Embedded Zero-tree Wavelet Packet Audio Coding with Psychoacoustic Modeling

Pao-Chi Chang (張寶基), Mei-Tai Chou (周美代), and Ching-Ming Huang (黃景明)

Department of Electrical Engineering
National Central University
Chung-Li, Taiwan
e-mail: pcchang@ee.ncu.edu.tw

Abstract

We propose a perceptual audio coding system that divides audio segments into 29 subbands via wavelet packet analysis, and utilizes a modified zero-tree coder based on the minimum masking thresholds generated by the psychoacoustic model.

The wavelet filter bank analysis-synthesis technique has been widely applied in many areas of digital signal processing, including audio and video coding. The embedded zero-tree wavelet (EZW) coding has shown its great performance in progressive image coding. In this work, we focus on high quality audio coding which delivers transparent perceptual quality. The segmented audio signal is divided into 29 subbands via wavelet packet analysis and then coded by a zero-tree coder with the modified algorithm based on the minimum masking thresholds which are generated by the psychoacoustic model. Compared with MPEG audio Layer II, the Masking-Embedded Zero-tree Wavelet Packet (M-EZWP) system we propose has better performance, especially in the case of very low bitrate. The perceptual transparent quality of monophonic audio can be achieved at about 40 Kbps. Furthermore, because of the embedded property, the M-EZWP system could be adjusted to various network conditions, including VBR and CBR transmissions.

Index Terms : Audio coding, Wavelet packet, Embedded zerotree, Psychoacoustic modeling

I. Introduction

High quality audio coding is indispensable in present multimedia applications, such as audio and video services over Internet or wireless communications. High quality, low bitrate, and affordable complexity are demanded for good and efficient audio compression. Subband-based coders are the most popular methods for current high-quality audio coding. They differ from each other in the ways of partitioning the frequency scale and of quantizing or coding information in each band [1][2][3][4][5]. For example, MPEG-1 audio coding uses filterbanks with 32 uniform subbands and bit allocation for scalar quantization by psychoacoustic model [6]. However, the uniform subband is not exactly coincided with the property of perceptual model. To avoid this defect, wavelet transform, a new analysis tool, is used. In 1995, Karelic and Malah proposed a wavelet-packet based zero-tree coder which was superior to MPEG Layer I [7]; however, the method took no account of the psychoacoustic model. In 1998, an audio coding system with adaptive wavelet packet decomposition and psychoacoustic modeling was designed by Srinivasan and Jamieson [8].

To combine the psychoacoustic model with a zero-tree coder, we propose a masking-embedded zero-tree wavelet packet (M-EZWP) system. Its structure including encoder and decoder is shown in Fig.1. The encoder consists of three major parts, filter banks, psychoacoustic model, and EZW coder. The input audio signal is decomposed into 29 bands by the filter banks of wavelet packets; minimum masking thresholds for each band are generated through psychoacoustic model; and then the wavelet coefficients in 29 bands are

EZW coded with masking thresholds embedded.

The general aspects of discrete wavelet transform (DWT) and the wavelet packets decomposition structure of our system are expounded in section II. The overview of psychoacoustic model is presented in section III. Then, the algorithm of masking-embedded zero-tree coding is detailed in Section IV. The simulation results are discussed in section V. Finally, conclusions are given in section VI.

II. Wavelet Packets

The discrete wavelet transform provides a set of building blocks for representing signals or functions [9]. The general formula is as following in which a signal $g(t)$ can be represented by expansion bases that are formed by scaling functions and wavelet functions.

$$g(t)=\sum_k c_{j_0}(k)2^{\frac{j_0}{2}}\varphi(2^{j_0}t-k)+\sum_k\sum_{j=j_0}^{\infty}d_j(k)2^{\frac{j}{2}}\psi(2^j t-k) \quad ...(1)$$

$\varphi(t)$: scaling function, $c_j(k)$: scaling coefficient

$\psi(t)$: wavelet function, $d_j(k)$: wavelet coefficient

where j_0 is the initial scale which may be zero or any integer, j is the scale of resolution, and k is the translation step.

The multiresolution property of wavelet transform is suitable for audio signal processing. The relationship between finer and coaser coefficients may be conducted as filtering formulas. If the sampling rate of a signal is higher than Nyquist rate, the scaling function will behave as an impulse function and the digital signal will be a good estimation of the high resolution coefficients. Thus, we may employ filter banks for a series of signal analyses.

The wavelet packet system, proposed by Coifman [10], allows a finer and adjustable resolution of frequencies at high frequencies. It also gives a richer structure that allows various adaptations to particular signals or signal classes.

To achieve efficient compression, the bandwidth of each subband should match that of critical bands as closely as possible. Therefore, in our system, the decomposition structure shown in Fig 2 is chosen [11]. The comparison of the bandwidth of each subband and that of each critical band is illustrated in Fig. 3. As for wavelet filters, M-EZWP system uses the biorthogonal 18-tap FIR filter, which has linear phase property [12].

Because the local characteristics of audio signals may be significantly different, it is better to divide the audio signal into short segments to get good compression. The segment consists of 1024 samples, which corresponds to 23.32 ms at the sampling rate of 44.1KHz. Input signals are fed in the filters continuously to solve the problems of delay and of the undesired edge effect.

III. Psychoacoustic Modeling

Human auditory system (HAS) [13] has many useful features in audio compression. Psychoacoustic model makes it practical via a series of mathematical formulas. By utilizing the psychoacoustic model, the minimum masking threshold of each subband is obtained and the distortion can be inaudible as long as it is below the minimum masking threshold.

M-EZWP system uses the same model as MPEG psychoacoustic model II. The noise-masking thresholds for the critical bands are calculated via a 1024-point FFT. The measure of tonality, whose value ranges from 0 to 1, is based on the predictability of the current frame from the past two frames. The spreading function describes the property of the ear response to mask noise at a frequency of neighboring a tone. Then, the “just masked” noise level, the minimum masking threshold, is calculated from the spreading function and the tonality index. The absolute threshold of hearing (ATH) and pre-echo control are also incorporated. Finally, the minimum threshold for each subband is extracted. Fig. 4 shows different audio segments of the flute signal from psychoacoustic model.

IV. Embedded Zero-tree Coding

Shapiro proposed the wavelet transform based embedded zero-tree image coding in 1993 [14]. He succeeded in designing a low bitrate and high quality image coding system. The "self-similarity" of wavelet coefficients is the key of making zero-tree algorithm efficient on coding significant maps. In the case of audio signal, wavelet packet coefficients of some subbands are highly correlated, too. Therefore, zero-tree coding has great potential to be used in audio coding. Furthermore, the perceptual characteristics also support the embedded property.

In the significant map decision of zero-tree algorithm, the coefficients are organized in a tree structure and classified into four types, POS,

NEG, IZ and ZTR. POS represents a positive significant coefficient while NEG is negative significant one. IZ stands for isolated zero which means itself is insignificant but with significant descendents. ZTR is a zero-tree root which indicates itself and all its descendents are insignificant. The threshold of each iteration is half of the each previous one. The algorithm terminates if the bit rate checked between two iterations reaches the target rate.

The first step of doing zero-tree scanning is to determine the tree structure of coefficients, i.e., the relationship of ancestors and descendents. Chain-tree, shown in Fig. 5(a), is the straightforward idea to build a tree from low frequency bands to high frequency bands. To search a better tree structure, we consider the harmonics of audio signal. According to the characteristics of most instruments, except for the lowest frequencies, if one coefficient is assumed insignificant, there is a relatively high probability that the coefficients in its harmonics are also insignificant. Therefore, the harmonic full-tree is set up as in Fig. 5(b), where all descendents are harmonics of ancestors. However, we observe that the energy of wavelet packets coefficients concentrates in lower, but not the lowest, bands. To make ZTR generated more easily, we separate the full-tree into four sub-trees and let high-energy band be the root of the sub-tree. Fig. 5(c) shows the harmonic sub-trees.

To cooperate with the psychoacoustic model, we make a modification in the significance decision rule of EZWP. As we mentioned in last section, each subband has a minimum masking threshold M_i. The wavelet packets coefficient C is significant only when it is larger than both threshold T and masking value M_i. The flow chart is shown in Fig. 6. Consequently, depending on the relation between T and M_i, the four symbols of zero-tree algorithm are determined as the two cases in Fig. 7.

Successive approximation quantization (SAQ) sequentially applies a sequence of thresholds to determine the significance. To all transform coefficients, the initial threshold T_0 is chosen such that $T_0 > \max \left|X_j\right|/2$, for all j. Dominant pass and subordinate pass are the two passes for SAQ. For each threshold, both dominant list and subordinate list are supposed to be scanned once. The dominant list contains the coordinates of those coefficients that have not yet been found to be significant. The subordinate list contains the magnitude of those coefficients that have been found significant. The dominant pass aims to determine the significant map. When POS or NEG occurs, its value is put to the subordinate list. Decoder sets its magnitude to 1.5T while receiving POS or NEG. The subordinate pass aims to determine the value list. The output is 1 when its value is on the upper half of uncertainty area. Otherwise, the output is 0. Decoder adds T/4 while getting 1, and subtracts T/4 while getting 0, respectively.

The transmission scheme of the significant maps and value lists for one frame is shown in Fig. 8. The order counts on the significance, which is amenable to achieve progression. Each frame has 4 sub-frames. Initial threshold is transmitted first. Then significant maps and value lists respectively follow. They are transmitted in the horizontal order while the transmission granularity can be a bit, a byte, or any unit. Since each sub-frame may have different statistics, the encoded length may vary. Thus we put an END-mark, which is an unique symbol, to each individual sub-frame at the end of each threshold scan to notify the decoder to accurately track the sequence with EZW maps. Both start-code and end-code are sent for synchronization. Finally, frame end code, which concludes a frame, is sent.

V. Simulation Results

We perform simulations in C language on Sparc 20 workstation. Each monophonic audio test signal is 5 seconds long is of CD quality. Namely, the sampling rate is 44.1 KHz with 16 bits per sample. The evaluation of audio quality is in accordance with segmental signal to noise ratio (SSNR) objectively or with listening test subjectively.

Depending on the way to terminate the zero-tree search, there exist two methods to control the bitrate of encoder. One is adjusting the masking values, and the other is giving a target bitrate. The first method is carried out by dividing the masking threshold by a value X, and the system is indicated as M-EZWP 1/X, where X can be any positive number. Since the masking thresholds calculated from the psychoacoustic model are based on FFT coefficients, which have poor time resolution, and are based on a population average, a smaller masking value may actually improve the quantization quality perceptually. The lower the masking threshold is adjusted, the more iterations that the zero-tree algorithm undertake, and the more accurately the signal is coded. Hence, the bitrate of each frame is changed with the masking thresholds of each frame . Fig. 9 shows an example of organ signal. It is found

that the bitrate changes with the signal energy. On the other hand, the second method is to terminate the zero-tree algorithm by bitrate check, which is similar to what the original EZW proposed. With the support of these two methods, both VBR with constant quality and CBR transmission schemes are surely achievable.

To find out the best tree structure of zero-tree search, we use the first method mentioned above to obtain the equal quality reconstruction and make a judgement according to the bitrates. The results of different masking conditions are separately listed in Table I and Table II. We can observe that the harmonic sub-tree structure yields the lowest bitrate for most instruments and it is unable to distinctly distinguish the winning one from the other two tree structures. In addition, the bitrate differences between chain-tree and full-tree are not so obviously as sub-tree is with them. Therefore, it is concluded that the harmonic sub-tree structure is the most efficient one.

Listening test was performed to compare the performance of M-EZWP with that of MPEG Layer II. Everyone who took the test and chose the preference was not told the playing sequence of audio signal. The tested audio got one score if it was chosen or half score if the listener failed to make a choice. The total scores was ultimately averaged to get the preference ratio. Then two different bitrate schemes were examined. The result is shown in Table III. It shows that the proposed system has almost the same perceptual quality with MPEG Layer II for its averaged ratio is near 50%.

VI. Conclusion

We have presented M-EZWP audio coding system that is based on wavelet packets subbands and the psychoacoustic model with an embedded zero-tree coder. The coding efficiency has been improved by the harmonic-based tree structure of subbands. Subjective listening test also shows that M-EZWP has almost the same quality as MPEG Layer II standard. The perceptual receptive quality is achieved at about 40 Kbps. In addition, the system provides two different bit-rate control schemes and can be easily adapted to CBR and VBR transmission channels.

References

[1] S. Boland and M. Deriche, “Audio Coding Using The Wavelet Packet Transform and A combined Scalar-Vector Quantization,” in *Proc. Int. Conf. Acoust., Speech, Signal Process. 1996*, pp. 1041-1044.

[2] X. Xiong and Z. Eryuan, “Digital Audio Codec Based on the Improved Optimization Algorithm of Adaptive Wavelets and Dynamic Bit Allocation Scheme,” *proceeding of ICSP’96*, pp. 1523-1526.

[3] P. Philippe, F. Moreau de Saint-Martin, M. Lever, and J. Soumagne, “Optimal Wavelet Packets for Low-Delay Audio Coding,” in *Proc. Int. Conf. Acoust., Speech, Signal Process. 1996*, pp. 550-553.

[4] He Dongmei, Gao Wen and Wu Jiangqin, “Complexity scalable audio coding algorithm based on wavelet packet decomposition,” *Proceedings of WCCC-ICSP 2000*, pp. 659-665.

[5] N. Ruiz, M. Rosa, F. Lopez, D. Martinez and R. Mata, “New algorithm for searching minimum bit rate wavelet representations with application to multiresolution-based perceptual audio coding,” *Proceedings of Pattern Recognition, 2000*, pp. 286 –289.

[6] ISO/IEC 11172-3:1993 Information technology - "Coding of moving pictures and associatedaudio for digital storage media at up to about 1.5 Mbit/s - Part 3: Audio".

[7] Y. Karelic and D. Malah, “Compression of High-Quality Audio Signals Using Adaptive Filterbanks and A Zero-Tree Coder,” *Electrical and Electronics Engineers in Israel*, 1995.

[8] P. Srinivasan and L. H. Jamieson, “High-Quality Audio Compression Using an Adaptive Wavelet Packet Decomposition and Psychoacoustic Modeling,” *IEEE Trans. on Signal Processing*, vol. 46, no. 4, pp. 1085-1093, April 1998.

[9] C. S. Burrus, R. A. Gopinath, and H. Guo, “Introdution to Wavelets and Wavelet Transforms,” 1998.

[10] R. R. Coifman and M. V. Wickerhauser, "Entropy-based algorithms for best basis selection," *IEEE Trans. Information Theory*, vol. 38, pp. 713-718, March, 1992.

[11] D. Sinha and A. H. Tewfik, “Low Bit Rate Transparent Compression using Adapted Wavelets,” *IEEE Trans. on Signal Processing*, vol. 41, no. 12, pp. 3463-3479, Dec. 1993.

[12] I. Daubechies, "Ten Lectures on Wavelets," no. 61 in CBMS-NSF Series in

Applied Mathematics, SIAM, Philadelphia, 1992.

[13] E. Zwicker and H. Fastl, Psychoacoustics, Facts and Models (Springer, Berlin, Heidelberg, 1990).

[14] J. M. Shapiro, "Embedded Image Coding Using Zerotrees of Wavelet Coefficients," *IEEE Trans. Signal Processing, Spec. Issue Wavelets Signal Processing*, vol. 41, pp. 3445-3462, Dec. 1993.

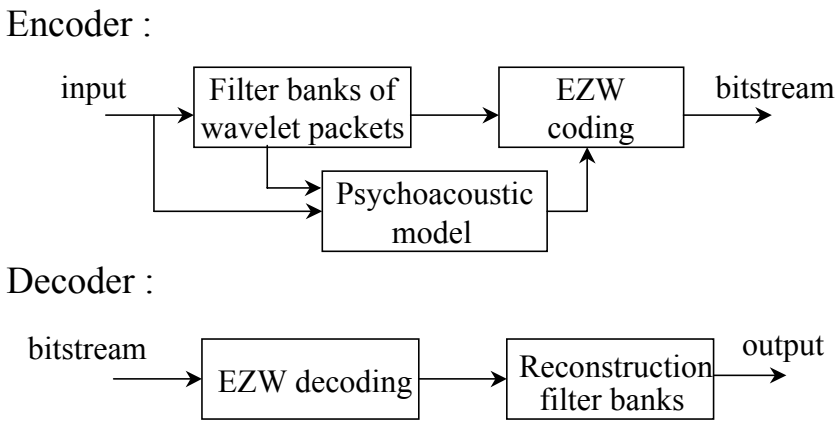


Fig.1 Block diagram of M-EZWP system

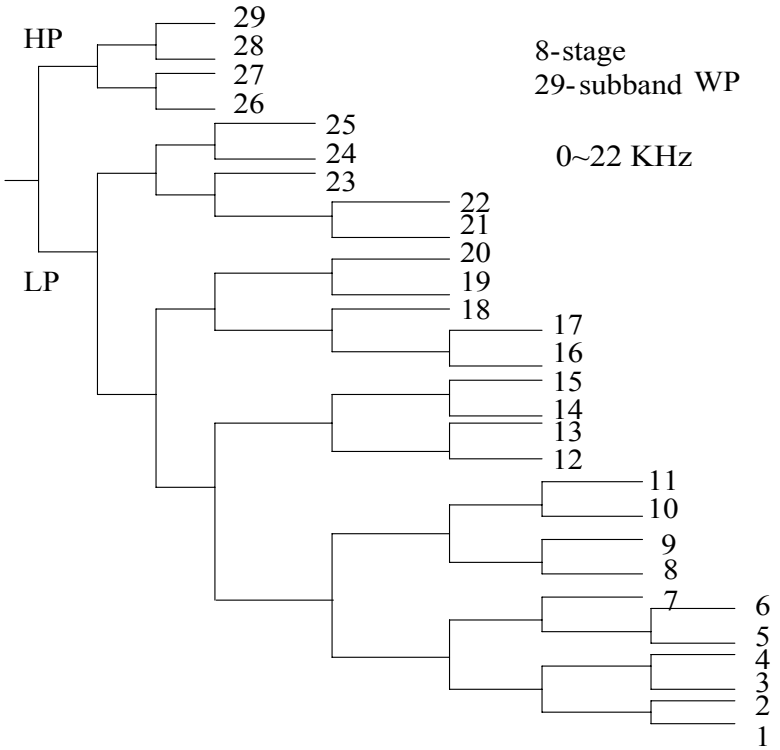


Fig. 2 Decomposition structure of M-EZWP system

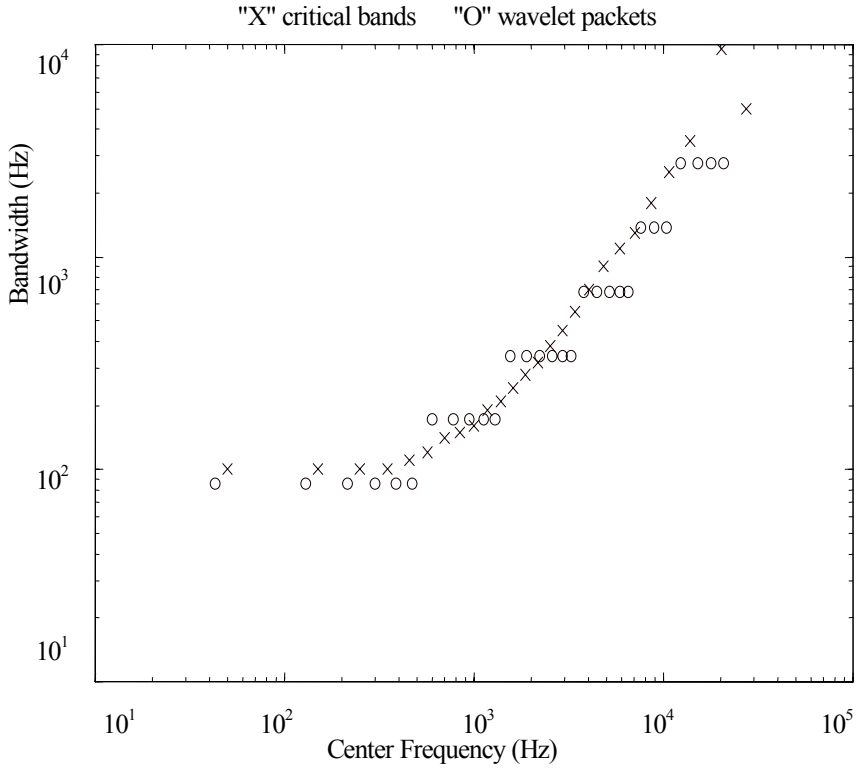


Fig. 3 Bandwidths of wavelet-packet subbands vs. critical bands

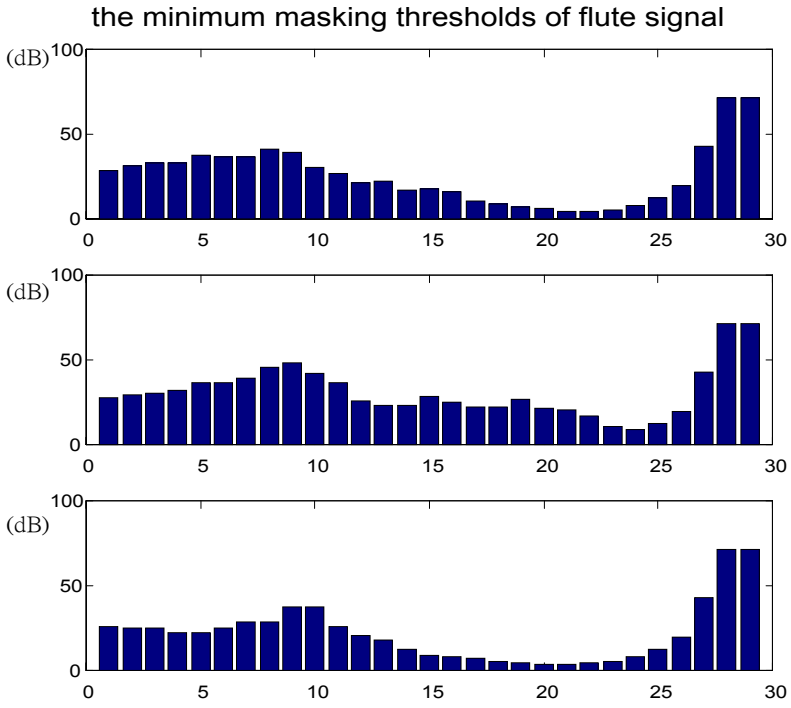
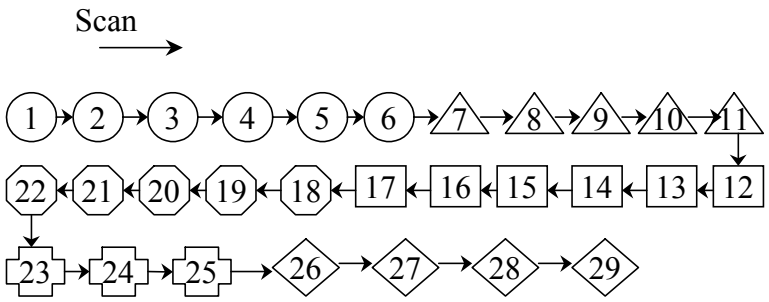
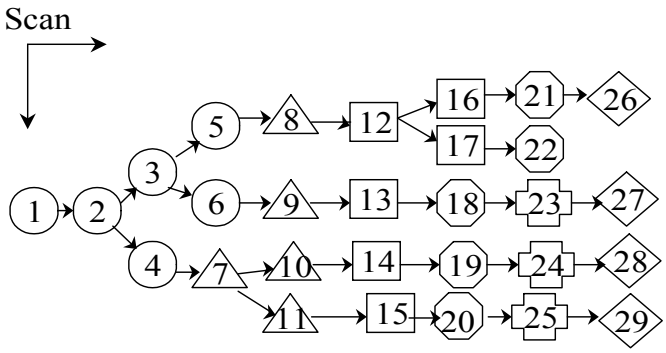


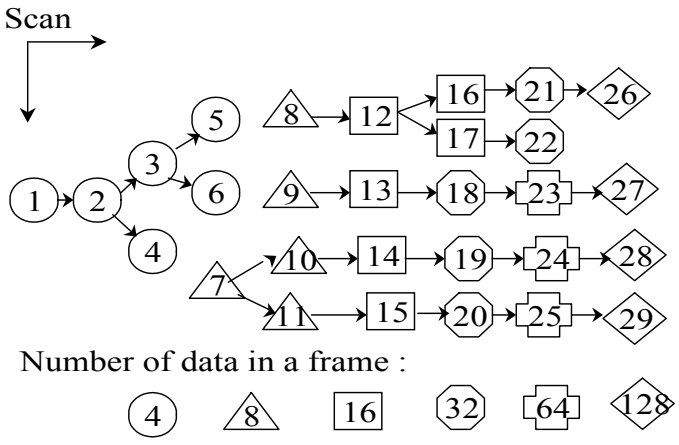
Fig. 4 Minimum masking thresholds for subbands of flute signal (2nd, 42th, 82th frame)



5(a)



5(b)



5(c)

Fig. 5 Three coefficient tree structures and scanning orders. (a) Chain-tree structure,

(b) Harmonic full-tree structure, (c) Harmonic sub-tree structure.

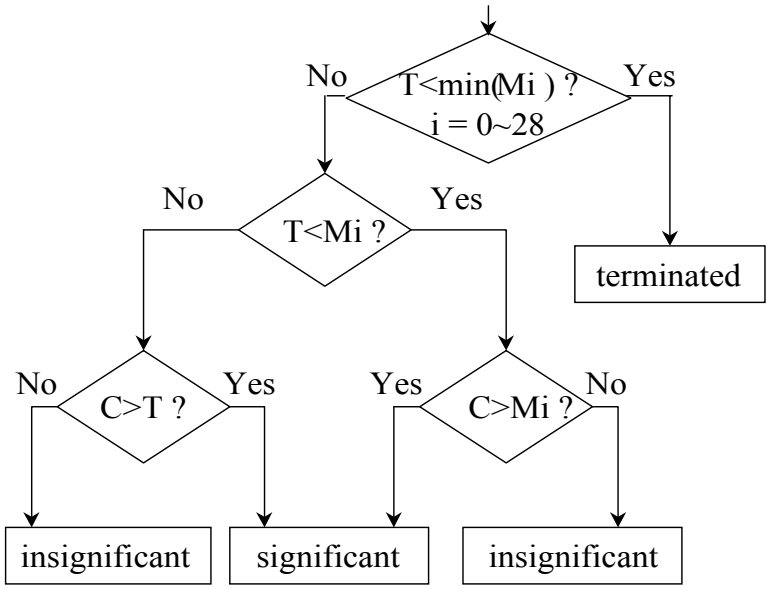


Fig. 6 Flow chart of masking combined significance decision rule

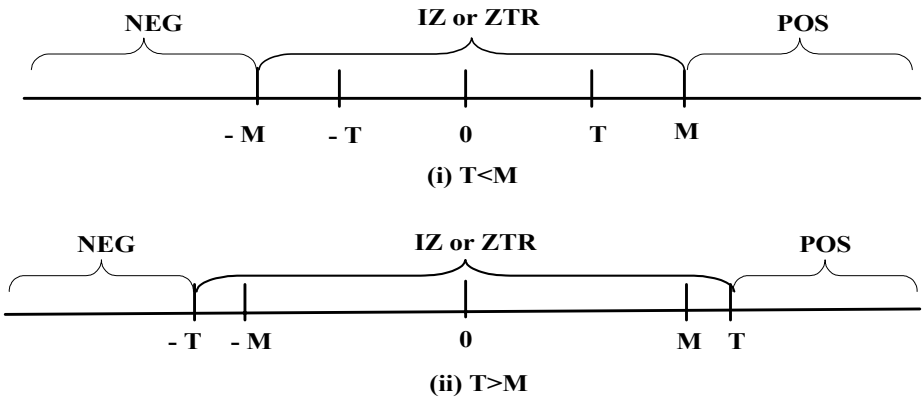


Fig. 7 Two cases in symbol encoding

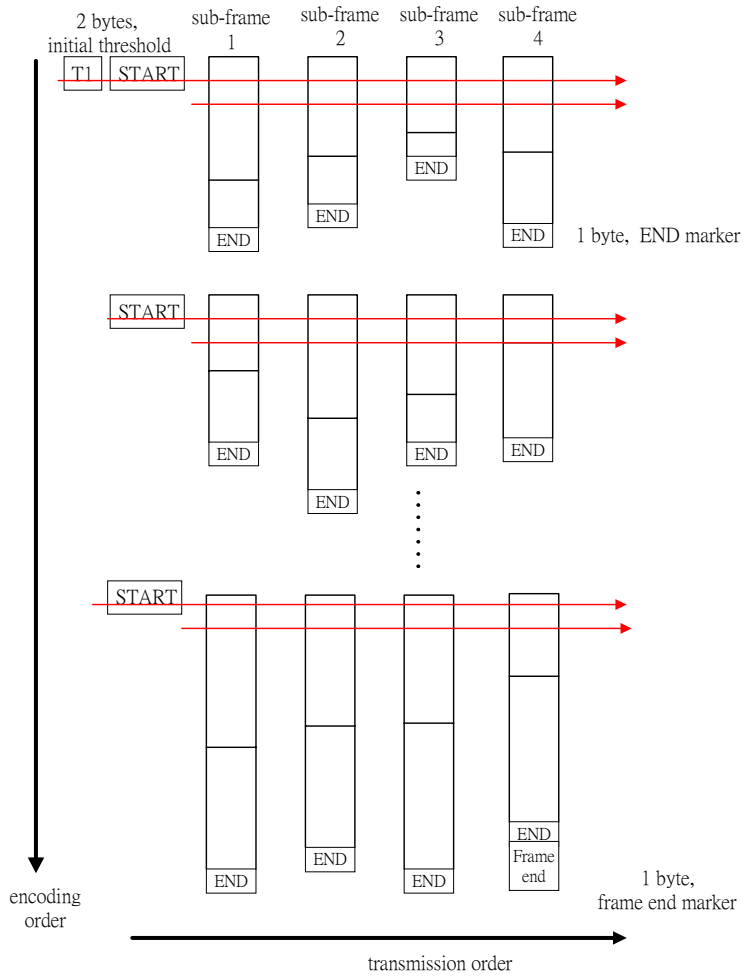


Fig. 8 Interleaving transmission of M-EZWP system

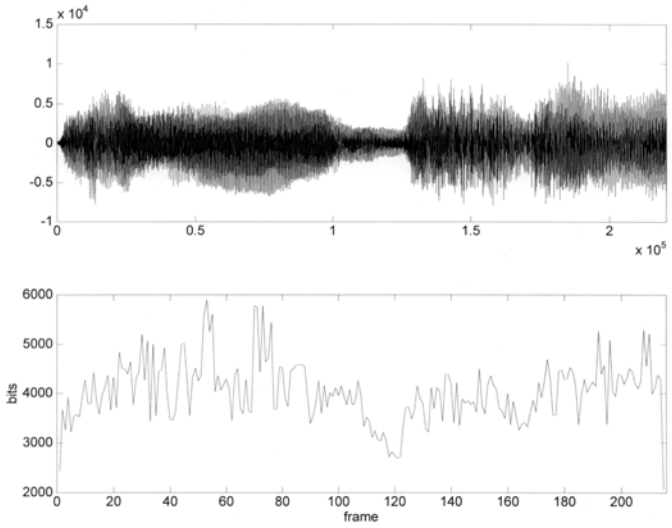


Fig. 9 Organ waveform and number of coding bits of each frame (M-EZWP 1/3)

Table I: bitrate comparison of various tree types (M-EZWP 1/5)

		chain-tree	full-tree	sub-tree
	SSNR (dB)	BR (Kbps)	BR (Kbps)	BR (Kbps)
Flute	22.59	137.74	133.41	129.15
Horn	28.77	161.23	171.37	166.92
Piano	26.44	190.84	191.14	184.73
Guitar	26.18	155.24	152.68	148.36
Harp	23.85	189.51	195.72	188.56
Violin	27.62	175.97	178.63	171.13
Cello	28.57	144.57	144.93	140.96
Organ	26.85	200.26	207.59	201.32
Trumpet	19.80	135.97	133.95	128.76
Chorus	23.23	197.81	199.22	191.40
Orchestra	20.52	208.18	208.46	200.73

Table II: bitrate comparison of various tree types (M-EZWP 1/4)

		chain-tree	full-tree	sub-tree
	SSNR (dB)	BR (Kbps)	BR (Kbps)	BR (Kbps)
Flute	20.71	122.39	114.56	107.49
Horn	27.13	148.53	154.01	132.06
Piano	23.81	179.59	173.48	158.66
Guitar	23.42	142.23	135.36	124.91
Harp	22.32	176.27	176.55	154.88
Violin	25.10	164.73	160.30	145.02
Cello	25.60	125.85	122.15	110.99
Organ	24.12	187.86	187.79	165.80
Trumpet	17.88	122.53	116.86	108.26
Chorus	20.65	186.82	179.59	164.11
Orchestra	18.03	194.99	185.41	171.56

Table III: preference ratios (%) of M-EZWP to MPEG audio Layer II

Preference ratio (%)	64Kbps	128Kbps
	M-EZWP 1/5	M-EZWP 1/10
Flute	50.50	62.50
Horn	41.67	45.83
Piano	45.83	41.67
Guitar	50.50	33.33
Lute	45.83	66.67
Harp	41.67	41.67
Violin	50.50	62.50
Cello	41.67	54.17
Organ	33.33	58.33
Trumpet	50.50	66.67
Chorus	33.33	50.50
Orchestra	33.33	33.33
Average	43.22	51.43