decomposition for each decimated block of 12 input samples. The time length of each block is $\tau = 32 \times 12 / f_s$, which is 8.7 ms at the sampling rate f_s = 44.1 kHz.

Next, the subbands are normalized by a scalefactor such that the maximum sample amplitude in each block is unity, then an iterative bit allocation procedure applies the masking thresholds to select an optimal quantizer from a predetermined set for each subband. Quantizers are selected such that both the masking and bit rate requirements are simultaneously satisfied. In each subband, scalefactors are quantized using 6 bits and quantizer selections are encoded using 4 bits. In Layer II, scalefactor information is further reduced by considering the properties of three adjacent 12-sample blocks and transmitting one, two, or three scalefactors.

The encoded bitstream consists of the quantized signal samples, bit allocation information, scalefactors, supplemented with a header, CRC code and ancillary data. A layer-II frame consists of 1152 samples: 3 groups of 12 samples from each of 32 subbands. The frames structure is shown in Figure 2.

There is no main header in an MP2 audio file. The frames are totally independent, each of them with its own header and audio information. So any part of the MP2 file could be cut and played correctly, and the decoder would start from the first complete frame found. A structure of the MP2 frame header is shown in Figure 3. A more detailed description is in [1] [9].

III DEVELOPMENT OF AN API FOR MPEG-1 LAYER II AUDIO BITSTREAMS

The MPEG-1 audio standard (mainly Layers II, III) is very popular for the storage and transmission of high quality audio signals. Many software and hardware implementations have been developed during recent years. However, when manipulating such compressed bitstreams, several tools or functionalities are still not available, or have only been partially studied and implemented. For our purposes (i.e. audio analysis and browsing), it is desirable to have direct access to particular data in the encoded bitstream without going through the full decoding process.

Within our research group, we have been developing an API (Application Interface Protocol) for MPEG-1 Layer II Bitstreams. It is a library of C++ functions. A summary of the functions is in Table 1. Layer II supports several compression schemes for different sampling frequencies and bit rates for one or two channel audio. Due to the high complexity and variability of the bitstream, this is a challenging programming task. At the moment, the API is able to process bitstreams for mono (one channel) recordings only, and it was tested on short files (several seconds long) only. It can be compiled under either Linux/Unix or Windows. After completion, we intend to implement this API in the applications described in the following section.

IV AUDIO FEATURE EXTRACTION

a) Scalefactor Derived Features

By definition, the scalefactors in the MPEG-1 encoded bitstream carry information about the maximum level of the signal in each subband within the group of 12 samples. Therefore the scalefactors can be used to estimate the volume (or short-term energy) or a coarse spectrum of the audio signal. So the scalefactors carry very useful information about the encoded signal. The time resolution of the curve obtained from sequence of the scalefactors is 8.7 ms (for fs = 44.1 kHz). Advantages of using scalefactors instead of coded samples are as follow:

- It is much easier to find the position of the scalefactors in the MPEG frame and decode them than to find and decode audio samples.
- The scalefactors are a very small part of the MPEG bitstream.
- Subsequently, detecting and processing the scalefactors is very straightforward and fast.

Applications where the scalefactors can be utilized are introduced below.

Table 1. API functions for MPEG-1 L-II bitstreams

Function	Description
`find_sync`	Give the position of the 12 bit sequence '11111111111' that defines the beginning of a frame (Figure 3).
`read_header`	Read data from the header starting from 'Frame sync' (Figure 3).
`read_alloc`	Display which of the subbands have bits allocated and the number of bits allocated per sample/ group of samples. Determines the size and structure of the audio data field in the frame, and also the size of the SCFSI field.
`read_scfsi`	Read 2-bits information for each of the subbands used. It determines whether 1, 2 or 3 scalefactors are used.
`read_sclf`	Read data from the scalefactor field. Correct values of the scalefactors are taken from look-up tables [1]
`read_sam`	Read audio data for each of subbands used, with 36 samples per subband. (First, all the function above have to be used)

b) Volume Contour

The volume contour (or amplitude envelope) of the signal can be estimated by summing relevant scalefactors over all subbands. We successfully applied audio features derived from the volume contour for silence detection and speech-music discrimination. In [10] [11], the scalefactors were used for silence detection together with black video frame detection for commercial breaks determination in broadcast television programmes. In [12], a speech–music discriminator was introduced. The proposed method is based on the observation of the modulation envelope of the band-limited signal. From the envelope, high-volume peaks are extracted. The width of the widest peak and average rate of peaks within a time interval of 4 seconds are chosen as features for the discriminator.

c) Subband Spectrogram and Spectral Features

Figure 4 compares spectrograms computed from FFT coefficients of the original signal and scalefactors (SCF-spectrum) from the encoded bitstream. Advantages of the SCF spectrum include fast computation and straightforward manipulation of the audio files. An MPEG compressed file is about 6 times smaller than original uncompressed audio file (PCM samples). The time resolution of a SCF spectrum is also sufficient for a majority of the audio signals.

A drawback of this approach is low resolution in the low frequency subband. In particular, the resolution in the first and second subbands is unsatisfactory because of the relatively large width of these subbands filters. One approach to increase this resolution is magnifying the spectrum at this area by computing the spectral coefficients (DCT, FFT) from the audio samples only for those particular subbands (similarly to Layer III compression scheme). Using this approach, frequency resolution increases by a factor of six for the same time resolution (for a group of twelve samples, six spectral coefficients are computed).

From such a spectrum, other spectral features can be derived such as band energy ratio, average energy in the subbands, spectrum centroid, spread etc. These features are useful to determine the characteristics and genre of audio, and to detect audio events that are relevant for audio segmentation (e.g. applause, car crash, sport programme highlights etc).

d) Voice Pitch Detection

The pitch (or fundamental frequency) of a voice is an important feature in speech processing, low-rate coding and synthesis. However, pitch information can be useful for audio characterization and segmentation (coarse speaker segmentation, gender discrimination, voice-unvoice discrimination, prosody analysis etc.)

An explanation of speech production theory and pitch determination is outside the scope of this paper. We note only that pitch information can be directly determined from MPEG encoded subband samples. Although pitch determination techniques vary in the way they perform their analysis, a majority of these methods require low pass filtering of the signal. The reason is that the pitch frequency of the human voice is below 900 Hz. The common range usually falls between 40-250 Hz for male voices and 120-500 Hz for female voices. In our case, the bandwidth of the first subband (690Hz @ $f_s = 44.1$ kHz) might be sufficient for a majority of pitch analyses.

e) Rhythm detection

Rhythm detection is very beneficial to the segregation of music signals from speech and other environmental sounds. We have successfully applied rhythm detection to speech-music discrimination [13], where our approach is based on long-term autocorrelation analysis. Only scalefactors are required for processing. After implementation of the rhythm feature into the previous model [12], the performance of the discriminator was found to increase significantly for certain types of audio signals.

V CONCLUSION

In this paper, we presented the MPEG-1 Audio bitstreams processing work which is based on the processing of the encoded bitstream. Future work includes the evaluation of the spectral features described in section 4.3 in terms of their usefulness for audio events discrimination, localization and voice pitch analysis using MPEG encoded samples. One such tool being developed will present a visual representation of the scalefactors over time, frequency and magnitude in order to evaluate their information content. This is being developed as a real-time C++ application under Linux and will be useful to provide a fast display of the spectral content of the MPEG-1 audio file.

REFERENCES

[1] ISO/IEC 11172-3, Coding of Moving Pictures and Associated Audio for Digital Storage Media a up to about 1.5 Mbit/s, Part 3: Audio, 1992.

[2] N. Patel and I. Sethi, "Audio Characterization for Video Indexing", *Proc. SPIE in Storage and Retrieval for Still Image and Video Databases*, Vol.2670: 373-384, San Jose, 1996.

[3] Y. Nakajima, et al. "A Fast Audio Classification from MPEG Coded Data", *Proc. ICASSP'99*, Phoenix, Arizona, May 1999.

[4] Y. Nakajima, H. Yanagihara, A. Yoneyama, and M.Sugano, "MPEG Audio Bit Rate Scaling on Coded Data Domain", *Proc. ICASSP'98,* Seattle, Washington, May 1998.

[5] P. De Smet, T.V. Stichele. "Subband Based MPEG Audio Mixing for Internet Streaming Applications", *Proc. ICASSP'01*, Salt Like City, Utah, May 2001.

[6] N. O'Connor, et al., "Físchlár: An On-line System for Indexing and Browsing of Broadcast Television Content", *Proc. ICASSP'01,* Salt Lake City, UT, pp. 418-421, May 2001.

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

[8] K. Brandenburg and G. Stoll, The ISO/MPEG-1 Audio: A Generic Standard for Coding of High-Quality Digital Audio, *J. Audio Eng. Soc*., Vol.42, No.10: 780-792, Oct. 1994.

[9] http://www.mp3-tech.org/programmer/frame_header.html.

[10] D. Sadlier, S. Marlow, N. O'Connor, and N. Murphy, "Automatic TV Advertisement Detection from MPEG Bitstream", *Proc. Int. Conf. on Enterprise Information Systems ICEIS 2001*, Setubal, Portugal, 7-10 July 2001.

[11] S. Marlow S, et al, "Audio and Video Processing for Automatic TV Advertisement Detection", *Proc ISSC 2001,* Maynooth, pp. 325-330, June 2001.

[12] R. Jarina, N. Murphy, N. O'Connor, and S. Marlow, "Speech-Music Discrimination from MPEG-1 Bitstream", In V.V. Kluev, N.E. Mastorakis (Ed.), *Advances in Signal Processing, Robotics and Communications*, WSES Press, pp. 174-178, 2001.

[13] R. Jarina, N. O'Connor, S. Marlow, and N. Murphy, "Rhythm Detection for Speech-Music discrimination in MPEG Compressed Domain", Accepted for *IEEE Int. Conf. DSP'02*, Santorini, Greece, July 2002.

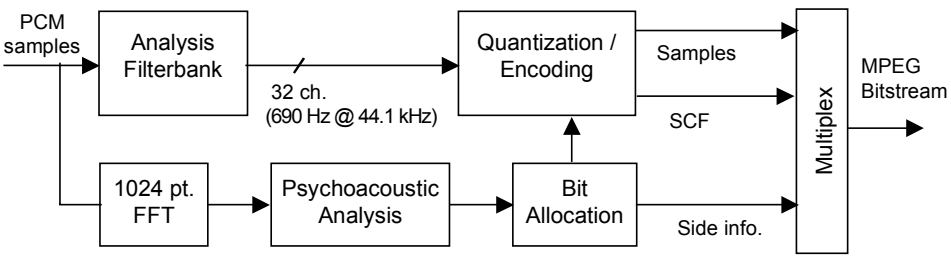

Figure 1. Block scheme of MPEG-1 Audio Layer II encoder

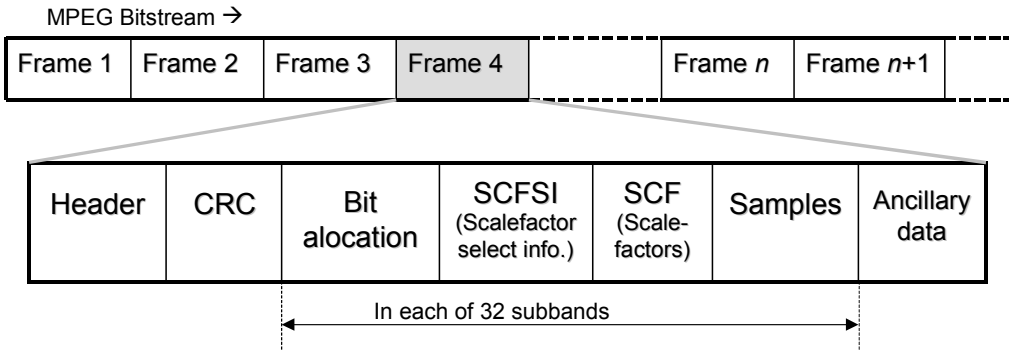

Figure 2. Structure of MPEG-1 Audio Layer II bitstream

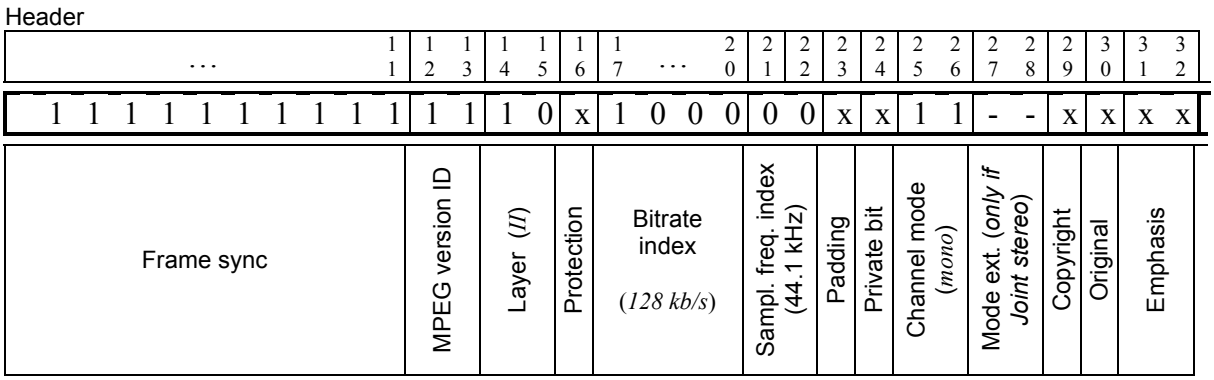

Figure 3. An example of the MPEG-1 L-II header (single (mono) channel, 128 kb/s, 44.1 kHz). 'x' means binary value depends on particular case, '-' means any binary value

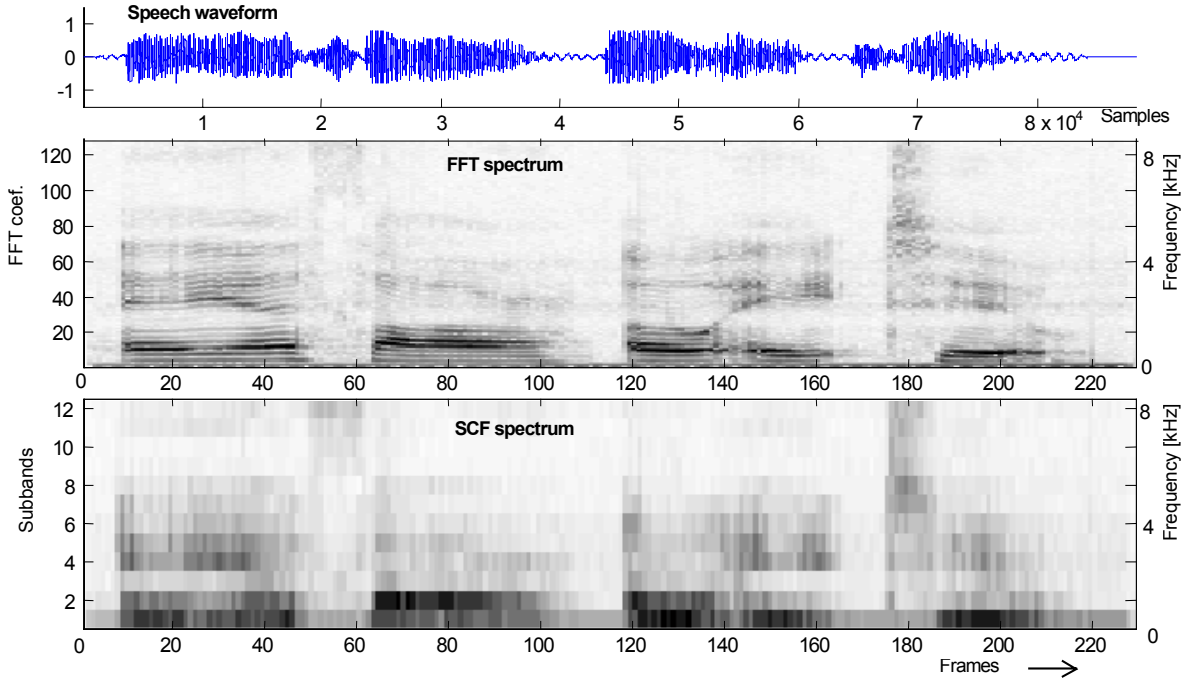

Figure 4. FFT spectrum and SCF spectrum (i.e. derived from scalefactors) of speech utterance.
(Female voice, spelling out "ASR082". fs = 44.1 samples/sec., 256-point FFT,
frame length is 8.7 ms, the first 12 subbands is displayed – up to 8.3 kHz)