SUBBAND CODING OF SPEECH AND AUDIO

Charles D. Creusere
Naval Air Warfare Center
China Lake, CA 93555
email: chuck@wavelet.chinalake.navy.mil

ABSTRACT

We discuss here the use of multirate filter banks for perceptually-weighted speech and audio compression. While our primary focus is on audio compression, we also review two recently proposed wideband speech coders that use filter banks to eliminate perceptually redundant information. Our goal is to examine and compare the time-frequency tradeoffs inherent in various coding algorithms as they try to exploit the masking properties of the human auditory system.

1. INTRODUCTION

One can make the argument that audio compression is the area in which multirate filter banks have been most successfully applied to date. This is evidenced by the fact that nearly every commercially available compression system designed to handle generic audio signals (as opposed to speech) uses some form of filter bank analysis. In some cases, the signal decomposition is performed using an overlapped transform which is alternately called a time domain aliasing cancellation (TDAC) filter bank, a modified discrete cosine transform (MDCT), or a Princen-Bradley filter bank [1]. In other compression systems, filter banks with fewer but better separated (i.e., having a longer impulse response) frequency bands are used. In the past, these have been called generalized or pseudo QMF banks [2]-[6], and while they do not provide perfect reconstruction even in the absence of coefficient quantization, their widespread adoption clearly indicates their suitability for audio compression [7], [8]. The MDCT and the pseudo-QMF bank offer good tradeoffs between time and frequency localization, and they both have highly efficient implementations built around a discrete transform. Figure 1 shows the efficient polyphase implementation of a cosine-modulated filter bank where the original N-th order lowpass prototype filter H(z) has been decomposed into 2M polyphase subfilters $G_k(z)$ such that

$$H(z) = \sum_{k=0}^{2M-1} G_k\left(z^{2M}\right)\cdot z^{-k}. \qquad (1)$$

The elements of the 2M×M transform matrix in Fig. 1 are given by

$$t_{kn} = 2\cos\left(\frac{\pi}{M}(k+0.5)\left(n-\frac{N}{2}\right)+\theta_k\right) \qquad (2)$$

where $\theta_k = (-1)^k \pi/4$, $k \in [0, M)$, and $n \in [0, 2M)$. In the case where N+1 = 2*m*M (*m* is an integer), fast implementations of (2) exist based on the fast Fourier transform (FFT) [9]. An MDCT with a 50% overlap is just a special case of (1) in which each of the polyphase subfilters has only one non-zero coefficient (i.e., the window weight). This becomes more clear if one rotates Fig. 1 90 degrees and notes that the 2M-length polyphase delay chain (including downsampling) is equivalent to a tapped-delay line which is clocked every M samples. Recent work in the area of perfect reconstruction cosine modulated filter banks has effectively unified the theories of lapped cosine transforms and pseudo-QMF banks [10]-[12], but a distinction between the two is often made depending on whether the coefficients which are grouped together for coding come from the same frequency band (subband coding) or from different frequency bands (transform coding). In addition, windowed FFTs and perfect reconstruction allpass filter banks have also been effectively employed for audio compression, but neither of these approaches has achieved the popularity of cosine-modulated implementations [13], [14].

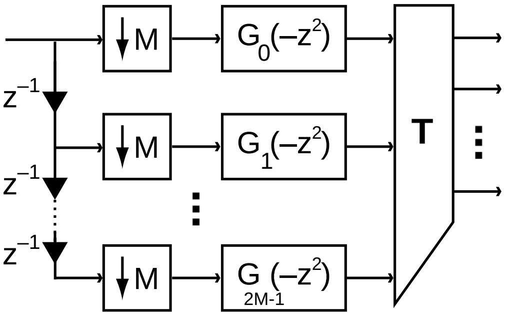

Figure 1: M-band cosine-modulated analysis filter bank

While efficient multirate filter banks are important, it is their unification with perceptual models of human hearing that has lead to their popularity in audio coding applications. Specifically, quantization noise can be appropriately localized in time and frequency by selecting the right decomposition; this property is essential for the coding algorithm to exploit tonal masking and absolute hearing thresholds [15]-[17]. While early work in narrowband (< 4kHz) transform-based speech coding was motivated primarily by coding gain [18], [19], a crude form of perceptual subband coding *was* proposed by Crochiere *et al* [20]. In our paper we survey a number of coding algorithms, focusing in particular on the interaction between the time-frequency decomposition and the perceptual coding. Our major emphasis is on generic audio coding since it is here that subband and transform methods have had the most impact. We do, however, also discuss candidates for the new ITU-T wideband speech coding standard since these too combine time-frequency decompositions with models of human hearing.

2. PERCEPTUALLY TUNED QUANTIZATION

Figure 2 shows the block diagram of a generic subband (transform) coding algorithm which can adapt its quantization (and possibly its decomposition) to optimize the perceived quality of its reconstructed audio. The dotted lines represent data exchanges which do not occur in all implementations. For example, all three MPEG 1 audio coders use a separate FFT to perform the frequency analysis required to do the bit allocation (indicating that path **a1** is active) while Dolby AC-3, Philip's digital compact cassette (DCC), and Sony's MiniDisc use only the output of their signal decompositions (path **a2**). The 'Perceptual Analysis' block computes the masking estimates which are required by the 'Bit Allocation' block to ensure that quantization errors in the reconstructed audio are inaudible. Based on this analysis of the signal, some audio coders also have the ability to alter their decompositions and the corresponding coefficient groupings (path **b**) to prevent the introduction of pre-echoes into the decoded audio. Note that the decoder simply inverts the operations of the encoder block by block to reconstruct an approximation of the input audio.

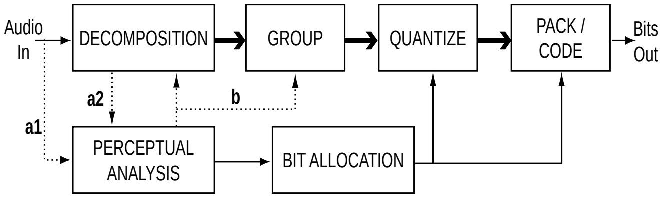

Figure 2: Generic audio encoder.

Perceptually transparent coding is accomplished primarily by exploiting the various masking properties of the human ear, specifically: the absolute threshold of hearing, simultaneous frequency masking, forward (temporal) masking, and backward masking. First, any frequency component of the signal whose power falls below the absolute threshold of human hearing need not be transmitted. This threshold is lowest between 2 and 4 kHz and goes up rapidly above 15 kHz. Next, if a small amplitude tonal signal occurs at the same time as a larger one of similar frequency, the smaller signal will be masked. This is called simultaneous masking and is specified in terms of critical bands which are defined on the bark scale [22]. These critical bands define the frequency resolution of the human auditory system-- from 0 to 500 Hz there are 5 uniform critical bands while above 500 Hz the width of each band expands by approximately 1/3 per octave. The effectiveness of the masking decreases by about 8 dB/bark for critical bands above the masker and 25 dB/bark for those below it, and it also depends strongly on the tonality of the input since pure tones mask each other much more effectively than noise-like signals. To estimate tonality, the Spectral Flatness Measure (SFM)-- basically the logarithm of the power spectrum's geometric mean divided by its arithmetic mean-- is generally employed [18]. Specifically, a ratio of the current SFM to the SFM of a maximally tonal input is used to compute the tonality of the current block of samples, and this tonality coefficient biases the masking threshold upward for highly tonal signals or downward for noise-like signals.

The final perceptual effects which must be considered in the design of the coding algorithm are forward and backward temporal masking. Forward masking occurs when the masking signal ends before the masked signal begins while backward masking is the exact opposite. Perceptual studies have shown that forward masking is the more effective of the two by a wide margin [22]. While most of the currently available coding algorithms claim to 'exploit' forward and backward masking, this statement is somewhat misleading. Explicitly, they exploit *simultaneous* masking to achieve bit rate reductions through adaptive bit allocation while implicitly exploiting forward masking to conceal the effects of time-frequency blocking on the quantized coefficients. In other words, if the masking signal contained within the block of coefficients ends prematurely, the quantization noise will still be concealed. The situation with backward masking, however, is entirely different since this phenomenon is highly localized around the leading edge of the masker. If blocks of coefficients representing a fixed time-frequency subdivision of the signal are jointly coded, then it is possible for pre-echo to be introduced into the reconstructed audio by the occurrence of a large masker in latter parts of a block. Thus, the goal of the coding algorithm is not so much to exploit backward masking as to compensate for its limitations. In fact, the entire motivation for using temporally adaptive transformations in the encoder (path **b** in Fig. 2) comes from the need for increased time localization of the quantization errors during sharp attacks (i.e., sudden increases in the short time power spectrum of the audio input). As one metric for comparison, we define the term 'temporal footprint' to be the extent to which quantization errors are localized in the time domain. For example, if an M-point MDCT is applied to the signal and quantization is performed, then the errors must be completely confined to 2M samples of the audio sequence (but most concentrated in the central M samples). In the next section, we assume that the temporal footprint for an M-point overlapped transform is 3M/2 samples to take into account the window shape. Note that this gives different values than in some previous work where the loss of temporal localization due to filtering effects was not considered (e.g., [7]).

Table 1: Comparisons of time-frequency tradeoffs

Coder	Temporal F.P.	Freq. Res.
MPEG– L1, L2	10.9 ms	690 Hz
MPEG– L3	8.7 ms, 22.2 ms	115 Hz, 44 Hz
MPEG– AAC	4.3 ms, 34.8 ms	172 Hz, 21,5 Hz
Dolby AC-3	4.3 ms, 8.7 ms	172 Hz, 86 Hz
Bell Labs (E)PAC*	4.3 ms, 34.8 ms	172 Hz, 43 Hz
Sony ATRAC*	8.7 ms, 34.8 ms	345 Hz, 86 Hz
ITU-T, AT&T	6 ms	109 Hz
ITU-T, PT	30 ms	22 Hz

3. CODING ALGORITHMS

Table 1 summarized the temporal footprints and frequency resolutions of the various algorithms discussed here. All of the wideband audio coders are assumed to be operating on data sampled at 44.1 kb/s; an input sampling rate of 16 kb/s is assumed for the two speech coders (the last two table entries). Systems using adaptive transforms have multiple entries, and the asterisks indicate that the temporal footprint and resolution may differ in the higher frequency bands.

3.1 Systems Using Pseudo-QMF Banks

MPEG Audio Layers 1 and 2 as well as Philip's DCC all use as their fundamental decomposition a 32-band pseudo-QMF bank based on a 512-tap lowpass prototype filter H(z) and implemented as shown in Fig. 1 [7], [8]. The Layer 1 coder is essentially identical to the DCC coder except that it uses separate a 512-point FFT to perform its spectral analysis of the input audio (i.e., path **a1** is active in Fig. 2) versus simply using the estimate produced by the subband decomposition itself (path **a2**). In either case, 12 samples (for Layer 2 this is the 'minimum' size) in each subband are grouped together for coding which implies that each time-frequency 'frame' corresponds to 384 input sample or 8.7 ms of audio at a sampling rate of 44.1 kHz. Furthermore, we also note that, in theory, as many as 511 samples (11.6 ms) of pre-echo can be introduced into the reconstructed audio by the filter bank itself since this is the amount of delay required to make the combined analysis/synthesis system zero phase [9]. In practice, one finds that even the harshest quantization seldom results in more than 50 samples of significant pre-echo (see [14]), giving a temporal footprint of approximately 10.9 ms for the complete system. A sharp audio attack in the second half of this time interval could result in 5 or 6 ms of pre-echo if an insufficient number of bits is used to code the frame. Also, this filter bank has a relatively coarse frequency resolution of 690 Hz per subband which implies that as many as six critical bands must be encoded together at the lowest frequencies (i.e., subband 0).

3.2 Systems Using Lapped Transforms

Audio coding systems using overlapped transforms are by far the most common-- a popularity which probably results from the ease with which their time-frequency footprints can be adapted. Three major wideband audio coding algorithms have been proposed which use exclusively lapped transforms: Dolby AC-3, Bell Lab's Perceptual Audio Coder (PAC), and the MPEG 2 advanced audio coder (AAC) [23]–[25]. In normal operation, AC-3 uses a TDAC decomposition with M = 256, implying that 512 samples are blocked together for processing to generate 256 distinct freqency coefficients (i.e., the overlap between blocks is 50%). The AC-3 decomposition uses both sine and cosine modulation functions and can be implemented very efficiently with an FFT. As previously noted, AC-3 does not implement a separate FFT for spectral analysis (i.e., path **a2** is active), but it does perform a highpass filtering operation directly on the audio to identify attacks. If an attack occurs during the second half of the 512 sample block, the block is split and two independent transforms (M = 128) are applied instead. Since each block of samples is coded separately, the temporal footprint is either 8.7 ms with 86 Hz frequency resolution or 4.3 ms with 172 Hz resolution, depending on the transform mode. Both PAC and MPEG 2 AAC are very similar to AC-3 in concept but more sophisticated. For example, they both normally use a 1024-point MDCT and switch to eight 128-point MDCTs when an attack is detected. Therefore, their temporal footprints are 34.8 ms with 21.5 Hz frequency discrimination when using the longer transform, and 4.3 ms with 172 Hz resolution using the shorter one. Clearly, the improved frequency resolution of AAC's and PAC's long block modes should result in higher bit rate reductions at times when the signal is relatively stationary while their short block modes provide them with a temporal footprint which is identical to that of AC-3 in situations where pre-echo is possible.

Both recent proposals for the ITU-T wideband speech standard (i.e., 7 kHz input bandwidth with bit rates of 16, 24, and 32 kb/s) also use overlapped transforms [26]. The AT&T proposal combines predictive coding with a 128-point MDCT resulting in a temporal footprint of 6 ms and a frequency resolution of 109 Hz (16 kHz input sampling rate). While this system has no difficulty with pre-echoes, its designers acknowledge that the short transform limits its rate-distortion performance for non-speech audio inputs. The PictureTel proposal, on the other hand, does not incorporate a speech generation model-- it relies entirely on auditory masking as discussed in Section 2. Here, a 320-point modulated lapped transform (MLT) is used, resulting in a temporal footprint of 30 ms and a frequency resolution of 22 Hz.

3.3 Hybrid Systems

A number of compression systems have also been proposed which use combinations of time-frequency analysis methods to implement their signal decompositions. The most notable of these are MPEG Audio Layer 3, Bell Lab's enhanced PAC (EPAC), and Sony's ATRAC for MiniDisc. All of these algorithms also use path **b** in Fig. 2 to adapt their decompositions so as to eliminate pre-echoes in the decoded audio. Layer 3 combines the fixed 32-band pseudo-QMF bank used in Layers 1 and 2 with either a 6 or an 18-point MDCT for increased frequency resolution within the subbands [7]. If the 6-point MDCT is used, then the temporal footprint of a coded block is about 8.7 ms (corresponding to 9 subband samples and including filter effects) while the frequency resolution is now 115 Hz (approximately the minimum width of a critical band). The 18-point transform, on the other hand, corresponds to blocks of 36 subband coefficients and results in a temporal footprint of approximately 22.2 ms with resolution of 44 Hz. Again, the decision on whether or not to switch between transforms is based on pre-echo considerations. In contrast, ATRAC uses a fixed, two stage QMF bank to nonuniformly partition the input audio into 3 subbands with frequency ranges of 0-5.5 kHz, 5.5-11 kHz, and 11-22 kHz [27]. The two lower frequency subbands are then transformed using either a 64 or 256-point MDCT, resulting in temporal footprints of 8.7 ms and 34.8 ms, respectively. The corresponding frequency resolutions are 345 Hz using the short transform and 86 Hz using the longer one. Because of the non-uniform nature of the initial QMF bank, the temporal footprint above 11 kHz can be either 2.2 or 17.4 ms (corresponding to M = 32 and M = 256, respectively) with frequency resolutions of 345 Hz and 43 Hz. Finally, EPAC uses a 1024-point MDCT during stationary periods but switches to a wavelet decomposition when an attack is detected [25]. Thus, its quantization noise can be more highly localized at higher frequencies-- exactly what is needed during an attack!

4. CONCLUSIONS

After examining a number of wideband audio and speech coders, one can only conclude that multirate filter banks have a had tremendous impact on the field. The key to their success is that they allow the coding algorithm to precisely control the localization of quantization errors in both time and frequency, thus facilitating psycho-acoustic error masking. While frequency localization varies considerably amongst the different algorithms, temporal localization is fairly consistent with minimum values of around 10 ms. This dichotomy is not surprising since the constraint dictating temporal error localization (backward masking) is far more stringent than the one dictating frequency error localization (simultaneous masking). Advanced audio coding algorithms are, however, clearly moving towards adaptive decompositions because they allow the encoder to achieve large bit rate reductions during stationary periods while effectively suppressing pre-echoes during attacks. Finally, the use of overlapped transforms in both of the ITU-T candidates clearly indicates that multirate filter banks still have a future in speech coding as well.

ACKNOWLEDGMENTS

The author would like to thank Dr. Nikil Jayant of Bell Laboratories and Joerg Kliewer of the University of Kiel, Germany for their invaluable comments and suggestions.

REFERENCES

[1] J.P. Princen and A.B. Bradley, “Analysis/synthesis filter bank design based on time domain aliasing cancellation,” *IEEE Trans. on Acoust., Speech, Signal Processing*, vol. ASSP-34, pp. 1153-61, Oct. 1986.

[2] J.H. Rothweiler, “Polyphase quadrature filters– a new subband coding technique,” Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, pp. 1280-83, 1983.

[3] P.L. Chu, “Quadrature mirror filter design for an arbitrary number of equal bandwidth channels,” *IEEE Trans. on Acoust., Speech, Signal Processing*, vol. ASSP-33, pp. 203-18, Feb. 1985.

[4] R.V. Cox, “The design of uniformly and nonuniformly spaced pseudo-quadrature mirror filters,” *IEEE Trans. on Acoust., Speech, Signal Processing*, vol. ASSP-34, pp. 1090-96, Oct. 1986.

[5] C.D. Creusere and S.K. Mitra, “A simple method for designing high-quality prototype filters for M-band pseudo QMF banks,” *IEEE Trans. on Signal Processing*, vol. 43, pp. 1005-7, April 1995.

[6] R.D. Koilpillai and P.P. Vaidyanathan, “A spectral factorization approach to pseudo-QMF design,” *IEEE Trans. on Signal Processing*, vol. 41, pp. 82-92, Jan. 1993.

[7] P. Noll, “MPEG digital audio coding,” *IEEE Signal Processing Magazine*, pp. 59-81, Sept. 1997.

[8] A. Hoogendorn, “Digital compact cassette,” *Proc. of the IEEE*, vol. 82, pp. 1479-89, Oct. 1994.

[9] P.P. Vaidyanathan, *Multirate Systems and Filter Banks*, Prentice-Hall, Englewood Cliffs, NJ, pp. 370-88, 1993.

[10] H.S. Malvar, “Modulated QMF banks with perfect reconstruction,” *Electronics Letters*, vol. 26, pp. 906-7, June 1990.

[11] T.A. Ramstad, “Cosine modulated analysis-synthesis filter bank with critical sampling and perfect reconstruction,” *Proc. IEEE Int. Conf. on Acoustics, Speech, Signal Processing*, pp. 1789-92, May 1991.

[12] R.D. Koilpillai and P.P. Vaidyanathan, “Cosine-modulated FIR filter banks satisfying perfect reconstruction,” *IEEE Trans. on Signal Processing*, vol. 40, pp. 770-83, April 1992.

[13] J.D. Johnston, “Transform coding of audio signals using perceptual noise criteria,” *IEEE J. Select. Areas Commun*., vol. 6, pp. 314-23, Feb. 1988.

[14] C.D. Creusere and S.K. Mitra, “Efficient audio coding using perfect reconstruction noncausal IIR filter banks,” *IEEE Trans. on Speech and Audio Processing*, vol. 4, pp. 115-23, March 1996.

[15] N. Jayant, J. Johnston, and R. Safranek, “Signal compression based on models of human perception,” *Proc. of the IEEE*, vol. 81, pp. 1385-422, Oct. 1993.

[16] P. Noll, “Wideband speech and audio coding,” *IEEE Communications Magazine*, pp. 34-44, Nov. 1993.

[17] P. Noll, “Digital audio coding for visual communications,” Proc. of the IEEE, vol. 83, pp. 925-43, June 1995.

[18] N.S. Jayant and P. Noll, *Digital Coding of Waveforms: Principles and Applications to Speech and Video*. Englewood Cliffs, NJ: Prentice Hall, 1984.

[19] R. Zelinski and P. Noll, “Adaptive transform coding of speech signals,” *IEEE Trans. on Acoust., Speech, Signal Processing*, vol. ASSP-25, pp. 299-309, Aug. 1977.

[20] R.E. Crochiere, S.A. Webber, and J.L. Flanagan, “Digital coding of speech in sub-bands,” *Bell Syst. Tech. J.*, vol. 55, pp. 1069-85, Oct. 1976.

[21] A. Gersho and R.M. Gray, *Vector Quantization and Signal Compression*. Boston, MA: Kluwer, 1992

[22] E. Zwicker and H. Fastl, *Psychoacoustics*. Berlin: Springer-Verlag, 1990, pp. 141-147.

[23] *Digital Audio Compression Standard (AC-3)*, published by the U.S. Advanced Television Systems Committee (ATSC), Document A/52, available from website http://www.atsc.org.

[24] M. Bosi et al., “ISO/IEC MPEG-2 audio multi-channel encoding,” *101th Audio Engineering Society Conv*., preprint 4382, Los Angeles, 1996.

[25] *Digital Signal Processing Handbook*, eds. V.K. Madisetti and D.B. Williams, IEEE-CRC, 1998, Chapter 42.

[26] From: ftp://standard.pictel.com/sq16_q20/codec_details.

[27] K. Tsutsui *et al*., “ATRAC: Adaptive transform acoustic coding for MiniDisc,” in *Conf. Rec., Audio Eng. Soc. Conv*., San Francisco, CA, Oct. 1992.