Basic Concepts of Audio Watermarking
( and an audio data hiding system done in CMLab-NTU-CSIE)
Selection of Different Approaches

- **Embedding Domain**
  - time domain
  - frequency domain
    - DFT, DCT, etc.

- **Modulation Method**
  - quantization-based
    - QIM, LSB
  - correlation-based (additive)
    - Spread Spectrum
Auditory Masking

- The effect by which one sound becomes inaudible in the presence of another sound.
- Temporal Masking
  - forward & backward
- Frequency Masking
  - related to critical bands
Critical Bands

- Human perception of frequency are modeled as a set of overlapping band-pass filters.

- Signals that lie within the same critical band are hard to separate for human ear.

### Approximate Critical Band Boundaries

<table>
<thead>
<tr>
<th>Band Number</th>
<th>Frequency (Hz)</th>
<th>Band Number</th>
<th>Frequency (Hz)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>50</td>
<td>14</td>
<td>1,070</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>11,625</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.
Possible Attacks in Different Environments

- Attacks in Digital Environment
  - common audio processing
    - denoising, equalization, resampling, etc.
  - compression
  - cropping, scaling

- Attacks in Analog Environment (DA/AD Conversion)
  - background noise
  - quantization error
  - amplitude modification
    - modification of the signal energy
    - filtering effects due to frequency responses of applied hardware
  - phase shift
Information Delivery Systems for Car Passengers without Networking Capabilities

Chun-Hsiang Huang, Po-Wei Chen, Ping-Yen Hsieh and Ja-Ling Wu
Introduction

- Watermarking-based copyright protection systems:
  - Have not been widely deployed and accepted yet.
    - The security of watermarking schemes?
    - Market timings & business models.
- Communications over acoustic channels by data-hiding schemes:
  - Have received more and more attentions.
  - Advantage
    - backward compatibility with analog audio broadcasting systems
  - Drawbacks
    - transfer distance
    - data transmission rate
  - Visual information delivery based on (in-)car radio systems.
Delivering Visual Information via Existing Car Radio System

Fig. 1. The proposed information delivery system based on car radio system
Delivering Visual Information via Existing Car Radio System

Advantages of this scheme:

- Broadcasted music does not have to be interrupted by annoying reports or advertisements.
- Passengers can actively decide whether he or she would like to receive the embedded information.
- The received information can be stored for time-shifted usage.
- New commercial values of audio broadcast channels and business models can be developed.
Embedding

- Host audio signals are divided into consecutive $n$-components frames, and their DCT coefficients are computed.
- Messages are represented by codewords in a $m$-element codebook.
  - Each codeword $C_j$ in a codebook is a pseudo-randomly generated binary sequence consisting of $l$ bits.
- The embedding process can be described by:
  \[ Y_i[\Delta + k] = X_i[\Delta + k] + a_i[\Delta + k] \cdot C_j[k], k = 1, \ldots, l \]
  $a_i$ is the scale factor that determines the fidelity of marked signal and takes the masking effect of each critical band into consideration.
  
  \[ a_i[k] = \bigcup_{b \in B} a_{i,b}[k] \]
  
  \[ a_{i,b}[k] = p \cdot \max_{k \in b} X_{i-1,b}[k] \]
Extraction

- The received DCT frame \( \tilde{Y}_i[k] \) is firstly divided by the scale factor \( \tilde{a}_i[k] \).
- Then, as in common spread-spectrum watermarking schemes, the correlation values between \( \tilde{Y}_i[k] \) and all the codewords are calculated.
- At last, the codeword corresponds to the highest correlation value is regarded as the embedded message.
Synchronization

- A two-stage synchronization scheme is adopted to prevent the out-of-sync problem.
  - Stage 1:
    Initially, all the $n$ possible offsets are iteratively tested by calculating all
    the sums of correlation values between the $s$ synchronizing codewords
    and the $s$ received coefficient frames shifted according to each offset.
    - The offset corresponds to the highest correlation value are used to
      rearrange the forthcoming audio signals.
  - Stage 2:
    A periodic synchronization operation is performed regularly each time
    an interval $t$ has elapsed.
    - $t$ is set as 1 second and the range of tested offset is limited to 4 samples
      before and after corresponding signal samples.
Experimental Results

- Five 20-second mono audio clips recorded from daily radio programs are used as the host audio signals.
- The sampling rate is set to 44.1 kHz.

Table 1: Audio clips adopted as host signals

<table>
<thead>
<tr>
<th>Audio Clips</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>SMNY</td>
<td>Symphony played by an orchestra</td>
</tr>
<tr>
<td>POP1</td>
<td>Chinese POP music soloed by a female singer</td>
</tr>
<tr>
<td>POP2</td>
<td>English POP music played by a band</td>
</tr>
<tr>
<td>Speech1</td>
<td>10-second traffic condition reports with subsequent music</td>
</tr>
<tr>
<td>Speech2</td>
<td>DJ talking with apparent background music</td>
</tr>
</tbody>
</table>
Experimental Results

- **Indoor Tests:**
  - $n = 512$, $m = 16$ or $32$, $l = 128$, $\Delta = 128$, and $p = 0.3$.
  - The achieved bit-rate is therefore 344 bps or 430 bps.

![Graph 1](image1.png)  ![Graph 2](image2.png)

**Fig. 2.** Detection performance for indoor tests where the bit-rate is 344 bps

**Fig. 3.** Detection performance for indoor tests where the bit-rate is 430 bps

<table>
<thead>
<tr>
<th>Audio Clips</th>
<th>SMNY</th>
<th>POP1</th>
<th>POP2</th>
<th>SPEECH1</th>
<th>SPEECH2</th>
</tr>
</thead>
<tbody>
<tr>
<td>SNR (dB)</td>
<td>20.9</td>
<td>17.2</td>
<td>19.6</td>
<td>28.8</td>
<td>26.6</td>
</tr>
</tbody>
</table>
Experimental Results

- **On-Road Tests:**

  ![Graphs showing experimental results](image)
  
  **Fig. 4** On-road experimental results when the bit-rate is 344 bps.
  
  **Fig. 5** On-road experimental results when the bit-rate is 344 bps and the crew remained silent throughout the experiment.
Experimental Results

- **Computation Time and Required Storage:**
  - Performed in real-time using a desktop computer with 3.4GHz CPU.
  - The extraction processes requires about 8.3 seconds and 11.5 seconds for the initial synchronization procedure.
  - Corresponding average buffer sizes in the receiver are 0.747M bytes and 1.035M bytes.
  - Subsequent extraction function can be performed in a real-time manner.
Discussions

- According to the experimental results, the proposed system does possess good robustness against noises.

- Due to the inconsistent traffic conditions, and the constraints on human power and research equipments, the experimental results cannot be easily generalized.
  - More real-world tests will be performed.

- The actual distortions caused by AM/FM broadcasting and car sound systems are not completely reproduced. (the effect of h(t)?)
  - Seek the opportunities for collaborative researches with campus radio stations or car sound system.

- New synchronization techniques and error correction code will be applied.