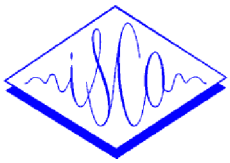
ISCA Archive
http://www.isca-speech.org/archive



Auditory-Visual Speech Processing
2007 (AVSP2007)
Hilvarenbeek, The Netherlands
August 31 - September 3, 2007

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] Chair for Pattern Recognition, University of Erlangen-Nuremberg, Erlangen, Germany
[2] Bavarian Archive for Speech Signals, Munich, Germany,
and Institute for Phonetics and Speech Processing, Munich, Germany
batliner@informatik.uni-erlangen.de

Abstract

In the German SmartWeb project, the user is interacting with the web via a PDA in order to get information on, for example, points of interest. To overcome the tedious use of devices such as push-to-talk, but still to be able to tell whether the user is addressing the system or talking to herself or to a third person, we developed a module that monitors speech and video in parallel. Our database (3.2 hours of speech, 2086 turns) has been recorded in a real-life setting, indoors as well as outdoors, with unfavourable acoustic and light conditions. With acoustic features, we classify up to 4 different types of addressing (talking to the system: On-Talk, reading from the display: Read Off-Talk, paraphrasing information presented on the screen: Paraphrasing Off-Talk, talking to a third person or to oneself: Spontaneous Off-Talk). With a camera integrated in the PDA, we record the user's face and decide whether she is looking onto the PDA or somewhere else. We use three different types of turn features based on classification scores of frame-based face detection and word-based analysis: 13 acoustic-prosodic features, 18 linguistic features, and 9 video features. The classification rate for acoustics only is up to 62 % for the four-class problem, and up to 77 % for the most important two-class problem "user is focussing on interaction with the system or not". For video only, it is 45 % and 71 %, respectively. By combining the two modalities, and using linguistic information in addition, classification performance for the two-class problem so far rises up to 85 %.

Index Terms: Multi-modal Human Machine Interaction, Off-Talk, audio, video

1. Introduction

In interactions with a communication partner — this can be another human or a machine — humans are not always focusing on this interaction itself. They can be distracted by other thoughts or by other people being present and interrupting. For a felicitous communication, it is pivotal that the communication partner can tell apart whether the partner focuses on the interaction itself or not. Depending on the modality, there are different identifiers for a (possible) missing focus of attention: looking away, speaking aside to a third person, private speech, i.e. speaking to one self, etc.

Gaze direction and/or head orientation in dyadic or multi-party conversation, esp. as indicators of attention and addressee, are dealt with in [1, 2, 3, 4, 5, 6]; further references are given in [7]. The fusion of gaze direction and/or head orientation with sound/speech is addressed in [1, 3, 5]. For the multi-party, human-human scenario of [1], a thorough analysis of gaze *direction* has been conducted. However, as it makes no prosodic differences whether the one or the other person is addressed, there is no detailed analysis of the audio-channel. In [3] additionally human-machine interaction occurs. Main differences observed in the audio channel are commands vs. conversation. The scenario in [5] is similar to the triadic scenario in SmartWeb. Here, from the audio channel the length of the speech segment is computed and combined with facial information.

A generic description of private speech is given in [8]; as for multi-modal human-computer interaction, cf. [9]. *Off-Talk*, i.e. speaking aside as a general phenomenon encompassing private speech, vs. *On-Talk*, i.e. addressing the communication partner, is dealt with in [10, 11, 12, 13, 14]. *Off-Talk* as a special dialogue act has not yet been the object of much investigation [15, 16] most likely because it could not be observed in human–human communication. (In a normal human–human dialogue setting, *Off-Talk* might really be rather self–contradictory, because of the 'Impossibility of Not Communicating' [17]. We can, however, easily imagine the use of *Off-Talk* if someone is speaking in a low voice not *to* but *about* a third person present who is very hard of hearing.)

We will use the terms *Off-Talk* for speaking aside (with the sub-classes *Read/Paraphrasing/Spontaneous Off-Talk: ROT/POT/SOT)* and *Off-View* for looking aside. Both phenomena are normally - but not always - signs for a missing focus of attention, i.e. for *Off-Focus*. If the focus of attention is the communication partner, i.e. *On-Focus*, we can normally observe *On-Talk* (the communication partner is addressed) and *On-View* (the communication partner is looked at). Note that *Off-View* is neither a sufficient nor necessary formal condition for *Off-Focus*: we can listen to our partner while looking away. Depending on the culture, this is sometimes necessary because extended eye contact can be considered as aggressive.

We report on a database collected within the German SmartWeb project [18]. The subjects had to communicate with a multi-modal dialogue system providing access to the web via a smart-phone. With a specific Situational Prompting Technique, we could avoid the sparse data problem and could elicit enough instances of missing focus of attention. Experimental results concerning the automatic classification of focus of attention using multiple, multi-modal knowledge sources (audio and video) will be given.

2. The SmartWeb Corpus

2.1. Experimental setup

The SmartWeb Video Corpus was build to gain visual information on the user's head and face during a simulated dialogue. The purpose of this corpus was to enable a module of SmartWeb to detect focus of attention with a video camera embedded into the mobile platform. Two audio tracks and one video track were recorded per session. Recording locations were selected among real life situations with acoustic and visual noise of varying degree. Also the tracks were recorded with resolutions and sample rates that could be expected from a scenario like a question and answering system using mobile networks.

The video track without embedded audio was recorded with a small fixed focus camera of a Nokia 6680 cell phone, which was carried by our subjects. The user could monitor the view of the camera over the display. Hence the majority of video tracks showed the head of the user roughly centered. Video was coded with H.263 directly on the cell phone. One audio track was received by the microphone of a headset and transmitted over a Bluetooth connection to the cell phone and afterwards via Wide-band Code Division Multiple Access WCDMA (UMTS) to a voice processing server located at the Bavarian Archive for Speech Signals, Munich (BAS). Due to the transmission technique this signal was recorded with 8 bits resolution and a sample rate of 8000 s/s with A-law coding following the protocol specification ITU-T G.711. (The other audio signal — not used in the experiments reported on in this paper — was received by a simple collar mounted microphone with 44100 s/s and 16 bit Pulse-Code-Modulation (PCM) coding.)

During the whole session the cell phone was connected to the voice processing server. The server is capable of recording speech, detecting end of speech by means of silence, and emitting arbitrary audio files. The sequence of recordings and acoustic prompts was determined by some XML-files which were in turn generated randomly for each session from a database. The prompts had been pre-recorded with a synthesized voice.

2.2. Data Collection

The data collection design aims at collecting multi-modal data in mobile, realistic environments for the automatic detection of the *On-Focus*/*Off-Focus* phenomenon. For this purpose the *Situational Prompting* technique (SitPro) ([19], [20]) was used as elicitation method[1]. This method integrates *script methods* with *interview techniques* and *speaker prompting* ([21], [22]) into so called *standard prompts*, *individualised prompts*, and *script prompts*. In a *standard prompt*, the human subject is told a characteristic topic of a subject area like soccer (team, group), navigation (public transport, pedestrians), community (restaurant), or information (tourist information, points of interest) to which she/he is supposed to pose a query. An *individualised prompt* is a prompt for which the subject provides his/her own topic. A *script prompt* simulates a three-turn conversation as frequently found in dialogues between human and machine. For this purpose an *instructor* with a female voice and an *operator* with a male voice representing the *Automated Prompting System* was simulated. A system prompt (the *instructor*) followed by a variable silence interval, a recording of the subject's utterance (the *caller*), and a possible system answer (the *operator*) are combined in a *prompt unit*. Six *prompt units* are bundled into an *action unit* as thematic episode. To gather a lot of *Off-Talk* utterances and motions of the subject's head, *SitPro* was used in a scenario with two subjects — the *caller* and the *companion* — and the *Automated Prompting System* APS on a mobile device (cf. Figure 1).

[1] Compared to Wizard-of-Oz experiments, the subject knows that the system is simulated, and system reactions are predetermined.

instructor → *caller*	You would like to know how long public transportation is available at night time.
caller → *operator*	How long is the underground running during the night?
instructor → *caller*	SmartWeb is going to display the answer. You can find the answer on card number eight. Please read it out aloud.
caller	THE REGULAR LINES ARE RUNNING UNTIL 2 A.M. AFTERWARDS THERE ARE NIGHT-LINES.

Table 1: Examples of read Off-Talk ROT (small capitalised)

The recording took place in various indoor and outdoor locations, for example in an office, a coffee bar or a park. As recording equipment, a hand-held unit with a built-in camera for video data recording was used. Two human subjects participated in each experiment. The speech and video data of the *caller* were recorded. His/her task was to interact with the *Automated Prompting System* in order to request information from the system on topics of interest, to read off information of the simulated display, to relay information to the second subject, the *companion*, and to answer the interposed *companion*'s questions. The instructor of the *Automated Prompting System* gives directions about the task, the situation and the topics, and the operator 'answers' the asked questions or gives feedback similar to the targeted SmartWeb system. Prior to the experiment, the *companion* receives a note with nine listed possibilities to distract the *caller* from the call task. For instance *Ask your partner to get informations about a nearby sightseeing*, *Tell your partner to hurry up* or *Ask your partner something concerning his most recent question*. Of course the *companion* may also invent her own interposing questions on the spot. Thus, a controlled triadic communication scenario was established in which the *companion* can only observe the *caller* but not the system's speech and display output.

To elicit *Spontaneous Off-Talk* (SOT), the *companion* had to disturb the interaction between the *caller* and the system with interposed questions. In order to get *Paraphrasing Off-Talk* (POT), the *caller* had the task to report to the *companion* what information he/she had found by interacting with the system. Under these two conditions, changing from *On-View* to *Off-View* was provoked because the *caller* had to move his/her head in the direction of the *companion* to react on his/her interruptions. *Read Off-Talk* (ROT) occurs when the *caller* reads a displayed text aloud. *Off-Talk* utterances which do not fit into these three categories, for instance thinking aloud, were labelled as *Other Off-Talk* (OOT). After a ten minute recording, the participants of the experiment changed their respective roles.

Examples of conversational snippets and their corresponding *Off-Talk* categories are given in tables 1, 2, and 3; note that all examples in the tables were translated from German into English.

instructor → *caller*	Now please ask the system how many times Bayern-München has already beaten the Hamburger SV.
caller → *operator*	How many times has Bayern-München already beaten the HSV?
operator → *caller*	Up to now Bayern-München has beaten the Hamburger SV 49 times.
caller → *companion*	BAYERN-MÜNCHEN HAS ALREADY BEATEN THE HSV FOUR TIMES.

Table 2: Examples of Paraphrasing Off-Talk POT (small capitalised)

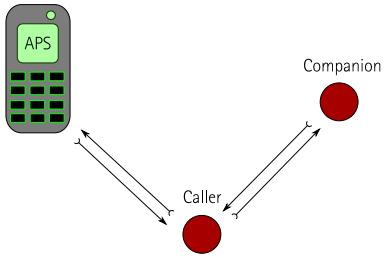


Figure 1: triadic communication situation

3. Annotation

Each word in the corpus has been manually annotated. The distribution of the labels is shown in Table 4. For all classification experiments *Other Off-Talk* was mapped onto *Spontaneous Off-Talk* (SOT). In Sect. 6, the fusion of the two modalities video and audio is investigated on the *utterance* or dialogue turn level (on average 10.8 words per utterance). Thus for each of the 2068 utterances, labels are calculated from the word level by a majority voting described in [23].

The manual annotation of the video recordings includes frame based labeling (7.5 frames per sec.) of the three classes *On-View* (79 %), *between On-/Off-View* (5 %), *Off-View* (14 %), and *No Face* (2 %) as well as the segmentation of faces with a surrounding rectangle[2] to train the face detector described in 4.3. *On-View* is defined as a face looking directly into the camera. Both eyes and the nose are in the image but can be partially occluded, for instance with a hand. Due to the coarse resolution of the images, gaze direction is not taken into account but only head orientation.

[2] automatic segmentation with the face detector of the OpenCV library plus manual segmentation of the On-View frames where the detector failed.

companion → *caller*	Do we actually have a ticket for the underground?
caller → *companion*	YES, I BOUGHT A TICKET.
companion → *caller*	Great. that's good to know.

Table 3: Examples of Spontaneous Off-Talk SOT (small capitalised)

Table 4: *Portion of labels for On-Talk (Not Off-Talk NOT), Read (ROT), Paraphrasing (POT), and Spontaneous Off-Talk (SOT)*

	% NOT	% ROT	% POT	% SOT
word	47.2	12.2	17.3	23.3
utterance	49.6	13.3	11.1	26.0

4. Three Feature Types

From the audio signal, prosodic information is extracted to classify the focus of attention. Part-of-speech (POS) labels are used additionally as linguistic information source. The video channel gives information on the user's gaze direction.

4.1. Prosodic Features

For spontaneous speech it is still an open question which prosodic features are relevant for the different classification problems, and how the different features are interrelated. We tried therefore to be as exhaustive as possible and used a redundant feature set leaving it to the statistic classifier to find out the relevant features and to do the optimal weighting of them. 100 relevant prosodic features modeling duration, energy, F0, jitter, shimmer, and the rate-of-speech were extracted word based but from different context windows. The context was chosen from two words before, and two words after, around a word; by that, we used a sort of "prosodic five-gram". For the computation of our features, we assumed 100 % correct word recognition and used forced alignment of the spoken word chain. Details on the prosodic features are given in [24]; this is a short account:

- length of filled/unfilled pauses before and after the word
- for energy, duration, and F0: a reference feature based on average values for all words in a turn
- for energy: maximum, mean, absolute value, normalized value, and regression coefficient with mean square error
- for duration: absolute and normalized
- for F0: minimum, maximum, mean, and regression coefficient with mean square error; relative position of onset, offset, minimum, maximum, mean on the time axis
- for jitter and shimmer: mean and variance
- a global rate-of-speech feature

In [14] it has been shown that duration features are highly important to discriminate On-Talk (i.e. Not Off-Talk NOT) vs. Read Off-Talk ROT. Duration is also described with features measuring the position of e.g. the maximum or minimum F0. Further, energy is important (higher for NOT). The other Off-Talk categories differ from NOT in lower energy, longer pauses, and a smaller range of F0. Additionally jitter and shimmer are important.

4.2. POS Features

Part-of-speech features are extracted to model linguistic characteristics of the spoken utterance. Again, a 100 % perfect speech recognizer is assumed. First, a POS flag is assigned to each word in the lexicon, cf. [25]. Six cover classes are used: AUX (auxiliaries), PAJ (particles, articles, and interjections), VERB (verbs), APN (adjectives and participles, not inflected), API (adjectives and participles, inflected), and NOUN (nouns, proper nouns). Then, each word of the utterance is described by $6 \times 5 = 30$ binary POS features, because a context of +/- two words is taken into account.

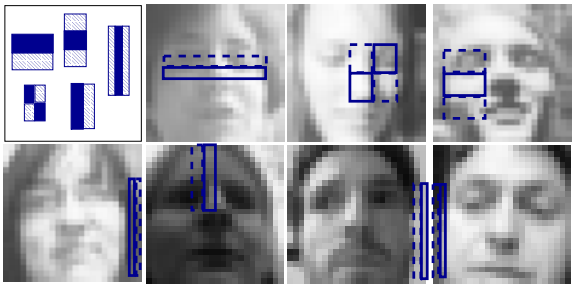


Figure 2: The 7 best features of the SmartWeb face detector. Top left: different shapes of Haar-Wavelets.

In Tab. 5 each On-/Off-Talk category is described by the distribution of POS categories. More content words and less function words are observed for ROT; the same trend can be observed for POT. Most function words are observed for SOT. The explanation is straightforward: the *caller* only reads words, that are presented on the display (and reports them to the *companion*). The words on the display are mostly content words: names of restaurants, cities, etc. Similar observations have also been made with another corpus [12]; features based on POS information are transferable to other domains which is not the case for bag-of-word features or keywords (”*SmartWeb*, show me ...”).

Table 5: *Word based evaluation of POS classes: percent occurrences for NOT, ROT, POT, and SOT*

	NOUN	API	APN	VERB	AUX	PAJ
NOT	21.2	5.4	4.2	7.1	9.1	52.9
ROT	**29.6**	5.7	**15.7**	7.2	7.3	34.5
POT	**27.3**	4.0	**10.5**	5.2	9.8	43.2
SOT	8.2	1.3	6.2	**10.5**	10.7	**63.1**
total	20.3	4.3	7.2	7.6	9.4	51.4

4.3. Face Detection

For the classification of *On-View/Off-View*, it is sufficient in our task to discriminate frontal faces from the rest. Thus, we employed a very fast and robust algorithm described in [26]. The face detection works for single images; no use of context information is implemented. The algorithm is based on simple Haar-like wavelets; all wavelets (up to scaling and translation) are shown in Fig. 2, top left. For each wavelet-feature, the light area is subtracted from the dark area (the dashed rectangle from the solid rectangle). From many possible features, wavelets containing complementary information are selected with the AdaBoost algorithm; a hierarchical classifier speeds up the classification. In this paper we use 176 $\times$ 144 grey-scale images, 7.5 per second; faces are searched in different sub-images, greater than half the image, and scaled to 24 $\times$ 24.

A classifier was trained using 9500 positive and 7500 negative samples from 60 speakers (additionally 485 faces plus 425 images containing landscape have been downloaded from the internet) using the OpenCV library[3]. The resulting face detector is based on 452 Haar-features; the best 7 are shown in Fig. 2 with random images (24 $\times$ 24) of the SmartWeb corpus in the background. Comparing the OpenCV default classifier based on 2913 features with our classifier trained on the SmartWeb data, the following results (discussed in [23]) are obtained: Our classifier detects only 80% of the faces of a control set with 375 German members of parliament, whereas the OpenCV classifier detects 99 %. However, the class-wise averaged recognition rate on the SmartWeb test set rises with the SmartWeb classifier from 81 to 88 %.

[3] http://sourceforge.net/projects/opencvlibrary/

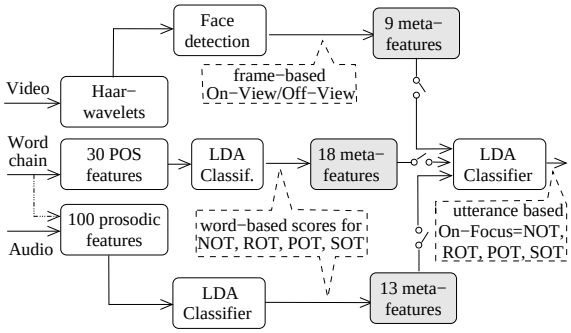


Figure 3: Utterance classification with meta-features

5. Fusion

.

For the multi-modal fusion, the classification of On/Off-View has to be combined with the classification of On-/Off-Talk. The target is an utterance based machine score for the four classes NOT (Not-Off-Talk, i.e. On-Talk), ROT (Read Off-Talk), POT (Paraphrasing Off-Talk), and SOT (Spontaneous Off-Talk), which have been manually annotated (Tab. 4). In the case of multi-modal classification, we refer to NOT as On-Focus; Off-Focus is subdivided in ROT, POT and SOT. In preliminary experiments described in [13, 14], we obtained good results for the classification of On/Off-Talk applying a word based prosodic analysis; for On-/Off-View, however, an image based classification makes more sense than e.g. analyzing an image averaged over all frames of a word, and is additionally quite efficient using the Viola-Jones algorithm. Further, we do not want to use a set of thresholds or rules to combine both modalities but want a classifier (”combiner”) to learn those decisions from the training data. Consequently, it makes sense to feed the ”combiner” with as much information as possible and to join the two steps *mapping onto the utterance level* and *fusion* in a single classification step based on ”meta-features” as illustrated in Fig. 3.

Using the word-based On-/Off-Talk recognition, 13 utterance-based meta-features are calculated: the number of words and the four word scores for NOT, ROT, POT, and SOT averaged over the whole turn. Further, the variation of each score is described with its maximum and minimum.

Similar utterance features are also calculated from the word-based POS classification. Here, additionally the percentage of each of 3 POS super sets — content words NOUN/API/APN, verbs VERB/AUX, function words PAJ, cf. Tab. 5 — is calculated per utterance. Together with the average, minimum and maximum linguistic word length (# graphemes) 18 linguistic meta-features are obtained.

From the frame based classification of On-/Off-View, nine

further utterance-based meta-features are calculated[4]: the number of frames, the proportion of On-View frames and this proportion separately for the 1st, 2nd, 3rd and 4th quartile of the utterance, in order to cope with situations, where the user e.g. does not look onto the display in the beginning or end of an utterance. Three further features are obtained by applying a morphological operation on the On-View contour: The frame based results are smoothed using three different time windows; this is important if, e.g., strong back light is the reason that a face is recognized only in every ith frame.

The utterance classification using an LDA-classifier as "combiner" is performed with combinations of 13, 18, or 9 meta-features (prosodic, linguistic, video).

6. Experimental Results

For the experiments, the data was divided into a training set and a test set. They comprise 58 vs. 37 speakers[5], 1130 vs. 748 utterances, and 13800 vs. 8400 words. All results are described with the class-wise averaged recognition rate (CL) which is the mean of all recalls. The *recall* of a class is the percentage of correctly classified elements given this class.

Table 6: *Confusion matrices using prosodic features (left) and face detection (right); % classified correctly*

	NOT	ROT	POT	SOT	NOT	ROT	POT	SOT
NOT	**64.8**	6.4	11.3	17.5	**69.7**	8.0	8.2	14.1
ROT	17.1	**62.2**	8.1	12.6	55.0	**12.6**	18.9	13.5
POT	18.4	10.3	**51.7**	19.5	12.6	4.6	**67.8**	15.0
SOT	8.7	4.3	16.1	**70.8**	18.6	8.7	42.3	**30.4**

The confusion matrices of the LDA-classifier resulting from separate evaluations of each modality are show in Tab. 6, left, for prosodic features, in Tab. 6, right, for features based on face detection, and in Tab. 7, left, using POS information. Obviously, it is difficult to detect POT using the audio channel or just the word chain; using the video-channel, a recall of 67.8 % is obtained for POT which correlates with Off-View. However, using solely video (Tab. 6, right) shows that the detection of ROT nearly always fails, and also the results for SOT are only little better than chance: it cannot be classified without using prosodic or linguistic information.

Table 7: *Confusion matrix using POS features (left) and a combination of 3 feature types (right): prosody (speaker normalized), POS, and video; % classified correctly*

	NOT	ROT	POT	SOT	NOT	ROT	POT	SOT
NOT	**62.5**	3.6	13.6	20.3	**79.7**	4.1	3.6	12.6
ROT	3.6	**67.6**	18.0	10.8	9.9	**73.0**	9.0	8.1
POT	23.0	8.0	**50.6**	18.4	9.2	8.0	**64.4**	18.4
SOT	21.2	2.5	13.0	**63.3**	8.7	3.7	15.5	**72.1**

In Tab. 8 classification rates are given for each feature type/modality and different combinations for the 2-class problem (On-Focus vs. Off-Focus) and for the 4-class problem (NOT, ROT, POT, SOT). "Pros. norm." indicates speaker normalized prosodic features (zero mean and variance 1). This optimistic case — knowing all the speaker's utterances in advance — shows how much improvement can be achieved using speaker adaptation. This way, in the 2-class case the classification with prosodic features rises from 68.6 to 76.6 % CL. With linguistic information (no adaptation required), 76.0 % CL are achieved, and with video information 70.5 %. Combining any two modalities, the classification rate rises up to 80.8 %. Using all 3 modalities, 84.5 % CL are obtained. 4 classes are discriminated with 72.3 %, no matter if speaker normalization is applied or not. The confusion matrix of the best constellation for the 4-class problem ("Pros. norm.", second last line in Tab. 8) is shown in Tab. 7, right. There is still high confusion between POT and SOT.

Table 8: *Classification of On-Focus vs. Off-Focus and On-Focus vs. ROT vs. POT vs. SOT using prosodic features, speaker normalized prosodic features, POS features, and face detection*

Pros.	Pros. norm.	POS	Video	CL in % 2-class case	CL in % 4-class case
•				68.6	55.3
	•			**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
•		•	•	83.8	72.3

7. Concluding Remarks

The phenomena that we addressed in this paper can be suppressed in dyadic human-machine interaction if some precautions are taken; for instance, a push-to-talk button and a strict system initiative can reduce *Off-Talk* and *Off-View* to a considerable extent: the dyadic setting in the SmartKom scenario (even without devices such as push-to-talk) yielded only some 6% *Off-Talk* words [12, 14]; this in turn constitutes the well-known sparse-data problem in real-life settings. However, especially in the more natural triadic and multi-party interaction settings, this is not possible or would result in a rather artificial interaction. We could successfully overcome this problem with the SitPro technique which resulted in more than 50% *Off-Focus*.

The transition of controlled, acted data with 'clean' recording settings onto more realistic scenarios 'in the open air' — this can be taken literally in the case of our SmartWeb data — results in unfavorable recording conditions: acoustic noise in the case of speech, and 'video noise' such as back-light, reduced brightness aso. This in turn prevents the use of sensitive techniques such as gaze tracking. Instead, we employed a rather simple and robust face detection algorithm. For speech, we so far used the spoken word chain; note, however, that our prosodic features are rather robust if used with output of speech recognition such as word hypothesis graphs. The same holds for POS features. Even if the video and audio cues do not always 'point towards the same direction' — *ROT* can trivially not be recognized with video information because the user has to face the system while reading, and *POT* is poorly recognized by using only audio information — a fusion of both channels

[4] slightly different values in comparison to [23] due to small changes of the alignment

[5] 4 of the 99 speakers were not used due to technical problems

and all three feature types yielded markedly better results than a uni-modal modelling.

8. Acknowledgements

This work was partially funded by the German Federal Ministry of Education and Research (BMBF) in the framework of SmartWeb, Grant 01 IMD 01 F, http://www.smartweb.dfki.de. The responsibility for the content lies with the authors.

9. References

[1] R. Stiefelhagen, J. Yang, and A. Waibel, "Modeling Focus of Attention for Meeting Indexing Based on Multiple Cues," *IEEE Transactions on Neural Networks. Special Issue on Intelligent Multimedia Processing, July 2002*, vol. 13, pp. 928–938, 2002.

[2] R. Stiefelhagen and J. Zhu, "Head orientation and gaze direction in meetings," in *CHI '02: CHI '02 extended abstracts on Human factors in computing systems*. New York, NY, USA: ACM Press, 2002, pp. 858–859.

[3] M. Katzenmaier, R. Stiefelhagen, and T. Schultz, "Identifying the Addressee in Human-Human-Robot Interactions Based on Head Pose and Speech," in *Proceedings of the 6th ICMI 2004*, 2004, pp. 144–151.

[4] N. Jovanovic and R. op den Akker, "Towards Automatic Addressee Identification in Multi-party Dialogues," in *Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue*, M. Strube and C. Sidner, Eds. Cambridge, Massachusetts, USA: Association for Computational Linguistics, April 30 - May 1 2004, pp. 89–92.

[5] K. van Turnhout, J. Terken, I. Bakx, and B. Eggen, "Identifying the intended addressee in mixed human-human and human-computer interaction from non-verbal features," in *ICMI '05: Proceedings of the 7th international conference on Multimodal interfaces*. New York, NY, USA: ACM Press, 2005, pp. 175–182.

[6] M. Rehm and E. André, "Where do they look? Gaze Behaviors of Multiple Users Interacting with an ECA," in *Intelligent Virtual Agents: 5th International Working Conference, IVA 2005*. Berlin, New York: Springer, 2005, pp. 241–252.

[7] D. Heylen, "Challenges Ahead. Head Movements and other social acts in conversation," in *Proceedings of AISB - Social Presence Cues Symposium*, 2005, pp. 45–52.

[8] M. K. Ahmed, "Private speech: A cognitive tool in verbal communication," in *The Language Programs of the International University of Japan Working Papers*, S. Kimura and M. Leong, Eds., vol. 5. International University of Japan, 1994.

[9] R. Lunsford, "Private Speech during Multimodal Human-Computer Interaction," in *Proceedings of the 6th ICMI*, Pennsylvania, 2004, p. 346, (abstract).

[10] D. Oppermann, F. Schiel, S. Steininger, and N. Beringer, "Off-Talk – a Problem for Human-Machine-Interaction," in *Proceedings of Eurospeech 2001*, Aalborg, 2001, pp. 2197–2200.

[11] R. Siepmann, A. Batliner, and D. Oppermann, "Using Prosodic Features to Characterize Off-Talk in Human-Computer-Interaction," in *Proceedings of the Workshop on Prosody and Speech Recognition 2001*, Red Bank, N.J., 2001, pp. 147–150.

[12] A. Batliner, V. Zeissler, E. Nöth, and H. Niemann, "Prosodic Classification of Offtalk: First Experiments," in *Proceedings of the 5th TSD*. Berlin: Springer, 2002, pp. 357–364.

[13] C. Hacker, A. Batliner, and E. Nöth, "Are You Looking at Me, are You Talking with Me – Multimodal Classification of the Focus of Attention," in *Proceedings of the 9th TSD*. Berlin: Springer, 2006, pp. 581–588.

[14] A. Batliner, C. Hacker, and E. Nöth, " To Talk or not to Talk with a Computer: On-Talk vs. Off-Talk," in *How People Talk to Computers, Robots, and Other Artificial Communication Partners*, ser. University of Bremen, SFB/TR 8 Report, K. Fischer, Ed., 2006, pp. 79–100.

[15] J. Alexandersson, B. Buschbeck-Wolf, T. Fujinami, M. Kipp, S. Koch, E. Maier, N. Reithinger, B. Schmitz, and M. Siegel, "Dialogue Acts in VERBMOBIL-2 – Second Edition," Verbmobil Report 226, Juli 1998.

[16] J. Carletta, N. Dahlbäck, N. Reithinger, and M. Walker, "Standards for Dialogue Coding in Natural Language Processing," Dagstuhl-Seminar-Report 167, 1997.

[17] P. Watzlawick, J. Beavin, and D. D. Jackson, *Pragmatics of Human Communications*. New York: W.W. Norton & Company, 1967.

[18] W. Wahlster, "SmartWeb: Mobile Applications of the Semantic Web," http://smartweb.dfki.de/Vortraege/SmartWeb-Wahlster-KI-2004-LNAI.PDF, 2004.

[19] H. Mögele, M. Kaiser, and F. Schiel, "SmartWeb UMTS Speech Data Collection. The SmartWeb Handheld Corpus," in *Proceedings of the 5th LREC*. Genova, Italy: ELRA, May 2006, pp. 2106–2111.

[20] M. Kaiser, H. Mögele, and F. Schiel, "Bikers Accessing the Web: The SmartWeb Motorbike Corpus," in *Proceedings of the 5th LREC*. Genova, Italy: ELRA, May 2006, pp. 1628–1631.

[21] D. Gibbon, R. Moore, and R. Winski, Eds., *Handbook of Standards and Resources for Spoken Language Systems*. Walter de Gruyter, 1997.

[22] S. Rapp and M. Strube, "An Iterative Data Collection Approach for Multimodal Dialogue Systems," in *Proceedings of the 3rd LREC*, Las Palmas, Canary Islands, Spain, May 2002, pp. 661–665.

[23] E. Nöth, C. Hacker, and A. Batliner, "Does Multimodality Really Help? The Classification of Emotion and of On/Off-Focus in Multimodal Dialogues - Two Case Studies." in *Proc. of the 49th International Symposium ELMAR-2007*, 2007, to appear.

[24] A. Batliner, K. Fischer, R. Huber, J. Spilker, and E. Nöth, "How to Find Trouble in Communication," *Speech Communication* , vol. 40, pp. 117–143, 2003.

[25] A. Batliner, M. Nutt, V. Warnke, E. Nöth, J. Buckow, R. Huber, and H. Niemann, "Automatic Annotation and Classification of Phrase Accents in Spontaneous Speech," in *Proceedings of Eurospeech99*, Budapest, 1999, pp. 519–522.

[26] P. Viola and M. J. Jones, "Robust Real-Time Face Detection," *Int. J. Comput. Vision*, vol. 57, no. 2, pp. 137–154, 2004.