Computational Paralinguistics in Everyday Environments

Björn W. Schuller

Imperial College London / UK
Machine Learning Group

University of Passau / Germany
Chair Complex & Intelligent Systems

audEERING GmbH / Germany
vs
**Speech Recognition**

- **1950**: single speaker, digits
- **1970**: 1000 words
- **1980**: several 1000 words
- **1990**: trained dictation
- **2000**: robust, million words
- **2010**: everyday usage

**Speaker Classification**

- **1960**: speaker identification
- **1997**: single speaker, emotion
- **2010**: few isolated states/traits
- **2020**: everyday usage?
Paralinguistics.

- **Speech Under Eating & Food**
  - 30 subjects, 6 food types, +ASR features

---

<table>
<thead>
<tr>
<th>Year</th>
<th>Category</th>
<th>Classes</th>
<th>%UA/AUC/CC</th>
</tr>
</thead>
<tbody>
<tr>
<td>2016</td>
<td>Deception</td>
<td>2</td>
<td>72.1</td>
</tr>
<tr>
<td></td>
<td>Sincerity</td>
<td>[0,1]</td>
<td>65.4+</td>
</tr>
<tr>
<td></td>
<td>Native Lang.</td>
<td>11</td>
<td>82.2</td>
</tr>
<tr>
<td>2015</td>
<td>Nativeness</td>
<td>[0,1]</td>
<td>43.3+</td>
</tr>
<tr>
<td></td>
<td>Parkinson’s</td>
<td>[0,100]</td>
<td>54.0+</td>
</tr>
<tr>
<td></td>
<td>Eating</td>
<td>7</td>
<td>62.7</td>
</tr>
<tr>
<td>2014</td>
<td>Cognitive Load</td>
<td>3</td>
<td>61.6</td>
</tr>
<tr>
<td></td>
<td>Physical Load</td>
<td>2</td>
<td>71.9</td>
</tr>
<tr>
<td>2013</td>
<td>Social Signals</td>
<td>2x2</td>
<td>92.7*</td>
</tr>
<tr>
<td></td>
<td>Conflict</td>
<td>2</td>
<td>85.9</td>
</tr>
<tr>
<td></td>
<td>Emotion</td>
<td>12</td>
<td>46.1</td>
</tr>
<tr>
<td></td>
<td>Autism</td>
<td>4</td>
<td>69.4</td>
</tr>
</tbody>
</table>
### Paralings.

<table>
<thead>
<tr>
<th>Year</th>
<th>Feature</th>
<th># Classes</th>
<th>%UA/*AUC/+CC</th>
</tr>
</thead>
<tbody>
<tr>
<td>2012</td>
<td>Personality</td>
<td>5x2</td>
<td>70.4</td>
</tr>
<tr>
<td></td>
<td>Likability</td>
<td>2</td>
<td>68.7</td>
</tr>
<tr>
<td></td>
<td>Intelligibility</td>
<td>2</td>
<td>76.8</td>
</tr>
<tr>
<td>2011</td>
<td>Intoxication</td>
<td>2</td>
<td>72.2</td>
</tr>
<tr>
<td></td>
<td>Sleepiness</td>
<td>2</td>
<td>72.5</td>
</tr>
<tr>
<td>2010</td>
<td>Age</td>
<td>4</td>
<td>53.6</td>
</tr>
<tr>
<td></td>
<td>Gender</td>
<td>3</td>
<td>85.7</td>
</tr>
<tr>
<td></td>
<td>Interest</td>
<td>[-1,1]</td>
<td>42.8+</td>
</tr>
<tr>
<td>2009</td>
<td>Emotion</td>
<td>5</td>
<td>44.0</td>
</tr>
<tr>
<td></td>
<td>Negativity</td>
<td>2</td>
<td>71.2</td>
</tr>
</tbody>
</table>
Paralings.

- Pseudo Multimodality

*MAE 8.4*
+CC 0.908*
%UA
Heart Rate
Skin Conductance 65.0
Facial Action Units 67.4
Eye-Contact

Inner Brow Raiser | Outer Brow Raiser | Brow Lowerer | Cheek Raiser | Lid Tightener | Upper Lip Raiser | Lip Corner Puller | Chin Raiser

- X -
Acoustic Robustness.
Features

Continuous Signals
- F0
- Energy
- Lin. Prediction
- Spectrum
- TF-Transform
- Formants
- Harmonicity
- Perturbation

Symbolic Information
- Events
- Semantics

Low-Level-Descriptors

Deriving

Filtering

 Chunking
- Extremes
- Means
- Percentiles
- Higher Moments
- Peaks
- Segments
- Regression

Deriving

Filtering

 TOKENIZING

 Vector Space

 Look-Up
Feature Robustness

Pitch Detection

- PDA in Time Domain
- PDA by Short Time Principle
- Determination of 1. Partial
- Analysis of Time Signal
- Correlation
- Analysis in Freq. domain
- Simplification of structure
- Maximum Likelihood
Feature Robustness

- **Pitch (FAU Aibo Corpus)**
  
  67.9% voiced frames, ~ 6% erroneous pitch (>10 % deviation)

<table>
<thead>
<tr>
<th>type</th>
<th># frames</th>
<th>percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>identical</td>
<td>574,485</td>
<td>93.67</td>
</tr>
<tr>
<td>small errors</td>
<td>452</td>
<td>0.07</td>
</tr>
<tr>
<td>voiced errors</td>
<td>8,804</td>
<td>1.43</td>
</tr>
<tr>
<td>unvoiced errors</td>
<td>1,877</td>
<td>0.30</td>
</tr>
<tr>
<td>octave errors ↓</td>
<td>23,498</td>
<td>3.83</td>
</tr>
<tr>
<td>octave errors ↑</td>
<td>239</td>
<td>0.03</td>
</tr>
<tr>
<td>other gross errors</td>
<td>3,923</td>
<td>0.63</td>
</tr>
</tbody>
</table>

~2.0% loss in recognition accuracy (duration features less affected)

End-2-End Learning

- Convolutional RNNs

<table>
<thead>
<tr>
<th>Arousal</th>
<th>CC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>.366</td>
</tr>
<tr>
<td>Deep CRNN</td>
<td>.686</td>
</tr>
</tbody>
</table>

End-2-End Learning

- **Example: AVEC 2016**

  ![Graph showing cell activations and prosodic feature over time](image)

  energy range (.77), loudness (.73), F0 mean (.71)

Timing

- **Gating**
  
  Implications for feature normalization, on-set detection, etc.

  One second suffices?

---

Timing

- **Learning Temporal Context**
  
  LSTM: Sequential Jacobian

"Context-Sensitive Learning for Enhanced Audiovisual Emotion Classification",
Bag-of-Audio-Words

Split Vector Quantisation
+ Histogram

openXBOW
# Features

Comparison on the RECOLA (AVEC 2016) task

<table>
<thead>
<tr>
<th></th>
<th>Arousal</th>
<th>Valence</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>CCC</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Valid/Test</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Functionals</td>
<td>.790/.720</td>
<td>.459/.402</td>
</tr>
<tr>
<td>BLSTM-RNN</td>
<td>.800/??</td>
<td>.398/??</td>
</tr>
<tr>
<td>CNN (e2e)</td>
<td>.741/.686</td>
<td>.325/.261</td>
</tr>
<tr>
<td>BoAW</td>
<td>.793/.753</td>
<td>.550/.430</td>
</tr>
<tr>
<td>BoAW+Fctls</td>
<td>.799/.738</td>
<td>.521/.465</td>
</tr>
</tbody>
</table>

Acoustic Robustness

- Additive Noise

<table>
<thead>
<tr>
<th>Accuracy [%]</th>
<th>-</th>
<th>NA</th>
<th>SA</th>
<th>NSA</th>
<th>NSA+FS</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>EMO-DB</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Clean Speech</td>
<td>74.9</td>
<td>-</td>
<td>79.6</td>
<td>-</td>
<td>80.4</td>
</tr>
<tr>
<td>Car Noise</td>
<td>60.5</td>
<td>72.1</td>
<td>75.1</td>
<td>76.3</td>
<td>77.3</td>
</tr>
<tr>
<td>Babble Noise</td>
<td>70.0</td>
<td>76.1</td>
<td>77.9</td>
<td>78.7</td>
<td>80.5</td>
</tr>
<tr>
<td>Babble+MINI</td>
<td>46.6</td>
<td>70.4</td>
<td>75.7</td>
<td>76.1</td>
<td>79.5</td>
</tr>
<tr>
<td><strong>eNTERFACE</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Clean Speech</td>
<td>54.2</td>
<td>-</td>
<td>61.4</td>
<td>-</td>
<td>62.8</td>
</tr>
<tr>
<td>Car Noise</td>
<td>38.5</td>
<td>48.3</td>
<td>51.8</td>
<td>56.7</td>
<td>59.7</td>
</tr>
<tr>
<td>Babble Noise</td>
<td>42.1</td>
<td>53.2</td>
<td>54.2</td>
<td>61.0</td>
<td>61.6</td>
</tr>
<tr>
<td>Babble+MINI</td>
<td>30.6</td>
<td>49.8</td>
<td>46.2</td>
<td>55.8</td>
<td>58.6</td>
</tr>
</tbody>
</table>

Acoustic Robustness

- Reverberation
  Matching to Acoustics (MA) Space (MS)

Acoustic Robustness

- **NMF Features**

  Emotion Challenge Task

<table>
<thead>
<tr>
<th></th>
<th>UAR [%]</th>
<th>C</th>
<th>CT_{RM}</th>
<th>RM</th>
<th>CTRV</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td>1.0</td>
<td>67.62</td>
<td>60.51</td>
<td>53.06</td>
<td>60.40</td>
<td></td>
</tr>
<tr>
<td>N30</td>
<td>1.0</td>
<td>65.48</td>
<td>52.36</td>
<td>50.23</td>
<td>56.02</td>
<td></td>
</tr>
<tr>
<td>N31I</td>
<td>1.0</td>
<td>65.54</td>
<td>53.10</td>
<td>50.36</td>
<td>56.33</td>
<td></td>
</tr>
<tr>
<td>IS + N30</td>
<td>0.5</td>
<td>67.37</td>
<td>49.15</td>
<td>51.62</td>
<td>56.05</td>
<td></td>
</tr>
<tr>
<td>IS + N31I</td>
<td>1.0</td>
<td>67.15</td>
<td>56.47</td>
<td>51.95</td>
<td>58.52</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>UAR [%]</th>
<th>C</th>
<th>CT_{RM}</th>
<th>RM</th>
<th>CTRV</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td>0.02</td>
<td>61.61</td>
<td>62.72</td>
<td>62.10</td>
<td>62.14</td>
<td></td>
</tr>
<tr>
<td>N30</td>
<td>0.2</td>
<td>53.57</td>
<td>65.61</td>
<td>54.87</td>
<td>58.02</td>
<td></td>
</tr>
<tr>
<td>N31I</td>
<td>0.5</td>
<td>54.50</td>
<td>66.54</td>
<td>56.20</td>
<td>59.08</td>
<td></td>
</tr>
<tr>
<td>IS + N30</td>
<td>0.05</td>
<td>65.13</td>
<td>66.26</td>
<td>60.39</td>
<td>63.93</td>
<td></td>
</tr>
<tr>
<td>IS + N31I</td>
<td>0.05</td>
<td>64.68</td>
<td>66.34</td>
<td>59.54</td>
<td>63.52</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>UAR [%]</th>
<th>C</th>
<th>CT_{RM}</th>
<th>RM</th>
<th>CTRV</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td>0.01</td>
<td>67.72</td>
<td>59.52</td>
<td>66.06</td>
<td>64.43</td>
<td></td>
</tr>
<tr>
<td>N30</td>
<td>0.05</td>
<td>66.73</td>
<td>67.55</td>
<td>52.66</td>
<td>62.31</td>
<td></td>
</tr>
<tr>
<td>N31I</td>
<td>0.2</td>
<td>65.81</td>
<td>64.61</td>
<td>63.32</td>
<td>64.58</td>
<td></td>
</tr>
<tr>
<td>IS + N30</td>
<td>0.005</td>
<td>67.64</td>
<td>62.64</td>
<td>66.78</td>
<td>65.69</td>
<td></td>
</tr>
<tr>
<td>IS + N31I</td>
<td>0.005</td>
<td>67.07</td>
<td>61.85</td>
<td>65.92</td>
<td>64.95</td>
<td></td>
</tr>
</tbody>
</table>

“Recognition of Non-Prototypical Emotions in Reverberated and Noisy Speech by Non-Negative Matrix Factorization”, JASP, 2012.
Acoustic Robustness

- **Multicondition**
  
  Feature Selection +
  
  Training

---

```
<table>
<thead>
<tr>
<th>[%] UAR</th>
<th>A</th>
<th>V</th>
<th>LOI</th>
<th>A</th>
<th>V</th>
<th>LOI</th>
</tr>
</thead>
<tbody>
<tr>
<td>clean</td>
<td>74.7</td>
<td>53.1</td>
<td>44.2</td>
<td>72.6</td>
<td>54.0</td>
<td>44.1</td>
</tr>
<tr>
<td>clean MC-FS</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>74.8</td>
<td>59.9</td>
<td>45.1</td>
</tr>
<tr>
<td>MCT mi.</td>
<td>74.4</td>
<td>56.9</td>
<td>49.5</td>
<td>76.5</td>
<td>61.6</td>
<td>50.0</td>
</tr>
<tr>
<td>MCT ma.</td>
<td>75.7</td>
<td>56.1</td>
<td>49.9</td>
<td>77.6</td>
<td>63.1</td>
<td>50.9</td>
</tr>
</tbody>
</table>
```

Acoustic Robustness

- **Feature Enhancement**
  
  Recurrent Denoising Autoencoder

Acoustic Robustness

CHiME15 noise: arousal

Acoustic Robustness

Smartphone noise: *arousal*

![Graph showing validation and test results for acoustic robustness.](image)

## Coding Robustness

- **Coding**
  - Matched Learning

<table>
<thead>
<tr>
<th>codec[kbit/s]</th>
<th>GEMEP</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th>CPSD</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SSNR</td>
<td>IS</td>
<td>PESQ</td>
<td>SSNR</td>
<td>IS</td>
<td>PESQ</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>G711&lt;sub&gt;64&lt;/sub&gt;</td>
<td>29.4</td>
<td>0.10</td>
<td>4.3</td>
<td>34.8</td>
<td>0.01</td>
<td>4.4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>G726&lt;sub&gt;40&lt;/sub&gt;</td>
<td>27.1</td>
<td>0.11</td>
<td>4.2</td>
<td>30.7</td>
<td>0.02</td>
<td>4.3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>G726&lt;sub&gt;32&lt;/sub&gt;</td>
<td>24.6</td>
<td>0.14</td>
<td>4.0</td>
<td>26.7</td>
<td>0.04</td>
<td>4.2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>G726&lt;sub&gt;24&lt;/sub&gt;</td>
<td>19.4</td>
<td>0.33</td>
<td>3.6</td>
<td>21.2</td>
<td>0.12</td>
<td>3.9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>G728&lt;sub&gt;16&lt;/sub&gt;</td>
<td>15.6</td>
<td>0.21</td>
<td>4.0</td>
<td>16.2</td>
<td>0.21</td>
<td>4.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GSM&lt;sub&gt;13&lt;/sub&gt;</td>
<td>10.2</td>
<td>0.40</td>
<td>3.4</td>
<td>11.6</td>
<td>0.40</td>
<td>3.4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>G7231&lt;sub&gt;6.3&lt;/sub&gt;</td>
<td>-2.2</td>
<td>1.44</td>
<td>3.4</td>
<td>-2.5</td>
<td>1.58</td>
<td>3.2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>G7231&lt;sub&gt;5.3&lt;/sub&gt;</td>
<td>-2.0</td>
<td>2.20</td>
<td>3.2</td>
<td>-2.4</td>
<td>2.24</td>
<td>3.1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LPC10&lt;sub&gt;24&lt;/sub&gt;</td>
<td>-3.1</td>
<td>29.08</td>
<td>2.4</td>
<td>-3.7</td>
<td>26.55</td>
<td>1.9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>codec2&lt;sub&gt;13&lt;/sub&gt;</td>
<td>-2.8</td>
<td>3.57</td>
<td>2.1</td>
<td>-3.1</td>
<td>3.05</td>
<td>2.1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>UAR[%]</th>
<th>Arousal</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th>Typicality</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th>Diagnosis</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>mi</td>
<td>ma</td>
<td>mu</td>
<td>mi</td>
<td>ma</td>
<td>mu</td>
<td>mi</td>
<td>ma</td>
<td>mu</td>
<td>mi</td>
<td>ma</td>
<td>mu</td>
<td>mi</td>
<td>ma</td>
<td>mu</td>
</tr>
<tr>
<td>clean</td>
<td>75.0</td>
<td>75.0</td>
<td>74.2</td>
<td>61.6</td>
<td>61.6</td>
<td>62.5</td>
<td>40.9</td>
<td>40.9</td>
<td>40.5</td>
<td>90.7</td>
<td>90.7</td>
<td>89.8</td>
<td>67.1</td>
<td>67.1</td>
<td>63.6</td>
</tr>
<tr>
<td>PCM&lt;sub&gt;128&lt;/sub&gt;</td>
<td>76.5</td>
<td>75.2</td>
<td>74.6</td>
<td>63.1</td>
<td>62.3</td>
<td>62.5</td>
<td>34.3</td>
<td>36.8</td>
<td>40.4</td>
<td>87.8</td>
<td>89.4</td>
<td>89.7</td>
<td>59.7</td>
<td>63.3</td>
<td>64.7</td>
</tr>
<tr>
<td>G711&lt;sub&gt;64&lt;/sub&gt;</td>
<td>75.7</td>
<td>74.7</td>
<td>74.3</td>
<td>61.0</td>
<td>59.9</td>
<td>63.0</td>
<td>33.1</td>
<td>37.4</td>
<td>39.7</td>
<td>88.3</td>
<td>89.5</td>
<td>89.9</td>
<td>60.7</td>
<td>63.7</td>
<td>64.4</td>
</tr>
<tr>
<td>G726&lt;sub&gt;40&lt;/sub&gt;</td>
<td>75.7</td>
<td>74.5</td>
<td>74.2</td>
<td>61.2</td>
<td>59.7</td>
<td>62.5</td>
<td>33.7</td>
<td>37.6</td>
<td>38.4</td>
<td>88.4</td>
<td>89.5</td>
<td>89.9</td>
<td>60.0</td>
<td>63.0</td>
<td>63.8</td>
</tr>
<tr>
<td>G726&lt;sub&gt;32&lt;/sub&gt;</td>
<td>75.8</td>
<td>74.7</td>
<td>74.5</td>
<td>61.6</td>
<td>59.2</td>
<td>62.5</td>
<td>34.0</td>
<td>35.6</td>
<td>40.3</td>
<td>88.4</td>
<td>88.7</td>
<td>88.8</td>
<td>59.7</td>
<td>63.1</td>
<td>64.2</td>
</tr>
<tr>
<td>G726&lt;sub&gt;24&lt;/sub&gt;</td>
<td>74.5</td>
<td>74.2</td>
<td>73.4</td>
<td>58.6</td>
<td>58.5</td>
<td>61.2</td>
<td>26.0</td>
<td>33.9</td>
<td>36.7</td>
<td>87.7</td>
<td>89.7</td>
<td>89.3</td>
<td>60.3</td>
<td>60.6</td>
<td>61.2</td>
</tr>
<tr>
<td>G728&lt;sub&gt;16&lt;/sub&gt;</td>
<td>75.7</td>
<td>74.9</td>
<td>74.5</td>
<td>62.1</td>
<td>59.5</td>
<td>62.5</td>
<td>32.1</td>
<td>38.3</td>
<td>40.3</td>
<td>87.8</td>
<td>89.4</td>
<td>88.8</td>
<td>59.7</td>
<td>62.3</td>
<td>64.2</td>
</tr>
<tr>
<td>GSM&lt;sub&gt;13&lt;/sub&gt;</td>
<td>76.0</td>
<td>75.4</td>
<td>73.4</td>
<td>60.1</td>
<td>58.3</td>
<td>61.2</td>
<td>34.0</td>
<td>35.7</td>
<td>36.7</td>
<td>88.3</td>
<td>88.7</td>
<td>89.3</td>
<td>61.4</td>
<td>63.6</td>
<td>61.2</td>
</tr>
<tr>
<td>G7231&lt;sub&gt;6.3&lt;/sub&gt;</td>
<td>75.9</td>
<td>73.9</td>
<td>73.9</td>
<td>61.9</td>
<td>59.5</td>
<td>63.7</td>
<td>31.0</td>
<td>36.7</td>
<td>36.9</td>
<td>87.2</td>
<td>88.2</td>
<td>89.3</td>
<td>58.1</td>
<td>62.8</td>
<td>63.5</td>
</tr>
<tr>
<td>G7231&lt;sub&gt;5.3&lt;/sub&gt;</td>
<td>75.2</td>
<td>74.8</td>
<td>73.0</td>
<td>59.5</td>
<td>60.2</td>
<td>64.4</td>
<td>30.7</td>
<td>35.9</td>
<td>34.8</td>
<td>87.3</td>
<td>88.8</td>
<td>89.0</td>
<td>58.6</td>
<td>61.3</td>
<td>63.5</td>
</tr>
<tr>
<td>LPC10&lt;sub&gt;24&lt;/sub&gt;</td>
<td>71.2</td>
<td>75.7</td>
<td>73.0</td>
<td>61.9</td>
<td>64.6</td>
<td>63.5</td>
<td>27.8</td>
<td>34.1</td>
<td>31.0</td>
<td>72.1</td>
<td>85.7</td>
<td>81.9</td>
<td>44.2</td>
<td>64.1</td>
<td>57.2</td>
</tr>
<tr>
<td>codec2&lt;sub&gt;13&lt;/sub&gt;</td>
<td>73.8</td>
<td>73.9</td>
<td>73.9</td>
<td>57.9</td>
<td>60.1</td>
<td>62.3</td>
<td>25.6</td>
<td>35.3</td>
<td>36.3</td>
<td>85.5</td>
<td>88.5</td>
<td>88.8</td>
<td>53.7</td>
<td>62.3</td>
<td>56.2</td>
</tr>
<tr>
<td>Avg.</td>
<td>75.1</td>
<td>74.7</td>
<td>73.9</td>
<td>60.8</td>
<td>60.2</td>
<td>62.7</td>
<td>31.1</td>
<td>36.1</td>
<td>37.4</td>
<td>86.3</td>
<td>88.7</td>
<td>88.6</td>
<td>57.8</td>
<td>62.7</td>
<td>62.2</td>
</tr>
</tbody>
</table>

Bandwidth Robustness

- **Channel**

Bandwidth Robustness

Linguistic Robustness.
Linguistic Robustness

- **Spoken Content Matching**

  Examples (LOSO)

<table>
<thead>
<tr>
<th>Model description</th>
<th>Acc. [%]</th>
<th>G 1</th>
<th>G 2</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td>EMO-DB</td>
<td>matched</td>
<td>57.2</td>
<td>46.9</td>
<td>48.9</td>
</tr>
<tr>
<td></td>
<td>mismatched</td>
<td>36.6</td>
<td>37.7</td>
<td>37.4</td>
</tr>
<tr>
<td>SUSAS</td>
<td>matched</td>
<td>64.6</td>
<td>60.3</td>
<td>60.7</td>
</tr>
<tr>
<td></td>
<td>mismatched</td>
<td>52.4</td>
<td>54.4</td>
<td>55.2</td>
</tr>
<tr>
<td>AVIC</td>
<td>matched</td>
<td>79.7</td>
<td>57.8</td>
<td>60.9</td>
</tr>
<tr>
<td></td>
<td>mismatched</td>
<td>49.2</td>
<td>51.3</td>
<td>50.1</td>
</tr>
</tbody>
</table>

"Emotion Recognition using Imperfect Speech Recognition,” Interspeech, 2010."
Linguistic Robustness

- **ASR Influence**
  - Salience
  - Emotion Challenge
  - 2-class Task


(INTERSPÆECH 2010)
Linguistic Robustness

• **Example: FAU Aibo**
  MFCC, polyphones, SC-HMM, full covariances
  Back-off bigrams
  Testing: \( E \succ A \succ N \succ M \)
  Training (AM): \( N \succ E \succ A \succ M \)

• **Explanation**
  *Sammon transformation:*
  High dispersion, neutral in the center
  Neutral words per turn

<table>
<thead>
<tr>
<th></th>
<th>Mother</th>
<th>Neutral</th>
<th>Emphat.</th>
<th>Anger</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>%</strong></td>
<td>44.2%</td>
<td>94.4%</td>
<td>56.7%</td>
<td>29.7%</td>
</tr>
</tbody>
</table>

Linguistic Robustness

- **Training and Adapting Models**

  * AM, LM, both*

  Word accuracy

  Significance

## Multilingual: 2/3 Covered?

<table>
<thead>
<tr>
<th>Language</th>
<th>% NS</th>
<th>Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mandarin</td>
<td>14.40</td>
<td>1</td>
</tr>
<tr>
<td>Spanish</td>
<td>6.15</td>
<td>2</td>
</tr>
<tr>
<td>English</td>
<td>5.43</td>
<td>3</td>
</tr>
<tr>
<td>Hindi</td>
<td>4.70</td>
<td>4</td>
</tr>
<tr>
<td>Arabic</td>
<td>4.43</td>
<td>5</td>
</tr>
<tr>
<td>Portuguese</td>
<td>3.27</td>
<td>6</td>
</tr>
<tr>
<td>Bengali</td>
<td>3.11</td>
<td>7</td>
</tr>
<tr>
<td>Russian</td>
<td>2.33</td>
<td>8</td>
</tr>
<tr>
<td>Japanese</td>
<td>1.90</td>
<td>9</td>
</tr>
<tr>
<td>Punjabi</td>
<td>1.44</td>
<td>10</td>
</tr>
<tr>
<td>German</td>
<td>1.39</td>
<td>11</td>
</tr>
<tr>
<td>Malay/Indonesian</td>
<td>1.16</td>
<td>14</td>
</tr>
<tr>
<td>Telugu</td>
<td>1.15</td>
<td>15</td>
</tr>
<tr>
<td>Vietnamese</td>
<td>1.14</td>
<td>16</td>
</tr>
<tr>
<td>Korean</td>
<td>1.14</td>
<td>17</td>
</tr>
<tr>
<td>French</td>
<td>1.12</td>
<td>18</td>
</tr>
<tr>
<td>Marathi</td>
<td>1.10</td>
<td>19</td>
</tr>
<tr>
<td>Tamil</td>
<td>1.06</td>
<td>20</td>
</tr>
<tr>
<td>Urdu</td>
<td>0.99</td>
<td>21</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Language</th>
<th>% NS</th>
<th>Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>Persian</td>
<td>0.99</td>
<td>22</td>
</tr>
<tr>
<td>Turkish</td>
<td>0.95</td>
<td>23</td>
</tr>
<tr>
<td>Italian</td>
<td>0.90</td>
<td>24</td>
</tr>
<tr>
<td>Cantonese</td>
<td>0.89</td>
<td>25</td>
</tr>
<tr>
<td>Thai</td>
<td>0.85</td>
<td>26</td>
</tr>
<tr>
<td>Gujarati</td>
<td>0.74</td>
<td>27</td>
</tr>
<tr>
<td>Polish</td>
<td>0.61</td>
<td>30</td>
</tr>
<tr>
<td>Pashto</td>
<td>0.58</td>
<td>31</td>
</tr>
<tr>
<td>Burmese</td>
<td>0.50</td>
<td>38</td>
</tr>
<tr>
<td>Sindhi</td>
<td>0.39</td>
<td>47</td>
</tr>
<tr>
<td>Romanian</td>
<td>0.37</td>
<td>50</td>
</tr>
<tr>
<td>Dutch</td>
<td>0.32</td>
<td>57</td>
</tr>
<tr>
<td>Assamese</td>
<td>0.23</td>
<td>67</td>
</tr>
<tr>
<td>Hungarian</td>
<td>0.19</td>
<td>73</td>
</tr>
<tr>
<td>Greek</td>
<td>0.18</td>
<td>75</td>
</tr>
<tr>
<td>Czech</td>
<td>0.15</td>
<td>83</td>
</tr>
<tr>
<td>Swedish</td>
<td>0.13</td>
<td>91</td>
</tr>
<tr>
<td>Balochi</td>
<td>0.11</td>
<td>99</td>
</tr>
</tbody>
</table>

Linguistic Robustness

- **Cross-Language Acoustics**
  
  Same language, within and across language family

<table>
<thead>
<tr>
<th>% UA</th>
<th>same L</th>
<th>within LF</th>
<th>across LF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arousal</td>
<td>94.0</td>
<td>66.3</td>
<td>62.7</td>
</tr>
<tr>
<td>Valence</td>
<td>81.7</td>
<td>61.9</td>
<td>54.6</td>
</tr>
</tbody>
</table>

Linguistic Robustness

- **Transfer Learning**

```
<table>
<thead>
<tr>
<th>PAIR-WISE (UAR)</th>
<th>EMODDB</th>
<th>SAVEE</th>
<th>EMOV0</th>
<th>Polish</th>
</tr>
</thead>
<tbody>
<tr>
<td>EMOV0 (German)</td>
<td>Baseline</td>
<td>55.8</td>
<td>56.4</td>
<td>71.9</td>
</tr>
<tr>
<td>KCCA SHLA^2</td>
<td>65.2</td>
<td>57.9</td>
<td>56.8</td>
<td>72.8</td>
</tr>
<tr>
<td>SAVEE (English)</td>
<td>Baseline</td>
<td>62.5</td>
<td>54.6</td>
<td>65.0</td>
</tr>
<tr>
<td>KCCA SHLA</td>
<td>71.9</td>
<td>59.5</td>
<td>69.4</td>
<td>70.2</td>
</tr>
<tr>
<td>EMOV0 (Italian)</td>
<td>Baseline</td>
<td>58.0</td>
<td>51.2</td>
<td>55.3</td>
</tr>
<tr>
<td>KCCA SHLA</td>
<td>67.3</td>
<td>58.5</td>
<td>71.9</td>
<td>64.3</td>
</tr>
<tr>
<td>Polish</td>
<td>Baseline</td>
<td>65.1</td>
<td>55.4</td>
<td>57.3</td>
</tr>
<tr>
<td>KCCA SHLA</td>
<td>71.1</td>
<td>61.9</td>
<td>56.1</td>
<td>58.3</td>
</tr>
</tbody>
</table>
```

Paralinguistic Robustness.
Only One Voice!

- **Multiple-Targets**
  There is just one Vocal Production Mechanism…

<table>
<thead>
<tr>
<th>% UA</th>
<th>Single</th>
<th>Multiple</th>
</tr>
</thead>
<tbody>
<tr>
<td>Likability</td>
<td>59.1</td>
<td>(+A,G,Cl) 62.2</td>
</tr>
<tr>
<td>Neuroticism</td>
<td>62.9</td>
<td>(+G,OCEA, Cl) 67.5</td>
</tr>
</tbody>
</table>
Model Switching

- **Model Selection**
  By: Age, Gender, Personality

4 Emotional Speech Corpora
AVIC, AEC
eINTERFACE, SUSAS

---

**Distribution matching**

**Split data**

---

<table>
<thead>
<tr>
<th>Train</th>
<th>Test</th>
<th>O</th>
<th>C</th>
<th>E</th>
<th>A</th>
<th>N</th>
<th>Gender</th>
<th>Age</th>
</tr>
</thead>
<tbody>
<tr>
<td>UAR</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(trMF → tsMF) 73.38</td>
<td>(trAY → tsAY) 74.11</td>
</tr>
<tr>
<td>ΔUAR</td>
<td>trHL</td>
<td>tsHL</td>
<td>0.07ns</td>
<td>-3.03</td>
<td>0.11ns</td>
<td>-0.77</td>
<td>0.75</td>
<td></td>
</tr>
<tr>
<td></td>
<td>trHL</td>
<td>tsL</td>
<td>-1.83</td>
<td>0.79</td>
<td>-3.47</td>
<td>-1.21</td>
<td>-3.66</td>
<td></td>
</tr>
<tr>
<td></td>
<td>tRH</td>
<td>tsH</td>
<td>0.00ns</td>
<td>-2.85</td>
<td>-0.02ns</td>
<td>-0.24ns</td>
<td>0.51ns</td>
<td>(trF → tsF) 2.47</td>
</tr>
<tr>
<td></td>
<td>tRL</td>
<td>tsL</td>
<td>-2.23</td>
<td>0.68</td>
<td>-2.61</td>
<td>-1.33</td>
<td>-3.34</td>
<td>(trA → tsA) -5.07</td>
</tr>
<tr>
<td></td>
<td>tRH</td>
<td>tsL</td>
<td>-5.39</td>
<td>-4.61</td>
<td>-10.28</td>
<td>-8.84</td>
<td>-9.88</td>
<td>(trF → tsM) -6.59</td>
</tr>
<tr>
<td></td>
<td>tRH</td>
<td>tsH</td>
<td>-7.19</td>
<td>-6.74</td>
<td>-6.78</td>
<td>-2.88</td>
<td>-4.21</td>
<td>(trA → tsY) 2.66</td>
</tr>
</tbody>
</table>

---

| UAR   |        |       |       |       |       |       | (trMF → tsMF) 63.16    | (trAY → tsAY) 62.41 |
| ΔUAR  | trHL   | tsHL  | 60.94 | 61.15 | 60.99 | 60.86 | 60.93                   |            |
| Rule  |        |       |       |       |       |       | (trMF → tsMF) 0.18ns   | 0.39       |

Higher-level Features

ND and D speech from Interspeech ComParE 2016

Holism: Vertical.

• **Cross-Task Self-Labelling**

<table>
<thead>
<tr>
<th>% UA</th>
<th>Likability</th>
<th>Emotion</th>
<th>Personality</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>57.2</td>
<td>68.9</td>
<td>66.4</td>
</tr>
<tr>
<td>Cross-Task Labelling</td>
<td>60.3</td>
<td>69.0</td>
<td>66.6</td>
</tr>
</tbody>
</table>

**Algorithm:** Cross-Task Labelling  
**Repeat for each task:**  
**Repeat until** $\mathcal{U} \in \{\}$:

1. (Optional) Upsample training set $\mathcal{L}$ to even class distribution $\mathcal{L}_D$
2. Use $\mathcal{L}/\mathcal{L}_D$ to train classifier $\mathcal{H}$, then classify $\mathcal{U}$
3. Select a subset $\mathcal{N}_{st}$ that contains those instances predicted with the highest confidence values
4. Remove $\mathcal{N}_{st}$ from the unlabelled set $\mathcal{U}$, $\mathcal{U} = \mathcal{U} \setminus \mathcal{N}_{st}$
5. Add $\mathcal{N}_{st}$ to the labelled set $\mathcal{L}$, $\mathcal{L} = \mathcal{L} \cup \mathcal{N}_{st}$

“Semi-Autonomous Data Enrichment Based on Cross-Task Labelling of Missing Targets for Holistic Speech Analysis”, ICASSP, 2016.
Holism: Next-Gen?

- Evolutionary Learning
- Reinforced Learning
- Analysis/Synthesis Gap

Evolving Deep CRNN w/ LSTM

Uncertainty Weighted Combination
More Data: The answer to it all?
New Data

- In the Wild

<table>
<thead>
<tr>
<th>Cultural Background</th>
<th>Age Group</th>
<th>Years Known the Other Participant</th>
<th>Self-Reported Familiarity Rating</th>
</tr>
</thead>
<tbody>
<tr>
<td>British</td>
<td>66</td>
<td>203</td>
<td>&lt;1</td>
</tr>
<tr>
<td>German</td>
<td>70</td>
<td>25</td>
<td>1</td>
</tr>
<tr>
<td>Hungarian</td>
<td>70</td>
<td>46</td>
<td>4</td>
</tr>
<tr>
<td>Serbian</td>
<td>72</td>
<td>46</td>
<td>3</td>
</tr>
<tr>
<td>Greek</td>
<td>56</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chinese</td>
<td>70</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

[http://db.sewaproject.eu](http://db.sewaproject.eu)
New Data

- Graz Real-Life Affect in the Street & Supermarket (GRAS²)

6 channel audio + video + eyetracking + EDA + temperature + 2x 3D motion

Ask for help
Gradually embarrassing:
denture adhesive
Anti-athlete’s foot cream
Efficient Labelling

- **Cooperative Learning in aRMT**
  0) Transfer Learning
  1) Dynamic Active Learning
  2) Semi-Supervised Learning

```
Labelled data          Add          Newly labelled

Train                  Confidence/Information

Model                  Class

Unlabelled data
```

Improvement of $UA \approx 5.0\%$

95.0% reduced

Efficient Labelling

"iHEARu-PLAY: Introducing a game for crowdsourced data collection for affective computing",

WASA, 2015.

https://ihearu-play.fim.uni-passau.de/
Group Assessment
Cultural Robustness
Multiple Microphones, “Chips-Bag“, ...?
Robust Gold Standard
Coupled ASR + CP?
Abstract

An increasingly long list of states and traits of speakers is being targeted for automatic recognition by computers including their age, emotion, health condition, or personality. However, hardly any of these have been encountered in “everyday” usage by the broad consumer mass up to now. This is certainly also owed to robustness issues, which shall be discussed here. Traditionally, these comprise speech enhancement, feature enhancement, feature space adaptation, or matched conditions training – mainly to cope with additive or convolutional noise. In addition, a number of further robustness issues mark this field of speech analysis, including interdependence of states and traits, potential subjectivity in the labels, phonetic content variation in the acoustic analysis, varying language and erroneous speech recognition in the linguistic analysis, and diversity of the cultural background of speakers. Finally, a number of hardly tackled issues remain such as the analysis of multiple speakers or in far field condition with multiple microphones. In the talk, an overview on these challenges and existing solutions is given. Then, required future research efforts will be named to help Computational Paralinguistics’ massive launch into the next generation dialogue systems and many other applications.