HIGH QUALITY AUDIO CODING SUITABLE FOR ISDN CHANNELS

Annie George and Bernt Ribbum

Department of Electrical and Computer Engineering
University of Wollongong
Australia.

ABSTRACT - In the last several years developments in signal processing technology and transmission technology has made it imperative that high quality audio coding algorithms be developed. A brief study is made into the various coders available for audio coding. OCF (Optimum Coding in the Frequency Domain) is an algorithm that allows audio source coding down to 64 kbits/s. This coder and the inclusion of it's basic principles into the MPEG - audio standard draft is discussed. A coder based on the principles of the OCF but for smaller bandwidth is being implemented by the authors.

INTRODUCTION

With the advent of the ISDN (Integrated Systems Digital Network) era, a wide variety of communication services are expected by customers. Transmission of high quality signals along these channels is necessary to cater for these services. In speech coding the goal is to get a clearly comprehensible and natural sounding speech; for this long term speech statistics can be used for constructing speech coders. In the case of audio coding, the situation is quite different. All types of acoustic signals will have to be coded and reconstructed without noticeable difference from the original. A brief study is done on the various algorithms for audio coding, with emphasis on the OCF (Optimum Coding in the Frequency Domain) coder. The OCF coder can be used where transmission over ISDN channels is required. An algorithm to code high quality audio, based on the OCF is being implemented by the authors. An international standard for high quality audio coding is being drafted by the ISO/MPEG (International Standards Organisation/Moving Pictures Experts Group). The third layer of this proposed standard incorporates the basic principles of the OCF algorithm. This paper takes a brief look at this layer.

HIGH QUALITY AUDIO CODING

All types of acoustic signals can be a part of an audio sequence. The statistics of these signals can be of any sort; the redundancy can vary from very high to very low levels. The coding algorithm for high quality audio signals must satisfy two important criteria (Eberlein, Gerhauser & Krageloh, 1990). The code must achieve two things:

- Reduce redundancy. This can be achieved by using an adaptive algorithm using time varying statistics of audio signals
- Reduce irrelevancy. This can be achieved by using psychoacoustical facts (e.g. masking effects)

The Major Compression Schemes

The major audio compression schemes can be broadly divided into subband coders and transform coders. There are advantages and disadvantages to the various coders as is discussed below.

(i) Subband coder

(a) Using the psychoacoustic principles, it is possible to calculate the just noticeable noise level in each band. To split the frequency into critical bands according to psychoacoustics, normally a filter tree is used (e.g. the Quadrature Mirror Filters - QMF). In this case each node of the tree splits the remaining frequency band into further subbands. The presence of so many filters in the filter tree makes the coder quite complex and also

results in a large delay. Another disadvantage is that the frequency resolution is inadequate to calculate the masking level.

(b) Another sect of subband coders uses polyphase filter banks. These filters are of moderate complexity and has a good resolution in the time domain. Another advantage is that delays are short. The frequency resolution of these coders are low. Due to this the critical bands do not correspond well to the frequency bands especially in the low frequency range. As a result of this the perceptual criteria cannot be easily used with this type of coder. This defect can be overcome with modifications, but will result in another disadvantage of increased complexity and lower time resolution.

(ii) Transform Coders

Transform coders are of moderate complexity. They have an excellent frequency resolution which is very useful when applying psychoacoustics; transform coders also have the potential for high redundancy reduction. Some of the algorithms use time domain aliasing cancellation (TDAC). These algorithms do not invariably suffer from blocking effects and low time resolution. The algorithm, when used with fully overlapping windows, can overcome the blocking effects (Eberlein, Gerhauser & Krageloh, 1990).

The OCF coder algorithm is that of a transform coder and uses MDCT (Modified Discrete Cosine Transform).

PRINCIPLES OF THE OCF CODER

The OCF coder works on the same mathematical background as the ATC (Adaptive Transform Coder). The OCF coder allows coding of high quality audio signals at bit rates of down to 64 kbit/s and this makes it suitable for use with the ISDN channels. Since it is a low complexity coder it is possible to implement in real time.

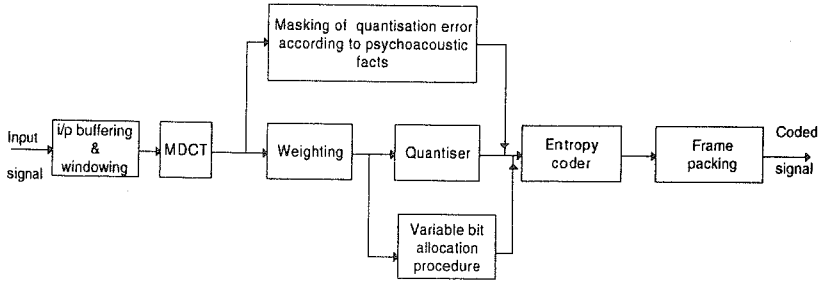

Figure 1 - Basic OCF coder algorithm

Input audio signal blocks are buffered and multiplied by an analysis window function. They are then transformed into the frequency domain using the MDCT.

MDCT is basically a DCT which is suitably modified to incorporate the algorithm. This involves subsampling of the values in the frequency domain. If the subsampling is included in the transform formula at the beginning, the formula would be (Jayant, Johnston & Shoham, 1991),

$$X_t(m) = \sum_{k=0}^{N-1} f(k)x_t(k)\cos\left(\frac{\pi}{2N}(2k+1+n)(2m+1)\right)$$

where N is the block length in time, n=N/2 is the block length in the frequency domain, $f(k)$ is a sine window, $f(k) = \sin(\frac{\pi}{N}(k+\frac{1}{2}))$, k=0,...,N-1, $x_t(k)$ is the kth sample of the tth block and $X_t(m)$ are the frequency domain values. The inverse transform is then given by,

$$y_t(p) = \sum_{m=0}^{n-1} X_t(m) \cos\left(\frac{\pi}{2N}(2p+1+n)(2m+1)\right)$$

The terms which occur due to subsampling in the frequency domain cancel each other in the synthesis stage, in a similar manner as in the QMF filters.

Varying the block length varies the frequency and time resolution. A large block will eventuate in an excellent frequency resolution but at the same time will also give low time resolution. On the other hand a small block will result in good time resolution and poor frequency resolution.

After the transformation into the frequency domain, the masking thresholds of the spectral values are derived (Brandenburg, 1987). These values are used in the psychoacoustic weighting and also to reduce the distortions to the minimum. An inner loop in the coding maintains the desired bitrate by selecting the proper quantiser step size. The entropy coded coefficients are multiplexed with some side information as required. The side information includes scale factors and quantiser step size.

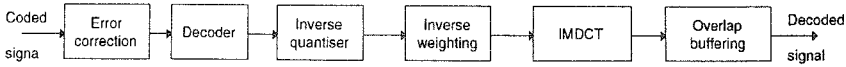

Figure 2 - Basic OCF decoder algorithm

The OCF decoder is much simpler as can be seen from Figure 2. The reconstruction of the frequency domain values are done first, by decoding the entropy coded data and multiplying them with the appropriate step size. As can be seen from the diagram, the inverse transformation, synthesis windowing and overlap add operations complete the decoding process.

THE ISO/MPEG - AUDIO DRAFT

The MPEG is part of the ISO-IEC JTC1/SC2/WG11 an organisation responsible for standardisation of representation of video and audio for information systems. They are currently in the process of drafting a standard for high quality audio coding. This standard is intended to specify the coded representation of high quality audio for storage media. The coder is compatible with existing PCM standards such as standard Compact Disc and Digital Audio Tapes. This standard is intended for sampling rates of 32, 44.1, and 48kHz (ISO/IEC JTC1 Audio draft 11172, 1992).

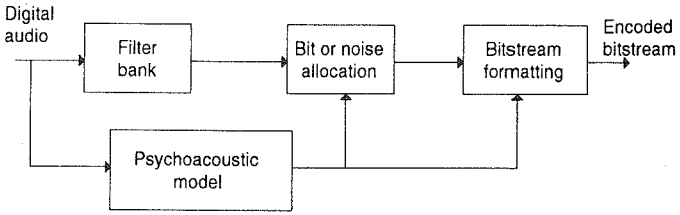

Figure 3 - primary parts of the encoder for the MPEG-audio algorithm
(from the ISO/IEC JTC1 Audio draft 11172, 1992)

The mapping from time to frequency domain is done with a filter bank in parallel with a psychoacoustic model. This model is used to guide the quantisation and the bit stream coding of the frequency mapped signals. The quantised signals with the side information are packed into a defined bit stream format.

The basic decoding algorithm is simple. The received signals are unpacked. Then the quantised and coded signals are converted back to the quantised sample values. The side information is used to determine the gain and bit allocation data. Finally the reconstructed samples are inverse transformed to the time domain.

The proposed standard has three layers. Each layer adds more to the compression ability and at the same time increases in complexity. The layer that is of interest here is the third layer. This layer uses each of the 32 bands from a polyphase filter bank as the signals that undergo MDCT. Entropy coding and sophisticated quantisers are included. The basic principles of the OCF coder are embedded in this layer. Since this coder can be used as a stereo coder at 64 kbits/sec for high quality applications, it would be of special interest to those working with ISDN.

THE ALGORITHM BEING IMPLEMENTED

The authors are currently working on a coder that would make use of lower bandwidth. A reduction in bitrate would result, although this reduction is yet to be quantified. Advantages of such a coder are

- A single ISDN channel can be used to send two audio streams which could be multiplexed.

-Audio signals can be transmitted along with video or text signals that are suitably modified.

-The computational complexity is reduced.

-More efficient use of the channel if CD quality audio is not required.

The algorithm for this coder is based on the OCF coder. Since the algorithm is an effective way of achieving a perfectly reconstructed signal, a frequency domain transform is required. MDCT is the transform used. The algorithm of this coder can also be said to be based on the ATC principles as it takes psychoacoustic facts into consideration for optimum use of the available bits.

Since the algorithm for this coder is very similar to the previously described OCF coder, a detailed description is omitted.

CONCLUSION

OCF and other similar coders that take psychoacoustic principles into consideration for coding, are gaining importance because they conform to the two important criteria for high quality audio coding. Redundancy and irrelevancy are kept to a low level due to the use of adaptive algorithm and the use of psychoacoustic facts. The other main advantage is that these coders are able to process signals that are compatible with the transmission requirements of ISDN. For a lower bitrate, similar methods can be applied to lower bandwidth signals. This could have several advantages for transmission of audio signals, and is currently being investigated by the authors.

REFERENCES

Brandenburg, K. (1987) *OCF - A New Coding Algorithm For High Quality Sound Signals,*Proc. ICASSP, pp. 141 -144.

Brandenburg, K. & Seitzer, D. (1988) *OCF: Coding High Quality Audio with Data Rates of 64 kbit/sec,* AES Preprints, vol. 2723, pp. 1-16.

Eberlein, E., Gerhauser, H. & Krageloh, S. (1990) *Audio Codec For 64 kbit/sec (ISDN Channel) - Requirements and Results,*Proc. ICASSP, pp. 1105 - 1108.

ISO/IEC JTC 1 *Information Technology - Coding of Moving Pictures and Associated Audio for Digital Storage Media up to about 1.5Mbits/s,* Draft 11172.

Jayant, N.S., Johnston, J.D., Shoham, Y. (1991) *Coding of Wideband Speech,* Proc. Eurospeech '91, pp. 373 - 379.

Jayant, N.S. (1990) *High Quality Coding of Telephone Speech and Wideband Audio,*Commun. Mag., January, pp. 10 - 20.

NEW AND IMPROVED PITCH DETERMINATION FOR THE IMBE VOCODER

T. S. Lim and M. S. Scordilis

Department of Electrical and Electronics Engineering
The University of Melbourne

ABSTRACT - A robust and accurate pitch determination algorithm of infinite resolution is presented in this paper. This method makes use of a hybrid of time domain and frequency domain pitch estimation techniques. For frame to frame analysis, this method was found to provide high accuracy in extracting the pitch period best representing the average pitch within a speech frame. It is a computationally efficient technique, particularly when used as part of the IMBE vocoder.

INTRODUCTION

In many speech synthesis and vocoder applications, one of the most important parameters used are the presence of voicing and the pitch period of the voiced parts of the signal. The human ear is very sensitive to pitch degradation and therefore accurate pitch evaluation is of paramount importance. Moreover, pitch determination is considered to be one of the most difficult tasks in speech processing. Since Dudley's attempt in 1939 to extract pitch by lowpass filtering (Schroeder, 1966), many techniques of automatic pitch extraction have been developed both for the time and for frequency domain (Hess, 1983).

Despite the large number of pitch estimation algorithms, the problem of a robust, reliable, and accurate pitch extraction still remains open. The complexity of pitch determination is due to the variability and irregularity in the nature of speech. In order to overcome the non-stationarity of the speech signal, short time analysis is often used. However, the wide range of values for the pitch period, as well as the changes in the state of voicing of the signal within the analysis frame (i.e., a mixture of voiced and unvoiced segments), lead to a crude average value or even wrong pitch estimation. In addition, a pitch estimate expressed as an integer multiple of the sampling interval, contains time quantization errors which may lead to audible distortion in speech coding application (Hess & Indefrey, 1984).

A pitch detection algorithm using a new similarity model was introduced by Medan, et al (1991) to overcome most of these problems in pitch processing. A more efficient, high resolution pitch estimation was also proposed, based on an IIR second order interpolator (Medan, 1991). After detailed simulation and observations it was found that these techniques do provide a robust and reliable pitch estimation. However, for high resolution (i.e., non-integer) pitch, required in speech coding applications, such as in the Multi-Band Excitation (MBE) Vocoder (Griffin & Lim, 1988), the accuracy of these methods is insufficient to overcome the problems introduced by the severe non-stationarity of the speech signals within an analysis frame.

In an effort to resolve this problem a hybrid method is proposed here and it combines time domain and frequency domain pitch refinement strategies. This hybrid method was found to be robust and reliable technique. It was used to accurately and efficiently estimate pitch in frame by frame analysis/synthesis of speech for the MBE vocoder.

PITCH DETERMINATION USING A SIMILARITY MODEL

Integer Pitch Determination

For short speech segments, voiced speech is of quasi-periodic nature, and although adjacent periods are rarely identical, they tend to be very similar. It has been shown (Medan, et al, 1991) that the normalized cross-correlation of adjacent speech segments provides an optimum criterion for determining their degree of similarity. The normalized cross-correlation over time interval τ is defined as: