Bilingual Acoustic Feature Selection for Emotion Estimation Using a 3D Continuous Model

Humberto Pérez Espinosa, Carlos A. Reyes García, Luis Villaseñor Pineda

National Institute of Astrophysics, Optics and Electronics
Tonantzintla, Puebla, México
1. Introduction
Some questions that arise when engaging in Speech Emotion Recognition:

- Is the way we express emotions social and cultural dependent?
- Is it possible to obtain objective measures from emotions?
- Are these measures language independent?
In this work:

- We look for acoustic features that allow us to estimate emotional states from speech regardless the spoken language (English / German)
- We study importance and the amount of information these features provide for each language
2. Emotion Model
Three-Dimensional Continuous Model

Three Emotion Primitives:
(H. Schlosberg, 1954)

- **Valence**: How negative or positive is an emotion
- **Activation**: Internal excitement of the individual
- **Dominance**: Degree of control that the individual intends to take on the situation
3. Emotional Speech Data
IEMOCAP database (C. Busso, 2008)

- Collected at SAIL lab at USC.
- Spoken in English
- 10 actors interacting in man/woman pairs
- Annotated with primitives: Valence – Activation – Dominance
- 1,819 speaker turns
VAM database (M. Grimm, 2008)

- Collected by Michael Grimm and the emotion research group at the Institut für Nachrichtentechnik of the Universität Karlsruhe
- Spoken in German
- Recordings of the German talk show “Vera am Mittag”
- Annotated with primitives: Valence – Activation – Dominance
- 947 utterances
4. Acoustic Features
<table>
<thead>
<tr>
<th>Feature Group</th>
<th>Type</th>
<th>Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>Voice Quality</td>
<td>Voice Quality</td>
<td>36</td>
</tr>
<tr>
<td>Elocution Times</td>
<td>Prosodic</td>
<td>125</td>
</tr>
<tr>
<td>Cochleagrams</td>
<td>Spectral</td>
<td>96</td>
</tr>
<tr>
<td>LPC</td>
<td>Spectral</td>
<td>111</td>
</tr>
<tr>
<td>Spectral Flux</td>
<td>Spectral</td>
<td>117</td>
</tr>
<tr>
<td>Energy Contour</td>
<td>Prosodic</td>
<td>129</td>
</tr>
<tr>
<td>F0 Contour</td>
<td>Prosodic</td>
<td>243</td>
</tr>
<tr>
<td>Spectral Max and Min</td>
<td>Spectral</td>
<td>468</td>
</tr>
<tr>
<td>Spectral Energy in Bands</td>
<td>Spectral</td>
<td>234</td>
</tr>
<tr>
<td>Spectral Roll off Point</td>
<td>Spectral</td>
<td>468</td>
</tr>
<tr>
<td>MFCC</td>
<td>Spectral</td>
<td>1,617</td>
</tr>
<tr>
<td>MEL Spectrum</td>
<td>Spectral</td>
<td>3,042</td>
</tr>
<tr>
<td>Probability of Voicing</td>
<td>Prosodic</td>
<td>117</td>
</tr>
<tr>
<td>Spectral Centroid</td>
<td>Spectral</td>
<td>117</td>
</tr>
</tbody>
</table>
Instance/Feature Selection

- Eliminate contradictory instances
- We already had found some important features
- Linear floating forward selection
- Aggregates the best evaluated attribute at each step
5. Results
Bilingual Performance

- Having identified the best acoustic feature sets we constructed individual classifiers to estimate each Emotion Primitive.

- We made experiments using only English, only German and Bilingual.

- Results of the learning experiments were obtained using Support Vector Machine for Regression (SMOreg) and evaluated by 10-Fold Cross Validation.

- Each feature group was evaluated separately using three metrics:
  - Share and Portion for showing the contribution of feature groups (Batliner, 2010)
  - Pearson’s Correlation Coefficient as main performance measure.
Dominance Correlation
## Cross-Lingual Performance

<table>
<thead>
<tr>
<th>Selection</th>
<th>Train/Test</th>
<th>Research Questions</th>
<th>Mean Corr</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cross-lingual</td>
<td>Mono-lingual</td>
<td>Are acoustic features that important for one language also important for another language?</td>
<td>0.649</td>
</tr>
<tr>
<td>Bi-lingual</td>
<td>Cross-Lingual</td>
<td>Are patterns identified to estimate the emotion primitives in one language useful to estimate the emotion primitives in another language?</td>
<td>0.376</td>
</tr>
<tr>
<td>Bi-lingual</td>
<td>Mono-lingual</td>
<td>Do we obtain better features by doing an acoustic features selection on bilingual data than doing it on monolingual data? Can we find features that provide complementary information for one language in this way?</td>
<td>0.645</td>
</tr>
<tr>
<td>Mono-lingual</td>
<td>Mono-lingual</td>
<td>Baseline</td>
<td>0.710</td>
</tr>
</tbody>
</table>
6. Conclusions
Conclusions

- Spectral analysis seems to be the most important for the three primitives:
  - Valence: LPC - MEL - MFCC – Spectral Flux
  - Activation: MFCC - Cochleagrams – LPC - Energy
  - Dominance: MFCC - Cochleagrams – Energy – LPC

- Emotional states can be estimated using a similar set of acoustic features for each of the two languages used in this work

- Patterns shown by these features are difficult to fit from one language to another without any adaptation

- It is possible to identify common patterns in both languages using a feature set that works for both of them
Conclusions

- Differences attributed to language may be magnified by other reasons
  - Acted emotions in a controlled environment (IEMOCAP) versus spontaneous emotions in an uncontrolled environment (VAM)
  - Number of instances in each database
  - Emotion diversity
7. Work in Progress
Creation of a spontaneous emotional speech database in Mexican Spanish:
- 28 Children playing a card game (Wisconsin Card Sorting Test)
- 2,500 utterances
- 6 emotion categories, 3 Primitives

Development of a fuzzy logic based method for emotion primitives interpretation:
- Estimation / Representation of emotion expressiveness level
- Estimation / Representation of emotions mixture
Thanks for your Attention

Contact: humbertop@inaoep.mx