PERCEPTUALLY MASKED AUDIO SUB-CHANNELS AND APPLICATIONS

TD Jackson, K Yates and FF Li Manchester Metropolitan University, UK

1 INTRODUCTION

Over the past 50 years, a large number of algorithms have been developed to imperceptibly embed information in host signals for various applications, predominantly, copyright protection and authentication.[1] Thus far, much of the work in this area has concentrated on embedding low capacity secure signatures, known as audio watermarks. The sensitive nature of the human auditory system (HAS), and limited bandwidth of an audio signal create a technical challenge for embedding large amounts of information. The information hiding process can be visualised as a triangle representing the trade-off between the three major concerns, imperceptibility, robustness and embedding data rate.

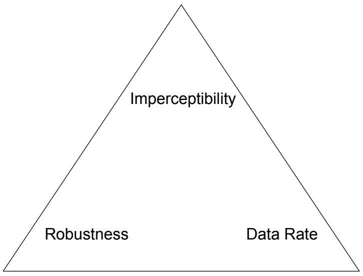

Figure 1. Triangle showing the trade-off between the three major concerns.

In recent years, psychoacoustic models have found prevalent applications in audio technology to achieve low bit-rate compression in the form of perceptual codecs. If perceptual coding algorithms are applied in conjunction with information hiding methods and the robustness concern of watermarking applications is abated, the embedding capacity is likely to increase dramatically, raising the possibility of hosting a second continuous channel. Such a scenario has been explored less. Developing related algorithms, exploiting potentials of and investigating applications for such techniques would be academically interesting and practically fruitful.

This paper explores the feasibility, proposes new techniques and discusses potential applications of hiding a subordinate audio signal in a high fidelity host audio signal without compromising perceived sound quality. Audio signals are converted to the Discrete Cosine Transform (DCT) domain. A psychoacoustic model is used to remove perceptually insignificant contents in the transformed domain. The freed frequency bins are used to host the subordinate signal. The algorithm is arranged so that the embedded subordinate signal is immediately below the perceptual masking threshold. As a result, the composite signal can be transmitted, stored and played back as a normal audio signal without resort to special equipment. When needed, the subordinate signal can be retrieved using a purpose designed decoder. Combining psychoacoustic models, audio compression and audio information hiding techniques, the proposed method creates a hidden virtual sub-channel within a normal audio channel. Potential applications include special sound effects, switchable foreign language channel, speech enhancement, hearing aids amongst other potential applications to be detailed in section 5.

2 BACKGROUND

2.1 Component Replacement

The possibility of using spectral component replacement as a method to transmit extra information was discussed by Ding[2]. However, more detailed investigation, practical experimentation and potential algorithms are required. Component replacement relies on the use of a psychoacoustic model to detect perceptually redundant spectral components on a frame by frame basis within the audio signal. These components can then be removed and replaced with information to form a sub-ordinate channel.

2.2 Perceptual model

The psychoacoustic model used in this paper is a Matlab implementation of 'Psychoacoustic Model One' from the MPEG standard BS ISO 11172:3 [3,4]. For a full description of the model the reader is referred to the aforementioned standard as only a brief overview will be given here. The model assumes 26 non-overlapping critical bands increasing in width with higher frequency and ultimately yields a minimum masking threshold, $LT_{min}(n)$, where n is a set of 32 equal width subbands. The power spectral density (PSD) of signal $x(t)$, is first obtained as in Equation 1,

$$X(k) = 10\log_{10}\left|\frac{1}{N}\sum_{t=0}^{N-1} h(t)x(t)\exp\left(\frac{-2jkt\pi}{N}\right)\right|^{2} \tag{1}$$

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

where $h(t)$ represents a Hann window shown in equation 2. $X(k)$ is then normalized by addition of a value Δ, to make the maximum value of $X(k)$ equal to 96 dB i.e.

$$\Delta = 96 - \mathrm{Max}[X(k)]. \tag{3}$$

From which, a set of rules is used to determine tonal and non-tonal components, for which individual masking thresholds are calculated. To limit computational complexity, these calculations are limited to a predefined width surrounding the actual masker component itself. Additionally, certain masker components are eliminated where several occur in close proximity and those lying below the auditory threshold. The remaining individual masking thresholds are summed with the threshold of hearing to produce a global masking threshold, $Ltg(k)$, from which the minimum masking threshold, $LT_{min}(n)$, is obtained as the minimum value of $Ltg(k)$ in subband n.

2.2.1 The Discrete Cosine Transform (DCT)

For our initial feasibility study, the component replacement is performed with the assistance of the Discrete Cosine Transform (DCT). The DCT is a Fourier related transform, first introduced in a seminal paper published in 1974[5]. The reader is referred to Rao and Yip's monograph for extensive treatment of the definition, mathematical background and potential applications of this transform[6]. The discrete cosine transform on a real number series is a linear, orthogonal, invertible and purely real transform that maps the original number series onto its frequency domain.

By definition, the DCT has a cosine basis function; is orthogonal and yields a purely real frequency domain representation of the original time series. The definitions read rather similar to the real part of DFT, but the difference between these two transforms stem from different assumptions: the DFT assumes that the time series is periodically continued with a period of N, whereas the DCT assumes that the series continued with its mirror image and then periodically continued with a period of *2N*. That is to say the DCT can be calculated from the real part of DFT on a double-length time series obtained by mirroring the original one. This gives one possible FFT based fast algorithm for the DCT. Such a fast DCT is typically achieved using pre- and post processors and an FFT routine.

There are several slightly different DCTs in the literature, the main ones being labelled DCT I-IV. The most suited, and therefore most commonly used in audio and video applications is the one known as DCT–II, simply refereed to as DCT by most authors, as it is throughout the remainder of this paper. There is also an MDCT (modified discrete cosine transform), a lapped version of the DCT.

The DCT $X_c(k)$ of a real series *x(n) of length N is defined by*

$$X_c(k) = c(k)\sum_{n=1}^{N} x(n)\cos\left(\frac{\pi(2n-1)(k-1)}{2N}\right); \qquad k \in [1, N] \tag{4}$$

and its inverse IDCT can be written as

$$x(n) = \sum_{k=1}^{N} c(k) X_c(k)\cos\left(\frac{\pi(2n-1)(k-1)}{2N}\right); \qquad n \in [1, N] \tag{5}$$

where

$$c(k) = \begin{cases} \sqrt{\frac{1}{N}}; & k = 1 \\ \sqrt{\frac{2}{N}}; & 2 \le k \le N \end{cases} \tag{6}$$

3 PROPOSED METHOD

The proposed method is operated in the DCT domain for the simplicity of signal manipulation. From a practical point of view, this also takes the advantage of low cost and real-time hardware.

3.1.1 Embedding

The algorithm for the embedding procedure is illustrated in Figure 2

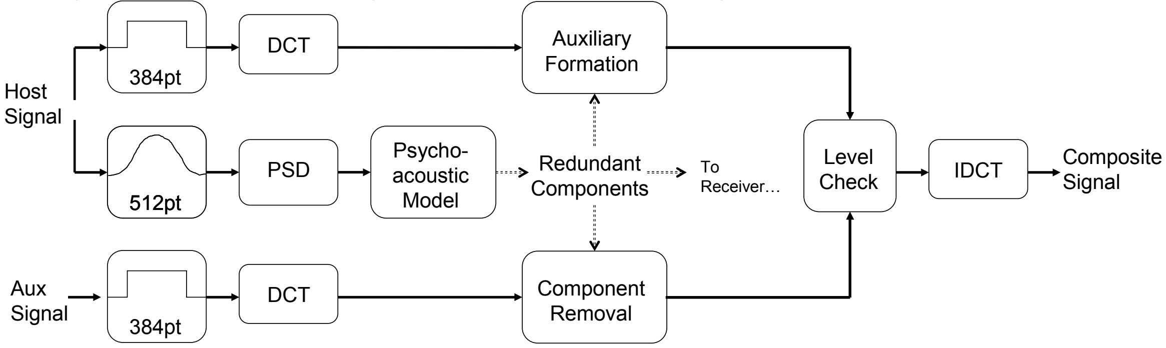

Figure 2 Embedding process

The detailed embedding method is described below and was published by the author[7]. To summarize, the host and subordinate audio signals are divided in time frames, each of which is analyzed by the psychoacoustic model to detect redundant subbands. Using the DCT, the subband contents are removed and replaced with frequency-shifted subband contents from the subordinate signal. The subordinate contents are adjusted so as to remain immediately below the masking threshold.

The host signal is represented by the discrete time series, *x(t)*. *X(k)* is obtained and normalized as in Equations 1 and 3, from which, $L_{sb}(n)$, the maximum sound pressure level in each subband, ***n***, is obtained by

$$L_{sb}(n) = Max\left[X(k)\right] \qquad X(k) \text{ in subband } n \qquad (7)$$

From the psychoacoustic model we then obtain, $LT_{min}(n)$, the minimum masking threshold for each subband i.e. the minimum value of *Ltg(k)* in subband *n*. From $L_{sb}(n)$ and $LT_{min}(n)$, the signal to mask ratio , $SMR_{sb}(n)$ is calculated by,

$$SMR_{sb}(n) = L_{sb}(n) - LT_{\min}(n) \qquad (8)$$

An index of perceptually irrelevant subbands, *PI(m)*, where *PI(m)* is the mth smallest integer in the set *PI* defined by,

$$PI = \{n \mid 1 \le n \le 32, SMR_{sb}(n) < 0\} \qquad (9)$$

and *M* = |*PI*|, the size of the set *PI*. The DCT of *x(t)* is obtained as per Equation 4, yielding $X_c(k)$. The irrelevant subbands of the host signal are removed by zeroing the appropriate coefficients of $X_c(k)$. To show this operation we introduce X_{sb} as an array of vectors $\mathbf{n}_1$ to $\mathbf{n}_{32}$, where $\mathbf{n}_i$, is the 12 DCT coefficients representing the i^{th} subband. Note that X_{sb} is purely an alternative method of indexing X_c and for all other intents and purposes, $X_{sb} \equiv X_c$.

$$X^*_{sb}(\mathbf{n}_i) = \begin{cases} [0,0,...,0] & i \in PI(m) \\ X_{sb}(\mathbf{n}_i) & i \notin PI(m) \end{cases} \qquad (10)$$

The modified DCT coefficients of the host signal are then transformed back to the time domain yielding, *x*(t)*. The DCT of the auxiliary signal, *y(t)*, is obtained as per Equation 4 yielding $Y_c(k)$. The coefficients need to be shifted such that the non-zero values of $Y_c(k)$ correspond to the zeroed values of $X_c(k)$. The bandwidth allocated for the auxiliary signal is determined by *M*.

$$Y^*_{sb}(\mathbf{n}_{PI(m)}) = Y_{sb}(\mathbf{n}_m) \qquad 1 \le m \le M \qquad (11)$$

where all other values remain as zeros and as previously stated, *M* is the size of *PI(m)*. The modified DCT coefficients of the auxiliary signal are then transformed back to the time domain yielding *y*(t)*. The spectral contents of the auxiliary signal are now contained exclusively within subbands *PI(m)* for *m = [1,2,...,M]*.

Before creating the composite ‘host plus auxiliary’ signal, a test is performed to determine whether the reformed auxiliary signal *y*(t)* can be masked by the modified host signal *x*(t)*. It is first necessary to calculate the PSD of *y*(t)* as in Equation 11.

$$Y^*(k) = \Delta + 10\log_{10}\left|\frac{1}{N}\sum_{t=0}^{N-1} h(t)\,y^*(t)\exp\left(\frac{-2jkt\pi}{N}\right)\right|^2 \qquad (12)$$

Where *h(t)* is as defined in equation 2 and *Δ* in equation 3.

In the above calculation, the signal, *y*(t)* is padded with 64 leading and trailing zeros to make the signal the correct length. The auxiliary signal to mask ratio is then calculated as

$$ASMR_{sb}(n) = L*_{sb}(n) - LT_{\min}(n) \tag{13}$$

adapted from Equation 8 and where *L** $L*_{sb}(n)$ is the largest value of *Y*(k)* in subband ***n***.

A multiplication factor, α, is then determined as

$$\alpha = 10^{-\left(\frac{\beta}{20}\right)} \tag{14}$$

where $\beta = max[ASMR_{sb}(n)]$. The composite signal is now given by,

$$s(t) = x*(t) + \alpha y*(t) \tag{15}$$

3.2 Detection and Extraction

For accurate recovery of the embedded channel, locations of the hidden frequency components are required to be transmitted to the decoder. The algorithm picks up these components re-assemble them and perform inverse DCT to obtain the output of the subordinate channel as illustrated in Figure 3. There are several possibilities as to how the location information may be transmitted to the decoder, some of which are discussed below.

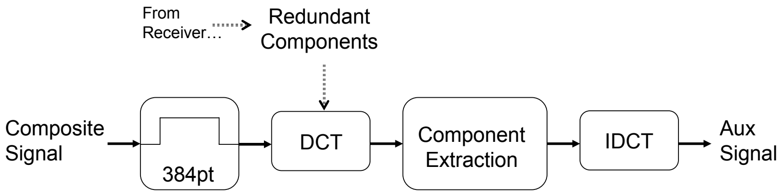

Figure 3. Sub channel decoding

The composite signal is converted to the DCT domain where using the received information the signal is split in two. The subband contents for the extracted subordinate signal are once again shifted back to their appropriate locations.

3.2.1 Recalculation of Masking Threshold

This method is based upon the assumption that the embedding process would not significantly alter the masking properties of the signal. Based on this assumption, the composite signal is analyzed as though it where the host signal, to identify the perceptually redundant subbands. These subbands should be the same ones identified at the embedding stage and therefore contain the subordinate audio. This may be seen as a somewhat naïve method, however, a preliminary investigation into exactly how effective it could be was carried out. At present, the decoder could correctly identify the hidden channel less than 40% of the time.

3.2.2 Alternative Methods

A further possibility is to use a traditional audio watermarking method to embed the subband locations. However the problem lies in the large amount of data required to be transmitted to the decoder leaving all the high capacity – low robustness algorithms available. The least significant bit method was used to carry the information to the decoder. Although successful in a controlled environment, any minor modification to the signal such as D/A and A/D would simply remove the information. For this kind of method to be successful, a more robust mechanism is sought.

Based on the results from the above, it is clear that an alternative approached is required. One approach is to use a hybrid method. If one or two subbands can be made such that it is guaranteed the decoder detects them as redundant then they can be made to contain the subband location information for the sub ordinate channel. This is an area of investigation to be considered for further work.

4 RESULTS

4.1 Composite Signal Quality Evaluation

Objective assessment of something that is inherently subjective, is technically challenging. Traditional measures such as the signal to noise ratio (SNR) *can* be produced but have little to offer as they do not relate to human perception. There are no universally applicable methods available. Subjective tests are the ultimate method in the assessment of any perceptually modified audio signals. A preliminary subjective test was carried out using an all male sample of seven subjects. Six audio clips of various musical styles were used as host signals to create a sub-channel containing telephone quality speech. Each clip was approximately 20 seconds in length. The listeners were asked to listen to two clips, repeated twice in an ABAB fashion, one being the host signal after undergoing the component removal stage, the other being the composite signal, the order being randomized, and asked to compare the quality on a scale of -3 to +3. The results were since normalized such that a negative value indicates the listener believed the composite signal to be of lesser quality, a positive value, of superior quality, and zero for no difference. The results are shown below in Table 1.

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

Table 1 Subjective testing results

4.2 Extracted Speech Signal Evaluation

Envelope distortion plays a determination role in the preservation of speech intelligibility[8], therefore it is used as a quality index for the proposed algorithm. Envelopes for both the original CD quality speech signal and the extracted speech signal were calculated. The signal is first rectified, followed by a low pass filter and a decimation stage. The decimation reduces the number of sample points to the minimum required by the nyquist theorem (i.e. sampling frequency equals twice the maximum frequency) after the low pass filtering. Figure 4 shows a comparison of two such envelopes.

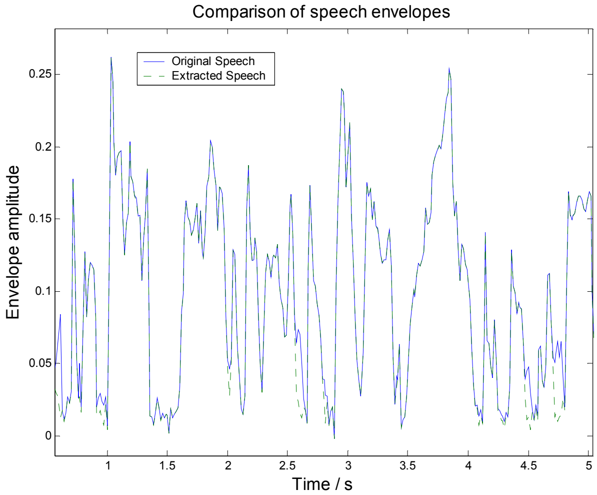

Figure 4 Comparison of speech signal envelopes.

5 POTENTIAL APPLICATIONS

5.1 Auxiliary audio channels

The use of multiple channels in communications, audio signal and reproduction systems is prevalent. They can be used to enhance services or provide extra functionality. For example, two channel and 5.1 multi-channel stereo are used to create a solid soundstage and the latter lends itself particularly well to improving dialogue intelligibility. Multi-channel schemes currently require physical multi-channel transmission bandwidth and storage space or special codecs. The former has inherent physical restraints whilst the latter implies increased demand of channel capacity, which typically results in bitstreams that require decoding before any usage is possible. These factors may incur some fundamental changes to existing equipment and systems from program production, dubbing, recording, and transmission to final reproduction. Creating an extra virtual sub-channel in a standard compatible audio signal has significant advantages and provides a cost-effective solution to additional functionality without resort to modification of existing systems. Such a channel could be used for a variety of applications such as foreign language dialogue, speech enhancement and even transmission of non-audio data such as text.

5.2 Speech enhancement

People with hearing problems often have great difficulty in understanding speech amongst background noise. This has proved to be a problem with TV programs which contain speech and background noise as the listener in general has no independent control over the volume of the speech and background noise. The sub-ordinate channel could be used to provide an independent speech signal which could be adjusted in volume free from background noise.

5.3 Transmission via air - hearing aid application

It has previously been suggested that a certain type of audio watermark could survive air transmission in a scheme proposed to detect bootleg recordings of live performances[9]. The scheme works by transmitting the perceptually masked signal during the performance. Undetectable to human ear, the bootlegger unknowingly records the signal which is statistically detectable by the owner of the material at a later date. If the method proposed in this paper was developed to sustain air transmission then it opens up new possibilities. An example of such an application could be a detecting device concealed in a hearing aid to extend the TV speech enhancement application described above.

6 CONCLUDING REMARKS

Research work so far has empirically and convincingly indicated the potential of a new method that can transmit one of more auxiliary signals by replacing spectral components of a host signal in a transformed domain. The attractiveness of this approach when compared with other alternatives is that it is completely compatible with a single channel system without equipment overheads. The method may be used to transmit audio for speech enhancement or to simultaneously transmit foreign language audio amongst other potential applications. The perceptual quality of the host audio signal is preserved as is the signals format, providing backwards compatibility.

There are many avenues ahead for this work to continue, of particular evidence and importance is to gain a thorough understanding of the system as a whole in order to able to optimize the procedure to both maximize the embedding capacity and to enable real-time encoding and decoding. Of further interest is to investigate into to the use of novel measuring techniques such as PEAQ[10] (Perceptual Evaluation of Audio Quality) in order to assess the quality of the composite signal. Such techniques can provide results similar to those from subjective testing but enable repeatability and reproducibility.

7 REFERENCES

1. I. J. Cox and M. L. Miller, “The first 50 years of the electronic watermarking”, *Journal of Applied Signal Processing*, Vol. 2, pp. 126-132. (2002).
2. D.Heping “Sub Channel below the perceptual threshold” in proceedings of IEEE International Conference on Acoustics Speech and Signal Processing 2003
3. F. A. P. Peticolas. MPEG Psychoacoustic Model 1 for MATLAB www.cl.cam.ac.uk/~fapp2/software/mpeg/
4. BS ISO 11172:3 Coding of Moving Pictures and Associated at up to about 1.5 Mbps – Part 3: Audio
5. N. Ahmed, T. Natarajan, and K. R. Rao, “Discrete cosine transform,” IEEE Trans. Comput., vol. C-23, pp. 90–93, Jan. (1974).
6. K. R. Rao and P. Yip, Discrete Cosine Transform: Algorithms, Advantages, and Applications. New York: Academic. (1990).
7. T.D. Jackson, F.F.Li and K.Yates “Hidden Auxiliary Media Channels in Audio Signals by Perceptually Insignificant Component Replacement” in Proceedings of IEEE International Conference on Multimedia and Expo. Amsterdam 2005.
8. T. Houtgast. and H. J. M. Steeneken “Envelope spectrum and intelligibility of speech in enclosures”, IEEE-AFCRL Speech Conference Proceedings, pp. 392-395. (1972)
9. R. Tachibana “Audio Watermarking for Live Performance”, in Proceedings of Security and Watermarking of Multimedia Contents V, SPIE vol. 5020, Santa Clara, 2003.
10. ITU-R recommendation BS. (Broadcast Service)1387 (PEAQ), "Method for Objective Measurement of Perceived Audio Quality"