Prosodic classification of offtalk: first experiments

Anton Batliner, Viktor Zeißler, Elmar Nöth, Heinrich Niemann

Angaben zur Veröffentlichung / Publication details:

Batliner, Anton, Viktor Zeißler, Elmar Nöth, and Heinrich Niemann. 2002. "Prosodic classification of offtalk: first experiments." In *Text, speech and dialogue: 5th International Conference, TSD 2002 Brno, Czech Republic, September 9–12, 2002*, edited by Petr Sojka, Ivan Kopeček, and Karel Pala, 357–64. Berlin [u.a.]: Springer. https://doi.org/10.1007/3-540-46154-X_51.

Prosodic Classification of Offtalk: First Experiments

Anton Batliner, Viktor Zeißler, Elmar Nöth, and Heinrich Niemann

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

Abstract. SmartKom is a multi-modal dialogue system which combines speech with gesture and facial expression. In this paper, we want to deal with one of those phenomena which can be observed in such elaborated systems that we want to call 'offtalk', i.e., speech that is not directed to the system (s peaking to oneself, speaking aside). We report the classification results of first experiments which use a large prosodic feature vector in combination with part-of-speech information.

1 Introduction

1.1 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 i nformation 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, several annotations are carried out. The recorded speech represents thus a special variety of non-prompted, spontaneous speech typical for man-machine-communication in general and for such a multi-modal setting in particular. More details on the system can be found in [13], more details on the recordings and annotations in [10,11].

1.2 Offtalk

The more elaborate an automatic dialogue system 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 elaborate automatic!dialogue systems like SmartKom, users behave more natural; thus, phenomena can be observed and have to be coped with that could not be observed in communications with very simple dialogue systems.

In this paper, we want to deal with one of these phenomena that we call ‘offtalk’. Offtalk is defined in [10] 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. Other terms are ‘speaking to oneself’, ‘speaking aside’, ‘thinking aloud’. 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 the annotation, two different types of offtalk are labelled: read offtalk (ROT) and other offtalk (OOT); every other word is via default annotated with the label NOT as ‘no offtalk’. 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 offtalk, 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. Offtalk as a special dialogue act has not yet been the object of much investigation [1,6] most likely because it could not be observed in human-human communication. (In a normal human-human dialogue setting, offtalk might really be rather self-contradictory, because of the “Impossibility of Not Communicating” [14]. We can, however, easily imagine the use of offtalk if someone is speaking in a low voice not *to* but *about* a third person present who is very hard of hearing.)

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. Offtalk 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 können Sie nur telefonisch reservieren*
user: *kannst du mich dahin verbinden <P> Aladdin*
system: *diese Funktion ist momentan leider nicht verfügbar <P> kann ich Ihnen auch anders helfen*
user: *<hm> <P> das<OOT> is’<OOT> doof<OOT> <hm> <P> <A> geh doch noch mal zurück <P> Aladdin <P> zurückgehen*
system: *einen Augenblick bitte*
user: *ja <P> +/wie sieht d=/+ wie sieht das aus mit dem Kino Schloß <P> dreiundzwanzig fünfundvierzig Armageddon <P> zwei Karten Aladdin <P> hoffentlich<OOT> nich’<OOT> wieder<OOT> telefonisch<OOT> doch<OOT>*
system: *in diesem Kino können Sie nur telefonisch reservieren*
user: *<hm> das<OOT> is’<OOT> natürlich<OOT> auch<OOT> doof<OOT> <P> gibt es noch eine andere Möglichkeit <P> sicher<OOT> nicht<OOT> <P> zurück*
system: *Reservierungen können nur im Kino StudioEuropa durchgeführt 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 fünfundvierzig <P> aha<OOT> Reservierung<ROT> bestätigen<ROT> da<OOT> kommen<OOT> wir<OOT> doch<OOT> schon<OOT> der<OOT> Sache<OOT> näher<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 offtalk, is more problematic; for a discussion we want to refer to [11]. Note, however, that the labellers listen to the dialogues while annotating; thus, they can 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 Material and Features Used

The material used for the classification task 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. Note that the recording is, at the moment, not finished yet; thus, this material represents only a part of the data that will eventually be available.

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 1, around the final syllable of a word or a word hypothesis, namely context 0 in Table 1; by that, we use so to speak a 'prosodic 5-gram'. A full acc ount of the strategy for the feature selection is beyond the scope of this paper; details and further references are given in [2]. Table 1 shows the 95 prosodic features used and their context. 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 [3,4]. The abbreviations can be explained as follows:
duration features 'Dur': absolute (Abs) and normalized (Norm); the normalization is described in [2]; 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 [2]; 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 logarithmized and normalized 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. [5]. 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 6 × 5, i.e., 30 PoS features, cf. the last line in Table 1.

Table 1. Ninety-five prosodic and 30 PoS features and their context.

features	**context size**				
	−2	−1	0	1	2
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		●			
API,APN,AUX,NOUN,PAJ,VERB	●	●	●	●	●

Table 2. Recognition rates in percent for different constellations; leave-one-out, offtalk vs. no-offtalk; best results are emphasized.

constellation	predictors	offtalk	no-offtalk	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

3 Classification

We computed a Linear Discriminant 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 [8]. We simply took an a priori probability of 0.5 for the two classes and did not try to optimize, for instance, performance for the marked classes. For classification, we used the jackknife (leave-one-out) method. The computation of the features was done with the spoken word chain (‘cheating’). Tab les 2 and 3 show the recognition rates for the two-class problem offtalk vs. no-offtalk 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) CL and the overall classification (*R*ecognition) *R*ate RR, i.e., all correctly classified cases, 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 2 and 3, we first computed a principal component analysis for the 5-gram- and for the uni-gram constellation, and used the resulting principal components PC with an eigenvalue $>$ 1.0 as predictors in a subsequent classification.

4 Interpretation

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 2 and 3. 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

Table 3. Recognition rates in percent for different constellations; leave-one-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

classification of boundaries and accents, that more predictors – ceteris paribus – yield better classification rates, cf. [3,4].

Now, we want to have a closer look at the nine PCs that model a sort of uni-gram 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 if we interpret Table 4. 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 'offtalk': if people speak aside or to themselves, they do this normally in lower voice and pitch. The most important difference between ROT and OOT is, however, not a prosodic, but a lexical one. This can be illustrated nicely by Table 4 where percent occurrences of PoS is given for the three classes ROT, OOT, and NOT. There are more content words CW in ROT than in OOT and NOT, especially NOUNs: 54.9 % compared to 7.2 % in OOT and 18.9 % in NOT. It is the other way round, if we look at the function words FWs, especially at PAJ (particles, articles, and interjections): very few for ROT (15.2 %), and most for OOT (64.7 %). The explanation is straightforward: the user only reads words that are presented on the screen, and these are mostly CWs – names of restaurants, cinemas, etc., which of course are longer than other word classes.

Table 4. PoS classes, percent occurrences for ROT, OOT, and NOT.

PoS	# of tokens	NOUN	API	APN	VERB	AUX	PAJ
ROT	277	54.9	8.3	17.0	1.8	2.9	15.2
OOT	529	7.2	2.5	10.8	9.3	5.7	64.7
NOT	9969	18.9	1.9	7.8	9.5	8.7	53.2

5 Concluding Remarks

Offtalk 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 these Wizard-of-Oz experiments. We have seen that a prosodic classification, based on a large feature vector – actually the very same that had been successfully used for the classification of accents and boundaries within the Verbmobil project, cf. [2] – yields good but not excellent classification rates. With additional lexical information entailed in the PoS features, classification rates went up. However, the frequency of ROT and OOT is rather low and thus, their precision is not yet very satisfactory; if we tried to obtain a very high recall for the marked classes ROT and OOT, precision would go down even more. Still, we believe that already with the used feature vector, we could use a strategy which had been used successfully for the treatment of speech repairs within the Verbmobil project, cf. [12]: there, we tuned the classification in such a way that we obtained a high recall at the expense of a very low precision for speech repairs. This classification could then be used as a sort of preprocessing step that reduced the search space for subsequent analyses considerably, from some 50.000 to some 25.000 instances. Another possibility would be an integrated processing with the A* algorithm along the lines described in [9], using other indicators that most likely will contribute to classification performance as, e.g., syntactic structure, the lexicon (use of swear words), the use of idiomatic phrases, out-of-sequence dialogue acts, etc. Eventually, experiments will have to be conducted that use word hypotheses graphs instead of the spoken word chain.

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. The responsibility for the contents of this study lies with the authors.

References

1. J. Alexandersson, B. Buschbeck-Wolf, T. Fujinami, M. Kipp, S. Koch, E. Maier, N. Reithinger, B. Schmitz, and M. Siegel. Dialogue Acts in VERBMOBIL-2 – Second Edition. Verbmobil Report 226, Juli 1998.
2. A. Batliner, A. Buckow, H. Niemann, E. Nöth, and V. Warnke. The Prosody Module. In: W. Wahlster, editor, *Verbmobil: Foundations of Speech-to-Speech Translations*, pp. 106–121. Springer, Berlin, 2000.
3. A. Batliner, J. Buckow, R. Huber, V. Warnke, E. Nöth, and H. Niemann. Prosodic Feature Evaluation: Brute Force or Well Designed? In: *Proc. 14th Int. Congress of Phonetic Sciences*, volume 3, pages 2315–2318, San Francisco, August 1999.
4. A. Batliner, J. Buckow, R. Huber, V. Warnke, E. Nöth, and H. Niemann. Boiling down Prosody for the Classification of Boundaries and Accents in German and English. In: *Proc. European Conf. on Speech Communication and Technology*, Vol. 4, pp. 2781–2784, Aalborg, September 2001.
5. A. Batliner, M. Nutt, V. Warnke, E. Nöth, J. Buckow, R. Huber, and H. Niemann. Automatic Annotation and Classification of Phrase Accents in Spontaneous Speech. In: *Proc. European Conf. on Speech Communication and Technology*, Vol. 1, pp. 519–522, Budapest, September 1999.

6. J. Carletta, N. Dahlbäck, N. Reithinger, and M. Walker. Standards for Dialogue Coding in Natural Language Processing. Dagstuhl-Seminar-Report 167, 1997.
7. N.M. Fraser and G.N. Gilbert. Simulating Speech Systems. *Computer Speech & Language*, 5(1):81–99, 1991.
8. W.R. Klecka. *Discriminant Analysis*. SAGE PUBLICATIONS Inc., Beverly Hills, 9 edition, 1988.
9. E. Nöth, A. Batliner, V. Warnke, J. Haas, M. Boros, J. Buckow, R. Huber, F. Gallwitz, M. Nutt, and H. Niemann. On the Use of Prosody in Automatic Dialogue Understanding. *Speech Communication*, 36(1–2), January 2002.
10. D. Oppermann, F. Schiel, S. Steininger, and N. Behringer. Off-Talk – a Problem for Human-Machine-Interaction? In *Proc. European Conf. on Speech Communication and Technology*, volume 3, pages 2197–2200, Aalborg, September 2001.
11. R. Siepmann, A. Batliner, and D. Oppermann. Using Prosodic Features to Characterize Off-Talk in Human-Computer-Interaction. In *Proc. of the Workshop on Prosody and Speech Recognition 2001*, 2001, pages 147–150, Red Bank, October 2001.
12. J. Spilker, A. Batliner, and E. Nöth. How to Repair Speech Repairs in an End-to-End System. In R. Lickley and L. Shriberg, editors, *Proc. ISCA Workshop on Disflueny in Spontaneous Speech*, pages 73–76, Edinburgh, September 2001.
13. W. Wahlster, N. Reithinger, and A. Blocher. SmartKom: Multimodal Communication with a Life-like Character. In *Proc. European Conf. on Speech Communication and Technology*, volume 3, pages 1547–1550, Aalborg, September 2001.
14. P. Watzlawick, J.H. Beavin, and Don D. Jackson. *Pragmatics of Human Communications*. W.W. Norton & Company, New York, 1967.