To Talk or not to Talk with a Computer: On-Talk vs. Off-Talk.

Anton Batliner, Christian Hacker, and Elmar Nöth

Lehrstuhl für Mustererkennung, Universität Erlangen–Nürnberg, Martensstr. 3,
91058 Erlangen, Germany
batliner,hacker,noeth@informatik.uni-erlangen.de
http://www5.informatik.uni-erlangen.de

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. By that, we can get more insight into the register 'Computer-Talk'. 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) found in these two databases.

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.

1 Introduction

As often, Shakespeare provides good examples to quote: in the passage from *Hamlet* above, we find two **'Asides'**, one for speaking aside to a third person and by that, not addressing the dialogue partners; the other one for speaking to oneself. Implicitly we learn that such asides are produced with a lower voice because when Hamlet addresses Guildenstern and Rosencrantz again, the stage direction reads *Aloud*.

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

University of Bremen, SFB/TR 8 Report No. 010-09/2006, Kerstin Fischer (Ed.):
How People Talk to Computers, Robots, and Other Artificial Communication Partners, 2006, p. 79-100
(Ms. format)

is, the less restricted is the behaviour of the users. In the early days, the users were confined to a very restricted vocabulary (prompted numbers etc.). In conversations with more elaborated automatic dialogue systems, users behave more natural; 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. In most cases, 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.

In this paper, we deal with this phenomenon **Speaking Aside** which we want to call **'Off-Talk'** following [1]. 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.; another term used in the literature is 'Private Speech' [2]. The default register for interaction with computers is, in analogy, called **'On-Talk'**. On-Talk is practically the same as Computer Talk [3]. However, whereas in the case of other (speech) registers such as 'baby-talk' the focus of interest is on the way **how** it is produced, i.e. its 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).

Off-Talk as a special dialogue act has not yet been the object of much investigation [4, 5] 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' [6]. 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.)

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 report classification results as well but want to focus on the prosody of On- vs. Off-Talk. To learn more about the phonetics of Computer-Talk, On-Talks vs. Off-Talk is a unique constellation because all other things are kept equal: the scenario, the speaker, 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.

In section 2 we present the two systems SmartKom and SmartWeb and the resp. databases where Off-Talk could be observed and/or has been provoked. Section 3 describes the prosodic and part-of-speech features that we extracted and used for classification and interpretation. In section 4, classification results and an interpretation of a principal component analysis are presented, followed by section 5 which discusses classification results, and by section 6 which discusses impact of single features for all databases.

2 Systems

2.1 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 [7] 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, 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 of hotels, restaurants, cinemas, etc., and maps 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 [8], more details on the recordings and annotations in [1, 9].

In the annotation, two different types of Off-Talk are labelled: *read Off-Talk* **ROT** and *other Off-Talk* **OOT**; every other word is via default annotated with the label **NOT** as *no Off-Talk*, i.e. On-Talk. If the user reads aloud words presented on the display, this is labelled as ROT; 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 these decisions are often not easy to make. Moreover, such a detailed labelling system would end up with a small number of tokens for each sub-type which anyhow had to be mapped onto higher classes in order to process them automatically.

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, ROT is given in capitals, and OOT boldface and recte.

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<OOT> is'<OOT> doof<OOT> <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<OOT>*

nich'<OOT> wieder<OOT> telefonisch<OOT> doch<OOT>
system: *in diesem Kino k"onnen Sie nur telefonisch reservieren*
user: *<hm> das<OOT> is'<OOT> nat"urlich<OOT> auch<OOT> doof<OOT> <P> gibt es noch eine andere M"oglichkeit <P> sicher<OOT> nicht<OOT> <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<OOT> Reservierung<ROT> best"atigen<ROT> da<OOT> kommen<OOT> wir<OOT> doch<OOT> schon<OOT> der<OOT> Sache<OOT> n"aher<OOT>*

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, 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, ROT is fairly easy to annotate: the labeller knows what is given on the display, and knows the dialogue history. OOT, 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 [9]. Note, however, that the labellers listened to the dialogues while annotating; thus, they could use acoustic information, e.g., whether some words are spoken in a very low voice or not. This is of course not possible if only the transliteration is available.

2.2 The SmartWeb System

In the SmartWeb-Project [10] – the follow-on project of SmartKom – a mobile and multimodal user interface to the Semantic Web is being developed. The user can ask open-domain questions to the system, no matter where he is: carrying a smartphone, he addresses the system via UMTS or WLAN using speech [11]. The idea is, as in the case of SmartKom, to classify automatically whether speech is addressed to the system or e.g. to a human dialogue partner or to the user himself. 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 take advantage of two modalities: speech-input from a close-talk microphone and the video stream from the front camera of the mobile

Table 1. *Cross-tabulation of On-/Off-Talk vs. On-/Off-View*

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

phone are analyzed on the server. In the video stream we classify **On-View** when the user looks into the camera. This is reasonable, since the user will look onto the display of the smartphone while interacting with the system, because he receives visual feedback, like the n-best results, 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[1]. In this paper, we concentrate on On-Talk vs. Off-Talk; preliminary results for On-View vs. Off-View can be found in [13].

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[2]) or acted (in our SW_{acted} recordings[3]). Besides *Read Off-Talk* (**ROT**), where the subjects read some system response from the display, the following categories of Off-Talk are discriminated: *Paraphrasing Off-Talk* ((**POT**) means, that the subjects report to someone else what they have found out from their request to the system, and *Spontaneous Off-Talk* ((**SOT**) can occur, when they are interrupted by someone else. We expect ROT to occur simultaneously with On-View and POT with Off-View. Table 1 displays a cross-tabulation of possible combinations of On-/Off-Talk with On-/Off-View.

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 (POT) and tells him that this is not her fault (SOT). The next system answer is correct and she reads it aloud from the screen (ROT). In

[1] In [12] On-Talk and On-View are analyzed for a Human-Human-Robot scenario. Here, face detection is based on the analysis of the skin-color; to classify the speech signal, different linguistic features are investigated. 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 becomes more difficult in an automatic dialogue system, since speech recognition is not solely based on commands.

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

[3] designed and recorded at our Institute

Table 2. *Three databases, words per category in %: On-Talk (NOT), read (ROT), paraphrasing (POT), spontaneous (SOT) and other Off-Talk (OOT)*

	# Speakers	NOT	ROT	POT	SOT	OOT [%]
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

the German example, Off-Talk is again labelled adjacent to the pertinent word, without blanks. The English translation contains only the wording; here, POT is given boldface and in italic, ROT in capitals, and SOT boldface and recte.

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

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.

2.3 Databases

All SmartWeb data has been recorded with a close-talk microphone and 8 kHz sampling rate. Recordings of the $\mathbf{SW_{spont}}$ data took place in situations that were as realistic as possible. No instruction regarding Off-Talk were given. The user was carrying a mobile phone and was interrupted by a second person. This way, a large amount of Off-Talk could be evoked. Simultaneously, video has been recorded with the front camera of the mobile phone. Up to now, data of 28 from 100 speakers (0.8 hrs. of speech) has been annotated with NOT (default), ROT, POT, SOT and OOT. OOT has been mapped onto SOT later on. This data consists of 2541 words; the distribution of On-/Off-Talk is given in Table 2. The vocabulary of this part of the database contains 750 different words.

We additionally recorded acted data ($\mathbf{SW_{acted}}$, 1.7 hrs.) to investigate which classification rates can be achieved and to show the differences to realistic data. Here, the classes POT and SOT are not discriminated and combined in *Other Off-Talk* (OOT, cf. SK_{spont}). First, we investigated the SmartKom data, that

Table 3. *100 prosodic and 30 POS features and their context*

	context size				
	-2	-1	0	1	2
95 prosodic features:					
DurTauLoc; EnTauLoc; F0MeanGlob			•		
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		•			
5 more in the set with 100 features:					
Jitter: Mean, Sigma; Shimmer: Mean, Sigma;			•		
RateOfSpeech			•		
30 POS-features:					
API,APN,AUX,NOUN,PAJ,VERB	•	•	•	•	•

have been recorded with a directional microphone: Off-Talk was uttered with lower voice and durations were longer for read speech. We further expect that in SmartWeb nobody using a head-set to address the automatic dialogue would intentionally confuse the system with loud Off-Talk. These considerations result in the following setup: The 17 speakers sat in front of a computer. All Off-Talk had to be articulated with lower voice and, additionally, ROT had to be read more slowly. Furthermore, each sentence could be read in advance so that some kind of "spontaneous" articulation was possible, whereas the ROT sentences were indeed read utterances. The vocabulary contains 361 different types. 2321 words are On-Talk, 1651 ROT, 2994 OOT (Table 2).

In the SmartKom ($\mathbf{SK_{spont}}$) database[4], 4 hrs. of speech (19416 words) have been collected from 92 speakers. Since the subjects were alone, no POT occurred: OOT is basically "talking to oneself" [14]. The proportion of Off-Talk is small (Table 2). The 16kHz data from a directional microphone was downsampled to 8kHz for the experiments in section 5.

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

3 Features used

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 the present paper, we confine our analysis to the word level to be able to map words onto the most appropriate semantic units later on. However, 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 information.

The spoken word sequence which is obtained from the speech recognizer is only required for the time alignment and for a normalization of energy and duration based on the underlying phonemes. In this paper, we use the transcription of the data assuming a recognizer 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 favor 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, that is, contexts -2 and -1, and two words after, i.e. contexts 1 and 2 in Table 3, around the current word, namely context 0 in Table 3; 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 [15]. Table 3 shows the 95 prosodic features used in section 4 and their context; in the experiments described in section 5, we used five additional features: global mean and sigma for jitter and shimmer (JitterMean, JitterSigma, ShimmerMean, ShimmerSigma), and another global tempo feature (RateOfSpeech). The six POS features with their context sum up to 30. The mean values DurTauLoc, EnTauLoc, and F0MeanGlob are computed for 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 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 [16, 17]. A detailed overview of prosodic features is given in [18]. The abbreviations of the 95 features can be explained as follows:

duration features 'Dur': absolute (Abs) and normalized (Norm); the normalization is described in [15]; the global value DurTauLoc is used to scale the mean duration values, absolute duration divided by number of syllables AbsSyl represents another sort of normalization;

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 normalized (Norm) values; the normalization is described in [15]; the global value EnTauLoc is used to scale the mean energy values, absolute energy divided by number of syllables AbsSyl represents another sort of normalization;

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; 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).

A Part of Speech (POS) flag is assigned to each word in the lexicon, cf. [19]. 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). For the context of +/- two words, this sums up to 6x5, i.e., 30 POS features, cf. the last line in Table 3.

4 Preliminary Experiments with a Subset of the SmartKom Data

The material used for the classification task and the interpretation in this chapter is a subset of the whole SmartKom database; it consists of 81 dialogues, 1172 turns, 10775 words, and 132 minutes of speech. 2.6% of the words were labelled as ROT, and 4.9% as OOT.

We computed a Linear Discriminant (LDA) classification: 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 [20]. We simply took an a priori probability of 0.5 for the two or three classes and did not try to optimize, for instance, performance for the marked classes. For classification, we used the leave-one–case-out (*loco*) method; note that this means that the speakers are seen, in contrast to the LDA used in section 5 where the leave-one-speaker-out method has been employed. Tables 4 and 5 show the recognition rates for the two–class problem Off-Talk vs. no–Off-Talk and for the three–class problem ROT, OOT, and NOT, resp. Besides recall for each class, the *CL*ass–wise computed mean classification rate (mean of all classes, unweighted average recall) CL and the overall classification (*R*ecognition) *R*ate RR, i.e., all correctly classified cases (weighted average recall), are given in percent. We display results for the 95 prosodic features with and without the 30 POS features, and for the 30 POS features alone – as a sort of 5-gram modelling a context of 2 words to the left and two words to the right, together with the pertaining word 0. Then, the same combinations are given for a sort of uni-gram modelling only the pertaining word 0. For the last two lines in Tables 4 and 5, we first computed a principal component analysis for the 5-gram- and for the

Table 4. *Recognition rates in percent for different constellations; subset of SmartKom, leave–one–case–out, Off-Talk vs. no-Off-Talk; best results are emphasized*

constellation	predictors	Off-Talk	no-Off-Talk	CL	RR
	# of tokens	806	9969	10775	
5-gram	95 pros.	67.6	77.8	72.7	77.1
raw feat. values	95 pros./30 POS	67.7	79.7	**73.7**	**78.8**
5-gram, only POS	30 POS	50.6	72.4	61.5	70.8
uni-gram	28 pros. 0	68.4	73.4	70.9	73.0
raw feat. values	28 pros. 0/6 POS 0	68.6	74.5	71.6	74.0
uni-gram, only POS	6 POS	40.9	71.4	56.2	69.1
5-gram, PCs	24 pros. PC	69.2	75.2	72.2	74.8
uni-gram, PCs	9 pros. PC 0	66.0	71.4	68.7	71.0

Table 5. *Recognition rates in percent for different constellations; subset of SmartKom, leave–one–case–out, ROT vs. OOT vs. NOT; best results are emphasized*

constellation	predictors	ROT	OOT	NOT	CL	RR
	# of tokens	277	529	9969	10775	
5-gram	95 pros.	54.9	65.2	71.5	63.9	70.8
raw feat. values	95 pros./30 POS	71.5	67.1	73.0	**70.5**	**72.6**
5-gram, only POS	30 POS	73.3	52.9	54.7	60.3	55.1
uni-gram	28 pros. 0	53.1	67.7	64.0	61.6	63.9
raw feat. values	28 pros. 0/6 POS 0	69.0	67.1	61.5	65.9	62.0
uni-gram, only POS	6 POS	80.1	64.7	18.2	54.3	22.1
5-gram, PCs	24 pros. PC	49.5	67.7	65.3	60.8	65.0
uni-gram, PCs	9 pros. PC 0	45.8	62.6	60.0	56.1	59.8

uni-gram constellation, and used the resulting principal components PC with an eigenvalue > 1.0 as predictors in a subsequent classification.

Best classification results could be obtained by using both all 95 prosodic features and all 30 POS features together, both for the two–class problem (CL: 73.7%, RR: 78.8%) and for the three–class problem (CL: 70.5%, RR: 72.6%). These results are emphasized in Tables 4 and 5. Most information is of course encoded in the features of the pertinent word 0; thus, classifications which use only these 28 prosodic and 6 POS features are of course worse, but not to a large extent: for the two–class problem, CL is 71.6%, RR 74.0%; for the three–class problem, CL is 65.9%, RR 62.0%. If we use PCs as predictors, again, classification performance goes down, but not drastically. This corroborates our results obtained for the classification of boundaries and accents, that more predictors – ceteris paribus – yield better classification rates, cf. [16, 17].

Now, we want to have a closer look at the nine PCs that model a sort of unigram and can be interpreted easier than 28 or 95 raw feature values. If we look at

the functions at group centroid, and at the standardized canonical discriminant function coefficients, we can get an impression, which feature values are typical for ROT, OOT, and NOT. Most important is energy, which is lower for ROT and OOT than for NOT, and higher for ROT than for OOT. (Especially absolute) duration is longer for ROT than for OOT – we'll come back to this result in section 6. Energy regression is higher for ROT than for OOT, and F0 is lower for ROT and OOT than for NOT, and lower for ROT than for OOT. This result mirrors, of course, the strategies of the labellers and the characteristics of the phenomenon 'Off-Talk': if people speak aside or to themselves, they do this normally in lower voice and pitch.

5 Results

Table 6. *Results with prosodic features and POS features; leave-one-speaker-out, class-wise averaged recognition rate for On-Talk vs. Off-Talk (CL-2), NOT, ROT, OOT (CL-3) and NOT, ROT, POT, SOT (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	-

In the following all databases are evaluated with an LDA-classifier and leave-one-speaker-out (*loso*) validation. 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 (unweighted average recall). In the 2-class task we classify On-Talk (NOT) vs. rest; for $N = 3$ classes we discriminate NOT, ROT and OOT (= SOT $\cup$ POT); the $N = 4$ classes NOT, ROT, SOT, POT are only available in SW_{spont}.

In Table 6 results on 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 uttered 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

Fig. 1. *ROC-Evaluation On-Talk vs. Off-Talk for the different databases*

speaker normalization (zero-mean and variance 1 for all feature components). Here, we assume that mean and variance (independent whether 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 on prosodic features are obtained for the acted data: 80.8 % CL-2 and even higher recognition rates for three classes, whereas chance would be only 33.3 % CL-3. Rates are higher for $\mathrm{SK}_{\mathrm{spont}}$ than for $\mathrm{SW}_{\mathrm{spont}}$ (72.7 % vs. 65.3 % CL-2, 60.0 % vs. 55.2 % CL-3).[5] For all databases results could be improved when the 100-dimensional feature vectors are normalized per speaker. The results for $\mathrm{SW}_{\mathrm{acted}}$ 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 $\mathrm{SW}_{\mathrm{spont}}$. Here, in particular the recall of ROT is significantly higher for $\mathrm{SK}_{\mathrm{spont}}$ (78 % vs. 57 %). In all cases a significant increase of recognition rates is obtained when linguistic and prosodic information is combined, e.g. on $\mathrm{SW}_{\mathrm{spont}}$ 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 reached. For $\mathrm{SW}_{\mathrm{spont}}$ 4 classes could be discriminated with up to 53.0 % CL-4. Here, POT is the problematic category that is very close to all other classes (39 % recall only).

Table 7. *Cross validation of the three corpora with speaker-normalized 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)

Fig. 1 shows the ROC-evaluation for all databases with prosodic features. In a real application it might be more "expensive" to drop a request that is addressed to the system than to answer a question that is not addressed to the system. If we thus set the recall for On-Talk to 90 %, every third Off-Talk word is detected in $\mathrm{SW}_{\mathrm{spont}}$ and every second in $\mathrm{SK}_{\mathrm{spont}}$. For the $\mathrm{SW}_{\mathrm{acted}}$ data, the Off-Talk recall is nearly 70 %; after speaker normalization it rises to 95 %.

[5] 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 favorable than in the case of SmartKom.

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 7. The diagonal elements show the *Train=Test* case, and in brackets the *loso* result from Table 6 (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; the database most related to SK_{spont} is SW_{spont}.

6 Discussion

As expected, results for spontaneous data were worse than for acted data (section 5). However, if we train with SW_{acted} and test with SW_{spont} and vice versa, the drop is just small. There is hope, that real applications can be enhanced with acted Off-Talk data. Next, we want to reveal similarities in the different databases and analyze single prosodic features. To discriminate On-Talk and OOT, all ROT words were deleted; for On-Talk vs. ROT, OOT is deleted. The top-ten best features are ranked in Table 8 for SW_{spont}, Table 9 for SW_{acted}, and Table 10 for SK_{spont}. For the case NOT vs. OOT the column CL-2 shows high rates for SW_{acted} and SK_{spont} with energy features; best results for NOT vs. ROT are achieved with duration features on SW_{acted}.

Most relevant features to discriminate On-Talk (**NOT**) vs. **OOT** (left column in Table 8, 9, 10) are the higher energy values for On-Talk, as well for the SW_{acted} data as for both spontaneous corpora. Highest results are achieved for SK_{spont}, since the user was alone and OOT is basically talking to oneself and consequently with extremely low voice. Also jitter and shimmer are important, in particular for SK_{spont}. The range of F0 is larger for On-Talk which might be caused by an exaggerated intonation when talking to computers. For SW_{acted} global features are more relevant (acted speech is more consistent), in particular the rate-of-speech that is lower for Off-Talk. Further global features are *EnTauLoc* and *F0MeanGlob*. Instead, for the more spontaneous SW_{spont} data pauses are more significant (longer pauses for OOT). In SK_{spont} global features are not relevant, because in many cases only one word per turn is Off-Talk (swearwords).

To discriminate **On-Talk** vs. **ROT** (right columns in Tables 8, 9, 10) duration features are highly important: the duration of read words is longer (cf. F0Max, F0Min). In addition, the duration is modeled with *Pos*-features: maxima are reached later for On-Talk.[6] Again, energy is very significant (higher for On-Talk). Most features show for all databases the same behavior but unfortunately there are some exceptions, probably caused by the instructions for the

[6] Note that these *Pos*-features are prosodic features that model the position of prominent pitch events on the time axis; if F0MaxPos is greater this normally simply means that the words are longer. These features should not be confused with POS, i.e. part-of-speech features which are discussed below in more detail.

Table 8. SW_{spont}*: Best single features for NOT vs. OOT (left) and NOT vs. ROT (right). Classification rate is given in* CL-2 *in %. The dominant feature group is emphasized. "•" denotes that the resp. values are greater for this type given in this column*

SW_{spont}	NOT	OOT	CL-2
EnMax	•		61
EnTauLoc	•		60
EnMean	•		60
PauseFill-before		•	54
JitterSigma	•		54
EnAbs	•		54
F0Max	•		53
ShimmerSigma	•		53
JitterMean	•		5
Pause-before		•	53

SW_{spont}	NOT	ROT	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

Table 9. SW_{acted}*: Best single features for NOT vs. OOT (left) and NOT vs. ROT (right)*

SW_{acted}	NOT	OOT	CL-2
EnTauLoc	•		68
EnMax	•		68
RateOfSpeech	•		65
F0MeanGlob	•		65
EnMean	•		63
ShimmerSigma	•		63
F0Max	•		61
EnAbs	•		61
F0Min		•	60
ShimmerMean	•		60

SW_{acted}	NOT	ROT	CL-2
DurTauLoc		•	86
*EnMax**Pos***		•	73
DurAbs		•	71
EnMean	•		71
*F0Max**Pos***		•	69
EnMax	•		69
DurAbsSyl		•	68
*F0On**Pos***	•		68
*F0Min**Pos***		•	65
RateOfSpeech	•		62

Table 10. SK_{spont}*: Best single features for NOT vs. OOT (left) and NOT vs. ROT (right)*

SK_{spont}	NOT	OOT	CL-2
EnMax	•		72
EnMean	•		69
JitterMean	•		69
JitterSigma	•		69
F0Max	•		69
ShimmerSigma	•		68
ShimmerMean	•		68
F0OnPos		•	67
EnAbs	•		66
EnNorm	•		61

SK_{spont}	NOT	ROT	CL-2
JitterMean	•		62
DurAbs		•	61
DurTauLoc	•		61
*F0Max**Pos***		•	61
EnTauLoc	•		69
*F0Min**Pos***		•	59
JitterSigma	•		59
EnMean	•		59
EnMax	•		58
F0Max	•		58

acted ROT: the global feature *DurTauLoc* is in SW_{acted} smaller for On-Talk, and in SW_{spont} and SK_{spont} smaller for ROT. Again, jitter is important in SK_{spont}.

To distinguish **ROT** vs. **OOT**, the higher duration of ROT is significant as well as the wider F0-range. ROT shows higher energy values in SW_{spont} but only higher absolute energy in SW_{acted} which always rises for words with longer duration.[7] All results of the analysis of single features confirm our results from the principal component analysis in section 4.

For all classification experiments we would expect a small decrease of classification rates in a real application, since we assume a speech recognizer with 100 % recognition rate in this paper. However, when using a real speech recognizer, the drop is only little for On-Talk/Off-Talk classification: in preliminary experiments we used a very poor word recognizer with only 40 % word accuracy on SK_{spont}. The decrease of CL-2 was 3.2 % relative. Using a ROC evaluation, we can set the recall for On-Talk to 90 % as above by higher weighting of this class. Then, the recall for Off-Talk goes down from $\sim$ 50 % to $\sim$ 40 % for the evaluation based on the word recognizer.

Using all 100 features, best results are achieved with SW_{acted}. The classification rates for the SK_{spont} WoZ data are worse, but better than for the SW_{spont} data since there was no Off-Talk to another Person (POT). Therefore, we are going to analyze the different SW_{spont} speakers. Some of them yield very poor classification rates. It will be investigated, if it is possible for humans to annotate these speakers, without any linguistic information. We expect further that classification rates will rise if the analysis is performed turn-based. Last but not least, the combination with On-View/Off-View will increase the recognition rates, since especially POT, where the user does not look onto the display, is hard to classify from the audio signal. For the SW_{spont} video-data, the two classes On-View/Off-View are classified with 80 % CL-2 (frame-based) with the Viola-Jones face detection algorithm [21]. The multimodal classification of the focus of attention will result in *On-Focus*, the fusion of On-Talk and On-View.

Table 11. SK_{spont}*: POS classes, percent occurrences for NOT, ROT, and OOT*

POS	# of tokens	NOUN	API	APN	VERB	AUX	PAJ
NOT	19415	18.1	2.2	6.6	9.6	8.4	55.1
ROT	365	56.2	7.1	18.1	2.2	2.2	14.2
OOT	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

The most important difference between ROT and OOT is not a prosodic, but a lexical one. This can be illustrated nicely by Tables 11 and 12 where percent occurrences of POS is given for the three classes NOT, ROT, and OOT (SK_{spont})

[7] In this paper, we concentrate on Computer-Talk = On-Talk vs. Off-Talk; thus we do not display detailed tables for this distinction **ROT** vs. **OOT**.

Table 12. SW_{spont}*: POS classes, percent occurrences for NOT, ROT, POT, and SOT*

POS	# of tokens	NOUN	API	APN	VERB	AUX	PAJ
NOT	2541	23.2	5.1	3.8	6.9	8.5	52.5
ROT	684	27.2	5.7	18.6	7.4	7.6	33.5
POT	1093	26.3	5.1	10.3	5.4	9.5	43.3
SOT	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

and for the four classes NOT, ROT, POT, and SOT (SW_{spont}). Especially for SK_{spont} there are more content words in ROT than in OOT and NOT, especially NOUNs: 56.2% compared to 7.2% in OOT and 18.1% in NOT. It is the other way round, if we look at the function words, especially at PAJ (particles, articles, and interjections): very few for ROT (14.2%), and most for OOT (63.9%). 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.

7 Concluding Remarks

Off-Talk is certainly a phenomenon the successful treatment of which 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 these Wizard-of-Oz 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 entailed in the POS features, classification rates went up.

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 a 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.

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.

Acknowledgments: 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.

References

1. Oppermann, D., Schiel, F., Steininger, S., Beringer, N.: Off-Talk – a Problem for Human-Machine-Interaction. In: Proc. Eurospeech01, Aalborg (2001) 2197–2200
2. Lunsford, R.: Private Speech during Multimodal Human-Computer Interaction. In: Proc. of the Sixth International Conference on Multimodal Interfaces (ICMI 2004), Pennsylvania (2004) 346 (abstract).
3. Fischer, K.: What Computer Talk Is and Is not: Human-Computer Conversation as Intercultural Communication. Volume 17 of Linguistics - Computational Linguistics. AQ, Saarbrücken (2006)
4. 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)
5. Carletta, J., Dahlbäck, N., Reithinger, N., Walker, M.: Standards for Dialogue Coding in Natural Language Processing. Dagstuhl-Seminar-Report 167 (1997)
6. Watzlawick, P., Beavin, J., Jackson, D.D.: Pragmatics of Human Communications. W.W. Norton & Company, New York (1967)
7. Fraser, N., Gilbert, G.: Simulating Speech Systems. CSL **5**(1) (1991) 81–99
8. Wahlster, W., Reithinger, N., Blocher, A.: SmartKom: Multimodal Communication with a Life-like Character. In: Proc. Eurospeech01, Aalborg (2001) 1547–1550
9. Siepmann, R., Batliner, A., Oppermann, D.: Using Prosodic Features to Characterize Off-Talk in Human-Computer-Interaction. In: Proc. of the Workshop on Prosody and Speech Recognition 2001, Red Bank, N.J. (2001) 147–150
10. Wahlster, W.: Smartweb: Mobile Application of the Semantic Web. GI Jahrestagung 2004 (2004) 26–27
11. 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. of the Seventh International Conference on Multimodal Interfaces (ICMI 2005), Trento, Italy (2005)
12. Katzenmaier, M., Stiefelhagen, R., Schultz, T.: Identifying the Addressee in Human-Human-Robot Interactions Based on Head Pose and Speech. In: Proc. of the Sixth International Conference on Multimodal Interfaces (ICMI 2004) . (2004) 144–151
13. Hacker, C., Batliner, A., Nöth, E.: Are You Looking at Me, are You Talking with Me – Multimodal Classification of the Focus of Attention. In: Proc. of the Ninth International Conference on Text, Speech, Dialogue, Berlin, Springer (2006) to appear
14. Batliner, A., Zeissler, V., Nöth, E., Niemann, H.: Prosodic Classification of Offtalk: First Experiments. In: Proc. of the Fifth International Conference on Text, Speech, Dialogue, Berlin, Springer (2002) 357–364
15. Batliner, A., Buckow, A., Niemann, H., Nöth, E., Warnke, V.: The Prosody Module. In Wahlster, W., ed.: Verbmobil: Foundations of Speech-to-Speech Translations. Springer, Berlin (2000) 106–121

16. Batliner, A., Buckow, J., Huber, R., Warnke, V., Nöth, E., Niemann, H.: Prosodic Feature Evaluation: Brute Force or Well Designed? In: Proc. ICPHS99, San Francisco (1999) 2315–2318
17. 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. of Eurospeech01, Aalborg (2001) 2781–2784
18. Batliner, A., Fischer, K., Huber, R., Spilker, J., Nöth, E.: How to Find Trouble in Communication. Speech Communication **40** (2003) 117–143
19. 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. of Eurospeech99, Budapest (1999) 519–522
20. Klecka, W.: Discriminant Analysis. 9 edn. SAGE PUBLICATIONS Inc., Beverly Hills (1988)
21. Viola, P., Jones, M.J.: Robust Real-Time Face Detection. Int. J. Comput. Vision **57**(2) (2004) 137–154