SUBBAND CODING OF DIGITAL AUDIO SIGNALS

by R.N.J. VELDHUIS, M. BREEUWER and R.G. VAN DER WAAL

Philips Research Laboratories, P.O. Box 80 000, 5600 JA Eindhoven, The Netherlands

Abstract
A subband coding method for high-quality digital audio signals is described. To achieve low bit rates at a high quality level, it exploits the simultaneous masking effect of the human ear. It is shown how this effect can be used in an adaptive bit allocation scheme. The method is capable of reducing the bit rate of a compact disk signal by a factor of seven. Results obtained with a low-complexity and a high-complexity system are discussed.

Keywords: adaptive bit allocation, digital audio, psycho-acoustics, simultaneous masking, source coding, subband coding

1. Introduction

Transmission and storage of high-quality digital audio signals are becoming important for the audio industry, for instance in the case of digital radio and of new applications for optical disks. The bit rate of a high-quality stereophonic digital audio signal, as it is recorded on a compact disk, is about 1.4 Mbit s$^{-1}$. For some transmission channels or storage media this is too high and therefore source coding is required. Since digital audio is associated with high quality, a perceptible loss of quality cannot be tolerated.

Source coding of audio signals at low bit rates generally introduces errors. This paper describes a coding method that attempts to keep coding errors inaudible by exploiting the simultaneous masking effect occurring in the human auditory system. This is the perceptive phenomenon that a weak signal, e.g. quantization noise, is masked (i.e. made inaudible) by a stronger signal, e.g. a pure tone in the audio signal. Simultaneous masking is briefly explained in Sec. 2.

Simultaneous masking is most effective if both masked and masking signal are in a rather narrow frequency band. This suggests the use of subband
coding, where the signal is first split up into frequency bands which are then quantized. The structure of the subband coding system is given in Sec. 3.

Quantization should be such that the quantization noise is masked by the audio signal. This is achieved by using uniform block companded quantization \(^1\)). The subband signals are split up into blocks. Each block is scaled to a unit level and then quantized by a uniform quantizer. Quantized data and scale factors are transmitted. In this manner the power of the quantization noise can be controlled by allocating a certain amount of bits to each quantizer.

In a subband coding system we can distinguish in-band masking, where both masked and masking signal are in the same subband, and out-of-band masking, where masking and masked signal are in different subbands. Both are exploited in the method described in this paper. Section 4 explains how the maximum power of the quantization noise that is masked, called the masked power, can be estimated for each subband.

Once the masked powers have been computed for all subbands, bits are allocated to the quantizers. Ideally the number of bits for each quantizer should be such that in each band the quantization noise is completely masked. However, the masked powers are signal dependent and therefore the number of bits needed to ensure complete masking varies in time. Here coding systems with a fixed bit rate are considered. Therefore the available bits must be divided over the subbands in such a way that the audible degradation of the output signal is minimal. This requires an adaptive bit allocation method, which is described in Sec. 5.

There is a trade-off between quality, bit rate and complexity. Complexity is largely determined by the splitting and merging subband filters. It can be kept low by keeping the number of subbands low and their minimum bandwidth high. At a fixed quality level, the lowest bit rate achievable with a ‘low-complexity’ system is higher than with more complex systems with more and narrower subbands. This is explained in Sec. 4. With a complex system a reduction in bit rate by a factor of seven can be obtained without a perceptible loss in quality. Results obtained with a simple and a complex system are discussed in Sec. 6. Section 7 summarizes the main results and gives some topics for future research.

2. Simultaneous masking

Simultaneous masking is the effect that a weak signal is made inaudible by a simultaneously occurring stronger signal. Masking has been discussed in great detail in refs 2 and 3. The use of masking in subband coding has been
Subband coding of digital audio signals

described in refs 4 and 5. In the next two paragraphs results from ref. 2 are repeated to illustrate the masking effect.

Consider as a test signal one pure tone. This is inaudible if its sound pressure level (SPL) is below the threshold of hearing. The threshold of hearing is a function of frequency, as is shown in fig. 1. In the presence of a second, stronger signal, the SPL above which the test signal is audible differs from the threshold of hearing. It is raised for frequencies close to the frequency of the stronger signal. This new threshold is called the masking threshold and the second, stronger signal is called the masking signal. The test signal is masked if its SPL is below the masking threshold. This is illustrated in fig. 1, where masking thresholds of narrow-band noise signals with a bandwidth of 90 Hz, centred at 1 kHz, at various sound pressure levels \( L_0 \) are shown.

The masking threshold, as a function of frequency, depends on the SPL and the spectrum of the masking signal. For instance, masking thresholds for a pure tone as a masking signal are depicted in fig. 2. Figure 3 shows masking thresholds of narrow-band noise signals with different centre frequencies \( f_m \). Masking thresholds of narrow-band noise signals have different shapes for centre frequencies below and above 500 Hz, as is illustrated by fig. 3. Furthermore, the masking thresholds of pure tones are lower and of a different shape than those of narrow-band noise signals, as illustrated by figs 1 and 2.

*) Figures 1, 2, and 3 are taken from ref. 2.
The masking thresholds for both pure tones as well as for narrow-band noise signals show an asymmetry around the frequency of the masking signal. Signals with frequencies higher than the frequency of the masking signal are better masked than signals with frequencies lower than the frequency of the masking signal.

So far, masking of a pure tone by another pure tone or by a narrow-band noise signal has been considered to illustrate the masking effect. What is needed is some kind of simple, analytical masking model that can be used in a source coding system. This will be introduced in Sec. 4.
The results presented in this section have been obtained from experiments performed on a large group of people\textsuperscript{2,3}). They are valid for young people with good hearing. Older people will generally have a higher threshold of hearing at the higher frequencies, say above 10 kHz.

3. Subband coding

It is clear from fig. 1 that masking is strongest for frequencies close to the frequency of the masking signal. This suggests that the masking phenomenon can be well exploited in a subband coding system. In such a system the encoder splits up the signal into frequency bands, called subbands, which are then quantized. The quantized subband samples, together with some additional data, are transmitted. The decoder reconstructs the subbands from the received quantized samples and the additional data. Then the subbands are merged into a replica of the original signal.

The splitting of the signal into subbands and the merging of the subbands into a replica of the original signal are carried out by decimating and interpolating filter banks, such as quadrature-mirror or conjugate quadrature-mirror filter banks\textsuperscript{7,8}). Because of the decimation, the sampling frequency of a subband signal equals twice the subband's bandwidth. Therefore the total sample rate after splitting is the same as the sample rate at the input. The ratio of the sampling frequency of the input signal of the filter bank and the sampling frequency of a signal in a subband is called the decimation factor of that subband. Because, as can be seen from fig. 3, the 'bandwidth' of the masking threshold increases with the frequency of the masking signal, the bandwidths of the subbands may increase with frequency.

In the process of quantization, quantization noise is added to the signals. If the filter banks have good frequency-separating properties, the quantization noise remains in the subband it was added to. It is masked by the audio signal if the signal-to-noise ratio is above a certain threshold. This implies that the quantizers must operate at a predetermined signal-to-noise ratio. This can be achieved with uniform block companded\textsuperscript{*}) quantizers\textsuperscript{1). In this type of quantizer the signal is first divided into blocks. Of these blocks the maximum absolute values, called peak values, are computed. By dividing the samples in the blocks by the peak values, they are scaled to a unit level. The scaled blocks are then quantized with a uniform quantizer. The signal-to-noise ratio expressed in dB is proportional to the number of bits used in the quantizer, so that the signal-to-noise ratio of a quantizer can be predetermined by allocating a certain amount of bits to it.

*) To 'compand' is a combination of the verbs to compress and to expand.
Figure 4 shows a diagram of a coding system with 20 subbands. As can be seen, quantized samples as well as coded peak values and side information to indicate the number of bits used for quantization are transmitted.

The numbers of samples in the blocks should be chosen sufficiently small so that changing statistics in the audio signal can be tracked. Otherwise it can no longer be assumed that the signal-to-noise ratio of the quantizer is constant. However, the number of samples in a block must not be too small, because then the bit rate increases as a result of the increased amount of side information. A maximum block length corresponding to 10–20 ms turns out to give good results. For instance, for the 20-band system of fig. 4 blocks of length 32, corresponding to 20 ms, have been chosen.

In Sec. 4 it is shown how for a given division of the signal into subbands the simultaneous masking effect described in Sec. 2 can be used to determine the masked powers in the subbands. How the final bit rate depends on the division into subbands is also explained. Section 5 shows how the number of bits allocated to each quantizer is computed.

4. A masking model for subband coding

In Sec. 2 results from psycho-acoustical literature have been presented to illustrate the simultaneous masking effect. These results cannot be used directly to compute masked powers in subbands as is needed in subband coding systems. For instance, in Sec. 2 masking of one pure tone by one other pure tone or by a narrow-band noise signal has been discussed. Music is a time-varying complex of tones combined with other noisy signals. Although
Subband coding of digital audio signals

some experiments have been performed in that direction\textsuperscript{9,10}, it is not known how the masking threshold of a complex of tones must be computed. Therefore it is unclear how, given a short-time music spectrum, masked powers in subbands must be computed.

As has been done in other audio source coding systems\textsuperscript{5,11} a masking model has to be assumed that provides answers to the above problems. It must be emphasized that this model is partly based on assumptions that have never been completely justified by experiments. The masking model chosen here is simple. More elaborate models will lead to better coding systems but with an increased complexity. As will be demonstrated in Sec. 6, good enough results can already be obtained with the simple model used here.

As a basis for the masking model the masking threshold of a pure tone is used. It is in this model assumed that, on a logarithmic scale, its shape is independent of the frequency of the tone and its SPL. This implies that for frequencies below 500 Hz the computed masking threshold will be too low. In the model used here the shape of the masking threshold of a pure tone is approximated by

\[
T(f_m, f) = \begin{cases} 
T_{\text{max}}(f_m) \left(\frac{f}{f_m}\right)^{28}, & f \leq f_m, \\
T_{\text{max}}(f_m) \left(\frac{f}{f_m}\right)^{-10}, & f > f_m.
\end{cases}
\]

In this expression \(f_m\) is the frequency of the masking signal and \(T_{\text{max}}(f_m)\) is the relative masking threshold at this frequency. \(T_{\text{max}}(f_m)\) depends on the frequency of the masking signal\textsuperscript{4}. The masking threshold of a tone is obtained by multiplying \(T(f_m, f)\) with the power of the tone. Figure 5 shows a stylistic approximation of the masking threshold according to the masking model of
For a frequency of 1000 Hz, $T_{max}(f_m)$ equals $-20$ dB). The model ignores the effects occurring around the frequency of the masking tone and its harmonics that can be observed in fig. 2, but it describes the masking threshold with sufficient accuracy.

Furthermore, it is assumed that masking is additive: the masking threshold for a signal containing more than one frequency component can be obtained by adding the masking thresholds of the components. That this is allowed can be concluded from results in ref. 9, although some of these results have not been confirmed in ref. 10. From an engineering point of view the assumption of additive masking is very attractive. Otherwise a straightforward computation of masked powers would not be possible.

The masking model used here is a simplification of reality. Coding systems based on it may show unexpected and unwanted effects. To avoid this, they are to be tested and optimized in extensive listening experiments.

It can be seen from fig. 1 that signals with a frequency lower than the frequency of the masking signal are hardly masked. Therefore only two kinds of masking are considered: in-band masking, which is masking within a subband, and masking of signals in subbands at higher frequencies. For both cases the masked power is computed as a function of the powers of the subband signals.

Firstly it is assumed that there is only one signal in the subband with index $i$. This subband ranges from frequencies $f_{i,l}$ to $f_{u,i}$. The signal power is $\sigma_{s,i}^2$. The quantization noise is assumed to have a flat spectrum in the subband. The worst-case situation for in-band masking occurs when the masking signal is a pure tone with a frequency $f_{u,i}$. In this case the power of the quantization noise in subband $i$ that is masked by a signal with power $\sigma_{s,i}^2$ in the same subband must be less than $\sigma_{s,i}^2 T(f_{u,i}, f_{i,l})$. This situation is illustrated in fig. 6.
Subband coding of digital audio signals

for a pure tone of 1000 Hz at an SPL of 70 dB. In this figure subband $i$ ranges from 830 to 1000 Hz.

The worst-case situation for the masking of noise in subbands at higher frequencies occurs when the masking signal in subband $i$ is a pure tone with a frequency $f_{i,t}$. In this case the power of the quantization noise in subband $j$ that is masked by a signal with power $\sigma_{s,i}^2$ in subband $i$ must be less than $\sigma_{s,i}^2 T(f_{i,t}, f_{u,j})$. This situation is illustrated in fig. 7 for a pure tone of 1000 Hz at an SPL of 70 dB. In this figure subband $i$ ranges from 1000 to 1330 Hz, and subband $j$ ranges from 1780 to 2180 Hz.

In this way the contribution of a subband to the masked power in all subbands can be computed. Because masking is assumed to be additive, the masked power in a subband can be obtained by adding all contributions.

The lowest achievable bit rate at a certain quality level depends on the division into subbands. The computations of the masked powers are based on worst-case assumptions. The real masked powers can be substantially higher. If the subbands are narrower the results of these computations will, on average, be closer to the real masked powers. This effect, however, is limited for in-band masking, because in reality the tops of the curves of figs 5, 6, and 7 are flatter than is depicted \(^4\), so that there is no point in decreasing the bandwidth of the subbands below a certain value. Because of the effect mentioned here, narrower subbands lead to higher masked powers and consequently the number of bits required for quantization can be lower.

The results of this section are only valid if the distribution of signal power over the subbands is stationary. In reality this is not true. Therefore, in a subband coding system, the masked powers must be computed periodically. As a consequence of this nonstationarity the number of bits needed to quantize each subband in such a way that the quantization noise is masked will also
vary in time. The allocation of bits to the quantizers in such a way that the coding system produces a fixed bit rate is discussed in Sec. 5.

5. Adaptive bit allocation

Before the subband signals are quantized they are divided into blocks. The blocks are arranged in an allocation window. An allocation window contains all subband samples during a period of time. This period is chosen in such a way that it contains one block of samples from the most decimated subband. This is in general the subband at the lowest frequency. An example of an allocation window for a 20-band system is shown in fig. 8. If the input sample frequency is 44100 Hz, subbands 1–8 have a bandwidth of 689 Hz, and subbands 9–20 have a bandwidth of 1378 Hz.

An allocation window contains $MD_{\text{max}}$ samples, where $M$ is the quantization block length, and $D_{\text{max}}$ is the maximum decimation factor. In the example of fig. 8, the block length and the maximum decimation factor both equal 32. The allocation window corresponds to a time duration $T_w$ given by

$$T_w = \frac{MD_{\text{max}}}{f_s}$$

where $f_s$ is the input sampling frequency. If the desired bit rate at the output is $R$, then $T_w R$ bits must be divided over the allocation window. A certain number of bits is reserved for peak values and side information. Good results have been obtained with peak values logarithmically quantized with 6 bits.
and 4 bits of side information per block indicating the number of bits per sample used for the block. Assume that $B_q$ bits are left for quantization of the subband samples. The following adaptive bit allocation procedure is used to divide these bits over the allocation window.

First for each block in the allocation window the peak value and the power are computed. The power in a block is obtained as the average of the squares of the samples. By using the masking model of Sec. 4, the masked powers are now computed for every block in the allocation window. This is done in such a way that a block only contributes to the masked power in blocks that lie within the time interval of the masking block. For instance, the block in subband 1 of fig. 8 contributes to the masked powers of all blocks in the allocation window, the block in subband 3 contributes to the masked power of all blocks in subbands 3–20, and the left-hand block of subband 9 only contributes to the masked powers in the left-hand blocks of subbands 9–20.

Assume that the number of blocks in the allocation window is $N$ and that they are numbered from 1 to $N$. The estimated masked power in the $i$th block is denoted by $\sigma^2_{m,i}$, and the peak value by $p_i$. If the samples in the block are uniformly quantized to $b_i$ bits, then the power of the quantization noise in the $i$th block is assumed to be given by $^{1)}$

$$\frac{1}{12} \left( \frac{2p_i}{2^{b_i}} \right)^2.$$ The noise-to-mask ratio in the same block is defined by

$$\frac{1}{12} \left( \frac{2p_i}{2^{b_i}} \right)^2 \frac{1}{\sigma^2_{m,i}}.$$ In the following it is sometimes more convenient to use a logarithmic noise-to-mask ratio $\gamma_i$ given by

$$\gamma_i = \log_2 \left( \frac{p_i}{\sigma_{m,i}} \right) - b_i.$$ The adaptive bit allocation procedure is such that the total noise-to-mask ratio, given by

$$\sum_{i=1}^{N} \frac{1}{12} \left( \frac{2p_i}{2^{b_i}} \right)^2 \frac{1}{\sigma^2_{m,i}}.$$
is minimized under the constraint
\[ \sum_{i=1}^{N} M b_i = B_q. \]

A further constraint is that all \( b_i \) must be integers with \( 0 \leq b_i \leq 15 \).

An elegant solution to this constrained integer minimization problem can be derived from the theory given in ref. 12. The result of this procedure for this case, as described below, is intuitively pleasing. Initially, it is assumed that all blocks are quantized with zero bits. The number of bits left to divide over the allocation window then is given by
\[ B = B_q, \]
and the logarithmic noise-to-mask ratio for the \( i \)th block by
\[ \gamma_i = \log_2 \left( \frac{p_i}{\sigma_{m,i}} \right). \]

Then the following three steps are repeated until \( B < M \), so that no more bits can be assigned to the blocks:
1. the block \( i \) with the highest logarithmic noise-to-mask ratio \( \gamma_i \) is searched,
2. then the number of bits per sample \( b_i \) for this block is increased by one, and
3. the number of bits left \( B \) is decreased by \( M \).

If this procedure is started with a sufficient number of bits, the result is that all noise-to-mask ratios will be approximately equal. The procedure can be refined by setting lower bounds to the noise-to-mask ratios of the subband. The reason for doing this is that once the noise-to-mask ratio in a block is below a certain threshold it is no longer necessary to add bits to the block because the quantization errors have become inaudible. In this way bits can be saved for other blocks.

Note that, if \( B_q \) is not a multiple of \( M \), a small number of bits are left over, and the bit rate will be a fraction less than the desired bit rate \( R \).

6. Results

The methods explained in this paper have been applied in two coding systems: a complex system splitting up the signal into 26 subbands,
Subband coding of digital audio signals

approximately one-third of an octave wide, and a simpler 20-band system of which the bandwidths were given in Sec. 5. In both systems the adaptive bit allocation method described in Sec. 5 has been used. Both systems have been designed for coding stereophonic 16-bit compact disc signals with a sample frequency of 44.1 kHz. Left and right channels are coded independently. With the 26-band system high-quality results can be obtained at bit rates of 200 kbit s$^{-1}$, which comes down to a reduction by a factor of seven. With the 20-band system similar results can be obtained at bit rates of 300 kbit s$^{-1}$, which comes down to a reduction by a factor of almost five. In listening experiments most of the processed pieces of music cannot be distinguished from the original source material. Physically, however, signal-to-noise ratios are measured between 15 and 30 dB, showing that the signal is heavily distorted.

The complexity of the systems is largely determined by the memory requirements of the filter banks. These are substantially higher for the 26-band system.

The filtering and coding delay is determined by the maximum decimation factor, the filter banks used, and the quantization block length. For the 26-band system the maximum decimation factor is 256 and the total delay can be as high as 800 ms. For the 20-band system the maximum decimation factor is 32 and a typical value for the total delay is 80 ms. Other types of filter banks and shorter quantization blocks may give lower values.

7. Conclusions

A subband coding method for high-quality digital audio signals has been described. It exploits a perceptive phenomenon called simultaneous masking. This phenomenon has been described with some examples. A masking model has been derived on which the coding algorithm is based. With this masking model the encoder computes the amount of quantization noise that is masked in each subband. An adaptive bit allocation algorithm distributes the bits over the subbands in such a way that the bit rate is fixed, and that the total noise-to-mask ratio is minimized. After decoding there is no, or very little, perceptible loss in quality, even at low bit rates.

Further improvements are possible. The masking model that is used can be extended in several ways. For instance, it can be adapted in such a way that it better describes masking at frequencies below 500 Hz. Temporal masking, which is masking of a signal occurring in time before or after the masking signal, can also be incorporated. A further reduction in bit rate can be achieved by exploiting redundancies in the quantized subband samples,
Raymond Veldhuis: Ir. degree (Electrical Engineering), University of Twente, Enschede, The Netherlands, 1981; Ph.D., University of Nijmegen, The Netherlands, 1988; Philips Research Laboratories, Eindhoven, 1982-. He has been working on restoration methods for lost samples in audio, speech, and television signals. Currently he is investigating source coding methods for high quality digital audio systems.

Acknowledgement

The authors would like to thank their colleague Gertjan Keesman for his help in developing the adaptive bit-allocation technique.

REFERENCES

5) G. Theile, M. Link and G. Stoll, Low bit rate coding of high-quality audio signals, AES Preprint 2431, 1987.

Authors

Raymond Veldhuis: Ir. degree (Electrical Engineering), University of Twente, Enschede, The Netherlands, 1981; Ph.D., University of Nijmegen, The Netherlands, 1988; Philips Research Laboratories, Eindhoven, 1982-. He has been working on restoration methods for lost samples in audio, speech, and television signals. Currently he is investigating source coding methods for high quality digital audio systems.

Marcel Breeuwer: Ir. degree (Electrical Engineering), Technical University, Delft, The Netherlands, 1982; Ph.D., Free University of Amsterdam, 1985; Philips Research Laboratories, Eindhoven, 1985--His Ph.D. work was on the effect of supplementing speech reading with auditory information. More recently he has been involved in source coding for high-quality digital audio signals. Currently he is working on video source coding.
Subband coding of digital audio signals

Robbert van der Waal: Ir. degree (Electrical Engineering), University of Twente, Enschede, The Netherlands, 1987; Philips Research Laboratories, Eindhoven, 1987-. He has been working on source coding methods for high-quality digital audio signals. Currently he is involved in the design of integrated analogue filters.