GfK Verein
Detecting Emotions from Voice
Respondents’ willingness to complete questionnaires declines

But it doesn’t necessarily mean that consumers have nothing to say about products or brands:

GfK Verein pursues two different ways to face this trend:

1. Analysis of encountered data: sentiment analysis of text, pictures, automatic logo detection, etc.

2. Data collection „on the fly“: inference of emotions from facial expressions and voice
Emotions happen on different levels each of which has its own measurement method.

**subjective perception**
- feelings
- cognitive assessments

**physiological reactions**
- peripheral or
- central nervous system

**expression patterns**
- facial expressions
- voice
- body language

self-report

heart rate, skin conductance, electrical brain activity (EEG), brain blood flow (fMRI), …

observational techniques
Analyzing a person’s facial expression is ideal for inferring emotional valence
But it is not possible to infer emotional arousal from a person’s face.

We aim on inference of:
• subtle emotional reactions
• and the continuous assessments which underlie all emotions („Appraisals“): Valence, and soon Novelty and Control

Emotional arousal can not be inferred from facial expression.
Voice analytics gives a 360-degree picture of a person’s opinion.

Voice analytics allows …

- content analysis (what is said)
- emotional analysis (who it is said)
- inference of emotional states without visual channels (e.g., telephone interviews),
- inference of additional emotional states, especially “arousal”,
- detection of additional states (e.g., interest, cognitive load, social dominance,…), and
- unburdens respondents.

Our vision:

Analyze – on a quantitative scale – open answers and comments about what really matters to consumers.
Emotional arousal indicates personally relevant events

According to Scherer (2005), emotions …
• can be seen as relevance detectors
• are produced by a sequence of appraisal checks

Montagrin & Sander (2016) suggest that relevant events
• elicit arousal,
• capture attention, and
• facilitate memory.

→ similar sentiment (positivity/negativity of content) in both examples
→ but different, plainly audible levels of personal relevance

Research cooperation to develop detector of emotional arousal from the voice
audEERING’s automatic voice analysis software typically follows a series of steps:

1a. Extraction of relevant segments

1b. Speech enhancement

2. Acoustic analysis

3. Features

4. Classification

Model training
The classification model is trained using machine learning algorithms and a training set of voice recordings.
Once the classification model is trained it can be applied on new recordings.
We collected voice recordings for model training in different domains, ad testing …
… and product concept testing, …

An electric toothbrush that enables you to detect and remove plaque in real time. While brushing, sensors in the brush identify plaque and alert you by ‘beeping’ to brush longer in that area so you won’t miss a spot.

An bristle free electronic toothbrush that cleans your entire mouth. It has a soft spongy, spherical head that automatically vibrates and rotates to clean your entire mouth – teeth, gums, tongue and cheeks. It is an easy way to get your whole mouth clean.

An electric toothbrush that takes its inspiration from modern car washes for a whole mouth clean. “Bristles” made of super soft flexible material are propelled by tiny water jets in the brush head causing the bristles to ‘woosh’ around all areas of your mouth delivering a unique polished, clean feeling.

An electric toothbrush that offers a spa-like brushing experience. Special features include a low vibration/noise handle and bristles that are heated to gently massage your gums. It will leave you feeling relaxed, pampered and squeaky clean every time.

An electric toothbrush leasing program which makes trying and upgrading easy and affordable. Order online and receive a brand-new brush in the mail. For a low monthly fee, this program allows you to upgrade to a new model anytime you want.
… and in different languages

<table>
<thead>
<tr>
<th>Recordings</th>
<th>(online)</th>
<th>(CLT)</th>
<th>Seconds (avg.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>German ad test</td>
<td></td>
<td>1187</td>
<td>21</td>
</tr>
<tr>
<td>German concept test</td>
<td></td>
<td>554</td>
<td>10</td>
</tr>
<tr>
<td>English concept test</td>
<td></td>
<td>2140</td>
<td>25</td>
</tr>
<tr>
<td>Spanish concept test</td>
<td></td>
<td>603 + 541</td>
<td>19</td>
</tr>
<tr>
<td>Chinese concept test</td>
<td></td>
<td>166 + 574</td>
<td>14</td>
</tr>
</tbody>
</table>
In the ad test, respondents were asked very detailed and guided, …

- Describe the TVC in own words
- Neutral baseline questions + 5 questions per TVC
- What do you like about the brand?
- What did you like about the TVC?
- What didn’t you like about the TVC?
- How did you feel during the exposure?
- What didn’t you like about the TVC?

In the ad test, respondents were asked very detailed and guided questions. For example, they were asked to describe the TVC in their own words, what they liked about the brand and the TVC, how they felt during the exposure, and what they didn’t like about the TVC. The survey included neutral baseline questions, which were always the same. Examples of responses included: ‘the music and the landscape’ and ‘always the same’.
Read the concept below, press the red button and talk to us for while about your reactions to this idea.

1. How useful or relevant is this idea to you?
2. What did you like about it...not like about it?
3. What would you say to a friend about it?
4. Overall, what did you think about it?

Examples:

‘don‘t interrupt’

‘don‘t trust the technology’
To gather information on the ‘real’ emotional content, we asked students to assess the ratings …

Each recording was rated several times by different raters, at least four times.

We recruited students with different native languages:

**German**
- psychology students
- rated all corpora (German, English, Spanish, Chinese)

**Chinese**
- students of humanities or social sciences
- rated only the Chinese corpus

**Spanish**
- students of humanities or social sciences
- rated only the Spanish corpus
… with respect to content, arousal, valence, and interest, using a browser-based annotation platform
Our raters perceive a broad range of emotional reactions with high agreement

Descriptive statistics for averaged arousal ratings of German annotators

<table>
<thead>
<tr>
<th>Test</th>
<th>Language</th>
<th>Min</th>
<th>Max</th>
<th>Mean</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>German</td>
<td>Ad Test</td>
<td>-5.83</td>
<td>5.00</td>
<td>-0.11</td>
<td>1.89</td>
</tr>
<tr>
<td>German</td>
<td>Concept</td>
<td>-6.80</td>
<td>6.60</td>
<td>0.64</td>
<td>2.31</td>
</tr>
<tr>
<td>English</td>
<td>Concept</td>
<td>-6.76</td>
<td>6.86</td>
<td>0.56</td>
<td>2.09</td>
</tr>
<tr>
<td>Spanish</td>
<td>Concept</td>
<td>-7.67</td>
<td>6.60</td>
<td>0.33</td>
<td>2.13</td>
</tr>
<tr>
<td>Chinese</td>
<td>Concept</td>
<td>-5.60</td>
<td>8.20</td>
<td>1.01</td>
<td>1.94</td>
</tr>
</tbody>
</table>

Inter-rater correlation for German raters

<table>
<thead>
<tr>
<th>Test</th>
<th>Language</th>
<th>Arousal</th>
<th>Valence</th>
<th>Interest</th>
</tr>
</thead>
<tbody>
<tr>
<td>German</td>
<td>Ad Test</td>
<td>0.54</td>
<td>0.63</td>
<td>0.43</td>
</tr>
<tr>
<td>German</td>
<td>Concept</td>
<td>0.71</td>
<td>0.68</td>
<td>0.61</td>
</tr>
<tr>
<td>English</td>
<td>Concept</td>
<td>0.69</td>
<td>0.67</td>
<td>0.60</td>
</tr>
<tr>
<td>Spanish</td>
<td>Concept</td>
<td>0.67</td>
<td>0.51</td>
<td>0.54</td>
</tr>
<tr>
<td>Chinese</td>
<td>Concept</td>
<td>0.70</td>
<td>0.47</td>
<td>0.54</td>
</tr>
</tbody>
</table>
Are there intercultural differences in the perception of arousal from voice?

**Probably not!**

Correlation between raters for Spanish / Chinese recordings

<table>
<thead>
<tr>
<th></th>
<th>DE</th>
<th>ES</th>
<th>CN</th>
</tr>
</thead>
<tbody>
<tr>
<td>DE</td>
<td>0.68</td>
<td>0.78</td>
<td>0.71</td>
</tr>
<tr>
<td>ES</td>
<td></td>
<td>0.63</td>
<td></td>
</tr>
<tr>
<td>CN</td>
<td></td>
<td></td>
<td>0.61</td>
</tr>
</tbody>
</table>
We used audERING‘s voice analysis solution to extract characteristic features from the voice

openSMILE
- processes over one thousand parameters
- that are used to build reliable decisions regarding a subject’s emotional state

<table>
<thead>
<tr>
<th>characteristics of speech prosody, e.g.,</th>
<th>advanced audio characteristics that relate to the way of articulation, e.g.,</th>
</tr>
</thead>
<tbody>
<tr>
<td>• variations and micro variations of pitch</td>
<td></td>
</tr>
<tr>
<td>• tone of voice</td>
<td></td>
</tr>
<tr>
<td>• modulations of loudness</td>
<td>• formant statistics (i.e., acoustic energy within a specific frequency spectrum)</td>
</tr>
<tr>
<td></td>
<td>• spectral timbre information</td>
</tr>
</tbody>
</table>
A state-of-the-art machine learning algorithm combines the extracted features and ratings into a classifier

- The model was trained using **all corpora** (German, English, Spanish, and Chinese) which are combined into one single model

- Model training was restricted to **German ratings** to make the annotations more comparable across all corpora

- Machine learning algorithm: **Long Short-Term Memory Neural Network**

- More than thousand characteristic **features** used for classification

- Feature normalization possible
What does our classifier hear in the examples?

The difference!

- 'vacillated'
  - 3.1
- 'the music and the landscape'
  - 1.2
- 'don’t interrupt'
  - 2.4
- 'nothing for the crowd'
  - 1.1
- 'always the same'
  - -2.8
- 'don’t trust the technology'
  - -2.6
We compared the classification results for emotional arousal to what humans hear?

Model validation was done separately from model training using a 5-fold cross validation:

- Several models have been trained using only four fifth of the speakers for training and the remaining fifth for validation.
- This was repeated five times (for each fifth) and for each language separately.

<table>
<thead>
<tr>
<th>Language</th>
<th>Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>EN</td>
<td>0.63</td>
</tr>
<tr>
<td>DE</td>
<td>0.60</td>
</tr>
<tr>
<td>CN</td>
<td>0.64</td>
</tr>
<tr>
<td>ES</td>
<td>0.56</td>
</tr>
</tbody>
</table>

To get an indication of how the model performs on completely new data:

- A new model was trained.
- Using German, Spanish and Chinese recordings for training.
- And English recordings for validation.

<table>
<thead>
<tr>
<th>Language</th>
<th>Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>EN</td>
<td>0.54</td>
</tr>
</tbody>
</table>
Although we focused on automatic detection of emotional arousal we trained classifier for valence and interest:

<table>
<thead>
<tr>
<th>Language</th>
<th>Valence</th>
<th>Interest</th>
</tr>
</thead>
<tbody>
<tr>
<td>EN</td>
<td>0.24</td>
<td>0.44</td>
</tr>
<tr>
<td>DE</td>
<td>0.30</td>
<td>0.35</td>
</tr>
<tr>
<td>CN</td>
<td>0.26</td>
<td>0.57</td>
</tr>
<tr>
<td>ES</td>
<td>0.31</td>
<td>0.43</td>
</tr>
</tbody>
</table>

- Classification accuracy for valence and interest is much lower than for arousal.
- Inter-rater correlations are lower for these dimensions compared to arousal.
  - Rating interest and valence is more difficult than arousal.
  - Perception of Valence and Interest, however, depends stronger on the content, i.e., whether the raters understand the language or not.
Because of culture-independent and sufficiently accurate results a commercial application has been released

We found that emotional arousal
- seems to be perceived interculturally and language independent
- and is automatically detectable quite well

We developed a robust classifier
- for emotional arousal in respondents voice
- in the context of ad and concept tests (and similar domains)
- usable for online tests and CLTs
- in (currently) German, English, Chinese, and Spanish

Release general model for application in commercial market research
GfK introduced voice analytics to capture insights on the innovation’s potential to improve life and emotionally connect with people.

Speech is the natural way to communicate, allowing us to express how we feel and what we think ...
A new concept performance score **Emotional Impact (E₃)** captures peoples’ emotional reactions to new product ideas.
Analyze not only respondents vocal expressions but also opinions ‘in the wild’?
Contact us!

Dr. Matthias Unfried
GfK Verein
Fundamental Research
Voice Analytics
0911 395-4514
matthias.unfried@gfk-verein.org

Dr. Anja Dieckmann
GfK Verein
Fundamental Research
GfK EMO Scan
0911 395-2033
anja.dieckmann@gfk-verein.org