Introduction to Audio Watermarking Schemes

Outline

• Problem model
• Previous work
• **Spread spectrum** information embedding
• Implementation parameters
• Simulation and experimental results
• Comparison to other techniques
• Application example: a localization and navigation system
• Conclusions and future work
Information Transmission

• Transmitting information from one device to another
  – Common mediums
    • Radio frequency
    • Optical connection
    • Cable connection
  – Transmitting acoustically
    • Hiding the information in audio signals, using a speaker as a transmitter and a microphone as a receiver
    • Characteristics
      – Low rate, limited range, but benefit from backward compatibility (through an acoustic channel)
Data Hiding Procedures for Acoustic Data Transmission

• Requirements
  – Fast
  – Capability to hide data into arbitrary signals, without prior knowledge of what they are
  – Why not using existing schemes?
    • Robust audio watermarking:
      Non-negligible delay between the host signal and the secret signal
Problem Model

[Host-blind decoding scenario]

- \( x(t) \): host audio signal
- \( m \): message symbol
- \( c_m(t) \): a codeword indexed by \( m \)
- \( f \): encoding function
- \( y(t) \) and \( y'(t) \): transmitted and received signal
- \( h(t) \): room impulse response
- \( w(t) \): noise
The encoder modifies the selected codewords based on the acoustic properties of the host signal.

\[ y(t) = x(t) + f(C_m(t), x(t)) \]

\[ y'(t) = y(t) * h(t) + w(t) \]

where \( * \) denotes the convolution operation.
Auditory Masking

• The effect by which one sound becomes inaudible in the presence of another sound

• Frequency masking:
  Auditory masking between frequency components that occurred simultaneous in time.
  related factors:
  – Frequency
  – Sound pressure level
  – Tone-like or noise-like characteristics of the masker and the marked signals
• Critical bands

  – Human perception of frequency are modeled as a set of overlapping band-pass filters of varying bandwidth

  – Signals that lie within the same critical band are hard to separate for human ear
## Critical Bands

### Approximate Critical Band Boundaries

<table>
<thead>
<tr>
<th>Band Number</th>
<th>Frequency (Hz) ¹</th>
<th>Band Number</th>
<th>Frequency (Hz) ¹</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>50</td>
<td>14</td>
<td>1,970</td>
</tr>
<tr>
<td>1</td>
<td>95</td>
<td>15</td>
<td>2,340</td>
</tr>
<tr>
<td>2</td>
<td>140</td>
<td>16</td>
<td>2,720</td>
</tr>
<tr>
<td>3</td>
<td>235</td>
<td>17</td>
<td>3,280</td>
</tr>
<tr>
<td>4</td>
<td>330</td>
<td>18</td>
<td>3,840</td>
</tr>
<tr>
<td>5</td>
<td>420</td>
<td>19</td>
<td>4,690</td>
</tr>
<tr>
<td>6</td>
<td>560</td>
<td>20</td>
<td>5,440</td>
</tr>
<tr>
<td>7</td>
<td>660</td>
<td>21</td>
<td>6,375</td>
</tr>
<tr>
<td>8</td>
<td>800</td>
<td>22</td>
<td>7,690</td>
</tr>
<tr>
<td>9</td>
<td>940</td>
<td>23</td>
<td>9,375</td>
</tr>
<tr>
<td>10</td>
<td>1,125</td>
<td>24</td>
<td>11,025</td>
</tr>
<tr>
<td>11</td>
<td>1,265</td>
<td>25</td>
<td>15,375</td>
</tr>
<tr>
<td>12</td>
<td>1,500</td>
<td>26</td>
<td>20,250</td>
</tr>
<tr>
<td>13</td>
<td>1,735</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

¹ Frequencies are at the upper end of the band.

---

Critical Bands

Real-time Imperceptible Acoustic Data Transmission

• Closely related to audio data hiding
  – Goal of both procedures is to imperceptibly add information to an audio signal

• Differences lie in
  – Objectives
    • Transmitting arbitrary information vs. simply detecting a hidden signal
  – Attacks
    • Room reverberations + D/A and A/D conversions vs. deliberate signal processing manipulations
• Schemes too slow
  – LSB embedding
  – QIM for phase coefficients:

(i). QIM: the message indexes one of the multiple interleaved quantizers applied to the host signal.

(II). a change in signal phase is inaudible, as long as the signal envelope remains the same or similar

– Adjusting relation between energies of different frames in time domain
• **Echo encoding:**
  -- short delay echoes are imperceptible to human, the information is embedded by adding one more more attenuated signal echoes at known delays.

  – Simple and real-time
  – Sensitive to the presence of additional echoes

• **SS data hiding scheme**
Sonic Watermarking

SS Information Embedding

\[ Y_i[k] = X_i[k] + f(C_m, X_{i-1}[k]) \]

Psycho-acoustically adjusting

A pseudo-random sequence of length N, drawn from U[0,1] distribution
• $X_{i-1}[k]$ and $X_i[k]$, $k = 1,2,\ldots, N$, represent the magnitudes of the Fourier transform coefficients of the two consecutive frames of the host signals.

• $U[0,1]$: uniform distribution over the interval $[0,1]$

• $C_m$: codeword indexed by $m$
f(C_m, X_{i-1}[k])=a[k]C_m[k]

\[ a_b[k] = p \cdot \max_{k \text{ belongs to } b} X_{i-1, b}[k] \]

[Encoding the data in real-time is desirable!]

- To make sure of inaudibility the control parameter \( p \) is computed based on the psycho-acoustic properties of the preceding frame of the host signal \( X_{i-1}[k] \).
The noise in band $b$ never exceeds a pre-defined fraction $p$ of the maximum amplitude of $X_{i-1}$ in the same critical band.

The host signal will mask the added signal within the same critical band.

$p$ controls the tradeoff between the sound quality and code reliability.

Code range: $[0, 1] \rightarrow [0, \alpha_b]$

Assumes that the frequency content of the host signal does not change significantly from one frame to the next, such that the scaling coeff. of the previous frame are reasonably valid for the current frame.
• Received signal
  – \( Y'[k] = Y[k] + W[k] \)
  \( Y'(k) \) is correlated with each codeword sequence \( C_j(k) \), \( k = 1, 2, \ldots, M \), and the codeword with the highest correlation is selected.

• Preprocessing
  – Whitening to negative effects on cross correlation due to non-linear scaling: HAS-based scaling degrades the desired correlation properties of the added pseudorandom sequences

  • \( Y'_{w,b}[k] = Y'_b[p[k]/\max Y'_b[k] \)
  • Code range changes accordingly

• Whitening: The magnitude of the Fourier transform coeff. In each critical band is \textit{normalized} such that the coeff, value lie in the range \([0,1]\).
• Correlation coefficients
  – The codeword corresponding to the highest correlation value is selected

• Synchronization
  : to ensure that the starting point of the decoding is synchronized with the beginning of the message
    – Each codeword is added to 3 consecutive frames
    – N length-N Fourier transforms staring at consecutive samples are taken
Implementation Parameters

- Hanning windowed DFT of overlapping 4096-sample frames
- # of symbols (codewords): M=2
- the length of codeword N=2048
- encoded into 20 sec pop music signal
- White Gaussian noise was added
Experimental Results: SNR is control parameter dependent

\[ \text{SNR} = 10 \log_{10} \left( \frac{E_{\text{host}} + E_{\text{code}}}{E_{\text{noise}}} \right) \]
Experimental Results (subjective test results)

- 0: no audible distortion
- Barely audible for $p<0.2$
- 10 listeners
Experimental Results (Error prob. vs distance-volume)

Empirical results for $p=0.2$
Experimental Results (SS vs Sonic watermarking schemes)

SS

Sonic watermarking
### Experimental Results (capacity vs. error rate)

<table>
<thead>
<tr>
<th>Method</th>
<th>Bits/Minute</th>
<th>Error Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sonic WM [1]</td>
<td>68</td>
<td>14.41%</td>
</tr>
<tr>
<td>Sonic WM [1]</td>
<td>45</td>
<td>8.89%</td>
</tr>
<tr>
<td>Spread-Spectrum</td>
<td>213</td>
<td>9.49%</td>
</tr>
<tr>
<td>Spread-Spectrum</td>
<td>64</td>
<td>1.87%</td>
</tr>
</tbody>
</table>
An Indoor Localization and Navigation System

- Reverberation time
  - 0.1 sec
- First source
  - Always code 1 in music
- Second source
  - Music only or music with code 2
Experimental Results

Applications
- Guiding in a shopping mall or an airport with music played
- Weather information, shop discounts or flight times