Digital watermarking method based on STFT histogram

Kotaro Sonoda (Nagasaki Univ, JP)
Aleksander Sek (Adam Mickiewicz Univ, PL)
**Data Hiding**

- **Embedding** is carried out in order to make the distortion of the Stego Signal from the cover signal under imperceptible.
- **Watermarking model** (main content is audio signal)
  - User can’t divide the message from Stego signal.
  - ex) Copy Control, fingerprint
- **Steganography** (main content is message)
  - User can’t notice the existence of message.
  - ex) Secret Communication,
Hidden message detection

Stego Signal, $s$

- Noise, DA/AD,
- Re-encoding, re-quantization,
- Mixing with other signal,
- Transmitted through air,
- Malicious attack by pirates

$s'$

Message Detection

- Requirements
  - As much message payload as possible \(\rightarrow\) Capacity
  - Resist against manipulations \(\rightarrow\) Robusteness
  - Inaudible distortion \(\rightarrow\) Imperceptibility
- In generally, Robustness, Capacity, and Inaudibility are related each other.
  - The most simple way to increase the robustness is changing the signal component gratefully to detect easily. \(\rightarrow\) The audio quality might be deteriorated.
General aim for audio data hiding

- Aim to detect the messages from a short segment (about 5s) cropped from anywhere of a piece of signal
Strategy on conventional methods

- Embed messages into short segments, repeatedly all over the signal

If the start point of analyzing segment is mismatched with one of the embedding sides, it is possible to fail to detect the messages.
• Embed messages into long segments in simultaneously
Changing Histogram form of time variant powers

If message is '0', the number of samples of histogram is fixed to satisfy
\[
\frac{h(2) + h(8)}{h(3) + h(7)} < th0
\]

If message is '1'
\[
\frac{h(2) + h(8)}{h(3) + h(7)} > th1
\]
Embedding (1)

Short Term Fourier Transform

Length of DFT

Time frame

Scaling in ERBs (Equivalent Rectangular Bandwidths)

<table>
<thead>
<tr>
<th>k+2</th>
</tr>
</thead>
<tbody>
<tr>
<td>k+1</td>
</tr>
<tr>
<td>k</td>
</tr>
</tbody>
</table>

### Time frame

Histogram Computation

L classes

### Pow.

Beta calculation

\[
\beta(n) = \frac{h(ll_n) + h(rr_n)}{h(lr_n) + h(rl_n)}
\]

\[
\begin{align*}
h(ll_n) &= \frac{L+1}{2} - 3n \\
h(lr_n) &= \frac{L+1}{2} - 3n + 1 \\
h(rl_n) &= \frac{L+1}{2} + 3n \\
h(rr_n) &= \frac{L+1}{2} + 3n + 1
\end{align*}
\]

\[\text{Pow.}\]
Embedding (2)

Normalizing $\beta(n)$ to $\beta'(n)$ according to the following role defined for one bit of payload:

$$\begin{cases} 
\beta'(n) < th_0 & \iff w(n) = 0 \\
\beta'(n) > th_1 & \iff w(n) = 1 
\end{cases}$$

Beta calculation:

$$\beta(n) = \frac{h(ll_n) + h(rr_n)}{h(lr_n) + h(rl_n)}$$

$$\begin{align*}
h(ll_n) &= \frac{L+1}{2} - 3n \\
h(lr_n) &= \frac{L+1}{2} - 3n + 1 \\
h(rl_n) &= \frac{L+1}{2} + 3n \\
h(rr_n) &= \frac{L+1}{2} + 3n - 1 
\end{align*}$$
Embedding (3)

Normalizing $\beta(n)$ to $\beta'(n)$ according to the following role defined for one bit of payload:

\[
\begin{align*}
\beta'(n) \leq th_0 & \iff w(n) = 0 \\
\beta'(n) > th_1 & \iff w(n) = 1
\end{align*}
\]

Replace coefficients with new value:

Time frame

Inverse STFT
Evaluation: Conditions

Imperceptibility
Annoyance grade of Stego signal (S) from Cover signal (C).

Robustness
Bit error rate of extracted message (M’) from Embedded message (M)

Attacks
- Mp3 conversion 128 kbps
- Noise addition in SN 36dB
- Bandpass filter 100Hz–6kHz, –12dB/oct.
- Time scale modification 4%
- Speed modification 10%
- Echo addition 100ms, –6dB
- Mp3 conversion twice

8 music sound files
- Each source is from SQAM DB
- 60 seconds data
- 44.1 kHz sampling rates
- Mono channel
### Tested sound files

EBU Sound Quality Assessment Material (SQAM)
Recordings for subjective tests

<table>
<thead>
<tr>
<th>Tested sound #</th>
<th>Ingredients</th>
</tr>
</thead>
<tbody>
<tr>
<td>27</td>
<td>Castanets</td>
</tr>
<tr>
<td>32</td>
<td>Triangles</td>
</tr>
<tr>
<td>35</td>
<td>Glockenspiel</td>
</tr>
<tr>
<td>40</td>
<td>Harpsichord</td>
</tr>
<tr>
<td>65</td>
<td>Orchestra</td>
</tr>
<tr>
<td>66</td>
<td>Wind ensemble</td>
</tr>
<tr>
<td>69</td>
<td>Pop music</td>
</tr>
<tr>
<td>70</td>
<td>Pop music</td>
</tr>
<tr>
<td>+</td>
<td>101 Harmonic Complex tone</td>
</tr>
<tr>
<td>parameter</td>
<td>setting</td>
</tr>
<tr>
<td>----------------------------------------</td>
<td>--------------------------------------</td>
</tr>
<tr>
<td>Length of one segment</td>
<td>25 ms x 200 frames (5 s)</td>
</tr>
<tr>
<td>Number of bits of a payload (N)</td>
<td>4, 8</td>
</tr>
<tr>
<td>Number of classes of histogram (L)</td>
<td>6*N+1 (25, 49)</td>
</tr>
<tr>
<td>Thresholds</td>
<td>th0 = 1/100, th1=100</td>
</tr>
</tbody>
</table>
Robustness against MP3 conversion

- At the condition embedding 4bits of payload in the frequency band below 19th ERB (1540 - 2200 Hz), it can be seen that embedded payload against MP3 still remains.
- Moreover the error rates grows up in the condition when MP3 conversion is done twice.

: □-solid:mp3o-4bit, ▽-solid :mp3o-8bit, □-dotted :mp3t-4bit, ▽-dotted :mp3t-8bit
Subjective Listening test (Double-blind Triple-stimulus with Hidden Reference)

Reference (no watermark)
Hidden reference (same to Ref) & Object (watermarked)
- be randomly assigned to Left and Right

For each Left and Right button's signal,
the subjects had to assess the audibility of potential distortions from Ref. Signal and grade from 1.0 (very annoying, totally distorted) up to 5.0 (Imperceptible, no distorted at all)
Results of Listening test
(3 men & 1 female)

- Although the grades are varying between subjects, almost of the average SDGs are graded greater than -1 (Perceptible, but not annoying) for the instrumental signals or ensemble signals.
- In the case of the SDG for harmonic complex tone, the watermark was more annoying for most of the subjects when it was applied to a high frequency band.

\[
SDG = \text{Grade}_{\text{Test}} - \text{Grade}_{\text{Reference}}
\]
Multiband embedding

- In single band embedding, we can embed only 2 or 4 bits in 5 seconds. (0.4 or 0.8 bps!!)

- Embedding multi bits into multi-bands between 1st and 20th ERB region
Multiband embedding
(Embedding multi bits into multi-bands between 1\textsuperscript{st} and 20\textsuperscript{th} ERB region)

\begin{tabular}{|c|c|c|c|c|c|c|}
\hline
 & castanets & & Wind ensemble & & Harmonic complex & \\
 & 27-mp3o & 27-mp3t & 66-mp3o & 66-mp3t & 101-mp3o & 101-mp3t \\
\hline
4 bits \times 3 (12 bits) & 0.00 & 2.78 & 0.00 & 0.00 & 0.00 & 0.00 \\
4 bits \times 4 (16 bits) & 0.00 & 10.42 & 0.00 & 0.00 & 0.00 & 6.25 \\
8 bits \times 2 (16 bits) & 8.33 & 37.50 & 8.33 & 22.92 & 0.00 & 12.50 \\
4 bits \times 5 (20 bits) & 3.33 & 16.67 & 0.00 & 3.33 & 0.00 & 15.00 \\
8 bits \times 3 (24 bits) & 16.67 & 43.06 & 6.94 & 20.83 & 25.00 & 25.00 \\
\hline
\end{tabular}

(BER[\%])

- Results showed that in the multi-band case the proposed method can embed up to 20 bits (4 bps) without decrease in robustness against MP3.
Conclusion

- Re-quantize powers of STFT coefficients, embedding in the form of histogram
- Proposed method resists the MP3 compression attacks in the case of embedding payload below 2kHz
- Proposed method doesn't produce audible distortions
- Multi-band embedding can embed up to 20 bits in 5 seconds without decrease in robustness against MP3