FACTA UNIVERSITATIS
Series: **Electronics and Energetics** Vol. 27, N° 3, September 2014, pp. 375 - 387
DOI: 10.2298/FUEE1403375D

USER-AWARENESS AND ADAPTATION IN CONVERSATIONAL AGENTS

Vlado Delić[1], Milan Gnjatović[1,2], Nikša Jakovljević[1], Branislav Popović[1], Ivan Jokić[1], Milana Bojanić[1]

[1]Faculty of Technical Sciences, University of Novi Sad, Serbia
[2]Graduate School of Computer Sciences, Megatrend University, Belgrade, Serbia

Abstract: *This paper considers the research question of developing user-aware and adaptive conversational agents. The conversational agent is a system which is user-aware to the extent that it recognizes the user identity and his/her emotional states that are relevant in a given interaction domain. The conversational agent is user-adaptive to the extent that it dynamically adapts its dialogue behavior according to the user and his/her emotional state. The paper summarizes some aspects of our previous work and presents work-in-progress in the field of speech-based human-machine interaction. It focuses particularly on the development of speech recognition modules in cooperation with both modules for emotion recognition and speaker recognition, as well as the dialogue management module. Finally, it proposes an architecture of a conversational agent that integrates those modules and improves each of them based on some kind of synergies among themselves.*

Key words: *conversational agent, user-awareness, adaptation, speech recognition, emotion recognition, speaker recognition, dialogue management*

1. INTRODUCTION

Context-awareness is certainly one of the most fundamental requirements for advanced conversational agents. Recognition and interpretation of the user's dialogue acts and dialogue management are always situated in a particular context. This is primarily due to the fact that many inherently present dialogue phenomena are context-dependent. Thus, nonlinguistic contexts shared between the user and the system (e.g., graphical displays) may influence the language of the user to a high extent with respect to frequency of "irregular" utterances (elliptical and minor utterances, utterances containing anaphora and exophora, etc.) [1]. In addition, the user's dialogue acts may fall outside the system's domain, scope and semantic grammar, or contradict his earlier dialogue acts. This is even more the case when we consider users in non-neutral emotional states. Forcing users to follow a preset grammar or interaction scenario is too restrictive, if possible at all, and

Received April 30, 2014
Corresponding author: Vlado Delić
Faculty of Technical Sciences, University of Novi Sad,
Trg Dositeja Obradovića 6, 21000 Novi Sad, Serbia
(e-mail: vlado.delic@uns.ac.rs)

would not be well accepted [2]. In such cases, the system needs a considerable amount of stored contextual knowledge to enable it to advance the conversation in spite of miscommunication and to maintain the dialogue's consistency.

However, the requirement for habitable natural language interfaces goes beyond pragmatics. Another reason relates to the technology. Speech recognition technology is still not accurate enough to deal with flexible, unrestricted language. In realistic settings, average word recognition error rates are 20–30%, and they go up to 50% for non-native speakers [3]. In certain conditions, speech recognition accuracies may degrade dramatically to an extent that systems become unusable even for cooperative users [4].

Researchers generally agree that conversational agents need to incorporate dialogue context models in order to maintain a consistent dialogue and overcome technical deficiencies. Yet, context is a complex construct and can be considered from different aspects. In this paper, we consider a restricted research question of how user-awareness may help in improving dialogue management. This paper summarizes some aspects of our previous work and presents work-in-progress. In the reported approach, we differentiate between two research lines:

- *User-awareness.* The system is user-aware to the extent that it recognizes the user and his/her emotional states that are relevant in a given interaction domain.
- *User-adaptation*. The system is user-adaptive to the extent that it dynamically adapts its dialogue behavior according to the user and his/her emotional state.

At the methodological level, these two lines of research are fundamentally different. The first line relates to a statistical approach to the research problems of automatic speech recognition (ASR), emotional speech recognition (ESR), and speaker recognition. Speech signal encodes not only information about the lexical content of the speaker's dialogue act, but also information about the speaker's voice characteristics that may be used for recognition of the speaker and his/her emotional state [5], [6]. The basic idea is to use data derived from both speech and language corpora, and apply automated analysis methods. Although speech/speaker/emotion recognition technologies have a common foundation, they are usually developed and applied separately. We build upon our previous work [7]-[13], and investigate the possibilities to combine these technologies rather than to apply them separately. Sections 2 and 3 discuss this in more detail.

The second research line relates to a representational approach to natural language processing and dialogue management. In previous work, we introduced a representational model of attentional information in human-machine interaction that provides a framework for more robust natural language understanding and designing adaptive dialogue strategies [2], [14]-[17]. Section 4 discusses the application of this model to designing user-adaptive conversational agents.

2. ACOUSTIC INFORMATION-BASED APPROACH TO USER-AWARENESS

2.1. Speech recognition

The task of automatic speech recognition is to translate spoken words into text. In order to accomplish this task, the reported speech recognizer exploits information about acoustic representations of phonemes, encapsulated in an acoustic model, and information about syntactic rules, encapsulated in a language model. The relation between words and phonemes is captured in a pronunciation dictionary where each word is segmented into at

least one sequence of phonemes. Since each phoneme has several acoustic representations, as a basic modeling unit we use a context dependent phone referred to as triphone.

The acoustic model is based on hidden Markov models and Gaussian mixture models. In order to reduce the model computational complexity and to achieve robust parameter estimation, similar states of triphones share parameters. The tree based clustering procedure presented in [18] is performed to find those similar states. The Gaussians are modeled using the full covariance matrix, since they obtain more accurate acoustic representation in comparison to models with diagonal covariance matrix [19]. However, in this variant the computational complexity of log likelihood is significantly increased. To overcome this problem, several approaches have been developed and applied [20], [21], [22]. The system uses feature vectors consisted of 15 mel-frequency cepstral coefficients (MFCC), normalized energy and their first derivatives. The feature vectors are extracted from 30 ms speech segments, every 10 ms. The training set for the acoustic model contains recordings of both scripted and spontaneous utterances produced by several dozen speakers, with a total duration of about 200 hours [23].

Language modeling is a special issue for highly inflected languages, since language models have to cover a range of grammatical categories (including tense, aspect, mood, case, etc.) and morphological derivations that involve the addition of prefixes and suffixes. In the currently predominant statistically-based approach to ASR, language models are trained on large text corpora. However, simple *N*-gram based language models do not suffice for morphologically more complex languages without significant modifications [24]. Our language model is a combination of 3 *N*-gram models. The first model is based on tokens (surface forms), the second on lemmata, and the third on classes [23]. The size of vocabulary causes data sparsity problems, resulting in the need for significantly greater language corpora, sufficient for obtaining a robust language model. The training set for the language model consists of text content from various newspapers, scientific articles and books, with a total volume of about 16 million words (178865 lemmata).

Splitting words into phoneme sequence is relatively simple for the Serbian language, due to the fact that it has phonemic orthography. However, there are some exceptions in word pronunciation (e.g. *dvanaest* is usually pronounced as *dvanajst*) and our phonetic inventory distinguishes stressed and unstressed variant of vowels, thus for mapping words into phones the system uses the pronunciation dictionary developed for speech synthesis [23].

The size of search space is determined by the following factors: the number of words which are expected to be recognized, the number of their pronunciation variants, and the number of hidden Markov model states in the acoustic model. For the real-time recognition, it is important to reduce the search space, which can be a significant problem for highly inflected languages, where many derived forms may exist for a single lemma. The standard way to cope with this problem is pruning, i.e., discarding the less probable hypotheses. For this purpose, a system should rely not only on an acoustic model, but also on a language model and information about word pronunciation. Our system uses a decoder based on the token-passing algorithm (a variant of the Viterbi algorithm in which the information about the path and score is stored at the word level instead of trellis state level). A detailed description of the decoder can be found in [25].

2.2. Emotion recognition

Emotional speech recognition is concerned with the task of identifying emotional states of the speaker automatically, based upon the analysis of his speech. Prosodic and spectral features are the most frequently ones used for this task, while the less frequently used features include voice quality features (e.g., harmonic-to-noise ratio, jitter, shimmer). Prosodic features, also referred to as paralinguistic features, include specific changes in pitch patterning, the energy of the voice signal, and changes in speech rate. The positions and bandwidth of formants, and a cepstral representation of the spectrum are usually selected as spectral features for emotional speech recognition. This is in line with the findings that the distribution of the spectral energy across the speech range of frequency is a possible measure of the emotional content of speech. In [11], we show that a feature set containing both the prosodic and the spectral features achieves high recognition accuracy (i.e., 91.5 %) of the basic emotional states (i.e., anger, joy, fear, sadness, and neutral). The feature vector was obtained by applying statistical functionals to the spectral/prosodic feature contours, where the most relevant functionals, ranked in the descending order, are: moments, extrema, and regression coefficients [12].

In many speech-based applications, it is beneficial to conceptualize the user's emotional states in a given interaction domain as positive or negative (e.g., for the purpose of detecting a frustrated or satisfied call-centre customer). Therefore, in our previous work, we also investigated the perspective of dimensional emotion models that describe emotional content in terms of valence (positive/negative emotion) and arousal (active/passive emotion). We conducted a comparative study of two acted emotion corpora to investigate possibilities for classification of discrete basic emotions in the valence-arousal space [26]. The first conclusion of this study was that the prosodic-spectral feature set proposed in [11] is almost equally effective in modeling emotions in the valence-arousal space as compared to modeling discrete emotional states. The second conclusion was that the discrimination of emotional states according to the arousal level is more successful than their discrimination according to the valence level [26].

Our research on acoustic information-based emotion recognition was primarily supported by the GEES corpus of emotional and attitude-expressive speech in Serbian [27]. It contains recordings of acted speech-based emotional expressions. Six drama students (3 female, 3 male) were engaged to produce emotionally colored utterances. They were given a set of textual entries (32 isolated words, 30 short sentences, 30 long sentences, and one passage of 79 words) and asked to express each entry in five emotional states (anger, joy, fear, sadness and neutral). The perception test demonstrated that the corpus contains acoustic variations that are indicative of emotional expression of the five target emotional states.

2.3. Speaker recognition

Our research on speaker recognition centers on a text-independent speaker recognition based on the feature set that contains mel-frequency cepstral coefficients (MFCC) and their first and second derivatives. The research was primarily supported by a corpus containing recordings of 121 native Serbian speakers (61 female, 60 male). Each speaker produced 14 audio recordings: one recording of the speaker uttering his/her first name and family name, two recordings of the speaker uttering a sequence of digits, and eleven recordings of the speaker uttering a sequence of syntactically unrelated words. To reduce

the dimensionality of the standard MFCC, we applied the technique of Principal Component Analysis (PCA). The reported experimental results [9] suggest that this technique is appropriate to reduce the dimensionality without reducing the recognition accuracy. The applied automatic speaker recognizer shows that already for a 14-dimensional PCA feature space, the recognition accuracy reaches the target value as in the 39-dimensional MFCC feature space.

MFCCs depend on the energy in an observed speech frame. Therefore the distribution of a speaker feature vectors depends on the lexical content and expressed emotions. To decrease the text dependency on the covariance matrices used for speaker modeling, we apply an algorithm of model elements weighting introduced in [10]. The basic idea may be formulated as follows: the importance of an element of the speaker model in the decision making processes decreases as its time variability increases. In accordance with this, an element of the speaker model that has the highest time variability will be assigned the smallest value. In real applications, it can be the case that, for some speakers, the automatic speaker recognizer has only one model determined during the training phase. Thus, the recognizer cannot observe the time variability of model elements. The time variability of speaker models depends primarily on the largest model elements. By applying a nonlinear function, such as the sigmoid function, on the largest model elements, the time variability of the speaker models is decreased, and consequently, the recognition accuracy is increased. Also, MFCCs depend on the assumed shape of auditory critical bands. When the MFCCs are determined under the assumption that the auditory critical bands have exponential shape based on the lower part of the exponential function, the automatic speaker recognizer shows more accurate performance than in the case when the rectangular or triangular auditory critical bands are applied [10].

If should be noted that emotional speech may significantly affect the accuracy of speaker recognition. However, not all emotions are equally critical for speaker recognition. Preliminary experiments conducted on the GEES database confirmed that, e.g., the emotion of anger changes the speaker's voice (i.e., timbre) to the greater extent than the emotion of sadness. In the next sections, we discuss how a combination of different knowledge sources may improve the recognition accuracy.

2.4. Interplay between speech, emotion and speaker recognition

Acoustic features and language information contained within the acoustic, pronunciation and language models may be efficiently combined and used for speech, emotion and speaker recognition [5]. High-level features, e.g., phones, idiolect, semantic, accent and pronunciation, reveal speaker characteristics, such as socio-economic status, language background, personality type, and environmental influence [6].

For speech recognition systems based on hidden Markov models in combination with Gaussian mixture models, numerous techniques have been developed for model adaptation to specific speaker and acoustic condition [28]-[31]. They can be grouped into two classes based on maximum a posteriori likelihood (MAP) and maximum likelihood (ML), respectively. A MAP-based adaptation interpolates the original prior parameter values with parameters obtained from the adaptation data, and thus the estimated parameters converge asymptotically to the adaptation domain as the amount of adaptation data increases [28]. However, in the case of sparse adaptation data, many model parameters remain unchanged [32]. ML-based methods assume that there is a set of linear transformation

which can map the existing model parameters into new adapted model parameters. Since they use linear transformation to map parameters, these methods are referred as to ML linear regression (MLLR). MLLR can be applied only to the Gaussian mean vectors or to both mean vectors and covariance matrices. A special case of MLLR where the mean vector and covariance matrix of a Gaussian have the same transformation matrix is called constrained MLLR. While the use of mean MLLR adaptation has the greatest positive impact, the use of variance MLLR adaptation may also bring a slight improvement in recognition accuracy [29]. The major advantage of MLLR over MAP adaptations is evident in the case of sparse adaptation data, where the same transformation can be applied to all Gaussians in the same acoustic class [32].

Alternatively, speaker adaptation can be achieved by transformation of features instead of model parameters. The common procedure is vocal tract length normalization [33], [34]. The basic idea is to find warp scales of the frequency axis for each speaker such that the spectrum fits to the spectrum of the universal speaker with a standard vocal tract length, and to apply that transformation on the used features. In this way, within-class scattering and the overlapping between classes are reduced. It is interesting to note that the constrained MLLR can be treated as feature transformation, and that it is commonly used for speaker adaptive training. Models trained in this way may achieve higher recognition accuracy [35]. Additionally, the accuracy of an ASR system can be improved by the adaptation of the language model in terms of reducing the search space and confusion between words [36], [37].

It is widely acknowledged that the speaker's emotional states affect the speech production system at several levels – from the higher levels of linguistic coding (word selection and sentence structure) to the lower levels of articulator movements (phoneme/word production). This, in turn, may significantly degrade the performance of ASR systems. In general, ASR performance and prosodic properties of an utterance are related. Variations in speaking style and speaking rate, relative to ASR training conditions, may have a negative impact on the performance of an ASR system [38]. Prosodic features reflect those variations, and some studies show that prosody itself is capable of re-ranking ASR hypotheses such as to separate the correctly recognized utterances from incorrectly recognized ones [39], [40]. It can be expected that an ASR system using acoustic models trained on neutral speech will have reduced performances in settings when it operates under the conditions of emotional speech. Reference [41] shows that training ASR models on neutral speech, and its subsequent adaptation on emotional speech samples, does have a positive impact on the recognition performance within such conditions.

In [11] and [42], we discuss how the same prosodic and spectral features can be employed for the purpose of speech recognition, emotion recognition and speaker recognition. Fig. 1. illustrates how knowledge from different sources is intended to be used in the reported speech processing module. The relationship between these technologies goes beyond prosodic and spectral features. In the next section, we discuss how emotion recognition can employ lexical and discourse information provided by an ASR system.

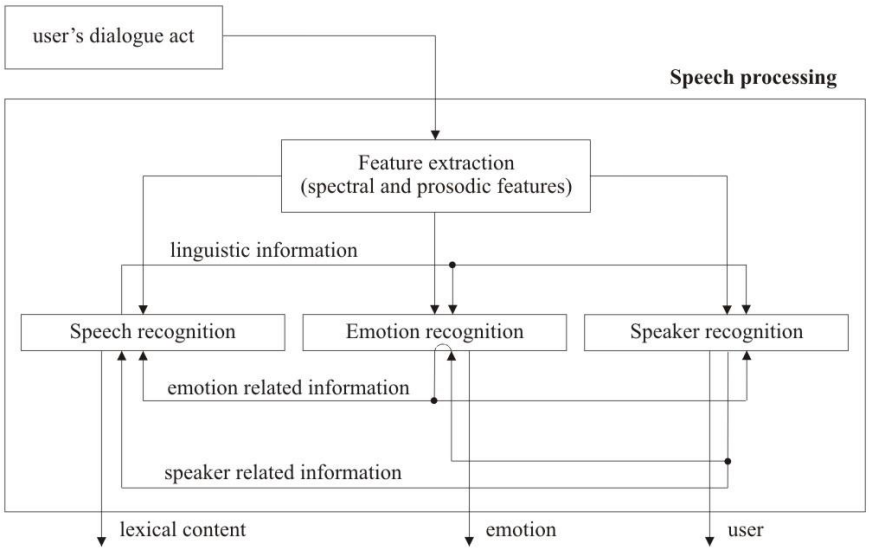

Fig. 1 Combining knowledge from different sources in the speech processing module

3. Emotion Recognition Based on Linguistic Information

Emotion recognition can be also based on lexical and discourse information [43], e.g., a semantic analysis of an output hypothesis of an ASR engine [44]. In line with this, one line of our research focuses on recognition and tracking of emotional states of the user from lexical information and other linguistic features. As part of previous work [1], a substantial refinement of the Wizard-of-Oz technique was proposed in order that a scenario designed to elicit affected speech in human-machine interaction could result in realistic and useful data. The NIMITEK corpus of affected speech in human-machine interaction was produced during a refined Wizard-of-Oz simulation. Ten healthy native German speakers participated in the study (7 female, 3 male, ages 18 to 27, mean 21.7). The corpus contains 15 hours of audio and video recordings. The number of the subjects' dialogue turns is 1,847, the average number of words per turn is 17.19 (with standard deviation 24.37), and the subjects' lexicon contains about 900 lemmata. The evaluation of the corpus with respect to the perception of its emotional content demonstrated that it contains recordings of emotions that were overtly signaled, and that the subjects' utterances are indicative of the way in which untrained, nontechnical users probably like to converse with conversational agents [1]. The transcribed version of the NIMITEK corpus was used to conduct a corpus-based examination of various linguistic features that may carry affect information [13]. For the purpose of this contribution, we illustrate the following linguistic features: key words and phrases, lexical cohesive agencies, dialogue act sequences, and negations.

The most obvious way of recognizing an emotional state is to detect key words and phrases in users' utterances. Examples from the NIMITEK corpus are given in Table 1. However, expressions of emotions are not necessarily limited to a single dialogue act, but

can also map over a range of mutually related dialogue acts. For example, the choice of lexical items made to create cohesion in the dialogue can signal an emotion-related state, both at the lexical level (e.g., repetitions), as well as at the semantic level (e.g., reformulations). Table 2 contains examples of repetitions and reformulations that signal negative emotional states. In contrast to this, another form of anaphoric cohesion in a dialogue is achieved by ellipsis-substitutions. The typical meaning of ellipsis-substitutions is *not one of co-reference – there is always some significant difference between the second instance and the first* [45]. To illustrate this, let us observe a typical example from the NIMITEK corpus: *Please do it! (Bitte tu das!)*. This utterance does not explicitly provide information what the system is expected to do, but contains an elliptical-substitution (verb *do*) which is used for signaling that the action the system performed is not the same as the action instructed by the user (indicated by the anaphoric reference *it*). In general, ellipsis-substitutions may signal a potential problem in communication.

Table 1 Examples of key words and phrases that relate to emotional states (adopted and adjusted from [13])

Emotional state	Examples of the subjects' key words and phrases
Annoyed	Sh*t (Sche*ße), stupid (blöd), Do what I say (Tu was ich sage), Oh … something like this I hate just like the plague. (Oh... so was hasse ich doch wie die Pest.)
Retiring	I don't understand it (Ich versteh' das nicht), It's not working at all (Das geht doch gar nicht).
Indisposed	I am going now (ich geh' gleich), Oh man (Oh man), God (Gott), I don't feel like doing any more. (Ich hab' kein' Bock mehr.)
Offending	You think, doll. (Denkst du, Puppe)
Satisfied	Super (Super), awesome (geil), I am good, am I not? (Bin gut, was?)

Table 2 Examples of lexical cohesive agencies that relate to negative emotional states (adopted and adjusted from [13])

Lexical cohesive agencies	Examples of the subjects' dialogue acts
Repetition	It just cannot be. It just … It just cannot be. (Das kann doch nicht sein. Das ist doch … das kann doch nicht sein.)
Reformulation	Not true at all. That's definitely wrong. (Gar nicht wahr. Das stimmt gar nicht.)
Ellipsis-substitution	Please do it.

Based on this study, a prototypical automatic annotator for recognition and tracking of the user's emotional states from linguistic information was implemented [13]. It should be noted that the emotional states in the NIMITEK corpus [1] were conceptualized within the data-driven 6-class emotion model ARISEN (annoyed, retiring, indisposed, satisfied, engaged, neutral). In addition, the subjects' expressions in the NIMITEK corpus often contain mixed emotions, and the human evaluators were allowed to assign more emotion labels to each subject's utterance. Thus, the automatic annotator was implemented to annotate mixed emotions, i.e., to attribute zero, one or more labels from the ARISEN model to each subject's utterance.

The results of the automatic annotation were compared with the results of the human evaluators. For the given 6-class emotional model ARISEN, the annotator showed the

following performance: 31.70% of subject emotional states were correctly, 34.35% of subject emotional states were not recognized, and 33.92% of subject emotional states were incorrectly recognized. Furthermore, the ARISEN model was down-sampled to a model that differentiates between 3 emotional states, i.e., negative (including annoyed, retiring and indisposed emotional states), neutral, and positive (including satisfied and engaged emotional states). For this 3-class problem, the annotator showed the following performance: 51.20% of subject emotional states were correctly recognized, 33.67% of subject emotional states were not recognized, and 17.26% of subject emotional states were incorrectly recognized. When interpreting these results, it should be kept in mind that the automatic annotation was based only on lexical information, while the human evaluators were influenced by prosody as well.

4. USER-ADAPTIVE DIALOGUE MANAGEMENT

The main idea underlying the conversational agent's adaptation is that its dialogue behavior is dynamically adapted according to the user and his emotional state. In this respect, the dialogue management module is the central component of the conversational agent. It consists of two components: dialogue context model and adaptive dialogue control [46].

4.1. Dialogue context model

Dialogue context model keeps track of information relevant to the dialogue. For the purpose of this contribution, it includes the following knowledge sources:

- lexical and propositional content of the user's dialogue act,
- attentional state,
- emotional state of the user,
- information about the user.

Among these sources, attentional state deserves further discussion. At the conceptual level, attentional state contains information about the dialogue entities that are most salient at any given point. Its purpose is twofold [2], [47]. First, it summarizes information from previous dialogue acts that are necessary for processing subsequent ones, and allows for processing spontaneously produced users' dialogue acts. This is an important characteristic of the system, not just because it enables a more natural dialogue, but also because forcing users to follow a preset grammar or interaction scenario is hardly acceptable for users in negative emotional states. The second purpose of attentional state is that it allows for predicting the dialogue behavior of the user, i.e., it forms the basis for expectations about the succeeding dialogue acts. This information is important both for automatic speech recognition, as a means of reducing a set of ASR hypotheses, and adaptive dialogue control, for taking initiative in a dialogue.

In [2], we introduced a representational model of attentional information in human-machine interaction that provides a framework for more robust natural language processing and dialogue management. This model integrates neurocognitive understanding of the focus of attention in working memory, the notion of attention related to the theory of discourse structure in the field of computational linguistics, and investigation of the NIMITEK corpus. To the extent that it is computationally appropriate, it was successfully adapted and applied in several prototypical conversational agents with diverse domains of interaction [14], including the dialogue management module reported in this paper.

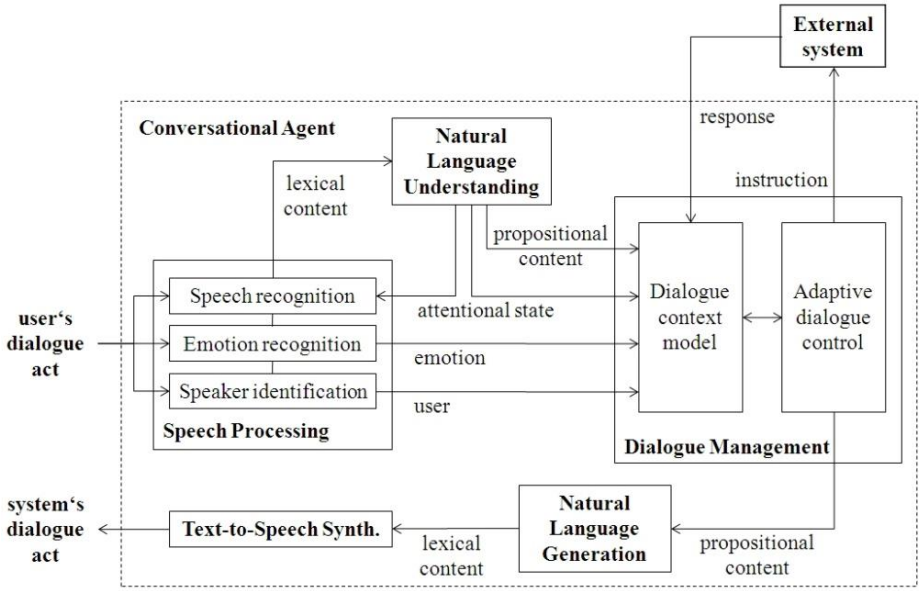

Fig. 2 The intended architecture of a conversational agent

4.2 Adaptive dialogue control

The dialogue control component implements dialogue strategies of the conversational agent. In general, a dialogue strategy involves deciding what to do next once the user's input has been received and interpreted, e.g., prompting the user for more input, clarifying the user's previous input, outputting information to the user, etc. [46]. We recall that the reported conversational agent is adaptive to the extent that it dynamically adapts its dialogue strategies according to the current user and his emotional state. Therefore, an adaptive dialogue strategy is specified by means of a set of rules that take information about the current dialogue context into account. We build upon previous work on emotion-adaptive dialogue strategies, and end-user design of adaptive dialogue strategies. It is important to note that the reported dialogue management module allows the end-user to design dialogue strategies. This makes two levels of adaptation possible. The dialogue behavior is not only dynamically adapted according to the current dialogue strategy, but also the dialogue strategy itself can be redefined by the user. For detailed discussion, the reader may consult [16], [17].

5. CONCLUDING REMARKS

This paper summarized some aspects of our previous work and presents work-in-progress on developing user-aware and adaptive conversational agents. The intended architecture of a conversational agent is given in Fig. 2. The speech recognition module and the dialogue management module (integrated with the natural language processing modules) are fully implemented, while emotion recognition and speaker recognition modules are implemented at a prototype level.

Current and future prospects of our research in this field include (but are not limited to): further investigation of the interplay between speech recognition, emotion recognition and speaker recognition, investigation of linguistic cues for early recognition of negative dialogue developments, further development of dialogue strategies for preventing and handling negative dialogue development, and investigation of more complex user models and alternative models of emotions.

Acknowledgement: *The presented study was performed as part of the project "Development of Dialogue Systems for Serbian and Other South Slavic Languages" (TR32035), funded by the Ministry of education, science and technological development of the Republic of Serbia.*

REFERENCES

[1] M. Gnjatović and D. Rösner, "Inducing Genuine Emotions in Simulated Speech-Based Human-Machine Interaction: The NIMITEK Corpus". *IEEE Transactions on Affective Computing*, vol. 1, no. 2, pp. 132-144, July-Dec. 2010, DOI: 10.1109/T-AFFC.2010.14

[2] M. Gnjatović, M. Janev, V. Delić, "Focus Tree: Modeling Attentional Information in Task-Oriented Human-Machine Interaction". *Applied Intelligence*, vol. 37, no. 3, pp. 305-320, 2012, DOI: 10.1007/s10489-011-0329-5

[3] D. Bohus and A. Rudnicky, "Sorry, I Didn't Catch That! An Investigation of Non-Understanding Errors and Recovery Strategies". In Recent Trends in Discourse and Dialogue, vol. 39 of Text, Speech and Language Technology, pp. 123–154, Springer, 2008.

[4] C.H. Lee, "Fundamentals and Technical Challenges in Automatic Speech Recognition". In Proc. of the 12th International Conference Speech and Computer, SPECOM 2007, pp. 25–44, Moscow, Russia, 2007.

[5] B. Schuller, G. Rigoll, M. Lang, "Speech Emotion Recognition Combining Acoustic Features and Linguistic Information in a Hybrid Support Vector Machine-Belief Network Architecture". In Proc. of ICASSP 2004, vol. 1, pp. I-577-580, 2004, DOI: 10.1109/ICASSP.2004.1326051

[6] T. Kinnunen and L. Haizhou, "An Overview of Text-Independent Speaker Recognition: From Features to Supervectors". *Speech Communication*, vol. 52, pp. 12-40, 2010, DOI: 10.1016/j.specom.2009.08.009

[7] V. Delić, M. Sečujski, N. Jakovljević, M. Gnjatović, I. Stanković, "Challenges of Natural Language Communication with Machines". Chap. 19 in DAAAM International Scientific Book 2013, pp. 371-388, 2013, DOI: 10.2507/daaam.scibook.2013.19

[8] N. Jakovljević, D. Mišković, M. Janev, M. Sečujski, V. Delić, "Comparison of Linear Discriminant Analysis Approaches in Automatic Speech Recognition". *Electronics and Electrical Engineering*, vol. 19, no. 7, pp. 76-79, 2013, DOI: 10.5755/j01.eee.19.7.5167

[9] I. Jokić, S. Jokić, Z. Perić, M. Gnjatović, V. Delić, "Influence of the Number of Principal Components used to the Automatic Speaker Recognition Accuracy". *Electronics and Electrical Engineering*, vol. 18, no. 7, pp. 83-86, 2012, DOI: 10.5755/j01.eee.123.7.2379

[10] I. Jokić, S. Jokić, V. Delić, Z. Perić, "Towards a Small Intra-Speaker Variability Models". *Electronics and Electrical Engineering*, vol. 20, 2014 *(in press)*.

[11] V. Delić, M. Bojanić, M. Gnjatović, M. Sečujski, S.T. Jovičić, "Discrimination Capability of Prosodic and Spectral Features for Emotional Speech Recognition". *Electronics and Electrical Engineering*, vol. 18, no. 9, pp. 51-54, 2012, DOI: 10.5755/j01.eee.18.9.2806

[12] M. Bojanić, V. Delić, M. Sečujski, "Relevance of the types and the statistical properties of features in the recognition of basic emotions in speech". *Facta Universitatis*, Series: Electronics and Energetics, vol. 27, 2014 *(in press)*.

[13] M. Gnjatović, M. Kunze, X. Zhang, J. Frommer, D. Rösner, "Linguistic Expression of Emotion in Human-Machine Interaction: The NIMITEK Corpus as a Research Tool". In Proceedings of the 4th Int. Workshop on Human-Computer Conversation, Bellagio, Italy, no pagination, 2008.

[14] M. Gnjatović and V. Delić, "A Cognitively-Inspired Method for Meaning Representation in Dialogue Systems". In Proc. of the 3rd IEEE Int. Conf. CogInfoCom-2012, Košice, Slovakia, pp. 383-388, 2012.

[15] M. Gnjatović and V. Delić, "Electrophysiologically-Inspired Evaluation of Dialogue Act Complexity". In Proc. of the 4th IEEE Int. Conf. CogInfoCom 2013, Budapest, Hungary, pp. 167-172, 2013.

[16] M. Gnjatović and V. Delić, "Cognitively-inspired representational approach to meaning in machine dialogue". *Knowledge-Based Systems*, DOI: 10.1016/j.knosys.2014.05.001, 2014.

[17] M. Gnjatović, "Therapist-Centered Design of a Robot's Dialogue Behavior". *Cognitive Computation*, Special issue: The quest for modeling emotion, behavior and context in socially believable Robots and ICT interfaces, Springer, DOI: 10.1007/s12559-014-9272-1 *(in press).*

[18] S. J. Young, J. Odell, P. C. Woodland, "Tree-based state tying for high accuracy acoustic modelling". In Proceedings of the Workshop on Human Language Technology, pp. 307-312, 1994, DOI: 10.3115/1075812.1075885

[19] N. Jakovljević, D. Mišković, E. Pakoci, T. Grbić and V. Delić, "Poređenje performansi nekoliko varijanata GMM u sistemima za prepoznavanje govora". In Proc. of 21th Telecommunications Forum, TELFOR 2013, Belgrade, Serbia, pp. 466-469, 2013.

[20] M. Janev, D. Pekar, N. Jakovljević, V. Delić, "Eigenvalues driven Gaussian selection in continuous speech recognition using HMMs with full covariance matrices". *Applied Intelligence*, vol. 33, no. 2, pp. 107-116, 2010, DOI: 10.1007/s10489-008-0152-9

[21] B. Popović, M. Janev, D. Pekar, N. Jakovljević, M. Gnjatović, M. Sečujski, V. Delić "A novel split-and-merge algorithm for hierarchical clustering of Gaussian mixture models". *Applied Intelligence*, vol. 37, no. 3, pp. 377-389, 2012, DOI: 10.1007/s10489-011-0333-9

[22] N. Jakovljević, Primena retke reprezentacije na modelima Gausovih mešavina koji se koriste za automatsko prepoznavanje govora, PhD thesis, University of Novi Sad, March 2014.

[23] V. Delić, M. Sečujski, N. Jakovljević, D. Pekar, D. Mišković, B. Popović, S. Ostrogonac, M. Bojanić, D. Knežević, "Speech and Language Resources within Speech Recognition and Synthesis Systems for Serbian and Kindred South Slavic Languages". In Proc. of the SPECOM 2013, Pilsen, Czech Republic, LNCS, vol. 8113, Springer, pp. 319-326, 2013, DOI: 10.1007/978-3-319-01931-4_42

[24] S. Ostrogonac, M. Sečujski, V. Delić, D. Mišković, N. Jakovljević, N. Vujnović Sedlar, *A Mixed-Structure N-gram Language Model*, Axon - inteligentni sistemi, Novi Sad, Serbia. International patent pening: PCT/RS2013/000009

[25] N. Jakovljević, D. Mišković, M. Janev, D. Pekar, "A Decoder for Large Vocabulary Speech Recognition". In Proc. of 18th International Conference on Systems, Signals and Image Processing, IWSSIP 2011, Sarajevo, Bosnia and Herzegovina, pp. 287-290, 2011.

[26] M. Bojanić, M. Gnjatović, M. Sečujski, V. Delić: "Application of dimensional emotion model in automatic emotional speech recognition". In Proc. of the 11th IEEE Int. Symp. on Intelligent Systems and Informatics, SISY 2013, Subotica, Serbia, pp. 353-356, 2013, DOI: 10.1109/SISY.2013.6662601

[27] S.T. Jovičić., Z. Kašić, M. Djordjević, M. Rajković, "Serbian emotional speech database: design, processing and evaluation". In Proc. of SPECOM 2004, St Peterburg, pp.77–81, 2004.

[28] J. Gauvain and C. H. Lee, "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains". *IEEE Trans. on Speech and Audio Process.*, vol. 2, no. 2, pp. 291-298, Apr. 1994, DOI: 10.1109/89.279278

[29] M.J.F. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition". *Computer speech & language*, vol. 12, no. 2, pp. 75-98, 1998, DOI: 10.1006/csla.1998.0043

[30] M.J.F. Gales and P.C. Woodland, "Mean and variance adaptation within the MLLR framework". *Computer Speech & Language*, vol. 10, no. 4, pp. 249-264, 1996, DOI: 10.1006/csla.1996.0013

[31] D. Povey and G. Saon, "Feature and model space speaker adaptation with full covariance Gaussians". In Proc. Interspeech 2006, paper 2050-Tue2BuP.14, 2006.

[32] M.J.F. Gales and S. Young, "The application of hidden Markov models in speech recognition". *Foundations and Trends in Signal Processing*, vol. 1, no. 3, pp. 195-304, 2008, DOI: 10.1561/2000000004

[33] N. Jakovljević, D. Mišković, M. Sečujski, D. Pekar, "Vocal tract normalization based on formant positions". In Proc. Inter. Language Technologies Conference IS-LTC 2006, Ljubljana, pp. 40-43, 2006.

[34] N. Jakovljević, M. Sečujski, V. Delić, "Vocal tract length normalization strategy based on maximum likelihood criterion". In Proc. EUROCON 2009, St. Petersburg, pp. 417-420, 2009, DOI: 10.1109/EURCON.2009.5167662

[35] G. Saon and J.T. Chien, "Large-vocabulary continuous speech recognition systems". *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 12-33. Nov. 2012, DOI: 10.1109/MSP.2012.2197156

[36] J.M. Lucas-Cuesta J. Ferreiros, F. Fernandez-Martinez, J.D. Echeverry, S. Lutfi, "On the dynamic adaptation of language models based on dialogue information". *Expert Systems with Applications*, vol. 40, no. 4, pp. 1069-1085, 2013, DOI: 10.1016/j.eswa.2012.08.029

[37] W. Kim, Language model adaptation for automatic speech recognition and statistical machine translation, PhD Thesis, Johns Hopkins University, 2005.

[38] L. ten Bosch, "Emotions: what is possible in the ASR framework". ITRW on Speech and Emotion, Northern Ireland, UK, pp. 189-194, 2000.

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

[40] D. Litman, J. Hirschberg, M. Swerts, "Predicting automatic speech recognition performance using prosodic cues". In Proc. of the 1st North American chapter of the Association for Computational Linguistics, NAAC, Seattle, pp. 218-225, 2000.
[41] B. Vlasenko, D. Prylipko, A. Wendemuth, "Towards robust spontaneous speech recognition with emotional speech adapted acoustic models". S. Wölfl (ed.), Poster and Demo Track of the 35th German Conference on Artificial Intelligence, KI-2012, Saarbrucken, Germany, pp. 103-107, 2012.
[42] B. Popović, I. Stanković, S. Ostrogonac, "Temporal Discrete Cosine Transform for Speech Emotion Recognition". In Proc. of the 4th IEEE Int. Conf. CogInfoCom 2013, Budapest, Hungary, pp. 87-90, 2013.
[43] C.M. Lee and S.S. Narayanan, "Toward detecting emotions in spoken dialogs". *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 2, pp. 293-303, 2005, DOI: 10.1109/TSA.2004.838534
[44] R. Müller, B. Schuller, G. Rigoll, "Enhanced Robustness in Speech Emotion Recognition Combining Acoustic and Semantic Analyses". In Proc. of the Workshop From Signals to Signs of Emotion and Vice Versa, Santorini, Greece, 2004.
[45] M. Halliday, *An Introduction to Functional Grammar*, Edward Arnold, London New York, Second edition, 1994.
[46] K. Jokinen and M. McTear, Spoken Dialogue Systems. Synthesis Lectures on Human Language Technologies, Morgan and Claypool, 2009.
[47] B. Grosz and C. Sidner, "Attention, intentions, and the structure of discourse". *Comput Linguist*, vol. 12, no 3, pp. 175-204, 1986.