Taking into Account the User's Focus of Attention with the Help of Audio-Visual Information: Towards less Artificial Human-Machine-Communication

Anton Batliner[1], Christian Hacker[1], Moritz Kaiser[2], Hannes Mögele[2], Elmar Nöth[1]

[1]Institute for Pattern Recognition, University of Erlangen-Nuremberg, Germany
[2]Bavarian Archive for Speech Signals, Munich, Germany
[2]Institute for Phonetics and Speech Processing, Munich, Germany

On−Focus
Where can I meet football−fans in Hannover ?
There are more than 100 football pubs in Hannover !
Off−Focus

- SmartWeb: Multimodal access to the semantic web
- Scenario handheld: User is interacting via a smart-phone
- Speech input is analysed on the server
- **No push-to-talk**
- Automatic recognition whether the user addresses the system (On-Focus) or talks to s.o. else (Off-Talk, Off-View)
- Analysis of prosody, linguistic info, and images of the camera integrated in the mobile phone

The SmartWeb Video Corpus

- 3.2 hours of speech, 2068 utterances (Bluetooth, UMTS, 8 kHz, 8 bit)
- 14 hours of video (H.263, camera of Nokia 6680 cell phone)
- Recording location: real life situations with varying degree of acoustic and visual noise
- Total # of speakers: 100; test set: 37

	On-View	Off-View
NOT (On-Talk)	**On-Focus, Interaction with the system**	*(unusual)*
ROT (Off-Talk)	Reading aloud from the display	—
POT (Off-Talk)	*(unusual)*	Reporting results from SmartWeb
SOT (Off-Talk)	Responding to an interruption	Responding to an interruption

Tab.1: Cross-tabulation of On-/Off-Talk vs. On-/Off-View

- **NOT**: Talking to the system, On-Talk (50 %)
- **ROT**: Read Off-Talk (13 %)
- **POT**: Paraphrasing Off-Talk (11 %)
- **SOT**: Spontaneous Off-Talk (26 %)

- Data Collection:
  - **Situational Prompting** technique (SitPro) with 2 subjects: the caller and the companion
  - Elicitation method based on standard prompts, individualised prompts, script prompts (simulating a conversation)
  - **Companion had to disturb the caller** to elicit POT
- Annotation of the data:
  - **Audio** (word based): NOT, ROT, POT, SOT.
  - Mapping to utterance level (dialogue turn)
  - **Video** (frame based): On-View, Off-V., No-Face
  - Semi-automatic segmentation of faces

Evaluation: Class-wise average recognition rate:
CL = Mean of recalls
2-class case: 0.5· (sensitivity + specifity)

Prosodic Features (word based)

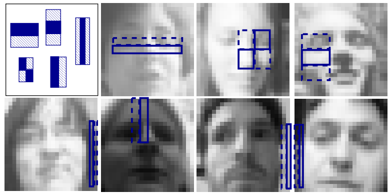
- **100 prosodic features** per word
  based on fundamental frequency, energy, duration, rate-of-speech, pauses, jitter, and shimmer
- **66 % CL** for On-Talk vs. Off-Talk
- **48 % CL** for NOT/ROT/POT/SOT
- POT is hard to recognise with prosody

Linguistic Features (word based)

- **30 features** describing the **part-of-speech** (POS) categories of ±2 words
- 6 POS cover classes:
  Nouns, verbs, auxiliaries, adjectives and participles (inflected/not inflected), PAJ (particles, articles, and interjections)
- Learning of POS sequences (**domain independent!**)

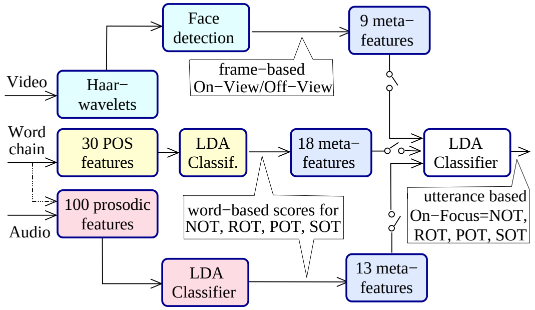
- Observation: Many nouns and adjectives for ROT; many PAJ for SOT
- **59 % CL** for On-Talk vs. Off-Talk
- **45 % CL** for NOT/ROT/POT/SOT

Face Detection (frame based)



- Classification of grayscale images (176 × 144, 7.5 per sec.) by applying the **Viola-Jones** algorithm
  (Haar-wavelets, looking for faces in plenty of sub-images scaled to 24 × 24, hierarchical classifier)
- Training with 18.000 images
- Selection of **425 features** with Adaboost
- Learning of perspective distortion, backlight, etc.
- **88 % CL** for On-View vs. Off-View
  (Default Open-CV classifier: 81 % CL)

Fusion



- Fusion of modalities
  - Mapping to the **sentence level**
  - Calculation of meta-features
- Calculation of **40 meta-features**
  - % frames On-View
  - % frames On-View after smoothing of the On-View contour
  - % frames On-view in the beginning of the turn
  - Av. word score for NOT, ROT, POT, SOT, resp.
  - Max. word score for NOT, ROT, POT, SOT, resp.
  - # frames, # words
  - % content words, % function words (PAJ)
  - Av. number of graphemes per word, etc.

Experimental Results

Pros.	POS	Video	CL in % 2-class case	CL in % 4-class case
•			**76.6**	62.4
	•		76.0	61.0
		•	70.5	45.1
•	•		**80.8**	68.4
•		•	79.7	66.8
	•	•	78.9	68.2
•	•	•	**84.5**	72.3

Tab.2: Classification of On-Focus vs. Off-Focus and On-Focus vs. ROT vs. POT vs. SOT

Conclusion

- Multimodal fusion for the classification of the focus of attention
- Classification with meta-features
- Markedly better results than uni-modal modelling
- Good performance, even if the underlying speech recogniser has low word accuracy:
  20 % WA → 72 % CL; 70 % WA → 82 % CL