Sound Detection through Transient Models using Wavelet Coefficient Trees

Michel VACHER, Dan ISTRATE and Jean-François SERIGNAT

CLIPS-IMAG Team GEOD, UMR 5524 CNRS-UJF-INPG, Grenoble (France)

CSIMTA, 19-22 septembre 2004
Outline

1. Application Context "H.I.S."
   - Surveillance System
   - Sound Corpus for Distress Recognition
   - Noisy Conditions in "H.I.S."

2. Transient Modeling

3. Detection Algorithms
   - Beginning of Sound
   - End of Sound

4. Sound Classification through GMM
   - Model Selection
   - Noise Attenuation
   - Results

5. Conclusion-Perspectives
Medical Telesurveillance

Various sensors:
- position
- infrared
- door contact
- physiology
- tensiometer
- heart rate
- microphones

Extracted informations through **Sound Analysis**:
- patient’s **activity**: door lock, phone...
- patient’s **physiology**: cough...
- **distress situation**: scream, glass breaking...
Sound Detection/Classification

- **Sound Acquisition**
  - Microphone
  - Sound Waveform

- **Sound Detection/Extraction**

- **Sound Classification**
  - C3
  - "RINGING PHONE"

**Application Context**
- Surveillance System
- Sound Corpus
- Noisy Conditions

**Transient Modeling**
- Detection Algorithms
  - Beginning of Sound
  - End of Sound

**Sound Classification**
- Model Selection
- Noise Attenuation
- Results

**Conclusion-Perspectives**
## Sound Corpus...

### Table: Sound Classes

<table>
<thead>
<tr>
<th>Sound Class</th>
<th>No of Frames</th>
<th>Duration of Each Sound</th>
</tr>
</thead>
<tbody>
<tr>
<td>Door Slap</td>
<td>47 398</td>
<td>140 ms-7.4 s</td>
</tr>
<tr>
<td>Breaking Glasses</td>
<td>9 338</td>
<td>330 ms-1.1 s</td>
</tr>
<tr>
<td>Ringing Phone</td>
<td>59 188</td>
<td>35 ms-10 s</td>
</tr>
<tr>
<td>Step Sound</td>
<td>3 648</td>
<td>1.4-5 s</td>
</tr>
<tr>
<td>Scream</td>
<td>17 509</td>
<td>370 ms-5.8 s</td>
</tr>
<tr>
<td>Dishes Sounds</td>
<td>7943</td>
<td>125 ms-1.35 s</td>
</tr>
<tr>
<td>Door Lock</td>
<td>605</td>
<td>24 ms-117 ms</td>
</tr>
</tbody>
</table>

- Frame Width = 16ms (256 samples at 16kHz)
- Variable Signal Duration, Ratio : 500/1
Noisy Corpus: 0, +10, +20, +40dB

- White Noise
- Experimental HIS Noise:
  - not stationary
  - recorded inside experimental apartment

Information from sound:
- Localisation (room)
- Identified sound class
- Detection time

In collaboration with TIMC laboratory

Room C311 (inside TIMC laboratory)
Discrete Wavelet Transform

- **Mother-Function:** *Daubechies Wavelet* $\psi$
  \[
  \left\{ \psi_{j,n}(t) = \frac{1}{\sqrt{2^j}} \psi \left( \frac{t - 2^j n}{2^j} \right) \right\}_{(j,n) \in \mathbb{Z}}
  \]

- **2048 samples**@ $f_s = 16kHz$ : **128ms windows**

![Daubechies Wavelet (5th coefficient)](image)

![Spectrum Modulus](image)
Wavelet and Spectrum

Daubechies Wavelet (5th coefficient)

Daubechies Wavelet (7th coefficient)

Spectrum Modulus

Time (samples)

Frequency (Hz)

Spectrum Modulus

Time (samples)

Frequency (Hz)
Frequency/Time Resolution

Figure: 3 Level Depth Tree of wavelet coefficients for N=2048 Samples, DWT Window Width=128 ms

- Time Resolution $\times$ Frequency Resolution $= \kappa$
- $\sum$ Energy of 3 Level Wavelet Trees
Wavelet Tree Energy

Figure: Sound Signal and Tree Energy

Tree Energy (summation of the 3 highest levels)

Chair fall in HIS noise occurring at t=10s
Proposed Detection Algorithm

Detection:

\[
\begin{align*}
\text{Energy and Statistical Analysis} & \quad \text{Threshold Determination} \\
{\square} & \quad {\blacksquare} \quad {\blacksquare} \quad {\blacksquare} \quad {\blacksquare} \\
\text{DWT} & \quad \text{Detection:} \\
\end{align*}
\]

"False" or "True"
### Beginning of Sound

- **ROC Curves**
  \[ \text{Missing Detection Rate} = f(\text{False Detection Rate}) \]

- **Equal Error Rate**: \( MDR = FDR \)

- **Detection EER**: 198 tests at each SNR level
  (99 noised sounds, 99 pure noise)

<table>
<thead>
<tr>
<th>Detection Method</th>
<th>SNR</th>
<th>HIS noise</th>
<th>White noise</th>
</tr>
</thead>
<tbody>
<tr>
<td>Several tree mean</td>
<td>0dB</td>
<td>6.7%</td>
<td>5.9%</td>
</tr>
<tr>
<td></td>
<td>( \geq +10dB )</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td></td>
<td>( \geq +10dB )</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td><strong>Filtered energy</strong></td>
<td>0dB</td>
<td>71.3%</td>
<td>19.2%</td>
</tr>
<tr>
<td>-conditioning</td>
<td>+10dB</td>
<td>45.2%</td>
<td>6.1%</td>
</tr>
<tr>
<td>~median filter-</td>
<td>+20dB</td>
<td>7.5%</td>
<td>6.1%</td>
</tr>
<tr>
<td></td>
<td>+40dB</td>
<td>6.1%</td>
<td>6.1%</td>
</tr>
</tbody>
</table>
Detection Delay for Sound Start

- Mean of detection delay for sound duration shorter than 2s for HIS noise (all sound classes)

<table>
<thead>
<tr>
<th></th>
<th>0dB</th>
<th>+10dB</th>
<th>+20dB</th>
<th>+40dB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Duration</td>
<td>23.6ms</td>
<td>13.9ms</td>
<td>9ms</td>
<td>5.5ms</td>
</tr>
</tbody>
</table>

- $\Delta t \leq 15\, \text{ms if } SNR \geq +10\, \text{dB}$

- **Safety Margin**: Systematic Advance of $25\, \text{ms}$
End of Sound Detection

1. **First Step**: Start of Sound Detection
2. **Time Inversion**
3. Detection on Time Inverted Signal using "First Step Algorithm"

![Graph showing a glass breaking at +40dB SNR, ending after 12,032 samples.](image)
End of Sound Detection

Figure: Extracted Cough Sound
A cough in HIS noise (SNR = 40dB)

- End of Signal is not lost
- Estimated Signal Duration = 1.1s
End of Sound Detection

Figure: Extracted Cough Noise at different SNR

- **SNR = 0dB**: End after 7,072 samples
- **SNR = 10dB**: End after 9,376 samples
- **SNR = 20dB**: End after 9,844 samples
End of Sound Detection: Results

1. Truncation at signal end: poor SNR in this part of signal
2. Extracted signal duration is shortened
3. Mean of the spread with real value (HIS noise and signals shorter than 2s)

<table>
<thead>
<tr>
<th></th>
<th>0dB</th>
<th>+10dB</th>
<th>+20dB</th>
<th>+40dB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Duration</td>
<td>560ms</td>
<td>433ms</td>
<td>335ms</td>
<td>10ms</td>
</tr>
</tbody>
</table>

4. Possible effects on classification:
   - No introduction of signal with only noise
   - Cut part of low amplitude
Gaussian Mixture Model

Training Step:

7 Sound Classes

K-MEANS
Expectation-Maximisation

7 GMM Models

Working Step:

Extracted Wave

Parameters Extraction
Maximum of Likelihood

Class Label
GMM Classification Method

1. Training Step: 7 Classes $\omega_k$
   - GMM Components for each Gaussian $m$
     - weight $\pi_{\kappa,m}$
     - average $\mu_{\kappa,m}$
     - covariance matrix $\Sigma_{\kappa,m}$

2. Working Step:
   - $x_i$: Parameter Vector for one Frame (16ms):
     - Likelihood Analysis: $p(x_i|\omega_k)$
   - $S = \{x_i; \ i = 1, n\}$ for the whole of Signal:
     - Geometrical Average:
       $$p(S|\omega_k) = \prod_{i=1}^{n} p(x_i|\omega_k)$$
   - Maximum Likelihood Criterion
## GMM Classification Method

**1. Training Step:** 7 Classes $\omega_\kappa$
- GMM Components for each Gaussian $m$
  - weight $\pi_{\kappa,m}$
  - average $\mu_{\kappa,m}$
  - covariance matrix $\Sigma_{\kappa,m}$

**2. Working Step:**
- $x_i$ Parameter Vector for one Frame (16ms):
  - Likelihood Analysis: $p(x_i|\omega_k)$
- $S = \{x_i; \ i = 1, n\}$ for the whole of Signal:
  - Geometrical Average:

\[
p(S|\omega_k) = \prod_{i=1}^{n} p(x_i|\omega_k)
\]

- Maximum Likelihood Criterion
16 MFCC Parameters

Extracted Signal:
- Frame Width = 16ms
- Time-lag between Windows = 8ms

Figure: Mel Frequency Cepstral Coefficients
Other Acoustical Parameters

1. ZCR: Zero Crossing Rate

2. RF: Roll Off Point

3. Centroid: Spectrum Energy Middle
Other Acoustical Parameters

1. **ZCR**: Zero Crossing Rate

2. **RF**: Roll Off Point

3. **Centroïd**: Spectrum Energy Middle
Other Acoustical Parameters

1. **ZCR:**
   - Zero Crossing Rate

2. **RF:**
   - Roll Off Point

3. **Centroïd:**
   - Spectrum Energy Middle
Model Selection (in Noiseless Conditions)

- Bayesian Information Criterion (BIC)
  \[ BIC_K = -2 \cdot L_K + \nu_K \cdot \ln(n) \]

- Number of components of the model: \( K \)
- Logarithmic maximum of likelihood: \( L_K \)
- Number of free parameters of the model: \( \nu_K \)
- Number of frames: \( n \)

Results for **door lock** class

<table>
<thead>
<tr>
<th>Number of Gaussian</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>BIC (16 MFCC)</td>
<td>11,000</td>
<td>10,752</td>
<td><strong>10,743</strong></td>
<td>10,757</td>
<td>13,400</td>
</tr>
</tbody>
</table>
Noise Attenuation by Wavelet Filtering

Extracted Signal:
- Frame Width = 8ms – 256 samples

Wavelet Filtering

Noise Estimation
@ t = t0 – 100ms
Noise Att. by Wavelet Filtering

1. **Noise Estimation** before Detection: $B_{max}^i, i \leq 7$

2. **Thresholds** on DWT Coefficients $K_i$:
   
   $\begin{align*}
   T_i &= 1.2 \times B_{max}^i \quad \text{for} \quad i \leq 2 \\
   T_i &= 0.9 \times B_{max}^i \quad \text{for} \quad i = 3 \\
   T_i &= 0 \quad \text{for} \quad i = 4 \ldots 7
   \end{align*}$

3. **Method F1**:
   
   $\begin{align*}
   K_i &= 0 \quad \text{if} \quad K_i < T_i \\
   K_i &= K_i \quad \text{if} \quad K_i \geq T_i
   \end{align*}$

4. **Method F2**:
   
   $\begin{align*}
   K_i &= 0 \quad \text{if} \quad K_i < T_i \\
   K_i &= K_i - (B_{max}^i / 10) \quad \text{if} \quad K_i \geq T_i
   \end{align*}$
Noise Estimation before Detection: $B_{\text{max}}^i, i \leq 7$

Thresholds on DWT Coefficients $K_i$:

\[
\begin{align*}
T_i &= 1.2 \times B_{\text{max}}^i \quad \text{for} \quad i \leq 2 \\
T_i &= 0.9 \times B_{\text{max}}^i \quad \text{for} \quad i = 3 \\
T_i &= 0 \quad \text{for} \quad i = 4 \ldots 7
\end{align*}
\]

Method F1:

\[
\begin{align*}
K_i &= 0 \quad \text{if} \quad K_i < T_i \\
K_i &= K_i \quad \text{if} \quad K_i \geq T_i
\end{align*}
\]

Method F2:

\[
\begin{align*}
K_i &= 0 \quad \text{if} \quad K_i < T_i \\
K_i &= K_i - (B_{\text{max}}^i/10) \quad \text{if} \quad K_i \geq T_i
\end{align*}
\]
Noise Detection

M. VACHER

Application Context
Surveillance System
Sound Corpus
Noisy Conditions

Transient Modeling
Detection Algorithms
Beginning of Sound
End of Sound

Sound Classification
Model Selection
Noise Attenuation
Results

Conclusion-Perspectives

Noise Estimation before Detection: $B_{max}^i, i \leq 7$

Thresholds on DWT Coefficients $K_i$:

$$
\begin{align*}
T_i &= 1.2 \times B_{max}^i \quad \text{for } i \leq 2 \\
T_i &= 0.9 \times B_{max}^i \quad \text{for } i = 3 \\
T_i &= 0 \quad \text{for } i = 4 \ldots 7
\end{align*}
$$

Method F1:

$$
\begin{align*}
K_i &= 0 \quad \text{if } K_i < T_i \\
K_i &= K_i \quad \text{if } K_i \geq T_i
\end{align*}
$$

Method F2:

$$
\begin{align*}
K_i &= 0 \quad \text{if } K_i < T_i \\
K_i &= K_i - (B_{max}^i/10) \quad \text{if } K_i \geq T_i
\end{align*}
$$

Noise Att. by Wavelet Filtering
Noise Estimation before Detection: $B^i_{\text{max}}$, $i \leq 7$

Thresholds on DWT Coefficients $K_i$:

$$
\begin{align*}
T_i &= 1.2 \times B^i_{\text{max}} \quad \text{for } i \leq 2 \\
T_i &= 0.9 \times B^i_{\text{max}} \quad \text{for } i = 3 \\
T_i &= 0 \quad \text{for } i = 4 \ldots 7
\end{align*}
$$

Method F1:

$$
\begin{align*}
K_i &= 0 \quad \text{if } K_i < T_i \\
K_i &= K_i \quad \text{if } K_i \geq T_i
\end{align*}
$$

Method F2:

$$
\begin{align*}
K_i &= 0 \quad \text{if } K_i < T_i \\
K_i &= K_i - (B^i_{\text{max}}/10) \quad \text{if } K_i \geq T_i
\end{align*}
$$
Classification Results

ECR: Error Classification Rate

Tab.: ECR for 16MFCC+ZCR+RF+Centroid in HIS noise presence (1577 tests for each SNR)

<table>
<thead>
<tr>
<th>Filtering</th>
<th>SNR [dB]</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0</td>
</tr>
<tr>
<td>Without</td>
<td>48.3</td>
</tr>
<tr>
<td>With F1</td>
<td>40</td>
</tr>
<tr>
<td>With F2</td>
<td>40.4</td>
</tr>
</tbody>
</table>

- Gain of **8%** in noise conditions
- ECR **10%** for "pure" sounds
Conclusion

1. Sound Extraction:
   - low delay after signal beginning
   - acceptable truncation at end of signal

2. Sound Classification:
   - ECR $\leq 20\%$ for 10dB SNR

3. Operational System under Realistic Conditions
Conclusion

1. Sound Extraction:
   - low delay after signal beginning
   - acceptable truncation at end of signal

2. Sound Classification:
   - $ECR \leq 20\%$ for 10dB SNR

3. Operational System under Realistic Conditions
Conclusion

1. Sound Extraction:
   - low delay after signal beginning
   - acceptable truncation at end of signal

2. Sound Classification:
   - $ECR \leq 20\%$ for $10\text{dB SNR}$

3. Operational System under Realistic Conditions
Perspectives

Actual Work:

1. Real-Time Tests in Experimental Conditions
2. Speech Recognition (Call for Help)
   - Segmentation between Sound and Speech
   - Speech-Corpus adapted to Distress Situations
   - Real-Time Speech-Recognition System

Possible Applications:

1. Multimedia Classification
2. Security Sound Surveillance
Perspectives

Actual Work:

1. Real-Time Tests in Experimental Conditions
2. Speech Recognition (Call for Help)
   - Segmentation between Sound and Speech
   - Speech-Corpus adapted to Distress Situations
   - Real-Time Speech-Recognition System

Possible Applications:

1. Multimedia Classification
2. Security Sound Surveillance
Thank you for your attention.
Daubechies Wavelets

- Mother-Function: *Daubechies Wavelet* $\psi$
  - with 6 vanishing moments
  - function with finite energy and fast decay
- scalings ($s$ factor) and translations ($u$ delay) of the mother wavelet function $\psi(t)$

\[
\left\{ \psi_{u,s}(t) = \frac{1}{\sqrt{s}} \psi \left( \frac{t - u}{s} \right) \right\}_{(u,s) \in \mathbb{R}}
\]

- DWT:

\[
\left\{ \psi_{j,n}(t) = \frac{1}{\sqrt{2^j}} \psi \left( \frac{t - 2^j n}{2^j} \right) \right\}_{(j,n) \in \mathbb{Z}}
\]

- 2048 samples@$f_s = 16kHz : 128ms$ windows
Fisher Discriminant Ratio

\[ FDR = \frac{\sum_{i=1}^{k} \sum_{j=1}^{k} (\bar{x}[i] - \bar{x}[j])^2}{\sum_{i=1}^{k} \text{Var}(x)[i]} \]

- The class: i,j
- \( \bar{x}[i] \) and \( \bar{x}[j] \) are parameter average for all classes
- \( \text{Var}(x)[i] \) is parameter variance for all classes
## FDR 2/2

<table>
<thead>
<tr>
<th>Parameter</th>
<th>FDR</th>
<th>Parameter</th>
<th>FDR</th>
</tr>
</thead>
<tbody>
<tr>
<td>MFCC1</td>
<td>2.72</td>
<td>MFCC11</td>
<td>2.88</td>
</tr>
<tr>
<td>MFCC2</td>
<td>16.07</td>
<td>MFCC12</td>
<td>3.20</td>
</tr>
<tr>
<td>MFCC3</td>
<td>10.33</td>
<td>MFCC13</td>
<td>1.48</td>
</tr>
<tr>
<td>MFCC4</td>
<td>10.02</td>
<td>MFCC14</td>
<td>3.61</td>
</tr>
<tr>
<td>MFCC5</td>
<td>2.01</td>
<td>MFCC15</td>
<td>3.26</td>
</tr>
<tr>
<td>MFCC6</td>
<td>2.91</td>
<td>MFCC16</td>
<td>4.41</td>
</tr>
<tr>
<td>MFCC7</td>
<td>3.36</td>
<td>ZCR</td>
<td>17.99</td>
</tr>
<tr>
<td>MFCC8</td>
<td>3.60</td>
<td>RF</td>
<td>16.70</td>
</tr>
<tr>
<td>MFCC9</td>
<td>0.53</td>
<td>Centroïd</td>
<td>23.75</td>
</tr>
<tr>
<td>MFCC10</td>
<td>3.34</td>
<td>Energy</td>
<td>2.54</td>
</tr>
</tbody>
</table>
Test Set

- Constitution of a everyday life sound data base
  - 3354 signals
  - CLIPS recording, RWCP CD and commercial film CD

- Detection test set:
  - 11 everyday life sounds (3 repetitions)
  - 3 environmental noises (white, real conditions and water noise)
  - 4 Signal to Noise Ratios (0, 10, 20, 40 dB)

- Classification test set:
  - 7 sound classes
  - 20 - 500 signals / class
Classification Test

Sound Classification according to training

- 7 GMM models
- One acoustical vector $x_i$
- Probability calculation for each class

\[
p(x_i \mid \omega_k) = \sum_{m=1}^{4} \pi_{k,m} \cdot \frac{1}{(2\pi)^{d/2} |\sum_{k,m}|^{1/2}} \cdot \exp(A_{i,k,m})
\]

\[
A_{i,k,m} = \left( -\frac{1}{2} (x_i - \mu_{k,m})^T \cdot \Sigma_{k,m}^{-1} \cdot (x_i - \mu_{k,m}) \right) \\
1 \leq k \leq 7
\]