Modeling affected user behavior during human-machine interaction

Bogdan Vlasenko, Ronald Böck and Andreas Wendemuth

Cognitive Systems, IESK, Otto-von-Guericke Universität,
Magdeburg, Germany
bogdan.vlasenko@ovgu.de

Abstract

Spoken human-machine interaction supported by state-of-the-art dialog systems is becoming a standard technology. A lot of effort has been invested for this kind of artificial communication interface. But still the spoken dialog systems (SDS) are not able to provide to the users a natural way of communication. Most part of the existing automated dialog systems is based on a questionnaire based strategy with sentence by sentence confirmation request. This paper addresses aspects of design and implementation of user behavior models in dialog systems for frustration detection and user intention recognition, aimed to provide naturalness of human-machine interaction. We overview our acoustic emotion classification, robust affected automatic speech recognition (ASR) and user emotion correlated dialog management. A multimodal human-machine interaction system with integrated user behavior model is created within the project ”Neurobiologically Inspired, Multimodal Intention Recognition for Technical Communication Systems” (NIMITEK). Currently the NIMITEK demonstration system provides a technical demonstrator to study user behavior modeling principles in a dedicated task, namely solving the game “Towers of Hanoi”. During communication with our demonstration system users are free to use natural language. By using natural language understanding and intention recognition modules the system provides task control management. To show the Spoken Dialog System performance uprating with the user’s behavior correlated dialog management we present results of the NIMITEK demonstrator’s usability test. After having analyzed the results of the usability test, we find out that our system provides more cooperative computer machine interaction and decreases interaction time required to complete the puzzle.

Index Terms: Emotion Recognition, User Behavior Adaptive Dialog Management.

1. Introduction

The importance of human behavior based dialog strategies in human-machine interaction (HMI) lies in existing limitations of automatic speech recognition technology. Current state-of-the-art Automatic Speech Recognition (ASR) approaches still cannot deal with flexible, unrestricted user’s language and emotional prosody colored speech [1]. Therefore, problems caused by misunderstanding a user who refuses to follow a predefined, and usually restricting, set of communicational rules seems to be inevitable.

In the domain of human-machine interaction [2], we witness the rapid increase of research interest in affected user behavior. However, some aspects of affected user behavior during HMI still turns out to be a challenge for developers of Spoken Dialog Systems (SDS). User behavior state analysis is one of the major challenges in the development of reliable human-machine interfaces. There are the universal categorical emotional states (anger, happy, sadness, etc.), prevalent in day-to-day communication scenarios. Recognizing such emotional states can help adjust system responses so that the user of such a system can be more engaged and have a more effective interaction with the system [3].

The primary aim of this paper is to present our implementation of adaptive dialog management in the NIMITEK [4] prototype spoken dialog system for supporting users while they solve the “Tower of Hanoi” puzzle. Further we provide results of the NIMITEK prototype SDS usability test.

2. Human-Machine Interface

Multimodal Human-Machine Interfaces (MHMI) have recently become a new feature for different applications [5]. We describe one possible MHMI architecture for a Spoken Dialog System, see Figure 1. For the last three years in a transdisciplinary cooperation [4] we developed the NIMITEK multimodal human computer interaction system. Humans employ several output modalities (mimics, speech, prosody) to communicate with a computer. Speech (in particular the prosody within speech) and mimics together serve as a multimodal emotion source. The current system has two independent acoustic and mimic based emotion classifiers. In this article we describe results of usability tests of the NIMITEK demonstrator with an acoustic based emotion classifier. Emotions are classified [6] into one of two possible emotion classes (neutral and anger).

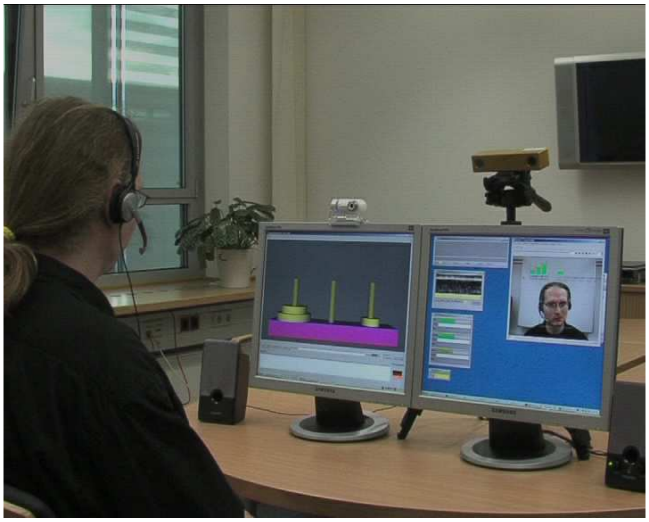

Figure 1: Multimodal human-machine spoken dialog system, NIMITEK Demonstrator.

Moreover, taking also into account the history of the previous system-user interaction and current user's emotional state, intentional levels are classified such as *cooperative, explorative or destructive.* The latter applies to users who wish to drive the system into dead ends, e.g. by deliberately mispronouncing words or giving contradictory commands. The NIMITEK system updates the emotional state of the user utterance per utterance. With the recognized commands and the emotional and intentional state, the task controller is driven.

3. User friendly spoken dialog management

In this section, we describe the spoken dialog system incorporated in the NIMITEK prototype system. Modeling of a user-machine interaction is represented in Figure 2. The emotion classifier uses three modalities: emotional prosody within spoken communication, literal meaning of user's utterances and user mimics. For the current usability test we test the NIMITEK prototype with speech based emotion classification. The output of the emotion classifier is the detected emotional state (neutral or anger) of the user. The possible textual meaning of the user's utterances is delivered also to the natural language understanding module. This module detects the command and forwards it:

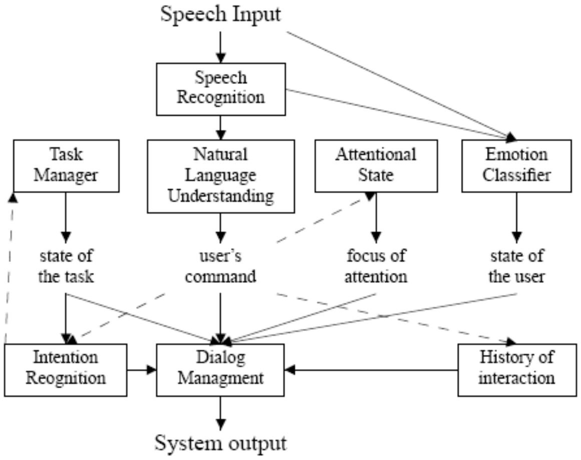

Figure 2: NIMITEK prototype Spoken Dialog System structure.

- *to the attentional state module for updating the focus of attention,*
- *to the history of interaction module to save current values of other interaction features and process context dependent users commands,*
- *to the intention recognition module for defining the user's intention based on his last command and current state of task,*
- *from intention recognition to the task manager module (including the graphical platform) for performing the detected command, update of the state of the task, and appropriate graphical display,*

A new entry is added to the history of interaction, containing: updated state of the task, detected command, current focus of attention, detected state of the user. For delimitation of types of frustrations: task related and communication incomprehension, we take into account the current state of the focus and history of interaction. When the user's game manipulations are far away from solving the ”Tower of Hanoi” task the system indicates a task related frustration. Then, if needed, the system provides a user support according to the current state of the interaction, emotional and intentional state of the user. Providing support to the user is represented in Figure 3.

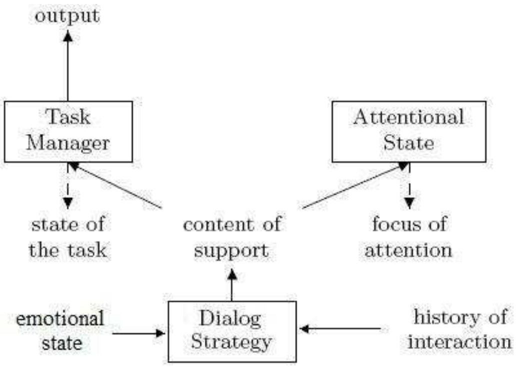

Figure 3: Providing support.

Our dialog management designed to support the user addresses the negative user state on two tracks: (i) to help a frustrated user to overcome problems that occur in the interaction, and (ii) to motivate a discouraged or apathetic user. The recognized intention of the user determines the direction of support: for a cooperative user, the next logical step is explained; for an explorative user, comprehensive coverage of possible steps is given; for a destructive user, the limitations of the next steps are explained. More technical details can be found in Gnjatović et al. [2, 7]. Generally, the support information may contain a proposed move, an audio message and an animation. In the case when support contains only an audio message or an animation, this information is delivered to the task manager module for appropriate display. If support contains also a proposed move, this information is sent: (i) to the task manager module for a performance of the proposed command and an update of the state of the task, (ii) to the attentional state module for an update of the focus of attention.

4. Affected Speech Processing

4.1. Databases

To train German monophones we used The Kiel Corpus of Read Speech. The Kiel Corpus is a growing collection of read and spontaneous German which has been collected and labeled segmentally since 1990. The Kiel Corpus comprises over four hours of labeled read speech of 26 female and 27 male speakers. The training set contain 2872 sentences.

Relative to an emotional HMM phoneme model, we decided for the popular studio recorded Berlin Emotional Speech Database (EMO-DB) [9]. 10 (5f) professional actors speak 10 German emotionally undefined sentences. For evaluation we used just ”anger” and ”neutral” samples. As result we have 205 phrases, which are marked as min. 60% natural and min. 80% assignable by 20 subjects. In average 96.2% accuracy is reported for a human perception test.

4.2. Wizard-of-Oz experiment

This research is essentially supported by the NIMITEK corpus of affected behavior in human-machine interaction collected within the reported research. It contains 15 hours of audio and video recordings produced during a Wizard-of-Oz experiment specially designed to induce emotional reactions. Technical details are reported in [7]. Ten healthy native German speakers (7 female, 3 male) in the age from 18 to 27 (mean 21.7) participated in the experiment. None of them had educational background or user experience related to state-of-the-art spoken dialog systems. Within this data aprox. 3 hours are related to the "Towers of Hanoi" game.

Gnjatović et al. [7] analyzed (all) 6798 commands from the NIMITEK corpus, enabling the system to process users commands of different syntactic forms: *elliptical commands, verbose commands (i.e., the commands that were only partially recognized by the speech recognition module), and context dependent commands*. We find out that users do not follow a predefined grammar.

4.3. Feature Extraction

We find out that it is appropriate to use the same features for speech and emotion recognition [10]. Speech input is processed using a 25ms Hamming window, with a frame rate of 10ms. We employ a 39 dimensional feature vector per each frame consisting of 12 MFCC and log frame energy plus speed and acceleration coefficients. Cepstral Mean Substraction (CMS) and variance normalization are applied to better cope with channel characteristics.

4.4. Affected Speech Recognition

For real time Automatic Speech Recognition (ASR) within the Spoken Dialog System, we used ATK and HTK [11]. Monophones ASR models are designed by training three emitting state Hidden Markov Models (HMM) with 16 Mixtures of Gaussians built for each phoneme. We are using a short version of German SAMPA which includes 36 phonemes. To reach a high performance on affected speech (as a most hard to recognize variety of spontaneous speech) recognition we applied MAP adaptation for the Kiel trained monophones HMM set with EMO-DB samples. Technical details are reported in [10].

To create an optimal language model we analyzed the NIMITEK corpus. We do not have enough material related to the "Tower of Hanoi" task in our corpus for bi-gram modeling. As a result we decide in favor of a "Tower of Hanoi" game related grammar based language model. For all possible commands constructions in the NIMITEK corpus we generate the grammar, which is afterwards being converted to a language model lattice. The garbage word model encapsulates possible Out of Vocabulary words.

4.5. Natural Language Understanding

There are few possible types of commands during playing the "Towers of Hanoi" game: *define a ring, define a direction of movement, actions with the ring and request for system support.* To handle context dependent commands (e.g., undo, move the next disk, etc.) the NIMITEK system processes history of interaction, previous focus of attention and current user's command.

4.6. Emotion Classification

For real time emotion classification within speech we used modified HTK and ATK. In our current version of the NIMITEK demonstrator we integrated a phoneme level emotion classifier.

We use a simple conceptual model of dynamic emotional state recognition on phoneme level analysis: the full list of 36 phonemes is modeled for neutral and anger emotion speaking style, independently. As a result 2 x 36 = 72 phoneme emotion models are trained [**?**]. In case of emotion challenge we have 72 phoneme emotion models for two emotional classes evaluation.

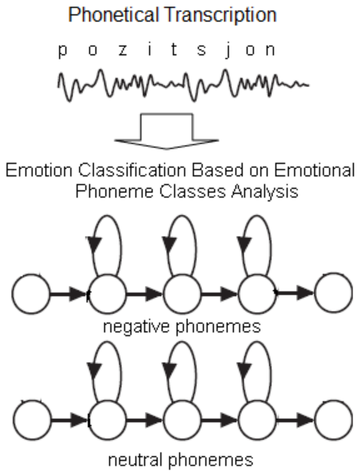

Figure 4: *Phoneme level emotion recognition.*

Emotional phonemes are modeled by training three emitting state HMM models with 16 Mixtures of Gaussians. There is not enough material in selected part of EMO-DB to train a robust monophone models. Hence, in contrast to the previous models [10] we are using Kiel trained monophones models as a background HMM model. The HTK toolkit was used for MLLR adaptation of background model on two phoneme emotion subsets: neutral and anger. Neutral and anger samples from EMO-DB are used for adaptation.

In case of emotion recognition we are using an ASR engine adapted for affective speech to recognize on word level as a start point. After this we are generating possible emotional phonetic transcriptions for sensible utterances by using an emotional phoneme set. In our case, two transcriptions for neutral and anger speaking styles are generated. To choose the most appropriate emotional state of the phonetic transcription for the recognized sentence the EM algorithm is used.

4.7. Results on acoustic emotion classification

Evaluation of phoneme level emotion classifier runs on the EMO-DB database in LOSO manner to address speaker independence (SI), as required by the NIMITEK demonstrator [10]. We achieved recognition rates of up to 97.3% for two classes emotion (neutral and anger), in comparison to 96.2% as a result of a human perception test for a two emotional classes subset. In case of spontaneous speech emotion classification performance will be lower.

5. NIMITEK spoken dialog system usability test

5.1. Evaluation description

For our experiments we established two types of SDS systems: first, a complex one with behavior based dialog management (DM) with emotion adaptive strategy and affective speech adapted ASR models. Second, the simple SDS has ASR models trained on a large neutral speech database. Otherwise, both systems are identical.

For the usability test we hired 8 students (4 female and 4 male). Half of them played the Tower of Hanoi game with complex SDS including behavior based DM strategy and the remaining testers used the simple SDS system with standard support, i.e. repeating the rules of the game or asking for repeating the command. The complex SDS varies the answers depending on the behavior of the user like asking for a specific peg or disk, repeating the rules, or giving general hints.

Also within the human-machine interaction users are able to follow the ASR output. When the garbage model was not able to encapsulate out of vocabulary words the user was able to see misrecognized system perceptible commands. We expect that users will try to adapt their commands vocabulary to contribute to the right system reaction.

5.2. Results

The experimental results of the system evaluation are presented in Table 1. During evaluation we recorded interaction time, number of utterances per interaction, and time and amount of utterance required for system perceptible commands vocabulary adaptation. Comparing the numbers of utterances which are necessary to solve the puzzle the behavior based DM system performs better. On average using the simple DM system the user needs ca. 18 utterances more to finish the game.

Table 1: Results of the usability test.

Trial	Behavior based DM				Simple DM			
	Utter	Time	Adapt		Utter	Time	Adapt	
			Utter	Time			Utter	Time
1.	34	05:43	1	00:00	44	05:40	1	00:00
2.	31	03:37	10	01:36	61	06:05	30	03:43
3.	34	02:44	10	01:04	81	11:48	10	01:51
4.	55	04:17	1	00:00	41	04:07	7	00:52
Mean	38.5	04:05	5.5	00:40	56.75	06:55	12	01:37

Considering the overall time which includes pauses and the system support, the behavior based SDS shows the better average results (04:05 vs. 06:55 minutes absolute talk time). Also with user behavior correlated SDS, users are more considerate to the ASR output. As a result they are adapting their commands vocabulary faster (00:40 vs. 1:37 minutes).

6. Conclusion

Within human-machine communication frustration situations, our SDS provides comprehensive help and exhaustive recommendation in context of the current state of the task. A User behavior based dialog system built upon emotion recognition within speech in combination with affected speech adapted ASR models decreases interaction time by 40%. During usability tests we find out that an affected speech adapted ASR model provides better spontaneous speech recognition performance in real applications. At the same time user behavior based dialog management stimulates the user for a more cooperative interaction with the computer. As a result the user's commands vocabulary adaptation time is decreased by 59%.

Emotions and intentions play a central role in human-machine communication. The research stimulation in the NIMITEK project helps to provide a close to natural way of human-machine interaction.

7. ACKNOWLEDGMENTS

The work has been conducted in the framework of the NIMITEK project (http://wwwai.cs.uni-magdeburg.de/nimitek), within the framework of the Excellence Program Neurowissenschaften of the federal state of Sachsen-Anhalt, Germany
(FKZ: XN3621A/1005M).

This project is associated and supported by the Transregional Collaborative Research Centre SFB/TRR 62 ”Companion-Technology for Cognitive Technical Systems” funded by the German Research Foundation (DFG).

8. References

[1] C.-H. Lee, “Fundamentals and technical challenges in automatic speech recognition,” in *SPECOM2007*, Moscow, Russia, 2007, pp. 25–44.

[2] M. Gnjatović and D. Rösner, “Adaptive dialogue management in the nimitek prototype system,” in *4th IEEE PIT'08*, Kloster Irsee, Germany, 2008, pp. 14–25.

[3] B. Schuller, D. Seppi, A. Batliner, A. Maier, and S. Steidl, “Towards more reality in the recognition of emotional speech,” in *IEEE ICASSP 2007*, vol. 2, Honolulu, Hawaii, USA, April 2007, pp. 941–944.

[4] A. Wendemuth, J. Braun, B. Michaelis, F. Ohl, D. Rösner, H. Scheich, and R. Warnemünde, “Neurobiologically inspired, multimodal intention recognition for technical communication systems (NIMITEK),” in *4th IEEE Tutorial and Research Workshop PIT 2008*, Kloster Irsee, Germany, 2008.

[5] M. Pantic and L. Rothkrantz, “Toward an affect-sensitive multimodal human-computer interaction,” in *Proccedings of the IEEE, vol. 91*, 2003, pp. 1370–1390.

[6] B. Schuller, B. Vlasenko, R. Minguez, G. Rigoll, and A. Wendemuth, “Comparing one and two-stage acoustic modeling in the recognition of emotion in speech.” in *ASRU 2007*, Kyoto, Japan, 2007.

[7] M. Gnjatović, “Adaptive dialogue management in human-machine interaction,” Ph.D. dissertation, Universität Magdeburg, 2009.

[8] A. P. Simpson, K. J. Kohler, and T. Rettstadt, *The Kiel Corpus of Read/Spontaneous Speech*. Arbeitsberichte, Nr. 32, (Universität Kiel) IPDS (Kiel), 1997.

[9] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss, “A database of german emotional speech,” *Interspeech*, pp. 1517–1520, 2005.

[10] B. Vlasenko and A. Wendemuth, “Processing affected speech within human machine interaction,” in *INTERSPEECH 2009*, Brighton, UK, 2009.

[11] S. Young, G. Evermann, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, *The HTK Book*, Cambridge University Engineering Department, 2002.