Annotation of expressive speech in the SmartKom project

Olga Dioubina
Department of Phonetics and Speech Communication
University of Munich, Germany
olga.dioubina@phonetik.uni-muenchen.de
10. September 2003
Speech corpus

1. Wizard-of-Oz dialogues: Simulated human-machine interaction;
2. Wizards: invisible, predefined vocabulary, manipulated voice;
3. Speech and visual data: 450 dialogues of approx. 4.5 minutes;
4. Speech: No posed or exaggerated emotions

Noticeable expression of the emotional or cognitive state of the speakers called User States.
Annotated by labelers with the help of different labels.
(For example observed during recognition failure.)
User state labels and prosodic accents
(Each of the labels is linked to a sound sample, just click on them.)
A not serious one

"Trust me when I say I'll be catching SOMETHING today."
Acoustic properties of user state accents

Oscillogram of the German utterance “Can one also eat there?” (listen) The label [Emphasis] was assigned to the word “dort” (listen) (+43 Hz/+2dB). The sentence was produced along with the facial expression of [Strong anger/Irritation].
Blend of **prosodic** and **user state** accents

Nein [Primary accent] [Emphasis].
<P> <A> hab’ ich nicht [Primary accent] [Strong Emphasis], weil ich würde normalerweise [Secondary accent] [Emphasis] gleich abhauen [Primary accent] [Emphasis], . . .

English: No, I don’t, because otherwise I’d ran away immediately. (listen)
Blend of *prosodic* and *user state* accents

81 of 84 words labeled with [Emphasis] were assigned with [Primary accent], [Secondary accent] or [Contrastive accent] by the prosodic group of labelers.

THEREFORE

**Action:** Extraction of 2979 labels of [Emphasis] in 280 WOZ dialogues in which 175 subjects had taken part.

**Result:** 94% of the same words were assigned with [Emphasis] or [Strong Emphasis] AND with [Primary accent], [Secondary accent] or [Contrastive accent].
Suggestion for annotation sequence

I. Nein [Primary accent]. \(<P> \ <A>\) hab’ ich nicht [Primary accent], weil ich würde normalerweise [Secondary accent] gleich abhauen [Primary accent], . . .

Prosodic accents should be seen by User State Labelers.

II. Nein [Primary accent] [Emphasis]. \(<P> \ <A>\) hab’ ich nicht [Primary accent] [Strong Emphasis], weil ich würde normalerweise [Secondary accent] [Emphasis] gleich abhauen [Primary accent] [Emphasis], . . .
**User state accents, prosodic accents and facial expressions in pitch and intensity**

<table>
<thead>
<tr>
<th>speaker</th>
<th>word</th>
<th>F0</th>
<th>A</th>
<th>PL</th>
<th>facial expression</th>
</tr>
</thead>
<tbody>
<tr>
<td>male</td>
<td>Film</td>
<td>+8</td>
<td>+3</td>
<td>PA</td>
<td>joy/gratification (weak)</td>
</tr>
<tr>
<td></td>
<td>Kino</td>
<td>+39</td>
<td>+3</td>
<td>PA</td>
<td>joy/gratification (weak)</td>
</tr>
<tr>
<td></td>
<td>Kino</td>
<td>+9</td>
<td>+1</td>
<td>PA</td>
<td>helplessness (weak)</td>
</tr>
<tr>
<td></td>
<td>Kino</td>
<td>-4</td>
<td>+2</td>
<td>PA</td>
<td>pondering/reflecting (weak)</td>
</tr>
<tr>
<td></td>
<td>das</td>
<td>+10</td>
<td>+5</td>
<td>PA</td>
<td>pondering/reflecting (weak)</td>
</tr>
<tr>
<td>fem.</td>
<td>ja</td>
<td>+10</td>
<td>+4</td>
<td>NA</td>
<td>helplessness (weak)</td>
</tr>
<tr>
<td>fem.</td>
<td>kann</td>
<td>-6</td>
<td>+5</td>
<td>no</td>
<td>pondering/reflecting (weak)</td>
</tr>
<tr>
<td>fem.</td>
<td>dort</td>
<td>+43</td>
<td>+2</td>
<td>PA</td>
<td>anger/irritation (strong)</td>
</tr>
<tr>
<td>fem.</td>
<td>bloede</td>
<td>+18</td>
<td>+5</td>
<td>PA</td>
<td>anger/irritation (weak)</td>
</tr>
<tr>
<td>fem.</td>
<td>bloed</td>
<td>+4</td>
<td>+1</td>
<td>PA</td>
<td>anger/irritation (weak)</td>
</tr>
</tbody>
</table>

Samples of 10 words which were assigned with [Emphasis] and their relationship to (F0) pitch in Hz, (A) intensity in dB, (PL) prosodic accents and facial expressions. [PA] means primary, [NA] – secondary accent.
Conclusion

1. **Prosodic accents** and **user state accents** are not clearly distinguished from each other if they are perceived from the speech signal only.

2. The prominence of **prosodic accents** and **user state accents** is supported by the increase of pitch and intensity.

3. **User state accents** are not clearly supported by emotional **facial expressions**. We suggest that speakers tend to express their emotional or cognitive state either by means of speech or by means of facial expressions.
THANK YOU VERY MUCH FOR YOU ATTENTION!

AND VERY MANY THANKS TO
Michiko Inoue, Sergei Mariev, Hartmut Pfitzinger, Susen Rabold, Uwe Reichel, Ulrich Reubold, Silke Steininger, Uli Tuerk and Karl Weilhammer
for their continuous support during this research.