Emotion Recognition in Speech under Environmental Noise Conditions using Wavelet Decomposition

J.C. Vásquez-Correa\textsuperscript{1}, N. García\textsuperscript{1}, J.R. Orozco-Arroyave\textsuperscript{1,2}, J.D. Arias-Londoño\textsuperscript{1}, J.F. Vargas-Bonilla\textsuperscript{1}, Elmar Nöth\textsuperscript{2}

\textsuperscript{1}Faculty of Engineering, University of Antioquia UdeA, Medellín, Colombia.
\textsuperscript{2}Pattern Recognition Lab., Friedrich-Alexander-Universität, Erlangen-Nürnberg, Germany

jesus.vargas@udea.edu.co
Introduction: Emotion recognition

Recognition of emotion in speech:

- Call centers
- Emergency services
- Psychologic therapy
- Intelligent vehicles
- Public surveillance
Introduction: Fear-type emotions
Introduction: Challenges

- Naturalness of databases (Acted, Natural, Evoked)
- Large set of features
- Acoustic conditions (Telephone, Background noise)
Introduction: Previous Work (2-class)

- Emotion recognition under AWGN noise
- Emotion recognition under GSM and wired-line telephone channel

<table>
<thead>
<tr>
<th>Condition</th>
<th>Original</th>
<th>Affected</th>
<th>KLT</th>
<th>logMMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>AWGN SNR=3dB</td>
<td>76.9%</td>
<td>71.3%</td>
<td>78.1%</td>
<td>74.7%</td>
</tr>
<tr>
<td>AWGN SNR=10dB</td>
<td>76.9%</td>
<td>74.7%</td>
<td>80.1%</td>
<td>76.7%</td>
</tr>
<tr>
<td>GSM channel</td>
<td>76.9%</td>
<td>77.8%</td>
<td>62.9%</td>
<td>70.6%</td>
</tr>
<tr>
<td>wired-line</td>
<td>76.9%</td>
<td>65.2%</td>
<td>59.0%</td>
<td>75.1%</td>
</tr>
</tbody>
</table>

**Table:** Emotion recognition Berlin database
Methodology

A new characterization approach based on wavelet packet transform for recognition of emotions in speech evaluated in non-controlled noise conditions.

- Log-energy
- Log-energy entropy
- MFCC
- Lempel-Ziv complexity
Methodology: Characterization

Wavelet decomposition **Voiced segments**

\[ x[n] \]

\[ W_1,0 \]

\[ W_2,0 \]

\[ W_3,0 \]

\[ W_4,0 \]

\[ W_5,0 \]

\[ W_6,2 \]

\[ W_6,3 \]

\[ W_6,4 \]

\[ W_6,5 \]

\[ W_3,1 \]

\[ W_4,1 \]

\[ W_5,1 \]

\[ W_6,3 \]

\[ W_6,4 \]

\[ W_6,5 \]

\[ W_4,2 \]

\[ W_4,3 \]

\[ W_4,6 \]

\[ W_4,7 \]

\[ W_5,5 \]

\[ W_5,6 \]

\[ W_3,2 \]

\[ W_3,3 \]

\[ W_3,4 \]

\[ W_3,5 \]

\[ W_3,6 \]

\[ W_3,7 \]

\[ W_2,1 \]

\[ W_2,2 \]

\[ W_2,3 \]

\[ W_1,1 \]

Wavelet decomposition **Unvoiced segments**

\[ x[n] \]

\[ W_1,0 \]

\[ W_2,0 \]

\[ W_3,0 \]

\[ W_4,0 \]

\[ W_5,0 \]

\[ W_6,2 \]

\[ W_6,3 \]

\[ W_6,4 \]

\[ W_6,5 \]

\[ W_3,1 \]

\[ W_4,1 \]

\[ W_5,1 \]

\[ W_6,3 \]

\[ W_6,4 \]

\[ W_6,5 \]

\[ W_4,2 \]

\[ W_4,3 \]

\[ W_4,6 \]

\[ W_4,7 \]

\[ W_5,5 \]

\[ W_5,6 \]

\[ W_3,2 \]

\[ W_3,3 \]

\[ W_3,4 \]

\[ W_3,5 \]

\[ W_3,6 \]

\[ W_3,7 \]

\[ W_2,1 \]

\[ W_2,2 \]

\[ W_2,3 \]

\[ W_1,1 \]
### Databases

<table>
<thead>
<tr>
<th>database</th>
<th># recordings</th>
<th>Speakers</th>
<th>Fs (Hz)</th>
<th>Naturalness</th>
<th>Emotions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Berlin</td>
<td>534</td>
<td>12</td>
<td>16000</td>
<td>Acted</td>
<td>Hot anger, Boredorm, Disgust, Anxiety/Fear, Happiness, Sadness, Neutral</td>
</tr>
<tr>
<td>Enterface05 (Audio-Video)</td>
<td>1317</td>
<td>44</td>
<td>44100</td>
<td>Evoked</td>
<td>Hot anger, Happiness, Disgust, Anxiety/Fear, Sadness, Surprise</td>
</tr>
</tbody>
</table>
# Experiments

<table>
<thead>
<tr>
<th>Experiment</th>
<th>Berlin DB</th>
<th>enterface05 DB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Multi-class</td>
<td>Anger, Disgust, Fear, Neutral</td>
<td>Anger, Disgust, Fear</td>
</tr>
<tr>
<td>2-class</td>
<td>(Anger, disgust, fear) vs Neutral</td>
<td>(Anger, disgust, fear, sadness) vs (Happiness, Surprise)</td>
</tr>
</tbody>
</table>

**Table:** Experiments performed
Methodology: Classification

Features voiced segments
  ↓
  GMM
  ↓
  GMM emotion k
  ↓
  Posterior probabilities

Features unvoiced segments
  ↓
  GMM emotion 1
  ↓
  Posterior probabilities
  ↓
  GMM emotion k
  ↓
  Posterior probabilities

Multi-class SVM (C, γ)
## Results: Original signals

<table>
<thead>
<tr>
<th>Segments</th>
<th>feat.</th>
<th>Class. task</th>
<th>Berlin DB</th>
<th>enterface05 DB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Voiced</td>
<td>120</td>
<td>multi-class</td>
<td>80.0 ± 11.6</td>
<td>57.7 ± 6.8</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2-class</td>
<td>89.9 ± 7.8</td>
<td>65.1 ± 4.6</td>
</tr>
<tr>
<td>Unvoiced</td>
<td>120</td>
<td>multi-class</td>
<td>62.5 ± 5.0</td>
<td>55.4 ± 6.8</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2-class</td>
<td>82.5 ± 8.6</td>
<td>64.6 ± 6.0</td>
</tr>
<tr>
<td>Fusion</td>
<td></td>
<td>multi-class</td>
<td>74.7 ± 11.9</td>
<td>61.6 ± 4.5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2-class</td>
<td>94.6 ± 5.1</td>
<td>69.2 ± 1.5</td>
</tr>
<tr>
<td>all signal</td>
<td>384</td>
<td>multi-class</td>
<td>84.3 ± 6.6</td>
<td>66.6 ± 4.2</td>
</tr>
<tr>
<td>openEAR [Eyben2012]</td>
<td></td>
<td>2-class</td>
<td>94.9 ± 4.1</td>
<td>68.6 ± 4.8</td>
</tr>
</tbody>
</table>

**Table:** Accuracy for original non-affected speech signals
Results: Original signals

<table>
<thead>
<tr>
<th>Segments</th>
<th>feat.</th>
<th>Class. task</th>
<th>Berlin DB</th>
<th>interface05 DB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Voiced</td>
<td>120</td>
<td>multi-class</td>
<td>80.0 ± 11.6</td>
<td>57.7 ± 6.8</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2-class</td>
<td>89.9 ± 7.8</td>
<td>65.1 ± 4.6</td>
</tr>
<tr>
<td>Unvoiced</td>
<td>120</td>
<td>multi-class</td>
<td>62.5 ± 5.0</td>
<td>55.4 ± 6.8</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2-class</td>
<td>82.5 ± 8.6</td>
<td>64.6 ± 6.0</td>
</tr>
<tr>
<td>Fusion</td>
<td></td>
<td>multi-class</td>
<td>74.7 ± 11.9</td>
<td>61.6 ± 4.5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2-class</td>
<td>94.6 ± 5.1</td>
<td>69.2 ± 1.5</td>
</tr>
<tr>
<td>all signal</td>
<td>384</td>
<td>multi-class</td>
<td>84.3 ± 6.6</td>
<td>66.6 ± 4.2</td>
</tr>
<tr>
<td>openEAR [Eyben2012]</td>
<td></td>
<td>2-class</td>
<td>94.9 ± 4.1</td>
<td>68.6 ± 4.8</td>
</tr>
</tbody>
</table>

**Table:** Accuracy for original non-affected speech signals
## Results: Original signals

<table>
<thead>
<tr>
<th>Segments</th>
<th>feat.</th>
<th>Class. task</th>
<th>Berlin DB</th>
<th>enterface05 DB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Voiced</td>
<td>120</td>
<td>multi-class</td>
<td>80.0 ± 11.6</td>
<td>57.7 ± 6.8</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2-class</td>
<td>89.9 ± 7.8</td>
<td>65.1 ± 4.6</td>
</tr>
<tr>
<td>Unvoiced</td>
<td>120</td>
<td>multi-class</td>
<td>62.5 ± 5.0</td>
<td>55.4 ± 6.8</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2-class</td>
<td>82.5 ± 8.6</td>
<td>64.6 ± 6.0</td>
</tr>
<tr>
<td>Fusion</td>
<td></td>
<td>multi-class</td>
<td>74.7 ± 11.9</td>
<td>61.6 ± 4.5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2-class</td>
<td>94.6 ± 5.1</td>
<td>69.2 ± 1.5</td>
</tr>
<tr>
<td>all signal</td>
<td>384</td>
<td>multi-class</td>
<td>84.3 ± 6.6</td>
<td>66.6 ± 4.2</td>
</tr>
<tr>
<td>openEAR [Eyben2012]</td>
<td></td>
<td>2-class</td>
<td>94.9 ± 4.1</td>
<td>68.6 ± 4.8</td>
</tr>
</tbody>
</table>

**Table:** Accuracy for original non-affected speech signals.  
**Previous Work:** 76.9%
### Table: Accuracy for original non-affected speech signals

<table>
<thead>
<tr>
<th>Segments</th>
<th>feat.</th>
<th>Class. task</th>
<th>Berlin DB</th>
<th>enterface05 DB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Voiced</td>
<td>120</td>
<td>multi-class</td>
<td>80.0 ± 11.6</td>
<td>57.7 ± 6.8</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2-class</td>
<td>89.9 ± 7.8</td>
<td>65.1 ± 4.6</td>
</tr>
<tr>
<td>Unvoiced</td>
<td>120</td>
<td>multi-class</td>
<td>62.5 ± 5.0</td>
<td>55.4 ± 6.8</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2-class</td>
<td>82.5 ± 8.6</td>
<td>64.6 ± 6.0</td>
</tr>
<tr>
<td>Fusion</td>
<td></td>
<td>multi-class</td>
<td>74.7 ± 11.9</td>
<td>61.6 ± 4.5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2-class</td>
<td>94.6 ± 5.1</td>
<td>69.2 ± 1.5</td>
</tr>
<tr>
<td>all signal</td>
<td>384</td>
<td>multi-class</td>
<td>84.3 ± 6.6</td>
<td>66.6 ± 4.2</td>
</tr>
<tr>
<td>openEAR [Eyben2012]</td>
<td></td>
<td>2-class</td>
<td>94.9 ± 4.1</td>
<td>68.6 ± 4.8</td>
</tr>
</tbody>
</table>
## Results: Original signals

<table>
<thead>
<tr>
<th>Segments</th>
<th>feat.</th>
<th>Class. task</th>
<th>Berlin DB</th>
<th>enterface05 DB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Voiced</td>
<td>120</td>
<td>multi-class</td>
<td>80.0 ± 11.6</td>
<td>57.7 ± 6.8</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2-class</td>
<td>89.9 ± 7.8</td>
<td>65.1 ± 4.6</td>
</tr>
<tr>
<td>Unvoiced</td>
<td>120</td>
<td>multi-class</td>
<td>62.5 ± 5.0</td>
<td>55.4 ± 6.8</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2-class</td>
<td>82.5 ± 8.6</td>
<td>64.6 ± 6.0</td>
</tr>
<tr>
<td>Fusion</td>
<td></td>
<td>multi-class</td>
<td>74.7 ± 11.9</td>
<td>61.6 ± 4.5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2-class</td>
<td>94.6 ± 5.1</td>
<td>69.2 ± 1.5</td>
</tr>
<tr>
<td>all signal</td>
<td>384</td>
<td>multi-class</td>
<td>84.3 ± 6.6</td>
<td>66.6 ± 4.2</td>
</tr>
<tr>
<td>openEAR [Eyben2012]</td>
<td></td>
<td>2-class</td>
<td>94.9 ± 4.1</td>
<td>68.6 ± 4.8</td>
</tr>
</tbody>
</table>

**Table:** Accuracy for original non-affected speech signals
Experiments: Environments

- Original non-affected speech signals
- Cafeteria babble noise
- Street noise

- KLT algorithm
- LogMMSE algorithm

SNR evaluated ranges from -3dB to 6dB
Results: Affected signals, 2-class (OpenEAR)

Berlin database

enterface05 database

Original
Noisy Caf
Noisy Street
KLT Caf
KLT Street
LogMMSE Caf
LogMMSE Street

Accuracy (%)
SNR (dB)
Results: Affected signals, M-class (OpenEAR)

Berlin database

enterface05 database

Accuracy (%) vs. SNR (dB)

- Original
- Noisy Caf
- Noisy Street
- KLT Caf
- KLT Street
- LogMMSE Caf
- LogMMSE Street
### Databases

<table>
<thead>
<tr>
<th>Database</th>
<th># Recordings</th>
<th>Speakers</th>
<th>Fs (Hz)</th>
<th>Naturalness</th>
</tr>
</thead>
<tbody>
<tr>
<td>Berlin</td>
<td>534</td>
<td>12</td>
<td>16000</td>
<td>Acted</td>
</tr>
<tr>
<td>Enterface05 (Audio-Video)</td>
<td>1317</td>
<td>44</td>
<td>44100</td>
<td>Evoked</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Segments</th>
<th>Classif task</th>
<th>enterface05 logMMSE</th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>openEAR</td>
<td>multi-class</td>
<td>66.9 ± 4.2</td>
<td>+0.3</td>
</tr>
<tr>
<td></td>
<td>2-class</td>
<td>68.8 ± 3.1</td>
<td>+0.2</td>
</tr>
</tbody>
</table>
Results: Affected signals, 2class (WPT)

Berlin database

enterface05 database

Accuracy (%) vs. SNR (dB) for different conditions and algorithms.
Results: Affected signals, 2-class (OpenEAR)

![Graph showing accuracy (%) vs. SNR (dB) for Berlin and enterface05 databases]

- **Berlin database**
  - Accuracy: 100%
  - SNR (dB): -3, -2, -1, 0, 1, 2, 3, 4, 5, 6

- **enterface05 database**
  - Original
  - Noisy Caf
  - Noisy Street
  - KLT Caf
  - KLT Street
  - LogMMSE Caf
  - LogMMSE Street

Accuracy (%) vs. SNR (dB) for Berlin database:
- Original: 100%
- Noisy Caf: 90%
- Noisy Street: 80%
- KLT Caf: 85%
- KLT Street: 80%
- LogMMSE Caf: 75%
- LogMMSE Street: 70%

Accuracy (%) vs. SNR (dB) for enterface05 database:
- Original: 100%
- Noisy Caf: 90%
- Noisy Street: 80%
- KLT Caf: 85%
- KLT Street: 80%
- LogMMSE Caf: 75%
- LogMMSE Street: 70%
1. A different scheme for feature extraction based on WPT is presented, it highlights the low frequency zone from the speech signal. Its performance it is acceptable for the 2-class problem when compared with a well established scheme as OpenEAR.

2. The use of WPT in low frequency bands must be evaluated more deeply in order to improve performance for Multi-class problem.

3. Other features calculated from the wavelet decompositions must be considered, specially for unvoiced segments.
4. New methodology seems to be more robust against non-controlled conditions. Although logMMSE algorithm outperforms KLT, performance for Speech Enhancement is not good enough. The affectation produced by the cafeteria babble noise is more critical than the produced by the street noise.

5. Evaluation of non-additive environmental noise must be addressed in the future.
Thanks!
Q?

jesus.vargas@udea.edu.co
Thanks!

Q?

jesus.vargas@udea.edu.co