WAVELET BASED PITCH EXTRACTION IN THE MPEG COMPRESSED DOMAIN

B.Anantharaman, K.R.Ramakrishnan, S.H.Srinivasan

Department of Electrical Engineering, Indian Institute of Science, Bangalore, India

ABSTRACT

A technique to extract pitch information directly from audio files encoded using MPEG/Audio coding standard is described. The technique works in the compressed domain and requires the MPEG/Audio file to be decoded only partially for extracting the encoded subband samples. The paper first proposes a method for extracting wavelet coefficients from the subband samples. The pitch interval is then estimated from the time interval between two successive maxima of the wavelet coefficients. It is shown that the computational complexity for compressed domain pitch extraction is less than 7% of the computations required for decoding the subband samples and finding the pitch.

1. INTRODUCTION

Thanks to the MPEG set of standards , which has become the enabling technology for many of today's multimedia applications, video libraries have become commonly available and grown in size. This has created a dire need for automatic and consistent indexing of the video in terms of its contents. The audio which accompanies the video contains enormous amount of information that can be used to index video for domain specific applications.

In applications where video need to be categorised in terms of gender or specific speaker, pitch is used as one of the important features. In this paper we describe a technique to extract pitch information from MPEG compressed audio bitstream. In a typical MPEG audio bitstream,the subband samples are quantised and allocated bits depending on a psychoacoustic model. In layer I and II &III, 12 and 36 samples respectively are taken from each of the subbands and packed into frames. We first extract wavelet coefficients by only partially decoding the required subband samples and use them to find out the pitch.

2. WAVELET BASED PITCH EXTRACTION

Mallat [5] has shown that if a wavelet function is chosen which is a derivative of smoothing function, then the local maximum of the wavelet coefficients indicate sharp variations in a signal whereas the local minima indicates slow variations in the signal. Further Mallat has shown that sharp changes occurring in a signal exhibit local maximum across several consecutive dyadic scales. This property has been used in [3] to find the instants at which glottal closure occurs, as glottal closures are accompanied by sharp variations in speech signal. Pitch is basically the time period between two successive glottal closures. The time instant at which the glottal closure occurs (onset of pitch) is found by detecting local maxima in the wavelet coefficients. Hence pitch can be found by calculating the time interval between two successive maxima.

Fig.1 shows the original speech signal with various levels of wavelet coefficients. It is observed that the onset of pitch(and hence the local wavelet maxima) is clearly trackable at the seventh level of wavelet coefficients than in the lower levels. Based on Mallat's Theory and the fact that the pitch lies in the low frequency range one can find the time interval between two successive maxima of the level seven wavelet coefficients. It has been reported in [4] that cubic spline wavelets outperform other wavelets in pitch extraction. Hence we have used cubic spline wavelets in our experiments. It has also been reported in [3] that wavelet based pitch extraction outperforms conventional methods like autocorrelation and cepstrum based pitch extraction.

3. EXTRACTION OF WAVELET COEFFICIENTS FROM MPEG ANALYSIS FILTER BANK

The filtering of a signal x[n] of length N by a filter h[n] of length N can be represented in matrix form by

$$\mathbf{Y} = \mathbf{H}\mathbf{X} \tag{1}$$

where

$$\mathbf{H} = \begin{bmatrix} h[0] & 0 & \cdots & 0 \\ h[1] & h[0] & \cdots & 0 \\ h[2] & h[1] & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ h[N-1] & h[N-2] & \cdots & 0 \\ 0 & h[N-1] & \cdots & h[1] \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & h[N-1] \end{bmatrix}$$

$\mathbf{Y} = [y[0], y[1], y[2].....y[2N-2]]^T$ is the output sequence,

$\mathbf{X} = [x[0], x[1], x[2].....x[N-1]]^T$ is the input sequence,

and $\{h[0], h[1], h[2].....h[N-1]\}$ are the impulse response coefficients. $\mathbf{H}$ is of size $2N \times N$

2001 IEEE International Conference on Multimedia and Expo
ISBN 0-7695-1198-8/01 $17.00 © 2001 IEEE

0-7695-1198-8/01/$10.00 (C) 2001 IEEE

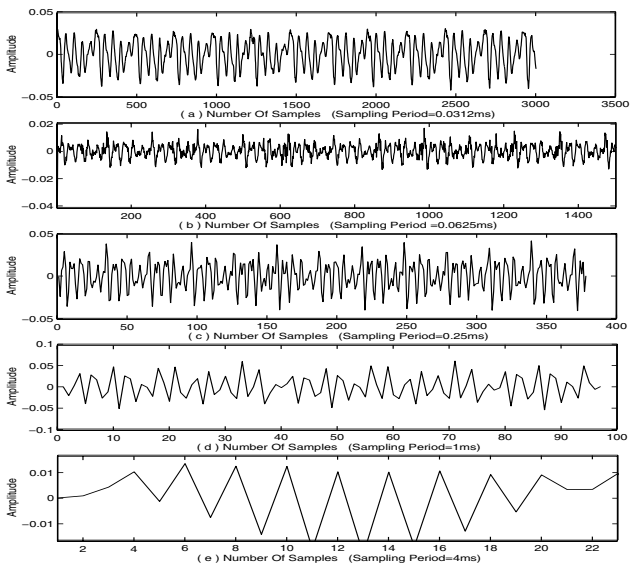

Fig. 1. *(a) Utterance of the word /a/. (b) Level One Wavelet Coeffs (c) Level Three Wavelet Coeffs. (d) Level Five Wavelet Coeffs (e) Level Seven Wavelet Coeffs.*

The filtering of a signal x[n] of length N by a filter h[n] of length N followed by decimation by factor M can be represented in a matrix form by

$$\mathbf{Y}_M = \begin{bmatrix} h[0] & 0 & \cdots & 0 \\ h[M] & h[M-1] & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ h[N-M] & h[N-M-1] & \cdots & 0 \\ 0 & h[N-1] & \cdots & h[1] \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & h[N-M+1] \end{bmatrix} \mathbf{X}$$

or $\mathbf{Y}_M = \mathbf{H}_M \mathbf{X}$ (2)

where $\mathbf{H}_M$ is a $\frac{2N}{M} \times N$ matrix derived from $\mathbf{H}$ by taking every M^{th} row. The input/output sequence of the i^{th} MPEG analysis filter in Fig.2 can be expressed as

$\mathbf{x}_i = \mathbf{H}_M^i \mathbf{X}\ (i = 1, 2...M)$. Hence the MPEG analysis filter bank can be represented by

$\mathbf{A} = \begin{bmatrix} \mathbf{H}_M^1 \\ \mathbf{H}_M^2 \\ \vdots \\ \mathbf{H}_M^M \end{bmatrix}$. The MPEG synthesis filter bank can be represented by $\mathbf{S} = [\mathbf{G}_M^1\ \mathbf{G}_M^2 .....\mathbf{G}_M^M]$ where, $\mathbf{G}_M^i (i = 1, 2...M)$ is a $2N \times \frac{N}{M}$ matrix and is obtained by taking every M^{th} column in the matrix representation of filtering operation as shown in(1) and $\mathbf{S}$ is of size $2N \times N$ [6]. In the case of MPEG filters, decimation factor M=32 and number of impulse response coefficients of each filter,N=512.

Extraction of the seventh level wavelet coefficients after synthesis is depicted in Fig.2 . This involves applying the lowpass and highpass wavelet filters to the decoded output

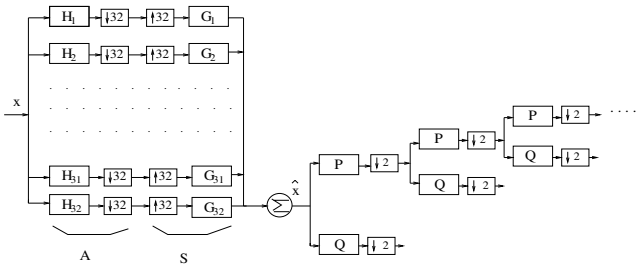

Fig. 2. *Extracting Wavelet Coefficients after MPEG decoding.A-MPEG analysis filter bank;S-MPEG synthesis filter bank;P-Lowpass Wavelet Filter;Q-Highpass Wavelet Filter.*

in a tree structured manner . The extraction of seventh level wavelet coefficients after synthesis can be represented in a compact form as shown in Fig.3

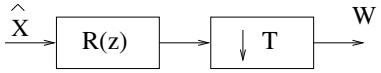

Fig. 3. *A branch of the wavelet tree decomposition for extracting the seventh level wavelet coefficients.*

where,

T=128.(for level seven)

R(z) = P(z)P(z^2)P(z^4)P(z^8)P(z^{16})P(z^{32})Q(z^{64})

P(z) is the Z transform of the lowpass wavelet filter

Q(z) is the Z transform of the highpass wavelet filter

Let r[n] represent the filter coefficients of R(z).The length of r[n] is made equal to that of the MPEG filters,by padding appropriate number of zeros.The operation in Fig.3 can be represented by

$$\mathbf{W} = \mathbf{R}_T \widehat{\mathbf{X}} \quad (3)$$

where $\mathbf{R}_T$ is the matrix representation of filtering of decoded samples $\widehat{\mathbf{X}}$ with filter R followed by decimation by T. When the decoded output $\widehat{\mathbf{X}}$ is written in terms of the input $\mathbf{X}$, (3) can be reduced to a form given by,

$$\mathbf{W} = \mathbf{R}_T \mathbf{S} \mathbf{A} \mathbf{X} \quad (4)$$

Let $\mathbf{W}_{sb}$ be the equivalent filters used in the subband domain,such that the seventh level wavelet coefficients are directly obtained from the output of the MPEG analysis filter bank. Then,

$$\mathbf{W} = \mathbf{W}_{sb} \mathbf{A} \mathbf{X} \quad (5)$$

From (4) and (5)

$$\mathbf{W}_{sb} = \mathbf{R}_T \mathbf{S} \quad (6)$$

3.1. Structure Of $\mathbf{W}_{sb}$

$\mathbf{W}_{sb}$ (6) is a $\frac{2N}{T} \times N$ matrix which consists of M, $\frac{2N}{T} \times \frac{N}{M}$ concatenated submatrices, which can be represented in the following form,

$$\mathbf{W}_{sb} = [\mathbf{WSb}_4^1\ \mathbf{WSb}_4^2, ... \mathbf{WSb}_4^M] \quad (7)$$

2001 IEEE International Conference on Multimedia and Expo
ISBN 0-7695-1198-8/01 $17.00 © 2001 IEEE
0-7695-1198-8/01/$10.00 (C) 2001 IEEE

Here $\mathbf{WSb}_4^i$ $(i = 1, 2...M)$ is a $\frac{2N}{T} \times \frac{N}{M}$ matrix and represents filtering operation followed by decimation by 4. Due to decimation by 4 version of the toeplitz structure shown by $\mathbf{WSb}_4^i$ $(i = 1, 2, ..M)$ matrices, convolution followed by decimation by 4 can be performed instead of matrix multiplication to get the level seven wavelet coefficients. Hence the method for calculating the seventh level wavelet coefficients from the output of the MPEG analysis filters is as follows:

Let $\mathbf{x}_i$ (i=1,2,. . . .32) each of arbitrary length (say N) represent the output of the MPEG analysis filters $\mathbf{H}_M^i$ respectively. Hence the wavelet coefficients $\mathbf{W}$ are obtained by

$W[n] = \sum_{i=1}^{32} \sum_{k=0}^{N-1} x_i[k] WSb^i[4n-k]$

where $\mathbf{WSb}^i = [\, WSb^i[0], WSb^i[1].....WSb^i[24]\,]$ represents a 25 length filter that can be inferred from $\mathbf{WSb}_4^i$. Since the contribution of MPEG analysis filters H_3 to H_{32} to the seventh level wavelet coefficients is expected to be small (because the frequency range spanned by the filters $H_3 - H_{32}$ is not included in the frequency range spanned by level seven wavelet coefficients) we can approximate by considering only the outputs of H_1 and H_2 and convolving with $\mathbf{WSb}^1$ and $\mathbf{WSb}^2$ respectively followed by decimation by 4 as shown in Fig. 4. Hence we have

$$\widehat{W}[n] = \sum_{i=1}^{2} \sum_{k=0}^{N-1} x_i[k] WSb^i[4n-k] \quad (8)$$

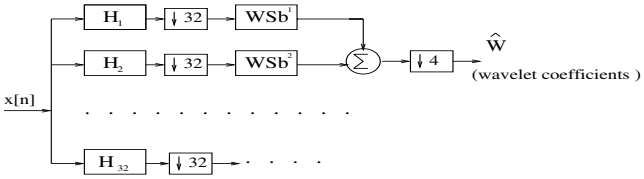

Fig. 4. *Using the first two MPEG filter bank outputs for calculating the seventh level wavelet coefficients.*

The coefficients extracted after synthesis (as shown in Fig.2) and from the output of the MPEG analysis filter bank(as shown in Fig.4) is shown in Fig.5.

However decimation by 4 as given in (8) is not performed as it has been observed that,clear pitch periods are obtained without decimation as shown in Fig.6.

3.2. Analysis of Computational Complexity

This analysis compares the number of multiplications required to extract the wavelet coefficients required for pitch extraction by a) Decoding the signal and subsequently finding the wavelet coefficients (Uncompressed domain processing) b) Directly from the subband output. (Compressed domain Processing)

Uncompressed domain pitch extraction :

Number of multiplications required for extracting the seventh level wavelet coefficients is equal to the number

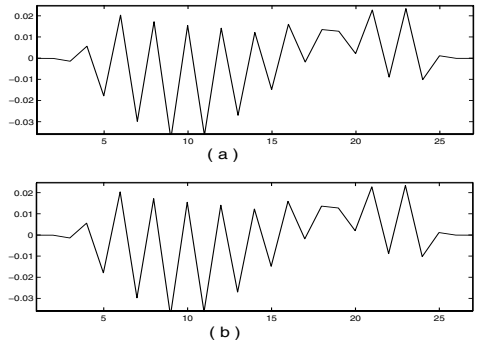

Fig. 5. *(a)Level Seven Wavelet Coefficients obtained after synthesis($\mathbf{W}$). (b).Level Seven Wavelet Coefficients obtained directly from the output of the MPEG Analysis Filter Bank($\widehat{\mathbf{W}}$);MSE between $\mathbf{W}$ and $\widehat{\mathbf{W}}$=4.4293e-09.*

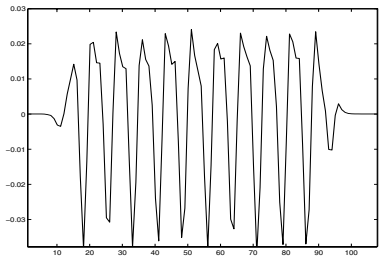

Fig. 6. *Level Seven Wavelet Coefficients containing 4 times the coefficients shown in Fig.5*

of multiplications required for synthesising the PCM samples and for extracting the seventh level wavelet coefficients from the synthesised output.

As the MPEG filter bank is cosine modulated number of multiplications required for synthesising 32 PCM samples is that of the prototype and the overhead which amounts to 2560 [512+(32 x 64)] multiplications[9]. Further the modulation overhead 2048 (32x64) multiplications can be drastically reduced by applying fast DCT,which amounts to just 80 multiplications[2] . Hence the total number of multiplications required for synthesising 32 PCM samples is 592 (512+80). Without loss of generality if we consider 1024 PCM samples (32 ms of a signal sampled at 32 KHz) for extracting pitch, the number of multiplications required for extracting the seventh level wavelet coefficients by using the Fast Wavelet Transform method is approximately 4032 (1024+512+256+128+64+32)2 . (The factor 2 arises from Cubic Spline wavelets which has 4 coefficients). Hence the total number of multiplications in the uncompressed domain processing amounts to 22976 (592 x 32 + 4032).

Compressed domain Pitch Extraction

As described we consider the output of the first two MPEG analysis filters. The seventh level wavelet coefficients are obtained by convolving the output of the first two

2001 IEEE International Conference on Multimedia and Expo
ISBN 0-7695-1198-8/01 $17.00 © 2001 IEEE
0-7695-1198-8/01/$10.00 (C) 2001 IEEE

MPEG analysis filters with that of $\mathbf{WSb}^i (i = 1, 2)$. If we consider 32 samples from each subband (which corresponds to 32ms of subband signal which would be sampled at a critical rate of 1KHz for a 32KHz signal), number of multiplications amounts to 1600 (32 x 25 x 2).

Hence the computational complexity for compressed domain processing expressed as a percentage of uncompressed domain processing is 6.96%

3.3. Algorithm

1) Decode blocks of 36 samples from each of the first two subbands in a frame, in case of layer II/ III(or decode blocks of 36 samples from each of the first two subbands by considering 3 frames at a time ,in case of layer I).Let $\mathbf{x}_1$ and $\mathbf{x}_2$ correspond to the subband samples thus decoded.i.e

$\mathbf{x}_1 = \{x_1[0], x_1[1], ...x_1[35]\}$

$\mathbf{x}_2 = \{x_2[0], x_2[1], ...x_2[35]\}$

2) Find $W[n] = \sum_{i=1}^{2} \sum_{k=0}^{35} x_i[k] W Sb^i[n-k]$.

3) The time interval between two successive maxima is found and the average of the time interval between successive maxima in a block gives the pitch period.

4. RESULTS

The following are the signals for which pitch was found using the algorithm in Section 3.3

a) Synthetic signal with pitch period = 5 ms.

Utterances of

b) '/a/ ' spoken by male speaker

c) '/a/ ' spoken by female speaker

d) 'audio signal ' spoken by female speaker.(after removal of silence and unvoiced segments)

All the above signals were encoded in MPEG,Layer I format. The pitch period over different blocks of frames for the above signals are given in Fig.7. Fig.7d compares wavelet method (shown in dark line) and autocorrelation method (shown in dashed line) of pitch extraction.

5. CONCLUSION

We have presented a method for extracting the level seven wavelet coefficients from the output of the MPEG analysis filter bank which are subsequently used for extraction of pitch. This method has been shown to be very efficient and can be used in the compressed domain for fast indexing and retrieval of multimedia documents without having to fully decode the compressed signal.

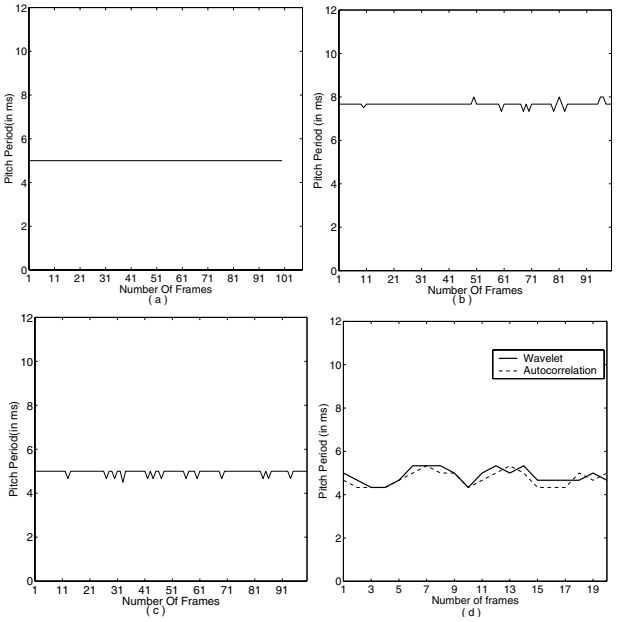

Fig. 7. *(a) Synthetic signal having pitch=5 ms. (b) /a/ spoken by male speaker . (c) /a/ spoken by a female speaker. (d) 'audio signal' spoken by female speaker.*

6. REFERENCES

[1] ISO/IEC 11172-3,"Information Technology- Coding of moving pictures and associated audio for digital storage media at up to 1.5Mbits/s.

[2] P.P.Vaidyanathan , Multirate Systems and Filter Banks,Englewood Cliffs,N.J :Prentice Hall,1993.

[3] S.Kadambe and G.F.Boudreaux-Bartels, "Application of the Wavelet Transform for Pitch detection of speech signals", IEEE Transactions on Information Theory,Vol 38,No.2 ,March 1992 .

[4] S.Kadambe and G.F.Boudreaux-Bartels, "A comparison of wavelet functions for pitch detection of speech signals," in Proc. Int. Conf. Acoust. Speech,Signal Processing, Toronto,Canada.May 1991, pp 449-452.

[5] Stephane Mallat and Sifen Zhong, "Characterization of signals from Multiscale Edges,"IEEE Trans. Pattern Analysis and Machine Intelligence ,Vol 4, No.7,July 1992.

[6] Chris A.Lanciani and Ronald W.Schafer, "Subband-Domain Filtering Of MPEG Audio Signals ",in Proc Int. Conf. Acoust. Speech,Signal Processing Vol. 2, 1999,pp 917-920.

[7] Davis Pan, "A Tutorial on MPEG/Audio Compression",IEEE Multimedia 1995.

2001 IEEE International Conference on Multimedia and Expo
ISBN 0-7695-1198-8/01 $17.00 © 2001 IEEE

0-7695-1198-8/01/$10.00 (C) 2001 IEEE