Are You Looking at Me, are You Talking with Me - Multimodal Classification of the Focus of Attention

Christian Hacker, Anton Batliner, and Elmar Nöth *

University of Erlangen-Nuremberg,
Chair for Pattern Recognition (Informatik 5)
Martensstraße 3, D-91058 Erlangen, Germany
hacker@informatik.uni-erlangen.de

Abstract. Automatic dialogue systems get easily confused if speech is recognized which is not directed to the system. Besides noise or other people's conversation, even the user's utterance can cause difficulties when he is talking to someone else or to himself ("Off-Talk"). In this paper the automatic classification of the user's focus of attention is investigated. In the German SmartWeb project, a mobile device is used to get access to the semantic web. In this scenario, two modalities are provided - speech and video signal. This makes it possible to classify whether a spoken request is addressed to the system or not: with the camera of the mobile device, the user's gaze direction is detected; in the speech signal, prosodic features are analyzed. Encouraging recognition rates of up to 93 % are achieved in the speech-only condition. Further improvement is expected from the fusion of the two information sources.

1 Introduction

In the SmartWeb-Project [1] 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 [2]. In this paper we present an approach to automatically classify 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 phone are analyzed on the server. In the speech signal we detect *On-Talk* vs. *Off-Talk* using prosodic information, that means, we investigate, whether people use different speech-registers when addressing a system (On-Talk) and when addressing a human dialogue partner. In this

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

paper, all linguistic information is neglected. In the video stream we classify *On-View* when the user's gaze direction is towards the camera. In deed, the users usually look onto the display of the smartphone while interacting with the system, because they receive visual feedback, like the n-best results, maps, or pictures. *Off-View* means, that the user does not look at the display at all.

After a short literature survey, recently recorded databases are described in Sect. 3. In this paper, acted and spontaneous speech is compared. Features to analyze On-Talk and On-View are described in Sect. 4; results of the classification are given in Sect. 5. A discussion of the results, an analysis of prosodic features and a motivation of the fusion of both modalities is given in Sect. 6.

2 Related Work

In Katzenmaier et al. [3] 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. Oppermann et al. [4] describe such a corpus collected in a Wizard-of-Oz experiment in the context of the SmartKom project (cf. SK_{spont} in Sect. 3). Unfortunately, only a small part of the data is labeled as Off-Talk. For this database, results of On-/Off-Talk classification using prosodic features and part-of-speech categories are given in [5]. It could be shown that the users indeed use different speech registers when talking to the system, when talking to themselves, and when reading from the display. Video information was not used, since in the SmartKom scenario the user was alone and nearly always looking onto the display while talking.

To classify the user's gaze direction, face-tracking algorithms like in [6] did not seem to be appropriate, since in our scenario the face ought to be lost, if the user does not look onto the display anymore. A very fast and robust detection algorithm to discriminate two classes (face/no face) is presented by Viola and Jones [7]. It is based on a large number of simple Haar-like features (cf. Sect 4). With a similar algorithm five facial orientations are discriminated in [8]. As features, different pairs of pixels in the image are compared.

3 Corpora

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[1]) or acted (in our

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

Table 1. *Three databases, words per category in %: On-Talk, read (ROT), paraphrasing (POT), spontaneous (SOT) and other Off-Talk (OOT)*

	# Speakers	On-Talk	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

SW_{acted} recordings). Besides *Read Off-Talk* (ROT), where the candidates had to read some possible system response from the display, the following categories of Off-Talk are discriminated: *Paraphrasing Off-Talk* (POT) means, that the candidates 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. 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, speech of 28 speakers has been annotated (0.8 hrs. of speech). This data consists of 2541 words; the distribution of On-/Off-Talk is given in Tab. 1. The vocabulary of this part of the database contains 750 different words. As for the video data, up to now 27 speakers recorded in different environment (indoor, outdoor, weak and strong backlight) have been annotated. These 4 hrs. video data that also contains non-speech segments consist of 76.1 % On-View, 15.5 % Off-View and 1.6 % without face; the rest was not well-defined.

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 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 (Tab. 1).

In SmartKom ($\mathbf{SK_{spont}}$), 4 hrs. of speech (19416 words) have been collected from 92 speakers. Since the candidates were alone, no POT occurred: OOT is basically "talking to oneself" [5]. The proportion of Off-Talk is small (Tab. 1).

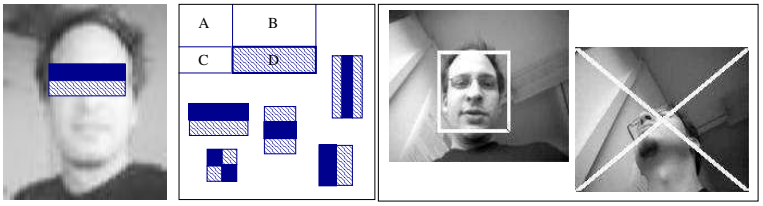

Fig. 1. *Face detection after Viola and Jones* [7]. *Left to right: The best feature, wavelet features and their calculation, results from our task (On-View and Off-View)*

4 Feature Extraction

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 syntactic and semantic classes, but only prosodic information. The spoken word sequence which is obtained from the speech recognizer in SmartWeb 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; lower accuracy decreases Off-Talk classification rates only to a small extend, as preliminary experiments have shown.

In most cases of the described acted data and in many cases of the other data, indeed, one can *hear* if the addressee is a machine or a computer; features could be loudness, accentuation, intonation, or rate-of-speech. For the automatic classification we use a highly redundant set of 100 prosodic features. These features are calculated for each word, and, additionally, for some of the neighboring words to encode information from the context. A short description of the features and abbreviations used in Tab. 2 and 3 is given in the following: 30 features are based on the energy (*Ene*) of the signal, e.g. maximum, minimum, mean (*Max*, *Min*, *Mean*), absolute value (*Abs*), or the position of the maximum (*MaxPos*). Further 25 features are calculated from the fundamental frequency f_0, e.g. *Max*, *Min* as above and the position of onset (beginning of voiced region), offset, and the extrema (*OnsetPos*, *OffsetPos*, *MaxPos* etc.). The reference point for all position-features is the end of the current word. 29 more features are calculated to characterize duration (*Dur*; *AbsSyl* is normalized with the number of syllables), and 8 to describe pauses (*Pau*) before and after the current word. Filled pauses contain non-words. Eight features are calculated for the whole turn, i.e., they have the same value for each word: 4 are based on jitter and shimmer, the rate-of-speech, and one feature is based on the f_0, energy and duration, respectively (*Global*). A detailed overview of prosodic features is given in [9].

For the classification of **On-View vs. 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 [7]. The face detection works for single images;

Table 2. *On-Talk vs. OOT: Best single features and classification rate* CL-2 *(average recall of the 2 classes) in %. The dominant feature group is emphasized. "* > 1*": values are greater for On-Talk*

SW_{spont}	On-T. / OOT	CL-2
EneMax	> 1	61
EneGlobal	> 1	60
EneMean	> 1	60
PauFilledBefore	< 1	54
JitterSigma	> 1	54
EneAbs	> 1	54
f_0*Max*	> 1	53
ShimmerSigma	> 1	53
JitterMean	> 1	53
PauBefore	< 1	53

SW_{acted}	On-T. / OOT	CL-2
EneGlobal	> 1	68
EneMax	> 1	68
RateOfSpeech	> 1	65
f_0*Global*	> 1	65
EneMean	> 1	63
ShimmerSigma	> 1	63
f_0*Max*	> 1	61
EneAbs	> 1	61
f_0*Min*	< 1	60
ShimmerMean	> 1	60

Table 3. *Best single features for On-Talk vs. ROT (cf. Tab. 2)*

SW_{spont}	On-T. / ROT	CL-2
EneGlobal	> 1	60
DurAbs	< 1	58
f_0*Max****Pos***	< 1	58
f_0*Onset****Pos***	> 1	57
DurGlobal	> 1	57
*EneMax****Pos***	< 1	56
EneMean	> 1	56
EneAbs	< 1	56
f_0*Offset****Pos***	> 1	55
f_0*Min****Pos***	< 1	53

SW_{acted}	On-T. / ROT	CL-2
DurGlobal	< 1	86
*EneMax****Pos***	< 1	73
DurAbs	< 1	71
EneMean	> 1	71
f_0*Max****Pos***	< 1	69
EneMax	> 1	69
DurAbsSyl	< 1	68
f_0*Onset****Pos***	> 1	68
f_0*Min****Pos***	< 1	65
RateOfSpeech	> 1	62

up to now, no use of context information is implemented. The algorithm is based on simple Haar-like wavelets; the most significant feature is shown in Fig. 1, left: The integral of the light area is subtracted from the integral of the dark area. All wavelets (up to scaling and translation) are shown in Fig. 1 in the middle. The integral of the quadrangle spanned by each pixel and the origin is calculated in advance. Then the area D can be easily computed from $(A + B + C + D) - (A + B) - (A + C) + A$. From many possible features, up to 6000 are selected with the ADABOOST algorithm; a hierarchical classifier speeds up the detection [7]. In this paper we use 176×144 grayscale images, 7.5 per second; faces are searched in different subimages, greater than half the image, and scaled to 24×24. Fig. 1, right, shows On-View and Off-View of a mobile phone user.

5 Experimental Setup and Results

In the following all databases are evaluated with the LDA-classifier and leave-one-speaker-out validation. All results are measured with the class-wise averaged

Table 4. *Results on audio (100 prosodic features) and video.* CL-*i is the average recall of i classes*

	audio			audio, normalized			video
	CL-2	CL-3	CL-4	CL-2	CL-3	CL-4	CL-2
SW_{spont}	65.3	55.2	42.0	66.8	56.4	49.8	76.5
SW_{acted}	80.8	83.9	-	92.6	92.9	-	-
SK_{spont}	72.7	60.0	-	74.2	61.5	-	-

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 vs. rest; for $N = 3$ classes we discriminate On-Talk, ROT and OOT (= SOT ∪ POT); the $N = 4$ classes On-Talk, ROT, SOT, POT are only available in SW_{spont}. First, we evaluated for each corpus the single best features (classifiers with 1 feature, each). To discriminate e.g. On-Talk and OOT, all ROT words were deleted. The top-ten best features can be found in Tab. 2; for the task On-Talk vs. ROT features are ranked in Tab. 3. The column CL-2 shows higher rates for SW_{acted} than for SW_{spont}, best results are achieved for On-Talk vs. ROT (SW_{acted}).

Tab. 4 shows results based on all 100 features for different databases. Again, best results are obtained for the acted data: 81 % CL-2 and even higher recognition rates for three classes, whereas chance would be only 33 % CL-3. For SK_{spont} higher rates are achieved than for SW_{spont}. All results could be improved when the 100-dimensional feature vectors are normalized per speaker (zero-mean and variance 1): Tab. 4, right, shows the results when 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. The results for SW_{acted} rise drastically to 93 % CL-3; for the other corpora a smaller increase can be observed. For SW_{spont} 4 classes could be discriminated with 50 % CL-4. Here, POT is the problematic category that is very closed to all other classes (35 % recall only). If we train with acted data and evaluate with SW_{spont}, we achieve 63 % CL-2, the other way round 86 % on SW_{acted}. The drop is surprisingly small, however, this does not hold for the 3-class task: rates for ROT are very low.

Fig. 2 shows the ROC-evaluation for all databases. 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 SW_{spont} and every second in SK_{spont}. For the SW_{acted} data, the Off-Talk recall is nearly 70 %; after speaker normalization it rises to 95 %.

As for the **On-View** classification, we evaluate our data frame based with a freely available classifier[2]. For On-View vs. {Off-View ∪ No-Face} 77 % CL-2 are achieved. For 6 of 27 speakers CL-2 was smaller than 60 %; the reason seems to be strong backlight. However, for 12 speakers recognition rates of more than 80 %

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

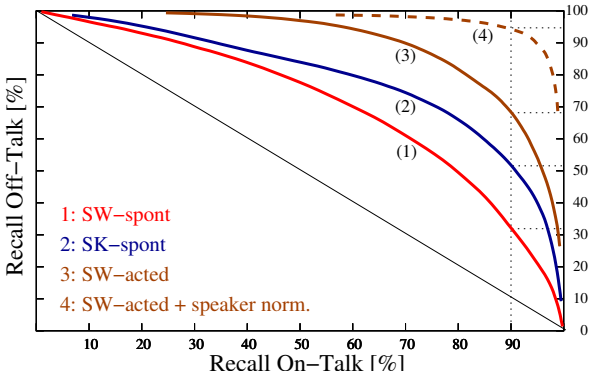

Fig. 2. *ROC-Evaluation On-Talk vs. Off-Talk for the different databases*

were achieves, for the seven best even 94 – 98 %. We expect, that classification rates will rise, if the results are averaged over words or sentences.

6 Discussion

As expected, results for spontaneous data were worse than for acted data (Sect. 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.

Most relevant features to discriminate **On-Talk vs. OOT** (Tab. 2) are the higher energy values for On-Talk, as well for the SW_{spont} data as for the acted data. Also jitter and shimmer are important. The range of f_0 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. Instead, for the more spontaneous SW_{spont} data pauses are more significant (longer pauses for OOT). To discriminate **On-Talk vs. ROT** (Tab. 3) duration features are highly important: the duration of read words is longer. In addition, the duration is modeled with *Pos*-features: maxima are reached later for On-Talk (e.g. caused by a continuation rise within asyndetic listing). Again, energy is very significant (higher for On-Talk). Most features show for both databases the same behavior but unfortunately there are some exceptions, probably caused by the instructions for the acted ROT: *DurGlobal* is in SW_{acted} smaller for On-Talk, and in SW_{spont} (and SK_{spont}) for ROT. To distinguish **ROT vs. OOT**, the higher duration of ROT is significant as well as the wider f_0-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.

For the SK_{spont} corpus, similarities with SW_{spont} could be observed for On-Talk vs. ROT. In the other cases, in particular jitter and shimmer become more important. Since OOT means "talking to oneself" (very low voice) in SK_{spont} the classification rate with energy increases.

Using all 100 features, best results are achieved with SW_{acted}. The classification rates for the SK_{spont} data are worse, but better than for the SW_{spont} data since there was no Off-Talk to another Person (POT). Thus, 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. Further, we expect, that classification rates will rise if the analysis is performed turn-based. Also the turn-based average from the video classifier is expected to result in more robust scores. Last but not least, the combination of both modalities 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. The multimodal classification of the focus of attention will result in *On-Focus*, the fusion of On-Talk and On-View. Additional linguistic features (bag-of-words or part-of-speech features) could further rise the accuracy.

7 Concluding Remarks

In this paper, a set of 100 prosodic features was analyzed; we classified from the audio signal whether the user speaks to the system or not. Very high classification rates up to 93 % are achieved for acted speech. A significant improvement was obtained by speaker normalization. Since On-View could be classified robustly from the video signal, a fusion of both modalities will increase recognition in the future. Further applications could be to control a car radio very robustly with On-Talk (acted speech is easy to learn), whereas most of the time the driver speaks to other occupants or to himself. For human-machine dialogues, e.g. with an avatar, additionally video information can be used. An application could be assisted living for the elderly, where the On-Talk module permanently listens for a potential command to control telephone, TV, and household appliances.

References

1. Wahlster, W.: Smartweb: Mobile Application of the Semantic Web. GI Jahrestagung 2004 (2004) 26–27
2. 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, Trento (2005)
3. Katzenmaier, M., Stiefelhagen, R., Schultz, T.: Identifying the Addressee in Human-Human-Robot Interactions Based on Head Pose and Speech. In: ICMI. (2004)
4. Oppermann, D., Schiel, F., Steininger, S., Beringer, N.: Off-Talk – a Problem for Human-Machine-Interaction. In: Proc. European Conf. on Speech Communication and Technology, Aalborg (2001)

5. Batliner, A., Zeissler, V., Nöth, E., Niemann, H.: Prosodic Classification of Offtalk: First Experiments. In: Proc. TSD, Berlin, Springer (2002) 357–364
6. Deutsch, B., Gräßl, C., Bajramovic, F., Denzler, J.: A Comparative Evaluation of Template and Histogram Based 2D Tracking Algorithms. In: Pattern Recognition, 27th DAGM Symposium , Berlin, Springer (2005) 269–276
7. Viola, P., Jones, M.J.: Robust Real-Time Face Detection. Int. J. Comput. Vision **57**(2) (2004) 137–154
8. Baluja, S., Sahami, M., Rowley, H.A.: Efficient Face Orientation Discrimination. In: IEEE International Conference on Image Processing, Singapore (2004)
9. Batliner, A., Fischer, K., Huber, R., Spilker, J., Nöth, E.: How to Find Trouble in Communication. Speech Communication **40** (2003) 117–143

Text, Speech and Dialogue
9th International Conference, TSD 2006, Brno, Czech Republic, September 11-15, 2006, Proceedings
Series: Lecture Notes in Computer Science , Vol. 4188
Sublibrary: Lecture Notes in Artificial Intelligence
Sojka, Petr; Kopecek, Ivan; Pala, Karel (Eds.)
2006, XV, 721 p., Softcover
ISBN-10: 3-540-39090-1
ISBN-13: 978-3-540-39090-9
Online version available

Online orders shipping within 2-3 days.

85,60 €

About this book

About this book

This book constitutes the refereed proceedings of the 9th International Conference on Text, Speech and Dialogue, TSD 2006, held in Brno, Czech Republic, in September 2006.

The 87 revised full papers presented together with 2 invited papers were carefully reviewed and selected from 175 submissions. The papers present a wealth of state-of-the-art research results in the field of natural language processing with an emphasis on text, speech, and spoken dialogue ranging from theoretical and methodological issues to applications in various fields and with special focus on corpora, texts and transcription, speech analysis, recognition and synthesis, as well as their intertwining within NL dialogue systems.

Written for:
Researchers and professionals

Keywords:
NLP
algorithmic learning
computational linguistics
dialogue
dialogue management
information extraction
information retrieval
knowledge discovery
natural language processing
semantic Web
semantic models
speaker verification
speech perception

speech processing
speech recognition
spoken language processing
statistical methods
text processing