To talk or not to talk with a computer: taking into account the user's focus of attention

Anton Batliner, Christian Hacker, Elmar Nöth

Angaben zur Veröffentlichung / Publication details:

Batliner, Anton, Christian Hacker, and Elmar Nöth. 2008. "To talk or not to talk with a computer: taking into account the user's focus of attention." *Journal on Multimodal User Interfaces* 2 (3-4): 171–86. https://doi.org/10.1007/s12193-009-0016-6.

Noname manuscript No.
(will be inserted by the editor)

To Talk or not to Talk with a Computer
Taking into Account the User's Focus of Attention

Anton Batliner · Christian Hacker · Elmar Nöth

Received: date / Accepted: date

Abstract If no specific precautions are taken, people talking to a computer can – the same way as while talking to another human – speak aside, either to themselves or to another person. On the one hand, the computer should notice and process such utterances in a special way; on the other hand, such utterances provide us with unique data to contrast these two registers: talking vs. **not** talking to a computer. In this paper, we present two different databases, SmartKom and SmartWeb, and classify and analyse On-Talk (addressing the computer) vs. Off-Talk (addressing someone else) — and by that, the user's focus of attention — found in these two databases employing uni-modal (prosodic and linguistic) features, and employing multimodal information (additional face detection).

Keywords focus of attention · Off-Talk · prosodic features · face detection · automatic classification

This work was funded by the German Federal Ministry of Education, Science, Research and Technology (*BMBF*) in the framework of the SmartKom project under Grant 01 IL 905 K7 and in the framework of the SmartWeb project under Grant 01 IMD 01 F. The responsibility for the contents of this study lies with the authors.

Anton Batliner
Chair of Pattern Recognition, Department of Computer Science, Friedrich-Alexander-University Erlangen-Nuremberg, Martensstr. 3, 91058 Erlangen, Germany
Tel.: +49 9131 85 27823
Fax: +49 9131 303811
E-mail: batliner@informatik.uni-erlangen.de

Christian Hacker
Elektrobit Automotive GmbH, Erlangen, Germany

Elmar Nöth
Chair of Pattern Recognition, Department of Computer Science, Friedrich-Alexander-University Erlangen-Nuremberg, Martensstr. 3, 91058 Erlangen, Germany

Ms., in: Journal on Multimodal User Interfaces, in press, available online at:
http://www.springerlink.com/content/121040/?Content+Status=Accepted

1 Introduction

Enter Guildenstern and Rosencrantz. […]
Guildenstern My honoured lord!
Rosencrantz My most dear lord! *[…]*
Hamlet […] You were sent for *[…]*
Rosencrantz To what end, my lord?
Hamlet That you must teach me *[…]*
Rosencrantz [Aside to Guildenstern] What say you?
Hamlet [Aside] Nay then, I have an eye of you! *[Aloud.]* If you love me, hold not off.
Guildenstern My lord, we were sent for.

In this passage from Shakespeare's *Hamlet*, we find two *'Asides'*, one for speaking aside to a third person, the other one for speaking to oneself. Implicitly we learn that such asides are produced with a lower voice: when Hamlet addresses Guildenstern and Rosencrantz again, the stage direction reads *Aloud*. And we can imagine that while speaking aside, Rosencrantz is turning towards Guildenstern, and Hamlet away from both — maybe towards the audience. Thus both speech characteristics and head/body orientation can play a role.

In interactions with a communication partner, 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 other 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, speaking to one self, etc.

Nowadays, the dialogue partner does not need to be a human being but can be an automatic dialogue system as well. The more elaborate such a system is, the less restricted is the behaviour of the users. In the early days, the users were confined to a very restricted vocabulary such as prompted numbers etc. In most systems still a push-to-talk (PTT) button has to be pressed before user interaction. In conversations with more elaborate automatic dialogue systems, users behave more naturally; thus, phenomena such as speaking aside can be observed and have to be coped with that could not be observed in communications with very simple dialogue systems. Normally the system should not react to these utterances, or it should process them in a special way, for instance, on a meta level, as remarks about the (mal–) functioning of the system, and not on an object level, as communication with the system. The monitoring of this phenomenon is most promising for applications where it is likely to occur: the driver of a car can address a navigation/information system or the co-driver; elderly people might tend to speak to themselves while alone in their flat. This has to be told apart from addressing a surveillance system by this very system itself because elderly people might not be able to operate a PPT button in a reliable way, cf. [16].

Asides can be on-topic or out-of-topic: the driver of a car can negotiate with the co-driver whether they should follow the advice just given by the information system (on-topic); or they can talk about their plans for next Sunday (out-of-topic). To detect out-of-topic vocabulary is a task, still too difficult for state-of-the art automatic dialogue systems, which have to keep the lexicon small by using only in-topic vocabulary. Thus, the system has to employ information on *how* has been spoken (prosody), *what* has

Ms., in: Journal on Multimodal User Interfaces, in press, available online at:
http://www.springerlink.com/content/121040/?Content+Status=Accepted

been spoken about (linguistics), and *where* the speaker is looking to (visual focus of attention).[1]

In this paper, we deal with this phenomenon *Speaking Aside* which we want to call *Off-Talk* following [23]. There *Off-Talk* is defined as comprising "every utterance that is not directed to the system as a question, a feedback utterance or as an instruction". This comprises reading aloud from the display, speaking to oneself (thinking aloud), speaking aside to other people which are present, etc. The default register for interaction with computers is, in analogy, called *On-Talk*. Here, *On-Talk* is practically the same as Computer Talk, cf. [12]. However, whereas in the case of other (speech) registers such as *baby-talk* the focus of interest is on the way *how* they are produced, i.e. their phonetics, in the case of Computer Talk, the focus of interest so far has rather been on *what* has been produced, i.e. its linguistics (syntax, semantics, pragmatics). This can be traced back to the different research traditions in psychology (baby-talk) and Natural Language processing (Computer Talk).

2 Related Work

Speaking to oneself ('self-directed speech') as a necessary component in children's development has been introduced by [24] as *egocentric speech* and elaborated on by [31]; it can be silent, 'inner' speech, or, if externalized, audible speech; an overview of this phenomenon which is nowadays called *private speech* can be found in [10]. [20] addressed senior subject's private speech interacting via speech and pen with a multi-modal map-based simulation task. The term *Off-Talk* has been introduced by [23] for phenomena that could be observed in a dyadic scenario where a single user is interacting with an automatic system [33]. The prosodic characteristics of this type of *Off-Talk* are described in [27, 9].

Speaking aside as a special *dialogue act* has not yet been the object of much investigation, cf. [1, 11], most likely because it could not be observed in those human–human communications which were analysed for dialogue act modelling. In a normal human–human dialogue setting, *Off-Talk* might really be rather self–contradictory, because of the 'Impossibility of Not Communicating', cf. [35]: automatically, each verbal production of a speaker will be taken by the dialogue partner as conveying some message. 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.

In the last years, a new research topic has emerged, namely multi-modal, multi-party interaction with other humans, for instance in meetings, or with both other humans and computers, for instance with information systems and/or embodied agents. Basically, matters are more complicated in such scenarios than in a dyadic, face-to-face scenario: several speakers can overlap, and light and audio conditions are often less favourable. Maybe because of these additional factors, so far, often rather coarse parameters have been employed such as head orientation in the video channel, and a binary decision of speech vs. non-speech in the audio channel. [28] address focus of attention using face tracking and estimating head poses; moreover, they predict focus from sound, i.e. focus of attention is triggered by participants who are speaking – no matter what they are saying. Speech is thus treated as a simple binary feature. In

[1] Note that in this example, gaze is of course not very promising: the driver should always focus the road ahead.

Ms., in: Journal on Multimodal User Interfaces, in press, available online at:
http://www.springerlink.com/content/121040/?Content+Status=Accepted

this scenario it makes no prosodic differences whether the one or the other person is addressed; consequently, there is no detailed analysis of the audio channel (speech), but only of the video channel. In [18] *On-Talk* and *On-View* (i.e. the speaker is looking at the communication partner) are analysed for a Human-Human-Robot scenario; face detection is based on the analysis of the skin-colour; to classify the speech signal, different linguistic features are investigated. Main differences observed in the audio channel are commands vs. conversation — a consequence of a low-complexity dialogue system. The assumption is that commands directed to a robot are shorter, contain more often imperatives or the word "robot", have a lower perplexity, and are easy to parse with a simple grammar. However, the discrimination of *On-/Off-Talk* will become more difficult in an automatic dialogue system, since speech recognition is not solely based on commands. [17] want to incorporate information on the addressee (word classes such as personal pronouns), dialogue history, and gaze direction. [25] investigate gaze direction in a gamble system. Gaze direction and/or head orientation in dyadic or multi-party conversations, esp. as indicators of attention and addressee, are dealt with in [29]. Further references to basic aspects of head movement in conversation are given in [15]. The scenario in [29] is similar to the triadic scenario in SmartWeb described below; from the audio channel the length of the speech segment is computed and combined with facial information.

Thus, so far only very coarse acoustic or linguistic parameters have been employed for detecting the focus of attention in multi-modal, multi-party interactions.

3 Outline

For automatic dialogue systems, a good classification performance is most important; the way how to achieve this could be treated as a black-box. In the present paper, however, we are not especially interested in classification and its fine-tuning but use these results as measures of goodness-of-fit and try to interpret the most salient features. To learn more about the phonetics of Computer-Talk, *On-Talk* vs. *Off-Talk* is a unique constellation because all other things are kept equal: the scenario, the speakers, the system, the microphone, etc. Thus we can be sure that any difference we find can be traced back to this very difference in speech registers – to talk or not to talk with a computer – and not to some other intervening factor. The same holds for the video channel and for head orientation as feature. For the experiments reported on in the following, we employ procedures from data mining and pattern recognition. The classifiers chosen are fast and reliable, with a possibly not best but competitive performance. A good classification means a good modelling of the phenomenon. We are as well interested in possible reasons for sub-optimal classification performance. We are investigating the fusion of knowledge sources within the same modality speech using prosodic and linguistic information, and across modalities using speech and video information.

The focus of this paper is to investigate whether feature extraction based on prosody, linguistic information and gaze direction is suitable for automatic classification of the user's focus of attention in two different scenarios, SmartKom and SmartWeb. Further, it will be found which information is important to classify which category of *Off-Talk*.

We will subcategorize *Off-Talk*, i.e. speaking aside, into the sub-classes *Read Off-Talk: READ*, *Paraphrasing Off-Talk: PARA*, and *Spontaneous Off-Talk: SPONT*, and

Ms., in: Journal on Multimodal User Interfaces, in press, available online at:
http://www.springerlink.com/content/121040/?Content+Status=Accepted

we will use *Off-View* for looking aside. *READ* means reading aloud what the system presents on the screen; *PARA* means that the users paraphrase to a third person present what they have been told by the system or have seen on the screen; *SPONT* describes any other type of *Off-Talk*, be this talking to oneself or to somebody else. If we do not tell apart *PARA*, *SPONT*, or other sub-categories from each other, we speak about the main class *Other Off-Talk: OTHER*. In a dyadic setting, *SPONT* is mostly speaking to oneself, in a triadic or multi-party setting, *SPONT* mostly means speaking to another communication partner. An overview of all *On-/Off-Talk* categories is given in Table 1; examples will be discussed in Section 4. Both *Off-Talk* and *Off-View* 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 observe *On-Talk* (the communication partner is addressed) and normally *On-View* (the communication partner is looked at).[2] 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 sometimes can be necessary because extended eye contact can be considered as aggressive. It depends on the context as well: when both communication partners are looking at a breath-taking landscape, it can be fully acceptable not to look at the partner while addressing him/her. In a closed-room setting, however, always looking away might be conceived as an impolite, even autistic trait.

In section 4 we present the two systems SmartKom and SmartWeb and the resp. databases where *Off-Talk* could be observed and/or has been provoked. For the comparison of *Off-Talk* with *On-Talk* in section 6 with the help of speech features, we first describe in section 5 the prosodic and part-of-speech features that were extracted and used for classification and interpretation. In section 7, we address the fusion of speech and head orientation information in SmartWeb. We show that the fusion within and across modalities contributes to classification performance and by that, to a better modelling of *Off-Talk* and *Off-Focus* — even if there is no straightforward correlation between *Off-Talk* and *Off-View*.

4 Systems

4.1 The SmartKom System

SmartKom is a multi–modal dialogue system which combines speech with gesture and facial expression. The speech data investigated in this paper are obtained in large–scaled Wizard-of-Oz experiments[3] within the SmartKom 'public' scenario: in a multi–modal communication telephone booth, the users can get information on specific points of interest as, e.g., hotels, restaurants, or cinemas. The user delegates a task, for instance, finding a film, a cinema, and reserving the tickets, to a virtual agent which is visible on the graphical display. This agent is called 'Smartakus' or 'Aladdin'. The user gets the necessary information via synthesized speech produced by the agent, and on the graphical display, via presentations of lists with points of interest, and maps

[2] When we are talking about the phenomenon, we use *On-Talk*, when we are talking about types and tokens in our databases, we use *ONTALK*, in analogy to the subtypes of *Off-Talk*.

[3] The role of the machine is played by a human operator while the users believe that they are communicating with a machine; after Baum, Lyman Frank: The Wonderful Wizard of Oz (1900), cf. [13]. The recordings took place at the Institute of Phonetics and Speech Communication, Ludwig-Maximilians-University, Munich.

Ms., in: Journal on Multimodal User Interfaces, in press, available online at:
http://www.springerlink.com/content/121040/?Content+Status=Accepted

of the inner city, etc. The dialogue between the system and the user is recorded with several microphones and digital cameras. Subsequently, annotations are carried out. The recorded speech represents thus a special variety of non–prompted, spontaneous speech typical for human–machine–communication in general and for such a multi–modal setting in particular. More details on the system can be found in [34,33], more details on the recordings and annotations in [23,27].

In the annotation, two different types of *Off-Talk* are labelled: *read Off-Talk READ* and *other Off-Talk OTHER*; every other word is via default annotated with the label *ONTALK* as *no Off-Talk*. If the user reads aloud words presented on the display, this is labelled as *READ*; it was decided not to tell apart all other types of *Off-Talk*, e.g., speaking aside to another person or speaking to oneself, because in this setting, such decisions are not easy to make; moreover, they would result in sparse data, i.e. only a few cases per class.

In the following example, a user wants to reserve two tickets for the cinema; first, the German original is presented, then an English translation. In the German example, <P> denotes pauses, +/ and /+, respectively, beginning and end of a repair, and <hm> a filled pause. *Off-Talk* is labelled adjacent to the pertinent word, without blanks. Most of the annotated phenomena are, however, filtered out of the German example. The English translation contains only the wording; here, *READ* is given in capitals, and *OTHER* boldface and regular.

system: *in diesem Kino k"onnen Sie nur telefonisch reservieren*
user: *kannst du mich dahin verbinden <P> Aladdin*
system: *diese Funktion ist momentan leider nicht verf"ugbar <P> kann ich Ihnen auch anders helfen*
user: *<hm> <P> das<OTHER> is'<OTHER> doof<OTHER> <hm> <P> <A> geh doch noch mal zur"uck <P> Aladdin <P> zur"uckgehen*
system: *einen Augenblick bitte*
user: *ja <P> +/wie sieht d=/+ wie sieht das aus mit dem Kino Schlo"s <P> dreiundzwanzig f"unfundvierzig <P> zwei Karten Aladdin <P> hoffentlich<OTHER> nich'<OTHER> wieder<OTHER> telefonisch<OTHER> doch<OTHER>*
system: *in diesem Kino k"onnen Sie nur telefonisch reservieren*
user: *<hm> das<OTHER> is'<OTHER> nat"urlich<OTHER> auch<OTHER> doof<OTHER> <P> gibt es noch eine andere M"oglichkeit <P> sicher<OTHER> nicht<OTHER> <P> zur"uck*
system: *Reservierungen k"onnen nur im Kino StudioEuropa durchgef"uhrt werden*
user: *aha na dann <P> gib mir doch mal das <P> Kino Europa <P> Kino Europa ah das sieht doch schon gut aus gut dann nehmen wir mal <P> <hm> <P> AmericanHistoryX zwoundzwanzig f"unfundvierzig <P> aha<OTHER> Reservierung<READ> best"atigen<READ> da<OTHER> kommen<OTHER> wir<OTHER> doch<OTHER> schon<OTHER> der<OTHER> Sache<OTHER> n"aher<OTHER>*

system: *For this cinema, only reservation by phone is possible.*
user: *Can you connect me, Aladdin?*
system: *I'm afraid, at the moment, this is not possible.*
user: **That's bad** *– well – please go back, Aladdin – go back!*
system: *One moment please!*
user: *Well, what about, what about the Castle Cinema, 10.45 pm, Armageddon – two tickets,*

Table 1 *Cross-tabulation of On-/Off-Talk vs. On-/Off-View. In the triadic scenario SmartWeb* OTHER *is mapped onto* SPONT*; in SmartKom (dyadic) no* PARA *and no* SPONT *occurred.*

	On-View	*Off-View*
ONTALK	On-Focus, Interaction with the system	(unusual)
READ	Reading from the display	(not possible)
PARA	(unusual)	Reporting results from the system (if triadic)
SPONT	Responding to an interruption (if triadic)	Responding to an interruption (if triadic),
OTHER	talking to one-self, swearwords	talking to oneself, swearwords

Aladdin – **I hope it's not again reservation by phone.**
system: *For this cinema, only reservation by phone is available.*
user: Again, that's bad. *Is there another possibility?* **I guess not!** *Go back!*
system: *Reservations are only possible for the Studio Europe.*
user: *Well, okay, Studio Europe, Studio Europe, that's fine, well, then let's take – uh – AmericanHistory, 10.45 pm,* **okay, CONFIRM RESERVATION, now we are coming to the point.**

At least in this specific scenario, *READ* is fairly easy to annotate: the labeller knows what is given on the display, and knows the dialogue history. *OTHER*, however, as a sort of wast-paper-basket category for all other types of *Off-Talk*, is more problematic; for a discussion we want to refer to [27]. Note, however, that the labellers listened to the dialogues while annotating; thus, they could use acoustic information, for instance, whether some words are spoken in a very low voice or not. This, of course, would not be possible if only the transliteration[4] was available. As there is no third communication partner, *Off-View* will not be modelled; if it occurs, it might be taken as spurious or as indication of considering/thinking.

4.2 The SmartWeb System

In the SmartWeb-Project, cf. [32] — the follow-up project of SmartKom — a mobile and multimodal user interface to the Semantic Web has been developed. The users can ask open-domain questions to the system, no matter where they are; carrying a smartphone, they address the system via UMTS or WLAN using speech, cf. [26]. Now the idea is, as in the case of SmartKom, to classify automatically whether speech is addressed to the system or to someone else. Thus, the system can do without any push-to-talk button and, nevertheless, the dialogue manager will not get confused. To classify the user's focus of attention, we employ information from two modalities: speech-input from a close-talk microphone and the video stream from the front camera of the mobile phone are analysed on the server. In the video stream, we classify *On-View* when the user looks into the camera. This is reasonable since the users normally will look onto the display of the smartphone while interacting with the system, because

[4] With 'transliteration' we denote the manual orthographic transcription of the utterances.

they receive visual feedback, like the n-best results of a query, maps and pictures, or even web-cam streams showing the object of interest. *Off-View* means that the user does not look at the display at all. We conceive *On-View* as looking onto the display vs. *Off-View* as looking away from it and 'binarise', i.e. operationalise this difference in head orientation with face detection: if a face is detected (head orientation towards the display), we assume *On-View*, if not (head orientation towards any other direction), we assume *Off-View*.[5]

For the SmartWeb-Project two databases containing questions in the context of a visit to a Football World Cup stadium in 2006 have been recorded. Different categories of *Off-Talk* were evoked (in the SW_{spont} database[6]) or acted (in our own SW_{acted} recordings[7], cf. Sect. 6.1). Besides *Read Off-Talk* (*READ*), where the subjects read some system response from the display, the following categories of *Off-Talk* are discriminated: *Paraphrasing Off-Talk* (*PARA*) means, that the subjects report to someone else what they have found out from their request to the system, and *Spontaneous Off-Talk* (*SPONT*) can occur, when they are interrupted by a third person present. We expect *READ* to occur simultaneously with *On-View* and *PARA* with *Off-View*. Table 1 displays a cross-tabulation of possible combinations of *On-/Off-Talk* with *On-/Off-View*, especially tailored for SmartWeb but taking into account SmartKom as well. Recording locations were selected among real-life situations with acoustic and visual noise of varying degree, e.g. in an office, a coffee bar, or a park. The system prompts were scripted, and the so-called *Situational Prompting Technique*, cf. [21], was used; in [6] more in-depth technical descriptions of recordings and the experimental design are given. Compared to Wizard-of-Oz experiments, the subject knows that the system is simulated, and system reactions are predetermined.

In the following example, only the user turns are given. The user first asks for the next play of the Argentinian team; then she paraphrases the wrong answer to her partner (*PARA*) and tells him that this is not her fault (*SPONT*). The next system answer is correct, and she reads it aloud from the screen (*READ*). In the German example, *Off-Talk* is again labelled adjacent to the pertinent word, without blanks. The English translation contains only the wording; here, *PARA* is given boldface and in italic, *READ* in capitals, and *SPONT* boldface and regular.

user: *wann ist das n"achste Spiel der argentinischen Mannschaft*
user: *nein <"ahm> die<PARA> haben<PARA> mich<PARA> jetzt<PARA> nur<PARA> dar"uber<PARA> informiert<PARA> wo<PARA> der<PARA> n"achste<PARA> Taxistand<PARA> ist<PARA> und<PARA> nicht<PARA> ja<SPONT> ja<SPONT> ich<SPONT> kann<SPONT> auch<SPONT> nichts<SPONT> daf"ur<SPONT>*
user: *bis wann fahren denn nachts die "offentlichen Verkehrsmittel*
user: *die<READ> regul"aren<READ> Linien<READ> fahren<READ> bis<READ> zwei<READ> und<READ> danach<READ> verkehren<READ> Nachtlinien<READ>*

[5] The realistic and sub-optimal light conditions in our scenario prevent us from using gaze direction as feature. We assume that de-synchronized gaze and head orientation do not occur too often; this is somehow corroborated by the good performance of the classification for multi-modal modeling described below.

[6] designed and recorded at the Institute of Phonetics and Speech Communication, Ludwig-Maximilians-University, Munich, cf. [6].

[7] designed and recorded at our Institute.

Table 2 *100 prosodic and 30 part-of-speech (POS) features and their context. Prosody is based on duration (Dur), energy (En), pitch (F0), pauses, jitter, and shimmer. POS categories are API (adjectives and participles, inflected), APN (adjectives and participles, not inflected), AUX (auxiliaries), NOUN (nouns, proper nouns), PAJ (particles, articles, and interjections), and VERB (verbs).*

word-based features for the actual word '0' and for two words to the left and right	**context size** -2	-1	0	1	2
100 prosodic features:					
DurTauLoc; EnTauLoc; F0MeanGlob; RateOfSpeech			•		
Dur: Norm,Abs,AbsSyl		•	•	•	
En: RegCoeff,MseReg,Norm,Abs,Mean,Max,MaxPos		•	•	•	
F0: RegCoeff,MseReg,Mean,Max,MaxPos,Min,MinPos		•	•	•	
Pause-before, PauseFill-before; F0: Off,Offpos		•	•		
Pause-after, PauseFill-after; F0: On,Onpos			•	•	
Dur: Norm,Abs,AbsSyl	•			•	
En: RegCoeff,MseReg,Norm,Abs,Mean	•			•	
F0: RegCoeff,MseReg	•			•	
F0: RegCoeff,MseReg; En: RegCoeff,MseReg; Dur: Norm		•			
Jitter: Mean, Sigma; Shimmer: Mean, Sigma			•		
30 POS-features:					
API,APN,AUX,NOUN,PAJ,VERB	•	•	•	•	•

user: *When is the next play of the Argentinian team?*
user: *no uhm* ***they only told me where the next taxi stand is*** **and not – well ok – it's not my fault**
user: *Until which time is the public transport running?*
user: THE REGULAR LINES ARE RUNNING UNTIL 2 AM AND THEN, NIGHT LINES ARE RUNNING.

5 Speech Features

The most plausible domain for *On-Talk* vs. *Off-Talk* is a unit between the word and the utterance level, such as clauses or phrases. In this section, we confine our analysis to the word level to be able to map words onto the most appropriate syntactic/semantic units later on. We do not use any deep syntactic and semantic procedures, but only prosodic information and a rather shallow analysis with (sequences of) word classes, i.e. part-of-speech (POS) information. A more in-depth linguistic modelling might provide more information; however, POS modelling is more robust because it is less dependent on the specific scenario. The spoken word sequence which is obtained from the speech recogniser is only required for the time alignment and for a normalisation of energy and duration based on the underlying phonemes. In this paper, we use the transliteration

(i.e. the orthographic transcription) of the data assuming a recogniser with 100 % accuracy.

It is still an open question which prosodic features are relevant for different classification problems, and how the different features are interrelated. We try therefore to be as exhaustive as possible, and we use a highly redundant feature set leaving it to the statistical classifier to find out the relevant features and the optimal weighting of them. For the computation of the prosodic features, a fixed reference point has to be chosen. We decided in favour of the end of a word because the word is a well–defined unit in word recognition, and because this point can be more easily defined than, for example, the middle of the syllable nucleus in word accent position. Many relevant prosodic features are extracted from different context windows with the size of two words before, i.e. contexts -2 and -1, and two words after, i.e. contexts 1 and 2, around the current word, namely context 0 in Table 2; by that, we use so to speak a 'prosodic 5-gram'. A full account of the strategy for the feature selection is beyond the scope of this paper; details and further references are given in [2].

Table 2 shows the 100 prosodic features and their context. The six POS features with their context sum up to 30 features. DurTauLoc is a local estimate of a global duration factor DurTau (which is speaker dependent and proportional to the reciprocal of the rate-of-speech), EnTauLoc is a local estimate of the global energy EneTau (average energy in the recordings of a speaker), and F0MeanGlob is the average fundamental frequency [2]. These features as well as the global tempo feature RateOfSpeech are estimated from a window of 15 words (or less, if the utterance is shorter); thus they are identical for each word in the context of five words, and only context 0 is necessary.

Note that these 130 features do not necessarily represent *the* optimal feature set; this could only be obtained by reducing a much larger set to those features which prove to be relevant for the actual task, but in our experience, the effort needed to find the optimal set normally does not pay off in terms of classification performance, cf. [3, 4]. A detailed overview of prosodic features is given in [5]; formulas and further references can be found in [2]. The abbreviations of the features can be explained as follows:

duration features 'Dur': absolute (Abs) and normalised (Norm); this normalisation is described in [2] and is based on duration statistics and on DurTauLoc; absolute duration divided by number of syllables AbsSyl represents another sort of normalisation;

energy features 'En': regression coefficient (RegCoeff) with its mean square error (MseReg); mean (Mean), maximum (Max) with its position on the time axis (MaxPos), absolute (Abs) and normalised (Norm) values; the normalisation is described in [2] and is based on energy statistics and on EnTauLoc; absolute energy divided by number of syllables AbsSyl represents another sort of normalisation;

F0 features 'F0': regression coefficient (RegCoeff) with its mean square error (MseReg); mean (Mean), maximum (Max), minimum (Min), onset (On), and offset (Off) values as well as the position of Max (MaxPos), Min (MinPos), On (OnPos), and Off (OffPos) on the time axis;[8] all F0 features are logarithmised and normalised as to the mean value F0MeanGlob;

length of pauses 'Pause': silent pause before (Pause-before) and after (Pause-after), and filled pause before (PauseFill-before) and after (PauseFill-after);

[8] Note that these position features are measured in msec.; strictly speaking, they are therefore rather duration features.

Ms., in: Journal on Multimodal User Interfaces, in press, available online at:
http://www.springerlink.com/content/121040/?Content+Status=Accepted

Table 3 *Three databases, words per category in %: ONTALK, READ, PARA, SPONT and OTHER*

	# Speakers	*ONTALK*	*READ*	*PARA*	*SPONT*	*OTHER*	[%]
SW_{spont}	28	48.8	13.1	21.0	17.1	-	
SW_{acted}	17	33.3	23.7	-	-	43.0	
SK_{spont}	92	93.9	1.8	-	-	4.3	

jitter, shimmer: global mean and sigma for micro-perturbations of F0 (jitter) and intensity (shimmer); calculated from all words of an utterance.

A *Part of Speech (POS)* flag is assigned to each word in the lexicon, cf. [8]. Six main 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). For the context of +/- two words, this sums up to 6x5, i.e., 30 binary POS features, cf. the last line in Table 2.

6 Comparing SmartKom with SmartWeb, Speech Only, Word-based

6.1 Databases

From the SmartKom (**SK_{spont}**) database, we use 4 hrs. of speech (20669 word tokens) from 92 speakers. The vocabulary contains 1800 types. Since the subjects were alone, no *PARA* occurred: *OTHER* is basically 'talking to oneself', cf. [9], but contains also spontaneous *Off-Talk*; *SPONT* is not annotated. The proportion of *Off-Talk* is small (Table 3). The 16kHz data from a directional microphone was downsampled to 8kHz for the experiments reported on in the following.

All SmartWeb data has been recorded with a close-talk microphone and 8 kHz sampling rate. The setting of **SW_{spont}** has been described above. For the experiments reported on in this section, data of 28 from 100 speakers (this part contains 0.8 hrs. of speech) have been used. The complete corpus with 3.2 hrs. of speech will be analysed in Section 7. The corpus has been annotated with *ONTALK* (default), *READ*, *PARA*, *SPONT* and *OTHER*. *OTHER* has been mapped onto *SPONT* later on. The 0.8 hrs. of speech consist of 5211 word tokens (750 different types); the distribution of *On-/Off-Talk* is given in Table 3.

We additionally recorded acted data (**SW_{acted}**, 1.7 hrs.) to find out which classification rates can be achieved and to show the differences to realistic data. The content of the acted data is based on the SmartWeb scenario described in Sect. 4.2. All queries to a fictive system, and spoken phrases to a fictive dialogue partner were pre-formulated sentences, together with detailed instructions on how to pronounce *On-Talk* and *Off-Talk*. Here, only the two *Off-Talk* classes *READ* and *OTHER* are discriminated, as in SK_{spont}. The corpus has been recorded after SK_{spont}, but before SW_{spont} and has been used for an initial SmartWeb demonstrator. The instructions, how to pronounce *On-/Off-Talk*, were based on observations from SK_{spont}. This way a corpus with similar properties as in SK_{spont} was produced, but with similar content as SW_{spont} – and with much more *Off-Talk* data. It was expected that this data would result in a classifier which clearly separates classes. In the ideal case, this classifier would also result in good classification rates for the spontaneous corpora SK_{spont} and SW_{spont}; when discussing Table 5, we will see that this aim has been achieved to a large extent.

Ms., in: Journal on Multimodal User Interfaces, in press, available online at:
http://www.springerlink.com/content/121040/?Content+Status=Accepted

The observations from SK_{spont} were the following: *Off-Talk* is produced with lower voice and durations are longer for *READ*. We further expect that in SmartWeb nobody using a head-set to address the automatic dialogue system would intentionally confuse the system with loud *Off-Talk*. These considerations result in the following setup for SW_{acted}: the 17 speakers sat in front of a computer. All *Off-Talk* had to be produced with lower voice and, additionally, *READ* had to be produced more slowly. Furthermore, each pre-formulated sentence could be read in advance so that some kind of '(pseudo-)spontaneous' production was possible, whereas the *READ* sentences were indeed read utterances. The vocabulary contains 361 different word types. 2321 words are *ONTALK*, 1651 *READ*, 2994 *OTHER* (Table 3).

Table 4 *LDA classification results with prosodic features and POS features; leave-one-speaker-out, class-wise averaged recognition rate for ONTALK vs. Off-Talk (CL-2), ONTALK, READ, OTHER (CL-3) and ONTALK, READ, PARA, SPONT (CL-4)*

	features	CL-2	CL-3	CL-4
SK_{spont}	100 pros.	72.7	60.0	-
SK_{spont}	100 pros. speaker norm.	74.2	61.5	-
SK_{spont}	30 POS	58.9	60.1	-
SK_{spont}	100 pros. + 30 POS	74.1	66.0	-
SW_{spont}	100 pros.	65.3	55.2	48.6
SW_{spont}	100 pros. speaker norm.	66.8	56.4	49.8
SW_{spont}	30 POS	61.6	51.6	46.9
SW_{spont}	100 pros. + 30 POS	68.1	60.0	53.0
SW_{acted}	100 pros.	80.8	83.9	-
SW_{acted}	100 pros. speaker norm.	92.6	92.9	-

6.2 Classification for Speech only

For automatic classification we employed a linear classifier which separates the classes (clusters in feature-space) using linear boundaries (e.g. plains in 3-dimensional feature space). We employed a Linear Discriminant Classifier (LDC) for all constellations: a linear combination of the independent variables (the predictors) is formed; a case is classified, based on its discriminant score, in the group for which the posterior probability is largest, cf. [19]. Validation was done with leave-one-speaker-out (*loso*), i.e., in turn one speaker was used for testing and all other speakers for training; this guarantees speaker-independence. All results are measured with the class-wise averaged recognition rate CL-N (N = 2, 3, 4) to guarantee robust recognition of all N classes, also for classes with small a priori probability. CL-N is the unweighted average recall,[9] i.e. for instance, CL-2 is the mean of sensitivity and specifity. In the 2-class task we classify *ONTALK* vs. rest; for $N = 3$ classes we discriminate *ONTALK*, *READ* and *OTHER* ($= SPONT \cup PARA$); the $N = 4$ classes *ONTALK*, *READ*, *SPONT*, *PARA* are only available in SW_{spont}.

In Table 4 results for the different databases are compared. Classification is performed with different feature sets: 100 prosodic features, 30 POS features, or all 130 features. For SW_{acted} POS-features are not evaluated, since all sentences that had to be

[9] The *recall* of a class is the percentage of correctly classified elements given this class.

produced were given in advance; for such a non-spontaneous database, POS evaluation would only measure the design of the database rather than the correlation of different *Off-Talk* classes with the 'real' frequency of POS categories. For the prosodic features, results are additionally given after speaker normalisation ('speaker norm.': zero-mean and variance 1 for each feature component). Here, we assume that mean and variance (no matter whether it is *On-Talk* or not) of all the speaker's prosodic feature vectors are known in advance. This is an upper bound for the results that can be reached with adaptation.

As could be expected, best results for prosodic features are obtained for the acted data: 80.8 % CL-2 and even higher recognition rates for three classes[10], whereas chance would be only 33.3 % for CL-3. Rates are higher for $\mathrm{SK_{spont}}$ than for $\mathrm{SW_{spont}}$ (72.7 % vs. 65.3 % CL-2, 60.0 % vs. 55.2 % CL-3).[11] For all databases results could be improved when the 100-dimensional feature vectors are normalised per speaker. The results for $\mathrm{SW_{acted}}$ rise drastically to 92.6 % CL-3; for the other corpora a smaller increase can be observed. The evaluation of 30 POS features shows about 60 % CL-2 for both spontaneous databases; for three classes lower rates are achieved for $\mathrm{SW_{spont}}$. Here, in particular the recall of *READ* is significantly higher for $\mathrm{SK_{spont}}$ (78 % vs. 57 %). In all cases a significant increase of recognition rates is obtained when linguistic and prosodic information is combined, e.g. on $\mathrm{SW_{spont}}$ three classes are classified with 60.0 % CL-3, whereas with only prosodic or only POS features 55.2 % resp. 51.6 % CL-3 are obtained. For $\mathrm{SW_{spont}}$, 4 classes could be discriminated with up to 53.0 % CL-4. Here, *PARA* is the problematic category that is very close to all other classes (39 % recall only). [12]

Table 5 *Cross validation of the three corpora with speaker-normalised prosodic features. Diagonal elements are results for Train=Test (leave-one-speaker-out in brackets). All classification rates in % CL-2*

		Test		
		SW_{acted}	SW_{spont}	SK_{spont}
Training	SW_{acted}	93.4 (92.6)	63.4	61.9
	SW_{spont}	85.2	69.3 (66.8)	67.8
	SK_{spont}	74.0	61.1	76.9 (74.2)

To compare the different prosodic information used in the different corpora and the differences in acted and spontaneous speech, we use cross validation as shown in Table 5. Such a cross-validation is a convenient way of finding out whether different databases are similar or not w.r.t. the features used: if classification performance breaks down when using different databases for training and testing, this is a proof for marked differences. The diagonal elements show the *Train=Test* case, and in brackets the *loso*

[10] For the CL-2 evaluation, a classifier with 2 classes A and B is trained, for the CL-3 evaluation a classifier with 3 classes A, B1, and B2 (with B = B1 ∪ B2). The results (in both cases: CL = average recall) for CL-3 can be higher when using a linear classifier, in particular in the extreme case, where B1 and B2 lie on opposite sides of A.

[11] The reason for this is most likely that in SmartKom, the users were alone with the system; thus *Off-Talk* was always talking to one-self – no need to be understood by a third partner. In SmartWeb, however, a third partner was present, and moreover, the signal-to-noise ratio was less favourable than in the case of SmartKom.

[12] All results are 'highly significant' since they are based on a large set of samples by using leave-one-speaker-out evaluation (20669 words in the case of $\mathrm{SW_{spont}}$). Using the Z-test for a proportion, an improvement of 1 percentage point (2 points in the case of $\mathrm{SW_{spont}}$ and $\mathrm{SW_{acted}}$) is significant at the 0.001 level.

Table 6 SK_{spont}: *Best single features for ONTALK vs. OTHER (left) and ONTALK vs. READ (right). The dominant feature group is emphasised. "•" denotes that the resp. feature values are greater for the class given in this column*

SK_{spont}	ON TALK	OTH ER	CL-2 [%]
EnMax	•		72
EnMean	•		69
JitterMean	•		69
JitterSigma	•		69
F0Max	•		69
ShimmerSigma	•		68
ShimmerMean	•		68
F0OnPos		•	67
EnAbs	•		66
EnNorm	•		61

SK_{spont}	ON TALK	RE AD	CL-2 [%]
JitterMean	•		62
DurAbs		•	61
DurTauLoc	•		61
*F0Max**Pos***		•	61
EnTauLoc	•		69
*F0Min**Pos***		•	59
JitterSigma	•		59
EnMean	•		59
EnMax	•		58
F0Max	•		58

Table 7 SW_{spont}: *Best single features for ONTALK vs. OTHER (left) and ONTALK vs. READ (right). The dominant feature group is emphasised. "•" denotes that the resp. feature values are greater for the class given in this column*

SW_{spont}	ONT TALK	OTH ER	CL-2 [%]
EnMax	•		61
EnTauLoc	•		60
EnMean	•		60
PauseFill-before		•	54
JitterSigma	•		54
EnAbs	•		54
F0Max	•		53
ShimmerSigma	•		53
JitterMean	•		53
Pause-before		•	53

SW_{spont}	ON TALK	RE AD	CL-2 [%]
EnTauLoc	•		60
DurAbs		•	58
*F0Max**Pos***		•	58
*F0On**Pos***	•		57
DurTauLoc	•		57
*EnMax**Pos***		•	56
EnMean	•		56
EnAbs		•	56
*F0Off**Pos***	•		55
*F0Min**Pos***		•	53

result from Table 4 (speaker norm.). The maximum we can reach on SW_{spont} is 69.3 %, whereas with *loso*-evaluation 66.8 % are achieved; if we train with acted data and evaluate with SW_{spont}, the drop is surprisingly small: we still reach 63.4 % CL-2. The other way round 85.2 % on SW_{acted} are obtained, if we train with SW_{spont}. This shows that both SmartWeb corpora are in some way similar; obviously, our instructions and the strategies chosen by our speakers while acting really mirrors spontaneous *Off-Talk* up to a large extent. The database most related to SK_{spont} is the other spontaneous database SW_{spont}. As expected, results for spontaneous data were worse than for acted data. However, if we train with SW_{acted} and test with SW_{spont} and vice versa, the drop is just small. Thus, there is hope that for real applications, the training set can be enhanced with acted *Off-Talk* data. For a rough estimation of *On-/Off-Talk*, the collection of acted data may even be sufficient and first of all significantly cheaper to produce.

6.3 Interpretation

Now we want to analyse single prosodic features by training 1-dimensional classifiers; this also reveals similarities in the different databases. We restrict ourselves to the two realistic corpora SK_{spont} and SW_{spont}, and refer to [7] for SW_{acted}. To discriminate

ONTALK from *OTHER*, all *READ* words were deleted; for *ONTALK* vs. *READ*, *OTHER* is deleted. A ranking of the best features (best classifiers based on only one feature) can be found in Table 6 and 7 for SK_{spont} and SW_{spont}. Most relevant features to discriminate *ONTALK* from *OTHER* (left column in Table 6, 7) are the higher energy values for *ONTALK* in both scenarios. Highest CL-2 is achieved for SK_{spont}, since the user was alone and *OTHER* is basically talking to oneself and consequently with extremely low voice. Jitter and shimmer are also important, in particular for SK_{spont}. The range of F0 (higher F0Max values) is larger for *ONTALK* which might be caused by an exaggerated intonation when talking to computers. For the SW_{spont} data — most probably due to hesitation phenomena — pauses are significant (longer pauses for *OTHER*). In SK_{spont} global features like EnTauLoc that are determined from a large context are not relevant, because in many cases only one word per turn is *Off-Talk* (swearwords).

To discriminate *ONTALK* from *READ*, (right columns in Tables 6, 7) duration features are highly important: the duration of read words (mostly content words, cf. Tables 8 and 9) is longer. Further duration features are the position of the maximum or onset of the fundamental frequency (reference point is here the end of the word). Again, energy is very significant (higher for *ONTALK* — Computer Talk is louder).

To distinguish *READ* vs. *OTHER* (not shown in the tables), the longer duration of *READ* is significant as well as the wider F0-range. *READ* shows also higher energy values in SW_{spont}.

Table 8 SK_{spont}*: POS classes, percent occurrences for ONTALK, READ, OTHER, and over all 20669 tokens. POS categories are API (adjectives and participles, inflected), APN (adjectives and participles, not inflected), AUX (auxiliaries), NOUN (nouns, proper nouns), PAJ (particles, articles, and interjections), and VERB (verbs).*

POS	# of tokens	NOUN	API	APN	VERB	AUX	PAJ
ONTALK	19415	18.1	2.2	6.6	9.6	8.4	55.1
READ	365	56.2	7.1	18.1	2.2	2.2	14.2
OTHER	889	7.2	2.6	10.7	8.9	6.7	63.9
total	20669	18.3	2.3	7.0	9.4	8.2	54.7

Table 9 SW_{spont}*: POS classes, percent occurrences for ONTALK, READ, PARA, SPONT, and over all 5211 tokens (subset of 28 speakers). POS categories are API (adjectives and participles, inflected), APN (adjectives and participles, not inflected), AUX (auxiliaries), NOUN (nouns, proper nouns), PAJ (particles, articles, and interjections), and VERB (verbs).*

POS	# of tokens	NOUN	API	APN	VERB	AUX	PAJ
ONTALK	2541	23.2	5.1	3.8	6.9	8.5	52.5
READ	684	27.2	5.7	18.6	7.4	7.6	33.5
PARA	1093	26.3	5.1	10.3	5.4	9.5	43.3
SPONT	893	8.1	1.5	5.7	11.5	10.3	62.9
total	5211	21.8	4.6	7.4	7.5	8.9	49.8

Ms., in: Journal on Multimodal User Interfaces, in press, available online at:
http://www.springerlink.com/content/121040/?Content+Status=Accepted

The most important difference between *READ* and *OTHER* is not a prosodic, but a lexical one. This can be illustrated nicely by Tables 8 and 9 where percent occurrences of POS is given for the three classes *ONTALK*, *READ*, and *OTHER* (SK_{spont}) and for the four classes *ONTALK*, *READ*, *PARA*, and *SPONT* (SW_{spont}). Especially for SK_{spont}, there are more nouns and adjectives (content words) in *READ* than in *OTHER* and *ONTALK*, especially NOUNs: 56.2% compared to 7.2% in *OTHER* and 18.1% in *ONTALK*. It is the other way round, if we look at the function words, cf. the PAJ column (particles, articles, and interjections): very few for *READ* (14.2%), and most for *OTHER* (63.9%); VERB and AUX display the same tendencies, albeit less pronounced. The explanation is straightforward: the user only reads words that are presented on the screen, and these are mostly content words – names of restaurants, cinemas, etc., which of course are longer than other word classes. For SW_{spont}, there is the same tendency but less pronounced. *PARA* contains many content words like *READ* but at the same time much more PAJ are observed.

Summing up, the following results have been discussed in this section: a very high classification rate of 92.6 % CL-2 has been obtained for acted data, whereas the same linear classifier results in only 61.9 and 63.4 % CL-2 on spontaneous data from the SmartKom and SmartWeb project. With spontaneous training data from SmartKom, up to 74 % CL-2 are reached on the SmartKom test data. The classification is worse for SmartWeb (68.1 % CL-2) since the users of the system were not alone, and the contrast between *On-Talk* and *Off-Talk* – in particular in terms of loudness or energy – is smaller. Energy is important to discriminate *ONTALK* from *OTHER*, duration is important to discriminate *ONTALK* from *READ*. The biggest difference between *READ* and *OTHER* are the POS-categories of the spoken words.

7 Utterance-based Fusion of Speech and Head Orientation in SmartWeb

7.1 Annotation

In the following, the fusion of the two modalities video and audio for the complete SmartWeb corpus SW_{spont} (3.2 hrs.) is analysed on the *utterance* or dialogue turn level. Using the word-based labels for the SmartWeb data, utterance labels are calculated from the word level by a majority voting described in [22], yielding 2068 utterances (on average 10.8 words per utterance). The distribution of the labels per word and per utterance is shown in Table 10; there is no marked difference.

Table 10 *Portion of labels for ONTALK, READ, PARA, and SPONT*

	% ONTALK	% READ	% *PARA*	% *SPONT*
word	47.2	12.2	17.3	23.3
utterance	49.6	13.3	11.1	26.0

The manual annotation of the video recordings includes frame based labelling (7.5 frames per sec.) of the 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

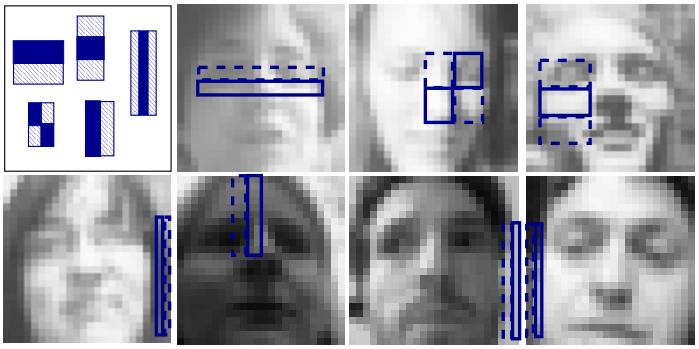

Fig. 1 The 7 best out of 452 features used by the SmartWeb face detector. Top left: different shapes of Haar-Wavelets

rectangle[13] to train the face detector described in 7.2. *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, operationalised as binary contrast using face detection: *On-View* vs. *Off-View*.

7.2 Detection of Head Orientation via Face Detection

In addition to the prosodic and POS features described above, we used features modelling head orientation. 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 [30]. The face detection works for single images; no use of context information is implemented. The algorithm is based on five Haar-like wavelets shown in Fig. 1, 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 (the 5 wavelets with arbitrary scaling and translation), the AdaBoost algorithm selects those wavelets containing complementary information; a hierarchical classifier speeds up the classification. In this paper we use 176 × 144 grey-scale images, 7.5 per second; faces are searched in different sub-images, greater than half the image, and scaled to 24 × 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[14]. The resulting face detector is based on 452 Haar-features; the seven best are shown in Fig. 1 with random images (24 × 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 [22]) 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 %.

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

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

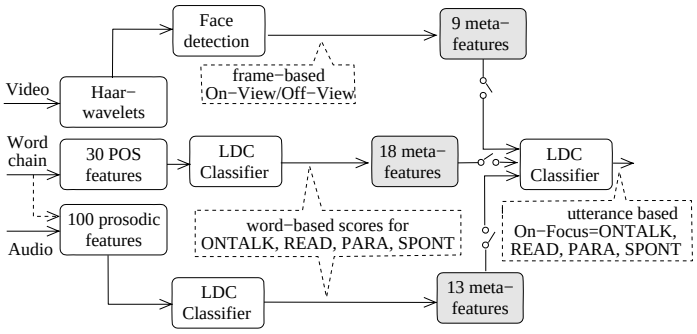

Fig. 2 Utterance classification with meta-features

7.3 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 *ONTALK*, *READ*, *PARA*, and *SPONT* which have been manually annotated (Tab. 10). In the case of multi-modal classification, we refer to *ONTALK* as *On-Focus*; *Off-Focus* is subdivided into *READ*, *PARA* and *SPONT*.

For the fusion of modalities, we do not want to use a set of thresholds or rules but want a classifier ('combiner') to learn those decisions from the training data. In general, there are two possible approaches: early fusion combines the modalities on the feature level, late fusion combines the decisions of unimodal classifiers. Early fusion is not possible in our case, since the face detector works image based[15] whereas the classifiers for prosody and POS are word based. For this paper, an approach towards decision fusion was developed which is based on meta-features; it combines the two steps *mapping onto the utterance level* and *fusion* as illustrated in Fig. 2. The meta-features are fed with detailed information from the word-level and image-level, and combine this information to a weak utterance-level decision (not a hard decision for one class, but several scores for all classes[16]) which serves as input to a linear classifier (LDC, cf. Chap. 6.2).

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 *ONTALK*, *READ*, *PARA*, and *SPONT* averaged over the whole turn. Further, the variation of each score is described with its maximum and minimum. Similar utterance-based 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. 9 — is calculated per utterance. Together with the average, minimum and maximum linguistic word length (# graphemes), 18 linguistic meta-features are obtained. These meta-features represent, so to speak, a condensed version

[15] For *On-/Off-View*, an image based classification makes more sense than for instance analysing an image averaged over all frames of a word, and is additionally quite efficient using the Viola-Jones algorithm.

[16] If – given a 2-class problem – a classifier decides for 51 % for class 1, we would probably falsify the overall result after fusion by using a hard decision (100 % class 1). Instead we use – in the case of prosodic and POS classification – 4 scores for *ONTALK*, *READ*, *PARA*, and *SPONT* as input to the fusion step.

of the prosodic and the lexical/semantic characteristics of the four classes by describing the word-level decisions within one utterance.

From the frame based classification of *On-/Off-View*, nine further utterance-based meta-features are calculated[17]: 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 for instance 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 recognised only in every ith frame. These meta-features describe the image-level decisions of a face detector.

The utterance classification using LDC as 'combiner' is performed with combinations of 13, 18, or 9 meta-features (prosodic, linguistic, and video).

7.4 Experimental Results: Fusion of Speech and Head Orientation

For the experiments, the data was divided into a training set and a test set. They comprise 58 vs. 37 speakers[18], 1130 vs. 748 utterances, and 13800 vs. 8400 words. All results are described with the class-wise averaged recognition rate CL-N (N = 2, 4), as described in Sect. 6.2.

Table 11 *Confusion matrices using prosodic features (left) and head orientation (right); % classified correctly*

	prosodic features				**head orientation**			
	ONTALK	*READ*	*PARA*	*SPONT*	*ONTALK*	*READ*	*PARA*	*SPONT*
ONTALK	**64.8**	6.4	11.3	17.5	**69.7**	8.0	8.2	14.1
READ	17.1	**62.2**	8.1	12.6	55.0	**12.6**	18.9	13.5
PARA	18.4	10.3	**51.7**	19.5	12.6	4.6	**67.8**	15.0
SPONT	8.7	4.3	16.1	**70.8**	18.6	8.7	42.3	**30.4**

The confusion matrices of the LDC resulting from separate evaluations of each modality are shown in Tab. 11 (left: for prosodic features, right: for features based on face detection), and in Tab. 12 (left: using POS information). Obviously, it is difficult to detect *PARA* using the audio channel or just the word chain; using the video-channel, a recall of 67.8 % is obtained for *PARA* which correlates with *Off-View*. However, using solely video (Tab. 11, right) shows that the detection of *READ* nearly always fails, and also the results for *SPONT* are only little better than chance: it cannot be classified without using prosodic or linguistic information.

In Tab. 13 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 (*ONTALK*, *READ*, *PARA*, *SPONT*); note that chance level for the 2-class problem is 50%, for the 4-class problem 25%. 'Pros. norm.' stands for speaker normalised features (zero mean and variance 1) as described in Sect. 6.2. This way, for the 2-class problem, the classification with prosodic features rises from 68.6 to 76.6 %

[17] slightly different values in comparison to [22] due to small changes of the alignment

[18] 4 of the 99 speakers were not used due to technical problems

Table 12 *Confusion matrix using POS features (left) and a combination of 3 feature types (right): prosody (speaker normalised), POS, and video; % classified correctly*

	POS features				**fusion**			
	ONTALK	*READ*	*PARA*	*SPONT*	*ONTALK*	*READ*	*PARA*	*SPONT*
ONTALK	**62.5**	3.6	13.6	20.3	**79.7**	4.1	3.6	12.6
READ	3.6	**67.6**	18.0	10.8	9.9	**73.0**	9.0	8.1
PARA	23.0	8.0	**50.6**	18.4	9.2	8.0	**64.4**	18.4
SPONT	21.2	2.5	13.0	**63.3**	8.7	3.7	15.5	**72.1**

CL-2. With linguistic information (no adaptation required), 76.0 % CL-2 are achieved, and with video information 70.5 %. Combining any two modalities, the classification rate rises up to 80.8 % CL-2. Using all 3 modalities, 84.5 % CL-2 are obtained. Four classes are discriminated with 72.3 % CL-4, no matter whether speaker normalisation is applied or not. The confusion matrix of the best constellation for the 4-class problem ('Pros. norm.', second last line in Tab. 13) is shown in Tab. 12, right. There is still some confusion between *PARA* and *SPONT*.

The experiments listed in Tab. 13 have shown that for multi-modal fusion, speaker normalisation (an approach that assumes that all the speaker's speech has been seen in advance) is not really necessary: for the 2-class problem, it is only .7 percent points better, and there is no difference at all for the 4-class problem. However, speaker normalisation or adaptation are still beneficial, if the underlying speech recogniser has a low word accuracy, e.g. in a noisy environment: up to now, all investigations are based on the assumption that a speech recogniser is available which has a word recognition accuracy close to 100 %. However, all results are also valid if the speech recogniser has a lower albeit more realistic word accuracy of only 70 %. In this case, 82 % CL-2 are achieved for the discrimination of *On-Talk* vs. *Off-Talk*; if the word accuracy drops to 20 % our system still reaches 72 % CL-2.

Table 13 *Classification of On-Focus vs. Off-Focus and On-Focus vs. READ vs. PARA vs. SPONT using prosodic features, speaker normalised prosodic features, POS features, and face detection*

Pros.	Pros. norm.	POS	Video	CL-2 in % 2-class case	CL-4 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

In future applications, further improvements could be possible by utilising additional information in the meta-classification step. This information could be the dialogue state in a system like the one described by [14]. *READ* is for instance more likely

to occur if complex information is shown on the display. Such a strategy mirrors the use of top-down knowledge and expectations in human-human interactions.

8 Concluding Remarks

Off-Talk is certainly a phenomenon whose successful treatment is getting more and more important, if the performance of automatic dialogue systems allows unrestricted speech, and if the tasks performed by such systems approximate those tasks that are performed within our experiments. We have seen that a prosodic classification, based on a large feature vector, yields good but not excellent classification rates. With additional lexical information encoded in the POS features, classification rates went up. Best is multi-modal classification, additionally taking into account video information.

Classification performance as well as the unique phonetic traits discussed in this paper will very much depend on the types of *Off-Talk* that can be found in specific scenarios; for instance, in an extremely noisy environment, talking aside to someone else might display the same amount of energy as addressing the system, simply because of an unfavourable signal-to-noise ratio. Under somehow favourable conditions, it might be possible not only to tell apart *On-Talk* from *Off-Talk* but also to differentiate types of *Off-Talk* with a reliable performance: for instance, *READ* tells the system that the user is concentrating on the interaction with the system, while a high percentage of *SPONT* might tell the system that at least for the moment, other topics might be more interesting for the user.

We have seen that on the one hand, Computer Talk (i.e. *On-Talk*) in fact is similar to talking to someone who is hard of hearing: its phonetics is more pronounced, energy is higher, etc. However we have to keep in mind that this register will most likely depend to some – even high – degree on other factors such as overall system performance: the better the system performance turns out to be, the more 'natural' the Computer-Talk of users will be, and this means in turn that the differences between *On-Talk* and *Off-Talk* will possibly be less pronounced.

The phenomena that we addressed in this paper can be suppressed in dyadic human-machine interaction if some pre-cautions 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, cf. [9,7]; 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. The sparse data problem could be solved by using the recording technique from [21] described in more detail in [6] 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 unfavourable recording conditions: acoustic noise in the case of speech, and 'video noise' such as back-light, reduced brightness and so on. 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

Ms., in: Journal on Multimodal User Interfaces, in press, available online at:
http://www.springerlink.com/content/121040/?Content+Status=Accepted

the same direction' — *READ* can trivially not be recognised with video information because the user has to face the system while reading, and *PARA* is poorly recognised by using only audio information — a fusion of both channels and all three feature types yielded markedly better results than a uni-modal modelling.

References

1. Alexandersson, J., Buschbeck-Wolf, B., Fujinami, T., Kipp, M., Koch, S., Maier, E., Reithinger, N., Schmitz, B., Siegel, M.: Dialogue Acts in VERBMOBIL-2 – Second Edition. Verbmobil Report 226 (1998)
2. Batliner, A., Buckow, A., Niemann, H., Nöth, E., Warnke, V.: The Prosody Module. In: W. Wahlster (ed.) Verbmobil: Foundations of Speech-to-Speech Translations, pp. 106–121. Springer, Berlin (2000)
3. Batliner, A., Buckow, J., Huber, R., Warnke, V., Nöth, E., Niemann, H.: Prosodic Feature Evaluation: Brute Force or Well Designed? In: Proc. ICPHS, pp. 2315–2318. San Francisco (1999)
4. Batliner, A., Buckow, J., Huber, R., Warnke, V., Nöth, E., Niemann, H.: Boiling down Prosody for the Classification of Boundaries and Accents in German and English. In: Proc. Eurospeech, pp. 2781–2784. Aalborg (2001)
5. Batliner, A., Fischer, K., Huber, R., Spilker, J., Nöth, E.: How to Find Trouble in Communication. Speech Communication **40**, 117–143 (2003)
6. Batliner, A., Hacker, C., Kaiser, M., Mögele, H., Nöth, E.: Taking into Account the User's Focus of Attention with the Help of Audio-Visual Information: Towards less Artificial Human-Machine-Communication. In: E. Krahmer, M. Swerts, J. Vroomen (eds.) Proceedings of AVSP 2007 (International Conference on Auditory-Visual Speech Processing, pp. 51–56. Hilvarenbeek (2007)
7. Batliner, A., Hacker, C., Nöth, E.: To Talk or not to Talk with a Computer: On-Talk vs. Off-Talk. In: K. Fischer (ed.) How People Talk to Computers, Robots, and Other Artificial Communication Partners, University of Bremen, SFB/TR 8 Report, pp. 79–100 (2006)
8. Batliner, A., Nutt, M., Warnke, V., Nöth, E., Buckow, J., Huber, R., Niemann, H.: Automatic Annotation and Classification of Phrase Accents in Spontaneous Speech. In: Proc. Eurospeech, pp. 519–522. Budapest (1999)
9. Batliner, A., Zeissler, V., Nöth, E., Niemann, H.: Prosodic Classification of Offtalk: First Experiments. In: Proceedings of the 5th TSD, pp. 357–364. Springer, Berlin (2002)
10. Berk, L.E.: Children's private speech: An overview of theory and the status of research. In: R.M. Diaz, L.E. Berk (eds.) Private speech. From social interaction to self-regulation, pp. 17–53. Erlbaum, Hillsdale, NJ (1992)
11. Carletta, J., Dahlbäck, N., Reithinger, N., Walker, M.: Standards for Dialogue Coding in Natural Language Processing. Dagstuhl-Seminar-Report 167 (1997)
12. Fischer, K.: What Computer Talk Is and Isn't: Human-Computer Conversation as Intercultural Communication, *Linguistics - Computational Linguistics*, vol. 17. AQ, Saarbrücken (2006)
13. Fraser, N., Gilbert, G.: Simulating Speech Systems. CSL **5**(1), 81–99 (1991)
14. Goronzy, S., Mochales, R., Beringer, N.: Developing Speech Dialogs for Multimodal HMIs Using Finite State Machines. In: Proc. ICSLP, pp. 1774–1777. Pittsburgh (2006)
15. Heylen, D.: Challenges Ahead. Head Movements and other social acts in conversation. In: Proceedings of AISB - Social Presence Cues for Virtual Humanoids, pp. 45–52. Hatfield, UK (2005)
16. Hönig, F., Hacker, C., Warnke, V., Nöth, E., Hornegger, J., Kornhuber, J.: Developing Enabling Technologies for Ambient Assisted Living: Natural Language Interfaces, Automatic Focus Detection and User State Recognition. In: Tagungsband zum 1. deutschen AAL (Ambient Assisted Living)-Kongress, pp. 371–375. VDE Verlag GMBH, Berlin, Offenbach (2008)
17. Jovanovic, N., op den Akker, R.: Towards Automatic Addressee Identification in Multiparty Dialogues. In: M. Strube, C. Sidner (eds.) Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue, pp. 89–92. Association for Computational Linguistics, Cambridge, Massachusetts, USA (2004)

Ms., in: Journal on Multimodal User Interfaces, in press, available online at:
http://www.springerlink.com/content/121040/?Content+Status=Accepted

18. Katzenmaier, M., Stiefelhagen, R., Schultz, T.: Identifying the Addressee in Human-Human-Robot Interactions Based on Head Pose and Speech. In: Proc. ICMI, pp. 144–151. State College, PA (2004)
19. Klecka, W.: Discriminant Analysis, 9 edn. SAGE PUBLICATIONS Inc., Beverly Hills (1988)
20. Lunsford, R.: Private Speech during Multimodal Human-Computer Interaction. In: Proc. ICMI, p. 346. Pennsylvania (2004). (abstract)
21. Mögele, H., Kaiser, M., Schiel, F.: SmartWeb UMTS Speech Data Collection. The SmartWeb Handheld Corpus. In: Proc. LREC, pp. 2106–2111. ELRA, Genova, Italy (2006)
22. Nöth, E., Hacker, C., Batliner, A.: 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, pp. 9–16. Zadar, Croatia (2007)
23. Oppermann, D., Schiel, F., Steininger, S., Beringer, N.: Off-Talk – a Problem for Human-Machine-Interaction. In: Proc. Eurospeech, pp. 2197–2200. Aalborg (2001)
24. Piaget, J.: Le langage et la pensée chez lenfant. Delachaux & Niestlé, Neuchâtel (1923)
25. Rehm, M., André, E.: Where do they look? Gaze Behaviors of Multiple Users Interacting with an ECA. In: Intelligent Virtual Agents: 5th International Working Conference, IVA 2005, pp. 241–252. Springer, Berlin, New York (2005)
26. Reithinger, N., Bergweiler, S., Engel, R., Herzog, G., Pfleger, N., Romanelli, M., Sonntag, D.: A Look Under the Hood - Design and Development of the First SmartWeb System Demonstrator. In: Proc. ICMI, pp. 159–166. Trento, Italy (2005)
27. Siepmann, R., Batliner, A., Oppermann, D.: Using Prosodic Features to Characterize Off-Talk in Human-Computer-Interaction. In: Proceedings of the Workshop on Prosody and Speech Recognition 2001, pp. 147–150. Red Bank, N.J. (2001)
28. Stiefelhagen, R., Yang, J., Waibel, A.: Modeling Focus of Attention for Meeting Indexing Based on Multiple Cues. IEEE Transactions on Neural Networks. Special Issue on Intelligent Multimedia Processing, July 2002 **13**, 928–938 (2002)
29. van Turnhout, K., Terken, J., Bakx, I., Eggen, B.: Identifying the intended addressee in mixed human-human and human-computer interaction from non-verbal features. In: Proc. ICMI, pp. 175–182. ACM Press, New York (2005)
30. Viola, P., Jones, M.J.: Robust Real-Time Face Detection. Int. J. Comput. Vision **57**(2), 137–154 (2004)
31. Vygotski, L.: Thought and language. M.I.T. Press, Cambridge, Mass. (1962). Original published 1934
32. Wahlster, W.: Smartweb: Mobile Application of the Semantic Web. GI Jahrestagung 2004 pp. 26–27 (2004)
33. Wahlster, W. (ed.): SmartKom: Foundations of Multimodal Dialogue Systems. Springer, Berlin, Heidelberg (2006)
34. Wahlster, W., Reithinger, N., Blocher, A.: SmartKom: Multimodal Communication with a Life-like Character. In: Proc. Eurospeech, pp. 1547–1550. Aalborg (2001)
35. Watzlawick, P., Beavin, J., Jackson, D.D.: Pragmatics of Human Communications. W.W. Norton & Company, New York (1967)

Ms., in: Journal on Multimodal User Interfaces, in press, available online at:
http://www.springerlink.com/content/121040/?Content+Status=Accepted