Automatic differentiation of form-function-relations of the discourse particle “hm” in a naturalistic human-computer interaction

Alicia Flores Lotz, Ingo Siegert and Andreas Wendemuth

Cognitive Systems Group, Otto-von-Guericke University Magdeburg, Germany
ingo.siegert@ovgu.de

Abstract: The development of speech-controlled assistance systems has gained more importance in this day and time. Application ranges from driver assistance systems in the automotive sector to daily use in mobile devices such as smartphones or tablets. To ensure the reliability of these Systems; not only the meaning of the pure spoken text, but also meta-information about the user or dialogue functions such as attention or turn-taking have to be perceived and processed. This further information is transmitted through intonation of words or sentences. In human communication, discourse particles serve to forward information, without interrupting the speaker. For the German language J.E. Schmidt [11] empirically discovered seven types of form-function-concurrences on the isolated DP “hm”. To also be considered in human-computer interaction, it is useful to be able to distinguish these different meanings of the DP “hm”.
In this paper we present an automatic classification-method to correlate a specific intonation curve to one of the seven form-function-prototypes. To verify the results of the classification algorithm we utilize a manual labeling.

1 Introduction

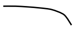
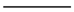
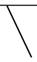
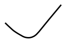
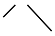
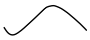
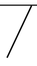
Particularly in the development of speech-controlled assistance systems, a more human-like interaction with the system needs to be obtained. The system should be capable of the adaption to the users’ individual skills, preferences and current emotional states [13]. In human-human-interaction (HHI) the behavior of the speaker is mainly characterized by semantic and prosodic cues such as short feedback signals. These signals transmit so called meta-information about certain dialogue functions as attention, understanding, confirmation or other attitudinal reactions [1]. Thus, these signals further the progress and coordination of human interaction. Moreover, they allow an interaction between speaker and interlocutor and the transmission of information about their behavioral or affective state without interrupting each other.
For a naturalistic interaction the interlocutor needs to be able to perceive and process this information. Looking at human-computer interaction (HCI), this means that the system must be able to detect those particles and interpret their meanings. In this paper a rule-based classification algorithm to evaluate the pitch contours of the DP “hm” is presented.

2 Discourse particles

In human communication, discourse particles (DPs) serve as independent small utterance units, occurring at communicative decisive points of the conversation and are a part of the previously mentioned feedback signals. They forward information, without interrupting the speaker and show the same intonation curve as whole sentences [11]. In German particular attention is paid

to monosyllabic words such as “ja”, “so”, “wie” or “hm”. The semantic content of these DPs is usually very small, which increases the importance of its prosody. Thus, by changing the intonation, an interjection can have several different meanings. Thereby the DP “hm” is among the most diverse interjections, also being referred to as a pure intonation carrier [4].
For the German language J.E. Schmidt [11] empirically discovered seven types of form-function-concurrences on the isolated DP “hm” pictured in Table 1. As the analysis was performed for HHI, it first has to be checked if the DPs also occur in HCI. An investigation on this assumption is made in [5, 12]. Both authors investigated the use of DPs within an HCI and have shown that DPs are used frequently by users talking to a machine. The only difference between an HHI and HCI usage of DPs is that in HCI is more focused on talk-organizing, or expressive functions.

Table 1 - Form-function relation of DP “hm” according to [11], the terms are translated into appropriate English ones. DP-5 can be seen as a combination of DP-7 and DP-3.

Name	idealized pitch-contour	Description
DP-1		attention
DP-2		thinking
DP-3		finalization signal
DP-4		confirmation
DP-5		decline
DP-6		positive assessment
DP-7		request to respond

To be able to make this distinction automatically, a classifier has to be developed that assigns the extracted pitch contour from the detected and measured DPs in the acoustic signal to one of the presented form-function-prototypes.

3 Dataset

The conducted study of this paper utilizes the LAST MINUTE Corpus [10], which is used to investigate HCI in case of dialog barriers and represents a naturalistic HCI. The setup of the corpus revolves around an imaginary journey to an unknown place “Waiuku”, which the subjects have won. The task of the subject is now to prepare the journey, equip the bags and select clothing using voice commands. More details on the design of the corpus can be found in [10] and [9].
To obtain the basic speech material, a sub-set of the corpus was utilized. In total a number of 274 DPs could be extracted from 25 hours of speech-material received from 56 subjects. Using a manual correction phase, unsuitable DPs were excluded leaving 259 DPs to server as raw material for the investigation of this paper.

4 Methods

In the following subsections the implementation of the classification-algorithm is presented. First the pitch-contours need to be extracted from the raw speech-material obtained in section 3. Then the classification of these courses takes place containing a pre-processing and analysis of the signals. Lastly a verification of the results is conducted. Figure 1 serves as structural flowchart depicting the single steps of the algorithm.

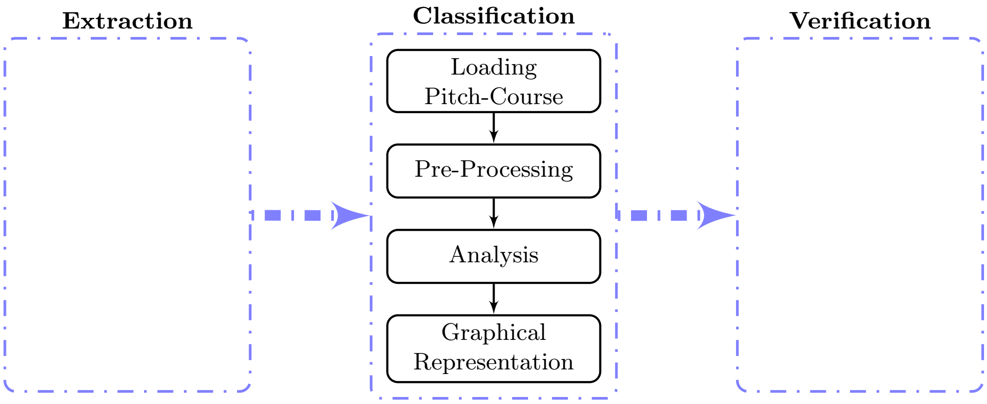


Figure 1 - Structural flowchart of the classification-algorithm

4.1 Pitch-extraction

First of all a reliable method to extract the pitch-contour is needed. Therefore, the computer software Praat [3] for analysis of speech in phonetics is used. To extract the pitch, the software uses an adapted autocorrelation method based on the division of the autocorrelation function of the windowed signal by the autocorrelation function of the window [2]. This modification makes the results more accurate, noise-resistant, and robust, than methods based on cepstrum or combs, or the original autocorrelation methods.

4.2 Pre-processing

Due to errors occurring during the extraction of the DPs from the raw speech material or the extraction of the Pitch-contour, a pre-processing of the pitch-measurements is necessary to ensure a reliable and robust classification and to reduce the complexity of the analysis algorithm. Common errors are the appearing of gaps and/or steps in the pitch-contour. To process these errors a pre-processing loop is developed.

The extracted pitch-contours are smoothed using a one dimensional median filter with a window size of five sample points (SPs). Contours with an insufficient number of SPs (less than 10 SP) can be neglected. The Signal is too short to ensure a reliable assignment of the prototypes by Schmidt (cf. Table 1).

Next an examination on gaps and steps is performed. The processing takes place alternately, beginning with the detection of gaps. In this research paper gaps are defined as an interruption of the pitch-contour. A distinction is made between short gaps, which can be seen as small signal interruptions over only a few SPs and significant gaps, which divide the pitch-contour into two sections. For the first type of gaps a linear interpolation for the missing SPs is sufficient to generate a continuous waveform. For the second type, the signal is split into two sections and the respectively longest part is maintained for classification. These two types can be distinct via the ratio between the length of the gap Δt_{gap} and the total length of the signal Δt_{total}. If the ratio exceeds a value of $\frac{1}{3}$ i.e. the gap defines one third of the signal length, it can be assumed that the signal needs to be separated, see Figure 2.

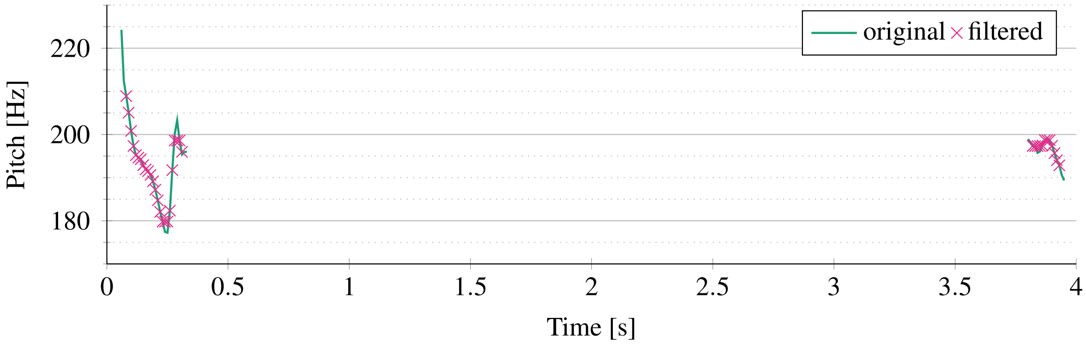


Figure 2 - Original and filtered signal of a measured signal with a gap.

The obtained pitch-counter is now examined for amplitude steps. An amplitude step corresponds with a rapid change in the pitch-value. This change can be up to 100Hz but is often split into two or three smaller steps with a lesser difference in the pitch-value, see Figure 3.

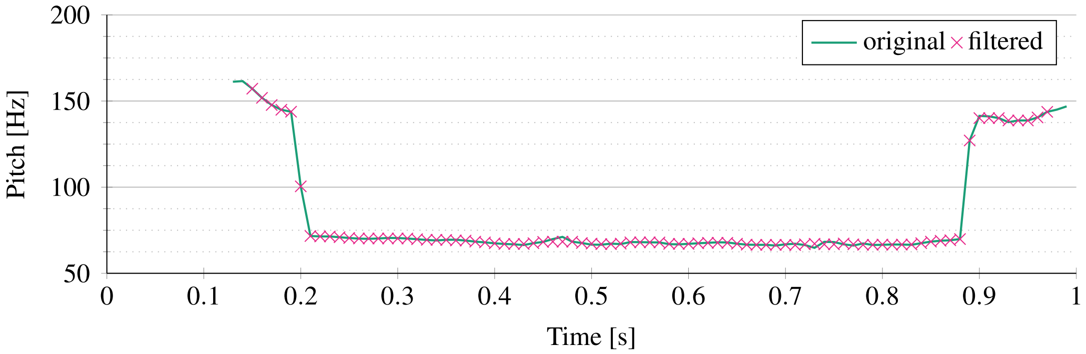


Figure 3 - Original and filtered signal of a measured signal with two steps.

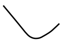
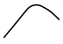
A problem is recognized in the differentiation between signals containing steps and strongly sloping ones. Therefore, two methods were developed. One considering the absolute distance $\Delta pitch$ between two consecutive SPs and the second regarding the relative ratio r_{step} between $\Delta pitch$ and the average change in pitch. It is assumed, that a step takes place either when $\Delta pitch > 10$ or $r_{step} > 3,5$ is satisfied. The further processing is performed analogously to the processing of gaps. The resulting signal sections need to be re-examined on their length to ensure a sufficient number of SPs. The precise details on the further functioning of the preprocessing loop can be found in [8].

4.3 Analysis

We distinguish between three types of prototype-courses: linear (DP-2, DP-3, DP-7), quadratic (DP-4, DP-5, DP-8, DP-9) and cubic ones (DP-1, DP-6). In addition, to the DPs mentioned in Table 1, two more formtypes (DP-8, DP-9) are considered, which were assumed frequently in a manual inspection of the pitch-contours (see Table 2). An evaluation of these additional formtypes, can be found in section 6.1.

For these two formtypes no statement about the functional use of the DP can be made, as the study only deals with the course of the Pitch-contour and not the identification of its meaning. To assign the course to one of the formtypes a regression of the pitch-contour is performed.

Table 2 - Additional formtypes based on a manual inspection of the pitch-contours.

Name	idealized pitch-contour
DP-8	
DP-9	

Depending on the coefficient of determination either a first, second or third order regression is performed. If the coefficient exceeds a value of 0.9 we assume that the regression function describes the real signal sufficiently, i.e. 90% of the pitch-values can be described using the regression function. The application of a higher order regression is not necessary. An exception is made for horizontal gradients with a low standard deviation (see [8]). Now considering the course of the regression function and its coefficient of determination a statement about the present form-function-prototype and its quality can be made.

Based on the slope of the first order linear regression line a tendency of the course can be determined. Depending on the tendency certain formtypes can be left out for the classification. It is possible to distinguish between rising, falling and horizontally extending pitch-contours. This allows an unambiguous assignment of a pitch-contour, sufficiently described by a linear regression function, to a linear prototype. In case of contours described by higher order regression functions the signal is divided into sections limited by their turning points. Based on length and slope of these sections the prototype can be assigned.

5 Evaluation

To ensure the reliability of the classification-algorithm a statement about how well the obtained prototypes match the original signals needs to be made. Moreover, the results need to be verified by conducting a manual labeling to attain an unprejudiced opinion on the pitch-contours.

5.1 Quality criterion for prototype assignment

To obtain a value for the measurability of the correspondence between the acquired form-function-prototype and the original pitch-contours, a quality criterion is determined. For this purpose the coefficient of determination has been found to be suitable. High values $(> 0,9)$ represent a good correspondence. Smaller values indicate a less reliable assignment of the prototype.

5.2 Course assessment of the discourse particles

To verify the results of the classification-algorithm a manual labeling of the pitch-contours is performed. For the investigation no information about the surrounding content or the speech material itself was given to the labelers. This permits an independent objective examination of the pitch-contours. In total 8 labelers were asked to assign either one of the seven form-function-prototypes by Schmidt, one of the two additional formtypes (DP-8, DP-9) or the option “no specification possible” to the pictured pitch-contour. Thereby the last option should only be considered if clearly none of the nine formtypes match.

6 Results

6.1 Classification-Algorithm

The pre-processing loop provides all appropriate pitch-contours to the analysis-algorithm. In this process 40 of the original 259 considered pitch-contours are discarded for classification, due to an insufficient number of SPs. The results of the 219 studied courses can be found in Table 3.

Table 3 - Resulting frequency distribution of the classification algorithm.

Label	**DP-1**	**DP-2**	**DP-3**	**DP-4**	**DP-5**	**DP-6**	**DP-7**	**DP-8**	**DP-9**
% Items	0.5	49.5	30.6	3.6	2.7	1.8	7.8	2.7	0.9

A majority of DP-2, DP-3 and DP-7 is detected. However, due to the small amount of available speech signals, only DP-2 and DP-3 define a clear majority. Usual for HCI is a decrease in partner-oriented signals, while the number of signals indicating a talk-organization, task-oriented or expressive function is increasing [5]. In that matter DP-3 is seen as partner-oriented and defines finalization signals. These not only include the termination of talk but additionally also expressions like "exhausted sigh" or exclamation of dismay, e.g. "oh dear". As the used LAST MINUTE Corpus implies the creation of a stressful situation, the increase of this formtype is not very surprising. In contrast to DP-3, DP-2 can be seen as talk-organizing like thinking or turn-taking/holding.

Looking at the additionally considered formtypes: Only a number of eight waveforms were assigned to DP-8 and DP-9. Leading to the assumption that these waveform are modifications of the prototypes according to Schmidt.

6.2 Manual labeling

The manual labeling was assessed via majority voting; if five or more labelers agreed on the assignment of a label, this assessment is used. Only for a number of 12 signals (4%), a majority could not be obtained, cf. Table 4. Analogously to Table 3 a majority of DP-2 and DP-3 can be seen. Overall a reduction of partner-oriented signals (DP-1, DP-3, DP-4, DP-6 and DP-7) can be detected.

Table 4 - Resulting frequency distribution of the manual labeling. The categories are according to Table 1, additionally used labels: NSP "no specification possible" and NM "no majority".

Label	**DP-1**	**DP-2**	**DP-3**	**DP-4**	**DP-5**	**DP-6**
% Items	/	53.0	23.7	5.0	4.6	0.9
Label	**DP-7**	**DP-8**	**DP-9**	**NSP**	**NM**	
% Items	3.2	2.3	1.4	5.5	0.5	

In total a consistency of the labeled functions and the classified prototypes was found in 79% of the assignments. Only a number of 46 measurements were assessed differently, including the previously mentioned 12 contours where the labelers disagreed on the assignment. For 13 more signals the assignment differed between DP-2 and DP-3/7. Thus only a disagreement in the slope of the course was found. It can be concluded that the labelers rather identify

optically perceived signals as a sloping than a linear prototype. Concerning the regression-order an increase of mismatch is found with an increase of order. In general this increase can be justified based on the different perception of the labelers and can be eliminated by an adjustment of the algorithm. Also taking the quality criterion into account a low coefficient of determination is frequent for a third order regression of the signal, as no higher order regression is conducted.
Further, the reliability of the labeling was established, using Krippendorff's Alpha (see [6]), to a value of 0.53, which indicates a moderate and appropriate reliability according to [7].

7 Conclusion

In this paper a rule-based classification algorithm has been developed, which correlates a specific intonation curve to one of the seven form-function-prototypes. As dataset, a sub-set of the LAST MINUTE corpus [10], which is used to investigate HCI in case of dialog barriers, is utilized. In order to unambiguously assign the signal to one of the prototypes according to [11], a pre-processing of the pitch-contour is required. During this pre-processing, typical measurement errors, such as gabs or steps, are being recognized and processed. Further the results of this pre-processing are approximated using regression analysis. A first, second or third order regression function is used, depending on the coefficient of determination. Depending on the contour of the regression function and its coefficient of determination a statement about the present form-function-prototype and the quality of the assignment can be made.
Using the developed classification algorithm, it is possible to assign each investigated DP of the given speech material to one of the seven form-function-prototypes. Thus it is possible to assign a functional meaning to the corresponding intonation curve of the DP. To verify the results of the classification algorithm we utilized a manual labeling with 8 labelers. In total a consistency of the results was shown in 79% of the given DPs. The reliability of the labeling was established using Krippendorff's Alpha to a value of 0.53.

References

[1] ALLWOOD, J., J. NIVRE and E. AHLSÉN: *On the Semantics and Pragmatics of Linguistic Feedback*. Journal of Semantics, 9(1):1–26, 1992.

[2] BOERSMA, P.: *Accurate Short-Term Analysis of the Fundamental Frequency and the Harmonics-To-Noise Ratio of a Sampled Sound*. Proceedings of the Institute of Phonetic Sciences, 17:97–110, 1993.

[3] BOERSMA, P.: *Praat, a system for doing phonetics by computer*. Glot International, 5(9-10):341–345, 2001.

[4] BOLINGER, D.: *Intonation and its Uses: Melody in Grammar and Discourse*. Stanford University Press, London, 1989.

[5] FISCHER, K., B. WREDE, C. BRINDÖPKE and M. JOHANNTOKRAX: *Quantitative und funktionale Analysen von Diskurspartikeln im Computer Talk*. International Journal for Language Data Processing, 20(1-2):85–100, 1996.

[6] KRIPPENDORFF, K.: *Content Analysis: An Introduction to Its Methodology*. SAGE Publications, Thousand Oaks, CA, USA, 3 ed., 2012.

[7] LANDIS, J. R. and G. G. KOCH: *The measurement of observer agreement for categorical data.* Biometrics, 33:159–174, 1977.

[8] LOTZ, A.: *Differentiation von Form-Funktions-Verläufen des Diskurs Partikels "hm" über unterschiedliche mathematische Herangehensweisen.* Masterarbeit, Otto-von-Guericke Universität Magdeburg, 2014.

[9] PRYLIPKO, D., D. RÖSNER, I. SIEGERT, S. GÜNTHER, R. FRIESEN, M. HAASE, B. VLASENKO and A. WENDEMUTH: *Analysis of significant dialog events in realistic human–computer interaction.* Journal on Multimodal User Interfaces, 8(1):75–86, 2014.

[10] RÖSNER, D., R. FRIESEN, M. OTTO, J. LANGE, M. HAASE and J. FROMMER: *Intentionality in Interacting with Companion Systems – An Empirical Approach.* In *Human-Computer Interaction. Towards Mobile and Intelligent Interaction Environments*, vol. 6763 of *LNCS*, pp. 593–602. Springer, Berlin, Heidelberg, 2011.

[11] SCHMIDT, J. E.: *Bausteine der Intonation.* In *Neue Wege der Intonationsforschung*, vol. 157-158 of *Germanistische Linguistik*, pp. 9–32. Georg Olms Verlag, Hildesheim, Germany, 2001.

[12] SIEGERT, I., D. PRYLIPKO, K. HARTMANN, R. BÖCK and A. WENDEMUTH: *Investigating the Form-Function-Relation of the Discourse Particle "hm" in a Naturalistic Human-Computer Interaction.* In BASSIS, S. and A. E. AMD FRANCESCO CARLO MORABITO (eds.): *Recent Advances of Neural Network Models and Applications*, vol. 26 of *Smart Innovation, Systems and Technologies*, pp. 387–394. Springer, 2014.

[13] WENDEMUTH, A. and S. BIUNDO: *A Companion Technology for Cognitive Technical Systems.* In *Cognitive Behavioural Systems*, vol. 7403 of *LNCS*, pp. 89–103. Springer, Berlin, Heidelberg, 2012.