An adaptive recovery strategy for handling miscommunication in human-machine interaction

Milan Gnjatović, Milana Bojanić, Branislav Popović, and Vlado Delić

Abstract — **This paper introduces an approach to handling miscommunication in human-machine interaction (HMI). Two aspects are considered: dealing with miscommunication (i.e., the recovery strategy) and detecting miscommunication (i.e., detecting changes of the user's behavior). The first part of the paper discusses the recovery strategy implemented in two prototype spoken dialogue systems. It is illustrated by several examples that were selected so that they and their combinations cover a wide range of different interaction situations caused by inaccurate speech recognition. The second part of the paper reports research in progress – an integrated classification of changes of the user's behavior from prosodic cues, linguistic cues, and cues from facial expressions.**

Keywords — **Spoken human-machine interaction, miscommunication, recovery strategies, user's behavior.**

I. Introduction

Miscommunication is a frequent and natural phenomenon in spoken communication and appears to be unavoidable [1]. In the context of spoken human-machine interaction (HMI), miscommunication can occur on different levels [2, p. 131]: on the conversational level (e.g., the user's utterance falls outside of the system's functionality), on the intentional level (e.g., the user's utterance falls outside of the system's semantic grammar), on the signal level (e.g., inaccurate speech recognition), etc. It is clear that the state-of-the-art automatic speech recognition approaches still cannot deal with flexible, unrestricted users' language (cf. [3]), although some of them implement barge-in capability, give a useful confidence scoring [4, p. 73], and enable features such as either detection or rejection of out-of-vocabulary speech to improve operational performance. Also, it is not reasonable to expect that users will always behave "cooperatively" and produce utterances that fall within the application's domain, scope and grammar. Still, left unmediated by better error awareness and recovery mechanisms, miscommunication *may severely limit the effectiveness and the naturalness of the interaction* [2, p. 124]. Therefore, in order to achieve a habitable language interface, there is an essential need for both models of interaction and dialogue strategies that support the user to overcome problems that occur due to miscommunication.

The aim of this paper is to introduce an approach to handling miscommunication in HMI. Our analysis of spontaneously produced utterances in HMI (cf. [5, p. 45-53]) shows that they take different syntactic forms, including: syntactically very simple utterances (e.g., elliptical or minor), ungrammatical constructions, context dependent utterances, affected speech, meta-language, etc. In [5], [6] and [7], we introduce and discuss a model of attentional state in HMI that facilitates processing of spontaneously uttered users' commands. The main advantage of this modeling is that, instead of predefining a grammar for accepted commands, we allow more flexible formulation of users' commands. The implementation of this model in the prototype spoken dialogue systems was demonstrated to work well for different syntactic forms of users' commands: elliptical commands, verbose commands (i.e., the commands that were only partially recognized by the speech recognition module), and context dependent commands.

In this paper, we extend this work. Two aspects of handling miscommunication in HMI are considered:

(1) Dealing with miscommunication – Section II discusses an adaptive recovery strategy for handling miscommunication that is based on the model of attentional state, and illustrates its implementations in two prototype spoken dialogue systems.

(2) Detecting miscommunication – Changes in both the user's behavior and his/her voice may also indicate troubles in communication, e.g., miscommunication. Section III introduces an approach to detection of miscommunication based on integrated detection from both signal based as well as content based recognition of changes in the user's behavior.

II. Dealing with miscommunication

The recovery strategy discussed in this section is implemented in two prototype systems: the NIMITEK system [5], [6] and the Contact system [4, p. 76]. The NIMITEK spoken dialogue system supports users while they solve problems in a graphics system. The dedicated task – that is also considered in the examples provided in

Milan Gnjatović (the corresponding author), Milana Bojanić, Branislav Popović and Vlado Delić are with the Department of Power, Electronics and Communications Engineering, Faculty of Technical Sciences, University of Novi Sad, Trg Dositeja Obradovića 6, 21000 Novi Sad, Serbia (381-21-4852533).

E-mail: milan.gnjatovic@alfanum.co.rs, bojanic.milana@yahoo.com, brvep@sezampro.rs, vdelic@uns.ac.rs

this section – is the *Tower of Hanoi* puzzle. The system Contact is intended to be used by the visually impaired. It reads aloud selected textual contents (e.g., news, articles, etc.) from various newspapers and websites over the telephone line.

A. The signal level

On the signal level, miscommunication is caused by inaccurate speech recognition. Non-recognition or misrecognition of parts of the user's utterance that carry information about the propositional content may significantly affect the interaction. The confidence scoring can be used as an indication of a problem in spoken natural language HMI [4]. The implemented recovery strategy encapsulates two ideas: (a) the conversation should be advanced in spite of miscommunication, and the system should support the user to overcome problems that occur due to miscommunication, and (b) support should be dynamically adapted according to the current state of the interaction.

In order to make these statements more clear, let us assume that a valid user's command is not correctly recognized by the automatic speech recognition module. Due to an actual error in speech recognition, the dialogue management module may interpret the command as a valid – although misinterpreted – command, an illegal command, a semantically incorrect command, or an unrecognized command. For each of these classes, the system provides, if needed, support.

For the purpose of illustration and without loss of generality, let us also assume that the user uttered "the second disk on the first peg". We discuss the system's behavior in several cases of inaccurate speech recognition.

Case 1: A part of the command is correctly recognized, and at least one phrase that relates to the propositional content can be derived, while the rest of the command is not recognized. For example, the textual version of the command outputted from the speech recognizer may be "<not recognized> on the first peg", where <*not recognized*> means either that the confidence scoring was under an acceptable level or that out-of-vocabulary words were detected. In this case, depending on the state of the task (i.e., which disk is currently selected), the interpreted move (i.e., moving the selected disk on the first peg) may be valid and thus performed, or illegal (i.e., not according to the rules of the puzzle). In the former case, if the performed move was not the move that the user instructed, he has a possibility to instruct an undo command. In the latter case, if the move was interpreted as illegal, appropriate Task-Support is provided by the system. The aim of Task-Support is to help the user to understand the given task, e.g., explaining the rules of the game, proposing the next correct move, etc.

Case 2: The command is not recognized, or a part of the command is correctly recognized, but no phrase that relates to the propositional content can be derived. For example, the textual version of the command outputted from the speech recognizer may be "<not recognized> disk <not recognized> peg". According to the definition of the dialogue strategy, when an unrecognized command is detected, Interface-Support is provided. The aim of Interface-Support is to help the user to formulate a command, e.g., the system may guide the user to formulate a command by stating iterative questions: which disk should be selected, where to move the selected disk, etc.

Case 3: A part of the command or the whole command is misrecognized. An example of the textual version of the command outputted from the speech recognizer may be "the second disk on the first disk". This will be classified as a semantically incorrect command. Another example is "the second disk on the second peg" which – depending of the current state of the interaction – may be classified as a valid command or an illegal command. Discussion of these examples is similar as above. If the command is classified as a semantically incorrect command or as an illegal command, the appropriate support is provided. In the case when the command is classified as valid, it is performed, although the user did not instruct that particular command. However, even then, if the performed move drew the state of the task back from the final solution of the puzzle, support will be provided. Only in the case when a misrecognized command is interpreted as a valid command that pushes the state of the task towards the final solution, the misrecognition will not be detected by the system. However, this is not a critical oversight, since it advances the conversation.

B. The conversational and intentional levels

Miscommunication is not only caused by word recognition errors (cf. [2]). On the conversational level, miscommunication occurs when the user's utterance falls outside of the application's domain and scope. These utterances are classified as unrecognized by the system. In such cases, the system provides Interface-Support, as discussed above.

On the intentional level, miscommunication occurs when the user's utterance falls outside of the system's semantic grammar (e.g., the user utters "revolve", which is not in the system's grammar, instead of "rotate", which is in the system's grammar). Here, we summarize some advantages of our approach to processing the commands that can reduce the level of miscommunication on the intentional level.

(1) The user can utter diverse phrases in order to address the same entity. For example, the focus instance smallest disk can be referred to as "smallest disk", "first disk", etc.

(2) The user can change the order of phrases in utterance. Given sets of phrases that may be used, the system automatically derives the propositional content from the user's command by detecting phrases that relate to certain entities in the domain of interaction. The order of phrases in utterance is not important. For example, the utterances "the smallest disk on the third peg" and "on the three – the smallest disk" are interpreted by the system in the same way.

(3) The user can use *wrapper* expressions (cf. [8]) that fall outside of the system's semantic grammar. Wrapper expressions do not relate to the propositional content of the

user's commands. For example, wrappers that represent expressions of politeness in the following user's utterances are given in italic: "The middle disk *please* on the number two" and "*I would like to* put the smallest disk on the three". To interpret the propositional content of the user's commands, the system derives only phrases that relate to entities in the domain of interaction, while it ignores wrappers. Finally, it should be mentioned that wrappers may carry affect information, so they are important for recognition and tracking of the user's emotional state from linguistic information, as discussed in Section III, Subsection B

III. Detecting miscommunication

Detecting miscommunication is also an important aspect of a recovery strategy. In [9, p. 103], it is discussed that non-understandings are generally easier to detect, misunderstandings may be detected given appropriate verification strategies, but misinterpretations may only become apparent at a later point in the dialogue, if at all. Our point of departure for detecting miscommunication is that the user's behavior reflects the state of the interaction [10, p. 198]. If there are no troubles in the communication, the user behaves neutral and is not engaged emotionally. Otherwise, the user's behavior changes accordingly (e.g., overt signaling of emotions, etc.). We combine three knowledge resources within an integrated classification of changes of the user's behavior: prosodic cues, linguistic cues, and cues from facial expressions. In contrast to the previous section that discusses the implemented recovery strategy, this section reports research in progress.

A. *Prosodic features*

To detect changes in the user's behavior, we consider some of the high-level prosodic features which were used in previous studies for emotion recognition [11], stress detection [12] and as cues to speech recognition failures [13]. They cover a wide range of acoustic changes. These features are:

- pitch (incl. minimum value, maximum value, range, mean, standard deviation),
- energy (incl. mean, standard deviation, range),
- speech rate (incl. duration of voiced segments and unvoiced segments).

Pitch, or fundamental frequency, depends on the tension of vocal folds and the subglottal air pressure, and, thus, carries information about the speaker's emotional state. Mean pitch value and its standard deviation may be used as good indicators of speech under stress when compared to neutral conditions [12]. For example, "angry" speech has an increase in mean pitch and wide range compared to neutral style. On the contrary, "sadness" typically has lower pitch mean and narrow range.

Energy, often referred to as intensity of the speech, is related to the arousal level of emotions in speech. The variation of energy of utterances can be used as a significant indicator for various speech styles [14]. Emotional states with high activation levels such as anger, surprise and happiness have higher energy, while sadness and disgust have lower energy.

The rate of speech may also indicate changes of the user's state. A speech segment is considered to be unvoiced if its fundamental frequency is zero. Segments with non-zero fundamental frequencies are voiced. Duration of unvoiced segments (i.e., pauses in the speech) and duration of voiced segments are parameters of changes in speaker's emotional state and also parameters of user's hyperarticulation and misrecognition by the system [14]. Also, it is useful to consider ratio of duration of voiced/unvoiced segments which has more discriminating power than these durations separately [15].

To address the inter-user variability, the proposed statistics are calculated at the utterance level and then normalized around the neutral voice of the user. At the start of the conversation, the system evaluates the user's neutral voice features to get information about the user's neutral speaking style. For example, one user might normally speak with a higher level of energy, while the other user might speak with a lower level of energy. The system tends to learn about the neutral speaking style of each user, in order to be able to recognize relative changes of his/her speaking style in the course of the interaction. More details on automatic emotion recognition in speech can be found in [16].

B. *Linguistic structures*

In addition to prosodic features, we consider also different linguistic features that may carry affect information. We primarily investigate a typology of users' utterances and sequences of users' utterances that indicate the emotional state of the user [17].

On the level of the user's utterance, emotional states may be recognized by detecting key words and phrases in user's utterances (e.g., "stupid", "super", "my God", "help", curse words, etc). Emotional expression may also map over a range of mutually related dialogue acts. Therefore, we consider also lexical cohesive agencies. Here we illustrate some of them:

Anaphoric cohesion. One form of anaphoric cohesion is ellipsis-substitution. A typical example is given in the following dialogue fragment:

USER: Can I **slide** the eight leftward?

SYSTEM: Yes.

USER: Then **do** it!

In his last turn, the user substitutes the verb *slide* with the more general verb *do*. The meaning of this substitution is not co-referential; it indicates that the system did not perform the instructed action and that there is a potential problem in the communication.

Lexical cohesion. Two forms of lexical cohesion often indicate problems in communication. On the lexical level, we consider repetitions. The user may choose to repeat his utterance when he believes that the system did not understand him, e.g., "Diagonally upward … I said diagonally upward". On the semantic level, we consider reformulations. For example, in the following sequence the user unsuccessfully tries to instruct the system to select a figure represented on the screen: "Next, we take the rhombus … A four-sided figure … The remained four-sided figure". He believes that his initial instruction falls outside of the system's semantic grammar, so he reformulates it in order to overcome the miscommunication.

Further lexical cues that we are going to consider are: sequences of questions, negation, and modal particles.

C. *Video features (Future research)*

Facial expressions are also an important resource for detecting changes in the user's behavior. Our future research includes two important questions that we aim to address. First, we investigate what is the appropriate set of features that we need in order to be able to track changes in the user's behavior, and what kind of information we get from such a set of features. At the moment, we primarily focus on lip motion tracking (cf. [18]). The second research question is how to fuse results of audio and video recognition in order to improve the accuracy and robustness of the recognition system.

IV. Conclusion

This paper introduced an approach to handling miscommunication in HMI. We considered two aspects of this approach: dealing with miscommunication and detecting miscommunication. With respect to the former aspect, the paper discussed the recovery strategy implemented in two prototype spoken dialogue systems. It addresses the miscommunication on the signal level, the conversational level, and the intentional level. The strategy is illustrated by several examples.

With respect to the latter aspect, we reported research in progress. Detection of miscommunication is based on the assumption that the user's behavior reflects the state of the interaction. Thus, we primarily considered the research questions of detecting negative changes in the user's behavior. We reported the research on an integrated classification of changes of the user's behavior from prosodic cues, linguistic cues, and cues from facial expressions.

Acknowledgments

The work described in this paper was supported in part by the Ministry for Science and Technological Development of the Republic of Serbia, within the project TR11001: "Human-Machine Speech Communication".

References

[1] M. McTear, I. O'Neill, P. Hanna, and X. Liu. Handling errors and determining confirmation strategies - an object-based approach. *Speech Communication*, 45(3):249–269, 2005.

[2] D. Bohus and A. Rudnicky. Sorry, I Didn't Catch That! An Investigation of Non-Understanding Errors and Recovery Strategies. In L. Dybkjær and W. Minker, editors, *Recent Trends in Discourse and Dialogue*, volume 39 of Text, Speech and Language Technology, pages 123–154. Springer, 2008. ISBN 978-1-4020-6820-1.

[3] C.-H. Lee. Fundamentals and Technical Challenges in Automatic Speech Recognition. Keynote paper at *XII International Conference "Speech and Computer" (SPECOM'2007)*, ISBN 6-7452-0110-x, pages 25–44. Moscow, Russia, 2007.

[4] V. Delić. A Review of R&D of Speech Technologies in Serbian and their Applications in Western Balkan Countries. Keynote paper at *XII International Conference "Speech and Computer" (SPECOM'2007)*, ISBN 6-7452-0110-x, pages. 64–83. Moscow, Russia, 2007.

[5] M. Gnjatović. *Adaptive Dialogue Management in Human-Machine Interaction*. Verlag Dr. Hut, München, ISBN 978-3-86853-189-3, 2009.

[6] M. Gnjatović and D. Rösner. Adaptive Dialogue Management in the NIMITEK Prototype System. In Lecture Notes In Artificial Intelligence (LNAI); Vol. 5078, *Proc. of the 4th IEEE Tutorial and Research Workshop Perception and Interactive Technologies for Speech-Based Systems (PIT'08)*, pages 14–25, Kloster Irsee, Germany, ISBN 978-3-540-69368-0, 2008.

[7] M. Gnjatović and D. Rösner. An approach to processing of user's commands in human-machine interaction. In *Proc. of the 3rd Language and Technology Conference (LTC'07)*, pages 152–156, Adam Mickiewicz University, Poznan, Poland, ISBN 978-83-7177-407-2, 2007.

[8] N. Campbell. On the Structure of Spoken Language. In *Proceedings of the 3rd International Conference on Speech Prosody 2006*, Dresden, Germany, 2006.

[9] M. McTear. Handling Miscommunication: Why Bother? In L. Dybkjær and W. Minker, editors, *Recent Trends in Discourse and Dialogue*, volume 39 of Text, Speech and Language Technology, pages 101–122. Springer, 2008. ISBN 978-1-4020-6820-1.

[10] A. Batliner, K. Fischer, R. Huber, J. Spilker, and E. Nöth. Desperately Seeking Emotions: Actors, Wizards, and Human beings. In *Proceedings of the ISCA Workshop on Speech and Emotion: A Conceptual Framework for Research*, pages 195–200, 2000.

[11] D. Ververidis and C. Kotropoulos. Emotional speech recogniton: resources, features and methods, *Speech Communication*, vol. 48, 2006, pp. 1162-1181.

[12] J. Hansen and S. Patil. Speech under stress: analysis, modeling and recognition, C. Müller (Ed.), *Speaker Classification I*, LNAI 4343, Springer-Verlag Berlin Heidelberg 2007, pp. 108-137.

[13] J. Hirschberg, D. Litman, and M. Swerts. Prosodic and other cues to speech recognition failures, *Speech Communication*, vol. 43, 2004, pp. 155-175.

[14] Z. Callejas and R. Lopez-Cozar. Influence of contextual information in emotion annotation for spoken dialogue systems, *Speech Communication*, vol. 50, 2008, pp. 416-433.

[15] M. Rajković, D. Rakić, S. Jovičić, M. Vojnović, and M. Đorđević. Intensity and Temporal Characteristics of Emotional Expressions in Serbian Spoken Discourse. (In Serbian). In *Proceedings of the TELFOR 2003*, 2003.

[16] M. Bojanić and V. Delić. Automatic emotion recognition in speech: possibilities and significance, *Electronics*, Vol. 13, No. 2, 2009, pages 35-40.

[17] M. Gnjatović, M. Kunze, X. Zhang, J. Frommer, and D. Rösner. Linguistic Expression of Emotion in Human-Machine Interaction: The NIMITEK Corpus as a Research Tool. In *Proc. of the Fourth International Workshop on Human-Computer Conversation*, Bellagio, Italy, 2008.

[18] Y. Lu, I. Cohen, X. Zhou, and Q. Tian, Feature Selection Using Principal Feature Analysis. In *Proc. of the ACM Multimedia 2007*, Augsburg, Germany, 2007.