Multimedia Technologies Summative Assignment

hzwr87

1 Describe the main difference between the Human "Earscape" model (Ref: Lecture note 2) and the Audio Noise Masking model (Ref: MP3). Explain how these two models are applied to optimize signal quantization in the audio digitalization process. [20 Marks]

The Human Earscape model defines what kinds of information that the human ear is capable of perceiving, and what level of sound pressure is required to enable us to perceive specific frequencies. The range of the earscape model is bounded at the low sound pressure end of the spectrum by the Miniumum Audability curve. This tells us that low frequency sounds can only be heard at higher sound pressure levels, whereas higher frequency sounds can be heard at comparatively lower sound pressure levels. The upper bound of this model is the Terminal Threshold, which is the sound pressure level as which the sound is not percieved to be any louder, and can cause pain and discomfort to the listener. This earscape model is defined for the average human listener and is used to allow compression algorithms to vary across different frequency levels. This model can tell us the appropriate frequency range for different types of content, and quantisation algorithms can therefore be optimised to remove information outside of these frequency ranges. For example, a voice recording file could have all frequencies outside of the typical human vocal range removed in order to simplify the audio digitialisation process and eventually reduce the file size.

The audio masking model in the MP3 standard takes advantage of a perceptual weakness of human hearing that occurs when a strong audio signal of a certain frequency obscures, or 'masks', weaker audio signals of similar frequencies. To take advantage of this, we can partition the audible spectrum into bands based on how capable the ear is of perceiving different frequencies. The purpose of this partitioning is to understand the size of the neighbourhood of frequencies that will be 'masked' by a strong audio signal. This can be taken advantage of in the signal quantisation process by first transforming the audio signal into the frequency domain, and breaking this down into subbands that represent the neighbourhoods in which masking will occur enables the MP3 algorithm to set the quantisation level to the maximum possible value such that the level of quantisation noise will be imperceptible to the listener. Masking can also occur temporally, as experiments have shown that it takes some time for us to be able to perceive a weak signal after hearing a strong signal at a similar frequency. The MP3 algorithm uses this to further quantise the signal, by determining whether a sound will be heard or not following a strong sound based on its frequency, sound pressure and the time between the two sounds.

2 MPEG/Audio Layer I encoding divides an input audio signal into 32 sub-bands forming the inputs for audio digitalization. Justify whether this procedure improves or degrades audio signal quantization. (Ref: MP3) [20 Marks]

This division is performed in MPEG layer 1 by a basic filter bank. It is intended to allow the encoder to vary the quantisation level across the spectrum of frequencies. This enables the algorithm to exploit the limited ability of the ear to resolve sounds within a band to hide quanitisation noise. The 32 subbands are supposed to represent the critical bands of human hearing. These subbands are of a constant width, whereas the critical bands of the ear are

not of constant width but instead vary as different frequency levels. Therefore, these 32 bands do not accurately reflect the ears critical bands. The bandwidth is too wide at the lower frequencies, which prevents the specific tuning of quantiser bits to the actual critical bands. Furthermore, this filter bank and its inverse are not lossless transformations, as even if no quantisation were to take place the inverse transform would not perfectly recover the original input signal. This means that this process adds noise to the input signal. Also, adjacent filter bands have a considerable frequency overlap, meaning that a signal at a single frequency can affect two adjacent filter bank outputs. From these 32 bands, the layer 1 algorithm groups together 12 samples from each of the bands, and allocates them a number of bits to be used for representation, given by a scale factor which attempts to maximise the resolution of the quantiser. These 32 groups of 12 samples are grouped into a frame. The masking that the layer 1 algorithm takes advantage of is entirely frequency based, with no consideration of temporal masking as discussed above. This is dealt with by the later layers of the algorithm.

The ability to vary the quantisation levels at different frequency ranges does improve the audio digitalisation process compared to applying the same quantisation level for the entire audio signal, which would lead to either a poor sound quality output or a large file size. However, due to the weaknesses of this first layer that are outlined above, the way that this is implemented means that the layer 1 algorithm doesn't get very close to approximating the true nature of human hearing. The subsequent layers of the MPEG algorithm aim to build on this to further improve the audio signal quantisation.

3 In your own words, explain the purpose of involving Discrete Cosine Transform (DCT) in MPEG/Audio Layer 3. Justify how this procedure improves the quality of audio digitalization in terms of both the audio signal preserved and file size produced. (Ref: MP3) [20 Marks]

The layer 3 algorithm is a far more refined and complex approach to audio digitalisation than the layer 1 and layer 2 algorithm. It attempts to resolve the filter bank deficiencies from layers 1 and 2 by making the filters have non-equal frequencies and by applying a Modified Discrete Cosine Transform. This MDCT subdivides the filter bank outputs further by expressing the audio signal as a list of coefficients of cosine waves. These coefficients can then be stored in a diagonal pattern and compressed. The use of the MDCT minimises the issues caused by the Discrete Fourier Transform of the previous layers, particularly around block edge effects, when data is quantised and then transformed back, the beginning and ending samples may not be coordinated with preceding and subsequent blocks, causing a certain level of noise. The MDCT removes such effects by overlapping frames by 50%. The transformation to the frequency domain also gives the masking and quantisation components more sizes of windows limiting the amount of output distortion that quantisation causes. Many of the coefficients in the frequency domain will be duplicates, particularly close to 0, enabling the signal to be easily compressed by techniques like run-length encoding and Huffman encoding. As this process is lossless, it enables a large reduction in the filesize while preserving the audio signal quality.

4 Justify the suitability of replacing the DCT function in MPEG/Audio Layer 3 with Wavelet transform, explaining any requirement in choosing Wavelet transform functions and what changes have to make in MPEG/Audio Layer 3 components to allow such a replacement. (Ref: Lecture note and MP3) [20 Marks]

A substantial difference between the wavelet transform and DCT is that the wavelet transform includes time resolution as well as frequency resolution. This allows wavelets to achieve a better representation of time within the audio file than a DCT is able to, as signals can be encoded relative to an adaptive time-frequency window. The advantages of this are clear, as it enables us to cut the singal of interest into parts that are temporally correlated

and analyse these separately. Layer 3 of the MPEG algorithm tries to use blocking in its MDCT to compensation for this weakness, but if wavelets were to be used instead, this would not need to be done.

Performing a wavelet transform is more computationally complex than performing a DCT, however it has been shown that performing a wavelet transform on an audio signal is of equivalent complexity to iteratively applying a standard filterbank as discussed in previous questions.

Choosing a wavelet to be used for a multi-purpose audio encoding algorithm such as MPEG requires it to be able to represent all forms of audio effectively, as classical music has very different characteristics to speech, and both are different to pop music. If it cannot do this, then encoding certain types of audio will work very well, whereas others will not perform very well. Therefore, the choice of wavelet is crucial. Haar wavelets and Daubechies wavelets are the most famous and common, but alternatives also exist. [1]

In terms of the components of layer 3, if a wavelet transform were to be implemented then the MDCT would be replaced with a wavelet transform, most likely also using a wavelet equivalent of the existing filter bank. Such wavelet filter banks provide perfect reconstruction. They are structured in trees, and output the results of both low pass and high pass filtering. The windowing component of the MDCT that is used to provide some level of time resolution is also not required if a wavelet transform was used, so can be removed.

5 In your own words, explain four differences between applying DCT and Wavelet transform to compress an image, and justify their difference in supporting image decompression and transmission. (Ref: Lecture note) [20 Marks]

The wavelet transform builds up a hierarchical representation of an image, making it well suited to use in JPEG. This allows more efficient network transfer, as a coarse representation of the image can be sent first, followed by a more detailed representation. If DCT were to be used, the image would need to be stored as a lower quality representation, as well as the image differences between several higher quality representations to reconstruct the image at different resolutions as required. Wavelets have the ability to perform multiresolution analysis allowing the support of the hierarchical and progressive modes of the JPEG algorithm without the need for any additional storage. This means that the wavelet transforms do not suffer as badly to transmission errors as if one of the layers is lost, the image can still be displayed at slightly lower resolution.

When DCT is used, each image block is transformed and quantised individually. This removes correlations between blocks, and limits its effectiveness in representing non-stationary objects. Wavelet based coding methods are much more capable of dealing with non-stationary objects thanks to the adaptive time-frequency window that they can make use of due to the fact that the wavelet transform includes a time resolution in addition to a frequency resolution.

The output of the wavelet transform correlates better to the way humans see the world, due to the fact that the wavelet basis functions match the Human Visual System (HVS) characteristics, leading to a superior image representation in the view of a human observer. They are also much more capable of avoiding so called “blocking artefacts” and mosquito noise that cause problems for DCT encoding, as a wavelet based encoding scheme does not need to convert the image into blocks and the lengths of its basis functions are variable. [4]

The main difference between the DCT and wavelet transform coefficients lies in the highpass bands. The highpass DCT bands provide higher frequency resolution, but lower spatial resolution. As a consequence of this, it provides a greater number of frequency bands, but the spatial information is more diffcult to recognise. By comparison, the wavelet subbands provide higher spatial resolution, and lower frequency resolution. As a result, the number of subbands is few, but the spatial resolution is superior.[2]

6 Level 4 students - Conducting a research in scalable video coding (SVC) in the H.264/AVC standard, write up your finding about how SVC is technically different from the original H.264/AVC standard, in terms of architecture, coding mechanism, functionalities, video representation and transmission. (Ref: Lecture note and SVC) [50 Marks]

SVC is an extension of the H.264/AVC that allows for partial transmission and decoding of a bit stream. The video that results has lower temporal or spatial resolution or reduced fidelity while retaining a reconstruction quality that is close to the original H.264/AVC. SVC provides network friendly scalability at bit-stream level with a moderate increase in decoder complexity compared to the original standard. This means that it is more suitable for video conferencing as well as for mobile to high definition broadcast, as well as professional editing.

The SVC design includes spatial and temporal scalability, which describe cases in which subsets of the bit stream represent the source content with a reduced picture size and frame rate. The design also includes fidelity scalability, enabling the substream to provide the same spatio-temporal resolution as the complete bit stream but with a lower fidelity. One SVC bit stream can therefore provide a wide variety of combinations of these basic scalability types. The fact that SVC supports spatial scalability means that the ratio of picture sizes in the complete bit stream and included substreams are not restricted to a particular value. Furthermore, these pictures of bit stream subsets may contain additional parts beyond the borders of the pictures of the complete bit stream, or may only represent a subarea of the complete bit stream. This relationship can be modified at any point. The fidelity scalability implicit in SVC also allows a low-complexity rewriting of a fidelity scalable bit stream into a singlelayer bit stream of the original standard which will give an identical output. Any SVC bit stream contains a subset bit stream that is compatible with a non scalable profile of the original standard and can be decoded by legacy decoders.

The architecture of SVC is layer based, much like the original H.264/AVC standard, being made up of the video coding layer (VCL) and the network abstraction layer (NAL). The VCL represents the coded source content, whereas the NAL formats the VCL representation and provides appropriate header information for use by transport layer protocols or storage media. the NAL contains both VCL NAL units, which contain coded video data, and non-VCL NAL units, which contain additional information. A set of NAL units that result in one decoded picture is known as an access unit, and several successive access units constitute a coded video sequence. The scalability that has been previously discussed is provided at the bit stream level. Discarding NAL units from a scalable stream produces a stream with reduced spatio-temporal resolution or fidelity. The NAL units that are absolutely required for the decoding of a specific spatio-temporal resolution and bit-rate are identified by the NAL header.

The encoding process is done with multiple representations of the video source, each with their own spatial resolution and fidelity. These representations are known as layers, and each have their own identifier. The layers in each access unit are encoded in increasing order of their identifiers, as already transmitted data of a previous layer can be employed to aid in the encoding of later layers. These layers that use data from other layers in their encoding process are known as enhancement layers, and the layers they are predicting from are known as reference layers. The number of layers present in the bit stream depends on the requirements of the application. A maximum of 128 layers is supported by SVC, up to 47 of which can be enhancement layers. Enhancement layers can be broken down into spatial enhancement layers, when the spatial resolution changes relative to the reference layer, and fidelity enhancement layers, when the spatial resolution is identical to the reference layer. The input images of each of the spatial and fidelity layers are split into macroblocks and slices. A macroblock is a method of representing a 16x16 area of luma samples and 8x8 samples of the 2 chroma components. These are organised as slices and can be parsed independently.

Within each layer, SVC follows the design of the original standard. The samples of each macroblock either use intrapicture prediction, which uses spatially neighbouring samples of previously coded blocks in the same picture, or interpicture prediction, which uses a spatially displaced region of a previously coded picture of the same layer. The residual representing the difference between the original and the prediction signal for a block is transformed with the decorrelating transform, the coefficients of which are scaled and quantised. These are then entropy encoded, along with other information about the macroblock, such as the displacement vector. These are differentially coded using the displacement vectors of neighbouring blocks as predictors. The residual is decoded using inverse scaling

and inverse transformation operations on the quantised transform coefficients. This residual is then added to the prediction signal, and the result is additionally processed before it is output, or possibly to be stored as a reference picture.

SVC also provides interlayer prediction, which uses the statistical dependencies between different layers to reduce the bitrate, and therefore improve the coding efficiency of the enhancement layers. These prediction methods can be chosen on a macroblock or submacroblock basis allowing an encoder to select the mode that selects the highest coding efficiency. For SVC enhancement layers an additional macroblock coding mode is provided, in which the the prediction signal is inferred for the reference layer without transmitting any additional information. When the colocated blocks of this reference layer are intracoded, the prediction signal is built up from the reconstructed intra signal of the reference layer. This is known as interlayer intra prediction. In the original H.264 standard, the displacement vectors of reference layer blocks can also be used as replacement for the normal spatial displacement vector predictor. Another method, the interlayer residual predictor, aims to reduce the bit rate required for transmission of macroblocks by using the residual signal of the colocated reference layer and the enhancement layer residual to get the difference between the original and the interpicture prediction signal and only encoding this difference.

Each spatial and fidelity enchancement layer can be decoded with a single motion compensation loop. For the reference layers, only the intracoded macroblocks and residual blocks that are required for interlayer prediction need to be reconstructed, and the corresponding displacement vectors must be decoded. This is an important feature of SVC, as only the complex operations of prediction and deblocking need to be performed for the target layer to be displayed. Temporal scalability can also be achieved by splitting the access units in to a temporal base and multiple temporal enhancement layers and ensuring that only access units of the same of coarser temporal layer are used for inter picture prediction.

To conclude, the above is a discussion of the features of SVC, but the most important differences between SVC and the original H.264/AVC standard are as follows: SVC enables the use of hierarchical prediction structures in order to provide temporal scalability and improve the coding efficiency and increase the effectiveness of quality and spatial scalable coding. SVC also provides new methods for inter-layer prediction, which improves the coding efficiency. The concept of key pictures gives us a method to efficiently control the drift for packet-based quality scalable coding with hierarchical prediction structures. Single motion compensation loop decoding for spatial and quality scalable coding provides a decoder complexity close to that of single-layer coding. Finally, SVC supports a modified decoding process that allows a lossless and low-complexity rewriting of a quality scalable bit stream into a bit stream that conforms to a nonscalable H.264/AVC profile.[3]

References

[1] Florian Bomers. *Wavelets in real time digital audio processing- Analysis and sample implementations.* PhD thesis, 2000.

[2] Himanshu M Parmar. Comparison of DCT and Wavelet based Image Compression Techniques. *International Journal of Engineering Development and Research*, 2(1):2321–9939, 2014.

[3] H. Schwarz, D. Marpe, and T. Wiegand. Overview of the Scalable Video Coding Extension of the H.264/AVC Standard. *IEEE Transactions on Circuits and Systems for Video Technology*, 17(9):1103–1120, sep 2007.

[4] Kamrul Hasan Talukder and Koichi Harada. Haar Wavelet Based Approach for Image Compression and Quality Assessment of Compressed Image. *IAENG International Journal of Applied Mathematics*, 36(February):8, 2010.