HIDDEN AUXILIARY MEDIA CHANNELS IN AUDIO SIGNALS BY PERCEPTUALLY INSIGNIFICANT COMPONENT REPLACEMENT

T. D. Jackson, F. F. Li and K. Yates

Department of Computing and Mathematics
Manchester Metropolitan University
Manchester, United Kingdom, M1 5GD

ABSTRACT

This paper proposes a method for the formation of an auxiliary media channel within a host signal. Using a psychoacoustic frequency masking model, perceptually insignificant subband components of the host audio signal are identified and removed. The auxiliary channel data are placed in the empty subbands in the host signal and scaled to a level below the audible threshold. An implementation is given along with results suggesting that the proposed method can effectively hide an auxiliary media channel in a normal audio signal without degrading the perceived sound quality.

1. INTRODUCTION

Data hiding and steganography techniques have attracted much attention in recent years [1]. For audio signals, frequency masking curves provide a means for the embedding of data below the perceptual threshold of the host signal [2, 3]. The feasibility of a "sub channel" making use of "an audio channel's capacity below the perceptual threshold" was postulated by Ding [4], however no tangible proofs and results were reported. Perceptually insignificant components of the host signal are first removed and then replaced with new data. Figure 1 shows how components of a signal lie below the perceptual threshold. This allows for a clean separation of the host and embedded signal. An auxiliary media channel (AMC) has many applications, they may be used for metadata, graphics or additional audio features such as foreign languages. These channels usually rely on additional bandwidth, additional storage space, and or special codecs. The proposed method differs from other audio steganograpy techniques that focus on secure covert communications to embed short utterances in longer signals[5]. The proposal is also substantially different from schemes such as Dolby Pro Logic in that the additional content is masked before decoding and can be used to carry various forms of data. A standard audio signal from which an auxiliary channel could be extracted when needed has significant advantages. This paper extends the idea presented in [4] for use with 'CD quality' audio signals, detailing a proposed implementation and example usage.

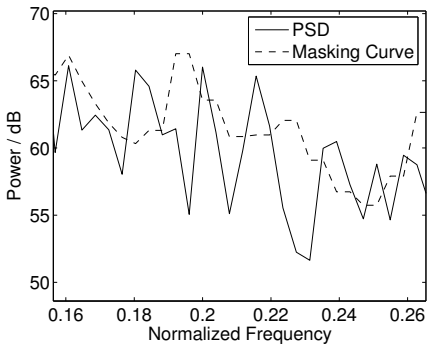

Fig. 1. Example Power Spectral Density (PSD) and Masking Curve. Where the PSD is below the masking curve identifies perceptually insignificant components

2. PROPOSED METHOD

Dividing a signal into subbands forms a basis for the implementation of perceptually insignificant component replacement. A psychoacoustic model can determine the perceptual significance of each subband. Those falling below the threshold of human hearing can be deemed insignificant and can be replaced. The Discrete Cosine Transform (DCT) is chosen for subband decomposition due to its potentially 'near FFT' fast performance and simplicity of manipulation, whilst a matlab implementation [6] of the readily available and widely accepted "Psychoacoustic Model 1" from the MPEG ISO/IEC 11172-3:1993 [7] is used to test the perceptual significance of each subband. The precise details of the model are too lengthy to be discussed in this paper and the reader is therefore referred to the standard. A schematic diagram for the implementation is shown in Figure 2.

The algorithm is performed on a frame by frame basis. A frame with a length of 384 with no overlap is fed to the DCT stages whilst a frame of length 512 with 128

0-7803-9332-5/05/$20.00 ©2005 IEEE

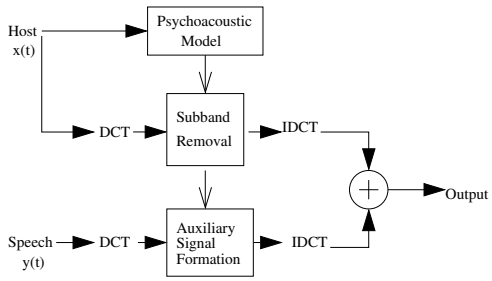

Fig. 2. Schematic overviewing the embedding process

sample overlap is used for the psychoacoustic model. Both signals are prefixed with 64 zeros, in order to provide synchronization the psychoacoustic model begins with sample 1 whereas the DCT begins with sample 65. The DCT is used to decompose the signals into subband components that allow removal and replacement. The host signal is represented $x(t)$ and the auxiliary signal $y(t)$, where t represents a discrete time series.

Stage 1 : Psychoacoustic Model. The data output from the psychoacoustic model are in terms of 32 equal width subbands n, where $n = \{1, 2, \ldots, 32\}$. The output gives the minimum masking level, $LT_{\min}(n)$ dB and the maximum Sound Pressure Level (SPL), $L_{sb}(n)$, normalized to a maximum per frame of 96 dB by addition of a value, Δ, such that $\Delta = 96 - \max(L_{sb}(n))$. The Signal to Mask Ratio, $SMR_{sb}(n)$, is then defined by Eq. 1.

$$SMR_{sb}(n) = L_{sb}(n) - LT_{\min}(n) \tag{1}$$

Thus, a negative value of $SMR_{sb}(n)$ indicates a perceptually insignificant subband. The set of perceptually insignificant subbands, R, is defined as,

$$R = \{n | SMR_{sb}(n) < 0\} \tag{2}$$

Stage 2 : Subband Decomposition. The DCTs of both $x(t)$ and $y(t)$ are obtained using Eq. 3.

$$X_c(k) = c(k) \sum_{t=0}^{N-1} x(t) \cos\left(\frac{\pi(2t+1)k}{2N}\right) \tag{3}$$

where $c(k) = \begin{cases} \sqrt{\frac{1}{N}} & k = 0 \\ \sqrt{\frac{2}{N}} & 0 < k \leq N-1 \end{cases}$

$X_c(k)$ is then reshaped into $X_{sb}(n, m)$ such that $k = 12(n-1)+m$ i.e. Each subband, n, comprises of 12 consecutive coefficients m from $X_c(k)$ where $m = \{1, 2, \ldots, 12\}$.

Stage 3 : Removal

The perceptually insignificant subbands are removed by zeroing their coefficients as in Eq. 4.

$$X^*_{sb}(n, m) = \begin{cases} 0 & n \in R \\ X_{sb}(n, m) & n \notin R \end{cases} \tag{4}$$

$X^*_c(k)$ is constructed by rearranging $X^*_{sb}(n, m)$ such that $k = 12(n-1) + m$.

Stage 4 : Auxiliary Signal Formation. The subband data of the AMC needs to be re-arranged such that its energy contents are now located in the empty spaces corresponding to the host signal as in Eq. 5. Figure 3 shows an example of the DCT coefficients of the auxiliary signal compared to that of the host signal with subbands removed.

$$Y^*_{sb}(R(n), m) = Y_{sb}(n, m) \qquad 1 \leq n \leq |R|, \forall m \tag{5}$$

where $R(n)$ represents the n^{th} smallest element in R. All other values of $Y^*_{sb}(n, m)$ are equal to zero. $Y^*_c(k)$ is constructed by rearranging $Y^*_{sb}(n, m)$ again with $k = 12(n-1) + m$. The IDCTs of both Y^* and X^* are obtained from Eq. 6 yielding y^* and x^*.

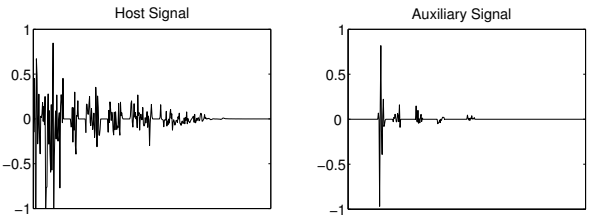

Fig. 3. Comparison of DCT for Host Signal with subbands removed and Auxiliary Signal. The x-axis represents the DCT coefficients.

$$y^*(t) = \sum_{t=0}^{N-1} c(k) Y^*_c(k) \cos\left(\frac{\pi(2t+1)k}{2N}\right) \tag{6}$$

Stage 5 : Correction for masking. To ensure the embedded AMC remains below the audible threshold, a correction factor, α, has to be applied, this value is obtained by comparison of the maximum SPL in each subband of $y^*(t)$, $L^*_{sb}(n)$, and minimum masking level of $x(t)$, $LT_{\min}$. $L^*(t)$ is then obtained by calculating the PSD of $y^*(t)$ according to Eq. 7.

$$Y^*(k) = \Delta + 10 \log_{10} \left| \frac{1}{N} \sum_{t=0}^{N-1} h(t) y^{*+}(t) \exp\left(\frac{-2jkt\pi}{N}\right) \right|^2 \tag{7}$$

where $y^{*+}(t)$ is equal to $y^*(t)$ padded with 64 zeros either side to account for application of the Hanning window

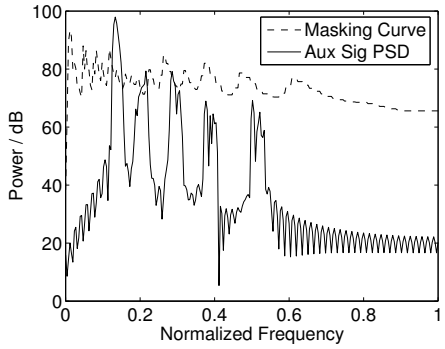

Fig. 4. Masking Curve

described by Eq. 8. $L^*_{sb}(n)$ is then taken as the largest value of $Y^*(k)$ in subband n.

$$h(t) = \frac{1}{2}\sqrt{\frac{8}{3}}\left(1 - \cos\left(\frac{2\pi t}{N}\right)\right) \quad 0 \leq t \leq N-1 \quad (8)$$

An auxiliary signal to mask ratio, $ASMR_{sb}(n)$, is determined by,

$$ASMR_{sb}(n) = L^*_{sb}(n) - LT_{\min}(n) \quad (9)$$

This allows the determination of the multiplication factor, α, given by,

$$\alpha = 10^{-\left(\frac{\beta}{20}\right)} \quad (10)$$

where $\beta = \max(ASMR_{sb}(n))$. Finally we obtain our encoded signal as,

$$s(t) = x^* + \alpha y^*(t) \quad (11)$$

Auxiliary Channel Extraction The subband locations and the values of α will be required at the decoding stage to enable successful extraction of the auxiliary media channel with no degradation in quality. It was suggested in [4] that component replacement would cause little difference to the masking curve such that it could be again calculated by the decoder, this is contested in [8] although to the best of the authors' knowledge, practical results have yet to be published for either hypothesis. A study into the feasibility of this method is too lengthy for inclusion in this paper. The extraction method presented in this paper is based on the assumption that the decoding data discussed above is transmitted to the receiver by any valid method.

To extract the AMC, the DCT of $s(t)$, $S_c(t)$ is obtained using Eq. 3 and rearranged into $S_{sb}(n,m)$. Based on the assumption that we know $R(m)$ and α, the auxiliary audio channel s_* can be determined as,

$$S^*_{sb}(n,m) = S_{sb}(R(n),m) \qquad 1 \leq n \leq |R|, \forall m \quad (12)$$

then all other values of $S^*_{sb}(n,m)$ are equal to zero. $S^*_c(k)$ is constructed by rearranging $S^*_{sb}(n,m)$ such that $k = 12(n-1)+m$. Taking the IDCT of $S^*_c(k)$ as in Eq. 6 and division by α yields the extracted auxiliary channel, $y^*(t)$.

3. RESULTS AND TEST PROCEDURE

The case chosen for this paper was to embed a telephone quality speech signal into a CD quality music host signal, an initial test was carried out with 6 music clips (of 20 seconds in length) to investigate the numbers of perceptually insignificant subbands. The results are presented below in Table 1.

Table 1. Example numbers of insignificant subbands

Music	Mean	Std Dev	Min	Max
Folk	16.2	1.8	8	25
Pop I	14.8	1.7	9	22
Rock	15.6	1.9	10	26
Pop II	15.1	1.6	11	24
Classical	18.4	1.8	14	24
Piano/Vocal	16.1	1.9	11	25

The required bandwidth to reproduce a telephone quality speech signal is 300-3400 Hz, which can be achieved by using the first 5 subbands (i.e. 0-3445 Hz). Using the same music clips, with R restricted to the five smallest values, a 20 second speech clip was embedded as an auxiliary channel. During the process, the value of α was recorded and the results are presented in table 2. As large as possible value of α is desirable as this represents the strength of the auxiliary signal relative to the host which will ultimately affect the robustness of the auxiliary channel. Comparison of insignificant subbands values against α values for each type of music shows that a larger number of the former does not relate to a larger value of the latter, i.e. more subbands does not mean greater AMC strength.

3.1. Subjective Testing

A small test group of 6 listeners was used to determine whether the embedding of the auxiliary channel had any perceptual effect. The listeners were presented with the clips of the sequences both with the perceptually insignificant subbands removed and replaced in a randomized order and asked whether they could distinguish the two, giving their answer in terms of how the second sequence compared to the first. The third example, Rock, was used as a control,

Table 2. α values

Music	Mean	Std Dev
Folk	0.03	0.12
Pop I	0.11	0.26
Rock	0.02	0.07
Pop II	0.11	0.23
Classical	0.01	0.03
Piano/Vocal	0.06	0.24

i.e. the two clips were identical. The results have been corrected to show how listeners compared the subband replaced signal to the empty subband signal and are presented in Table 3. a zero indicates no difference, 1 for slight difference, 2 for different and 3 for significantly different. Negative values indicate that the listener believed the encoded to be worse, positive values that the encoded sounded better. The asterisk marked values indicate a difference was heard when in fact the two clips were identical.

Table 3. Subjective Testing Results

Music	Listener					
	1	2	3	4	5	6
Folk	0	-1	-1	0	1	0
Pop I	0	0	2	1	-2	-1
Rock	0	0	0	1*	1*	1*
Pop II	1	0	-2	-2	-2	0
Classical	-2	-2	3	2	-1	1
Piano/Vocal	0	0	0	-1	1	2

As the auxiliary signal data was made to remain below the masking curve it is assumed that the difference between signal with subbands removed and that with subbands replaced would be inaudible. Results indicating that some listeners identified the encoded version as sounding superior suggest that listeners may have in fact had difficulty in distinguishing the two, hence it may be the case that listeners are able to detect minor differences but cannot reliably conclude whether one is preferable over the other.

4. DISCUSSIONS AND CONCLUSIONS

This paper has identified a method for the formation of an auxiliary media channel in a standard audio signal based on a method described as perceptually insignificant component replacement. An algorithm for the implementation of such a method has also been described in detail. It has been shown how the channel can be used to carry a reduced bandwidth speech signal. The DCT was used for subband removal and replacement due to its 'near FFT' performance and simplicity of manipulation. The well known Psychoacoustic Model 1 was used to identify the perceptually insignificant subbands. Analysis of several pieces of music showed that on average 14–18 subbands could be regarded as perceptually insignificant, with the minimum always exceeding the five subbands required to carry the telephone quality speech signal. However, the number of available subbands was not the only criterion as the signals with fewer available subbands allowed a greater strength of the embedded signal compared to the host. Subjective testing of the embedding stage confirmed that whilst the auxiliary channel data remained below the masking curve then any differences would not be audible. Backwards compatibility is maintained as the auxiliary media channel does not affect performance of the encoded signal on equipment without the decoding capabilities.

Acknowledgment

The EPSRC is acknowledged for providing the PhD research studentship that allows this work to take place.

5. REFERENCES

[1] I. J. Cox and M. L. Miller, "The first 50 years of electronic watermarking," *Journal of Applied Signal Processing*, no. 2, pp. 126–132, 2002.

[2] S. Poomdaeng, S. Toomnark, and T. Amornraksa, "Digital watermarking using psychoacoustic model," in *International Technical Conference on Circuits/Systems Computers and Communications Phuket Thailand*, July 2002, pp. 872–875.

[3] M. D. Swanson, B. Zhu, A. H. Tewfik, and L. Boney, "Robust audio watermarking using perceptual masking," *Signal Processing*, vol. 66, no. 1, pp. 337–355, May 1998.

[4] D. Heping, "Sub-channel below the perceptual threshold," in *ICASSP*, 2003.

[5] K. Gopalan and S. Wenndt, "Audio Steganography for Covert Data Transmission by Imperceptible Tone Insertion," *WOC 2004, Banff, Canada* July 8 10, 2004.

[6] F. A. P. Petitcolas, "MPEG psychoacoustic model I for MATLAB," www.cl.cam.ac.uk/ fapp2/software/mpeg/

[7] ISO/IEC 11172-3:1993

[8] J. Chou, K. Ramchandran, D. Sachs and D. Jones, " Audio Data Hiding with Application to Surround Sound, " in *ICASSP*, 2003.