Multimedia/Audio Coding Formats

Noor Afshan Fathima
Graduate Student: Dept. of Computer Science (India Online)
Illinois Institute of Technology, Chicago

***Abstract*—A research study summarization of techniques/algorithms of 3 different types of Multimedia/Audio coding formats. To successfully understand content representation format for storage or transmission of digital media/audio over the network.**

Keywords— container, coding formats, MP4, OGG Vorbis, Opus.

I. Introduction

The data produced/captured by humans can be stored in the form of audio, image, video etc. This data has to be efficiently stored and at need transmitted across the network reaching the masses who use variety of devices. Therefore technologies are built to seamlessly represent this data on every kind of device available.

The majority of data(media) is compressed, which means it has been altered to take up less space on any device being used. A Codec is used to compress and decompress this data interpreting the media file and determining how to play it on the device. A Container is a bundle of media files. It can contain a video, an audio codec and also subtitles.

II. OVERVIEW OF TYPES.

In this study we are looking at three different types of multimedia/audio coding formats.

A. *H.264 or MPEG-4 Part 10, Advanced Video Coding (MPEG-4 AVC)*

The MPEG-4 format is used to share media on the web. Video, audio, subtitles and still images tracks are compressed separately. It supports wide variety of codecs. It uses the following techniques.

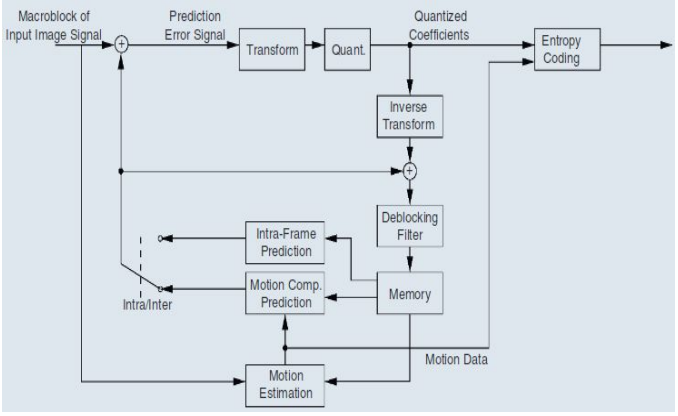

fig 1. MP4 Block Diagram

1. *Intra Prediction:*

Intra prediction means that the samples of a macroblock are predicted by using only information of already transmitted macroblocks of the same image. In H.264/AVC, two different types of intra prediction are possible for the prediction of the luminance component Y. The first type is called INTRA_4×4 and the second one INTRA_16×16.

2. *Transform, Scaling, and Quantization:*

H.264/AVC uses spatial transform coding of the prediction residual. The transformation is applied to 4X4 blocks, and instead of providing a theoretical inverse DCT formula to be approximated by each implementer within specified tolerances, a separable integer transform with similar properties to a 4X4 DCT is used.

A quantization parameter (QP) is used for determining the quantization of transform coefficients in H.264/AVC. It can take on 52 values. The quantization step size is controlled logarithmically by QP . Each increase of six in QP causes a doubling of the quantization step size, so each increase of one in QP increases the step size by approximately 12%.

The quantized transform coefficients of a block generally are scanned in a zigzag fashion and transmitted using entropy coding methods. The 2X2 DC coefficients of the chroma component are scanned in raster-scan order.

3. *Entropy Coding:*

In H.264/AVC, two methods for entropy coding are supported. These are called context-adaptive variable-length coding (CAVLC) and context-adaptive binary arithmetic coding (CABAC). CABAC has higher complexity than CAVLC, but has better coding efficiency.

4. *Inverse Transform:*

All inverse transform operations in H.264/AVC can be implemented using only additions and bit-shifting operations of 16-bit integer values.

5. *De-blocking filter:*

One particular characteristic of block-based coding is the accidental production of visible block structures. Block edges are typically reconstructed with less accuracy than interior pixels and “blocking” is generally considered to be one of the most visible artifacts with the present compression methods. For this reason, H.264/AVC defines an adaptive

in-loop deblocking filter, where the strength of filtering is controlled by the values of several syntax elements. H.264/MPEG-4 AVC deblocking is adaptive on three levels: On slice level, On block edge level, On sample level.

Error Robustness and Network Friendliness:
For efficient transmission in different environments, the seamless and easy integration of the coded video into all current and future protocol and network architectures is important. Therefore, both the VCL and the NAL are part of the H.264/AVC standard. The VCL specifies an efficient representation for the coded video signal. The NAL defines the interface between the video codec itself and the outside world.

B. Vorbis :

Vorbis is a free and open-source software project. It produces an audio coding format and a codec for lossy audio compression. It is used in conjunction with Ogg container format.

The Vorbis bit stream specification consists of four packet types, which occur consecutively. The first three are headers. The header size is unlimited.
The identification header: Identifies the bit stream as Vorbis and gives the version in use. It includes audio characteristics required for further interpretation such as sampling rate and channel number.
The comment header: Includes tags consisting of user comments and a vendor string. Tags themselves may be user-defined.
The codec setup header: Setup components include 'modes', 'mappings', 'floors', 'residues' and 'codebooks', all of which have specific roles in the decoding process.

I. Encoding Techniques:

In general, encoding involves separating the input audio into frames which are compressed to form packets. An overlapping transform called the Modified Discrete Cosine Transform (MDCT), which is a type of discrete Fourier transform, is used to convert time domain data to the frequency domain for further processing.
The MDCT involves time-domain aliasing cancellation (TDAC) which cancels errors resulting from the IMDCT by overlapping (using a %50 overlap) and adding windows. The decoder synthesizes audio frames from the packets and reassembles them to approximate the original audio stream.

II. Decoding Techniques:

Decode packet type flag: First the decoder must verify that a given packet contains audio data by inspecting its type flag.
Decode mode number: The 'mode number' indicates the current frame size, window type, transform type and mapping number.
The frame size is a power of 2 between 64 and 8192, and can be either 'short' or 'long'. Short windows are used near attack transients in order to limit artifacts associated with the MDCT. The window taper varies for long windows depending on whether the previous and subsequent frames are short or long.
The transform type is "always type 0, the MDCT, in Vorbis I".
The mapping number contains a description of the channel coupling scheme and a list of 'sub-maps' which bundle sets of channel vectors.
Decode window shape: The window shape for a given long frame is decoded.
Decode the floor: The floor "vector is a low-resolution representation of the audio spectrum for the given channel in the current frame, generally used akin to a whitening filter." During encoding this data is extracted from the log spectrum of the audio stream using either floor configuration type 0 or 1.
Type 0: "uses Line Spectral Pair (LSP, also alternately known as Line Spectral Frequency or LSF) representation to encode a smooth spectral envelope curve as the frequency response of the LSP filter. This representation is equivalent to a traditional all-pole infinite impulse response filter as would be used in linear predictive coding; LSP representation may be converted to LPC representation and vice-versa."
Inverse channel coupling of residue vectors: The bit rate is lowered during encoding by eliminating redundancies between channels.
Two mechanisms exist for channel coupling:

1. Channel interleaving via residue backend type 2
2. Cartesian to square polar mapping.

The inverse process is performed during decoding. For encoder quality settings equal or greater than six, channel coupling is lossless.

C. Opus:

Opus is designed to handle a wide range of interactive audio applications, including Voice over IP, videoconferencing, in-game chat, and even live, distributed music performances. It scales from low bitrate narrowband speech at 6 kbit/s to very high quality stereo music at 510 kbit/s. Opus uses both Linear Prediction (LP) and the Modified Discrete Cosine Transform (MDCT) to achieve good compression of both speech and music.

Encoding Techniques:

The Opus encoder operates on frames of either 10 or 20 ms, which are divided into 5 ms subframes.

VAD:

The Voice Activity Detector (VAD) generates a measure of speech activity by combining the signal-to noise ratios (SNRs) from 4 separate frequency bands. In each band the background noise level is estimated by smoothing the inverse energy over time frames. Multiplying this smoothed inverse energy with the subband energy gives the SNR.

HP Filter:

A high-pass (HP) filter with a variable cutoff frequency between 60 and 100 Hz removes low frequency background and breathing noise. The cutoff frequency depends on the SNR in the lowest frequency band of the VAD, and on the smoothed pitch frequencies found in the pitch analysis, so that high pitched voices will have a higher cutoff frequency.

Pitch Analysis:

The pitch analysis begins by prewhitening the input signal, The whitening makes the pitch analysis equally sensitive to all parts of the audio spectrum. The whitened signal is then downsampled in two steps to 8 and 4 kHz, to reduce the complexity of computing correlations. The candidate's correlation value is then compared to a threshold.

Noise Shaping:

The noise shaping compares the input and output speech signals and feeds to the input of the quantizer. Opus uses warped noise shaping filters at higher complexity settings as the frequency-dependent resolution of these filters better matches human hearing. Separating the noise shaping from the linear prediction also lets us select prediction coefficients that minimize the bit rate without regard for perceptual considerations.

Pulse Coding:

The integer-valued excitation signal which is the output from the NSQ is entropy coded in blocks of 16 samples.

LSF Quantization:

The LSF quantizer consists of a VQ stage with 32 codebook vectors followed by a scalar quantization stage with inter-LSF prediction.

Entropy Coding:

The quantized parameters and the excitation signal are all entropy coded using range coding.

Stereo Prediction:

In Stereo mode, Opus uses predictive stereo encoding where it first encodes a mid channel as the average of the left and right speech signals. Next it computes the side channel as the difference between left and right, and both mid and side channels are split into low- and high-frequency bands. Each side channel band is then predicted from the corresponding mid band using a scalar predictor. The prediction-residual bands are combined to form the side residual signal S, which is coded independently from the mid channel M.

References

[1] https://en.wikipedia.org/wiki/H.264/MPEG-4_AVC

[2] ftp://ftp.tnt.uni-hannover.de/pub/papers/2004/CASM_2004_MN.pdf

[3] https://tools.ietf.org/html/rfc6716#page-132

[4] http://opus-codec.org/

[5] https://xiph.org/vorbis/doc/Vorbis_I_spec.html