Corchado, J. M. et al. (eds.) (2014). *Highlights of Practical Applications of Heterogeneous Multi-Agent Systems. The PAAMS Collection: PAAMS 2014 International Workshops, Salamanca, Spain, June 4-6, 2014. Proceedings*. (pp. 167-178). (Communications in Computer and Information Science; 430). Springer International Publishing
DOI: http://dx.doi.org/10.1007/978-3-319-07767-3_16

A proposal for processing and fusioning multiple information sources in multimodal dialog systems

David Griol, José Manuel Molina, Jesús García-Herrero

Computer Science Department
Carlos III University of Madrid
Avda. de la Universidad, 30, 28911 - Leganés (Spain)
{david.griol,josemanuel.molina,jesus.garciaherrero}@uc3m.es

Abstract. Multimodal dialog systems can be defined as computer systems that process two or more user input modes and combine them with multimedia system output. This paper is focused on the multimodal input, providing a proposal to process and fusion the multiple input modalities in the dialog manager of the system, so that a single combined input is used to select the next system action. We describe an application of our technique to build multimodal systems that process user's spoken utterances, tactile and keyboard inputs, and information related to the context of the interaction. This information is divided in our proposal into external and internal context, user's internal, represented in our contribution by the detection of their intention during the dialog and their emotional state.

Keywords: Multimodal Systems, Conversational Agents, Fusion Techniques, Dialog management, User Modeling

1 Introduction

Research on multimodal interaction has grown considerably during the last decade, as a consequence of the development of innovative input interfaces, as well as the advances in research fields such as speech interaction and natural language processing [1, 2]. However, multimodal fusion has not evolved at the same rate, which has lead to minor advances at the different possibilities of combining input modalities [3, 4].

Multimodal dialog systems [5–7] are dialog systems that process two or more combined user input modes. According to [8], fusion of input sources in these systems must be approached in a global way: from the point of view of the architecture of a multimodal system as a whole, then, from the point of view of multimodal dialog modeling, and finally from an algorithmic point of view.

The architectural perspective focuses on necessary features of an architecture to allow usability in the integration of a fusion engine. Most of the current systems have been developed following the basis for multimodal interaction defined

by important projects like Smartkom [7]. Smartkom's interaction metaphor was based on the idea that the user delegates a task to the virtual communication assistant which is visualized as a life-like character. Among the input modalities considered there were spoken dialog, graphical user interfaces, gestural interaction, facial expressions, physical actions, and biometrics. In the output, it provided an anthropomorphic user interface that combined speech, gesture, and facial expressions.

Multimodal dialog modeling refers to the module of the multimodal system that controls the interaction: the dialog manager. This module decides the next action of the multimodal system [9–11], interpreting the incoming semantic representation of each input modality in the context of the dialog. In addition, it resolves ellipsis and anaphora, evaluates the relevance and completeness of user requests, identifies and recovers from recognition and understanding errors, retrieves information from data repositories, and decides about the next system's response. Fusion techniques in multimodal dialog systems are usually integrated in the dialog manager [12].

Finally, the algorithmic perspective studies logic and algorithms used to integrate data coming from different input recognizers into an application-usable result. Fusion of input modalities can be achieved at a number of different levels of abstraction, as well as considering increasing levels of complexity. Multi-sensor data fusion can be performed at four different processing levels, according to the stage at which the fusion takes place: signal level, pixel level, feature level, and decision level [13].

In this paper we propose a general-purpose approach to cost-efficiently develop an adapt a multimodal dialog system. The main objective is to reduce the effort required for both the implementation of a new system and the adaptation of systems to deal with user's specific features, a new task or modality. Our proposal follows an architecture that integrates several modules dealing with input modalities, as speech or visual and tactile interaction, and also the context of the interaction. We differentiate between two types of context: *internal* and *external*. The former describes the user state, modeled in our proposal by the user's intention during the dialog and the user's emotional state, whereas the latter refers to the environment state (e.g. location and temporal context).

We also propose a multimodal fusion methodology that is integrated in the dialog manager of the system. This module takes the input information sources into account to generate and encode a single input used for the selection of the next system action.

2 Proposal for developing multimodal dialog systems

The general architecture used for the development of multimodal applications can be separated in four different components: input modalities and their recognizers, output modalities and their respective synthesizers, the integration committee, and the application logic [8]. Indeed, using multimodality efficiently implies a clear abstraction between the results of the user's input analysis, the

processing of this input, answer generation and output modalities selection. As Figure 1 shows, this clear separation is achieved with help of the integration committee, responsible for management of all input and output modalities.

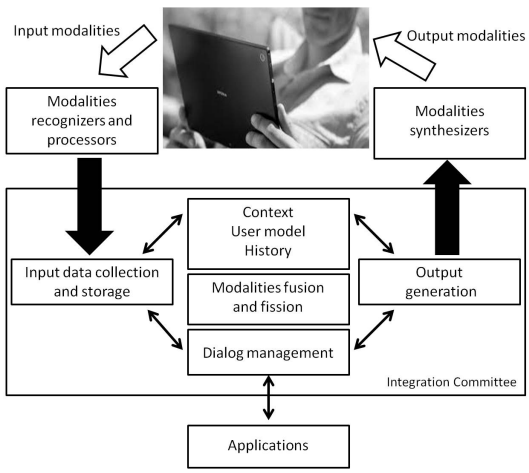

Fig. 1. General architecture for the generation of multimodal dialog systems

The integration committee can itself be separated in five different subcomponents. First, input modalities are collected into the input data collection and storage module, which is in charge of identifying and storing input data. The Modalities fusion and fission module manages input data prepares it for processing by the application logic. When the fusion and fission engines reach an interpretation, it is passed to the dialog management module.

Figure 2 describes the process for adapting the general architecture presented in Figure 1 by introducing the key points of our proposal. A spoken dialog system integrates five main tasks to deal with user's spoken utterances in natural language: automatic speech recognition (ASR), natural language understanding (NLU), dialog management (DM), natural language generation (NLG), and text-to-speech synthesis (TTS).

Speech recognition is the process of obtaining the text string corresponding to an acoustic input [14]. It is a very complex task as there is much variability in the input characteristics, which can differ depending on the linguistics of the utterance, the speaker, the interaction context and the transmission channel. Linguistic variability involves differences in phonetic, syntactic and semantic components that affect the voice signal. Inter-speaker variability refers to the big difference between speakers regarding their speaking style, voice, age, sex or nationality.

Once the conversational agent has recognized what the user uttered, it is necessary to understand what he said. Natural language processing is the pro-

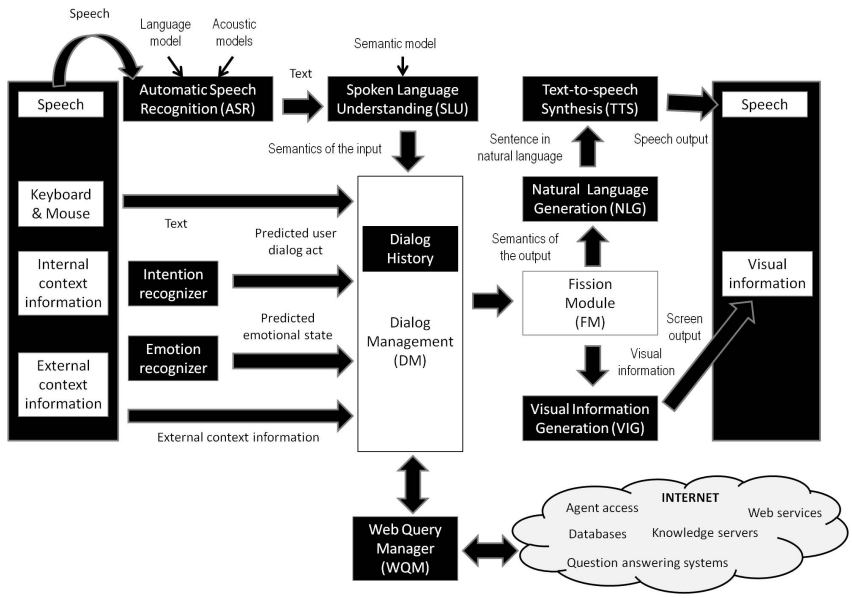

Fig. 2. Proposed framework for the generation of multimodal dialog systems

cess of obtaining the semantic of a text string [15, 16]. It generally involves morphological, lexical, syntactical, semantic, discourse and pragmatical knowledge. Lexical and morphological knowledge allow dividing the words in their constituents distinguishing lexemes and morphemes. Syntactic analysis yields a hierarchical structure of the sentences, while semantic analysis extracts the meaning of a complex syntactic structure from the meaning of its constituents. In the pragmatic and discourse processing stage, the sentences are interpreted in the context of the whole dialog.

There is not a universally agreed upon definition of the tasks that a dialog manager has to carry. Traum and Larsson [17] state that dialog managing involves four main tasks: i) updating the dialog context, ii) providing a context for interpretations, iii) coordinating other modules and iv) deciding the information to convey and when to do it. Thus, the dialog manager has to deal with different sources of information such as the NLU results, database queries results, application domain knowledge, and knowledge about the users and the previous dialog history [11].

Natural language generation is the process of obtaining texts in natural language from a non-linguistic representation. The simplest approach consists in using predefined text messages (e.g. error messages and warnings). Finally, a text-to-speech synthesizer is used to generate the voice signal that will be transmitted to the user.

As explained in the introduction section, a multimodal dialog system involves user inputs through two or more combined modes, which usually complement spoken interaction by also adding the possibility of textual and tactile inputs provided using physical or virtual keyboards and the screen. In our contribution, we want also to model the context of the interaction as an additional valuable information source to be considered in the fusion process.

With regard to external context, our proposal is based on additional agents used to capture and provide this information to the spoken conversational agent. Regarding internal context, our proposal merges the traditional view of the dialog act theory, in which communicative acts are defined as intentions or goals, with the recent trends that consider emotion as a vital part for social communication. To do so, we contribute a user state prediction module based on an intention recognizer and an emotion recognizer.

Finally, we also propose a statistical methodology that combines multimodal fusion and dialog management functionalities. To do this, a data structure is introduced to store the information provided by the user's inputs and the context of the interaction. This information is coded taking into account the confidence measures provided by the modules that capture and process the different information sources. This data structure is taking into account in a classification process whose result allows the selection of the next system response. The following subsections describe the different methodologies proposed to develop the main modules of the multimodal dialog system.

2.1 Modeling user's intention

Research in techniques for user modeling has a long history within the fields of language processing and dialog systems. The main purpose of a simulated user in this field is to improve the usability of a dialog system through the generation of corpora of interactions between the system and simulated users [18]. Two main approaches can be distinguished to the creation of simulated users: rule based and data or corpus based. In a rule-based simulated user the researcher can create different rules that determine the behavior of the system [19]. The main objective of data-based techniques is to automatically explore the space of possible dialog situations and learn new potentially better dialog strategies [20].

The statistical technique that we propose to model user's intention is described in [21]. The proposed technique carries out the functions of the ASR and SLU modules, i.e., it estimates user's intention providing the semantic interpretation of the user utterance in the same format defined for the output of the SLU module. A data structure, that we call *User Register* (UR), contains the information provided by the user throughout the previous history of the dialog. For each time i, the proposed model estimates user's intention taking into account the sequence of dialog states that precede time i, the system answer at time i, and the objective of the dialog $\mathcal{O}$. The selection of the most probable user answer U_i is given by:

$$\hat{U}_i = \arg\max_{U_i \in \mathcal{U}} P(U_i | UR_{i-1}, A_i, \mathcal{O})$$

The information contained in UR_i is a summary of the information provided by the user up to time i. That is, the semantic interpretation of the user utterances during the dialog and the information that is contained in a user profile (e.g., user's name, gender, experience, skill level, most frequent objectives, additional information from previous interactions, user's neutral voice, and additional parameters that could be important for the specific domain of the system). We propose to solve the previous equation by means of a classification process, which takes the current state of the dialog (represented by means of the set $UR_{i-1}, A_i, \mathcal{O}$) as input and provides the probabilities of selecting the different user dialog acts. Figure 3 shows the described process followed by the proposed intention recognizer and its interaction with the rest of modules of the multimodal dialog system.

2.2 Modeling user's emotional state

Although emotion is receiving increasing attention from the dialog systems community, most research described in the literature is devoted exclusively to emotion recognition [22] and not to the use of this valuable information in the fusion and dialog management processes. Emotions change people voices, facial expressions, gestures, and speech speed. They can also affect the actions that the user chooses to communicate with the multimodal system.

Our emotion recognition method, based on the previous work described in [23], firstly takes acoustic information into account to distinguish between the emotions which are acoustically more different, and secondly dialog information to disambiguate between those that are more similar. We were interested in recognizing negative emotions that might discourage users from employing the system again or even lead them to abort an ongoing dialog. Concretely, we considered three negative emotions: anger, boredom, and doubtfulness, where the latter refers to a situation in which the user uncertain about what to do next).

The proposed emotion recognizer employs acoustic information to distinguish anger from doubtfulness or boredom and dialog information to discriminate between doubtfulness and boredom, which are more difficult to discriminate only by using phonetic cues. This process is shown in Figure 3. The first step for emotion recognition is feature extraction. The aim is to compute a list of 60 features from a speech input which can be relevant for the detection of emotion in the users' voice [23]. The second step of the emotion recognition process is feature normalization, with which the features extracted in the previous phase are normalized around the user neutral speaking style. Once we have obtained the normalized features, we classify the corresponding utterance with a multilayer perceptron (MLP) into two categories: *angry* and *doubtful_or_bored*. If the utterance is classified as *doubtful_or_bored*, it is passed through an additional step in which it is classified according to two dialog parameters: depth and width.

2.3 Acquiring and processing external context

External contextual information is usually measured by hardware or software-based sensors (such as GPS and monitoring programs), or provided by the users. Typically, sensors rely on low level communication protocols to send the collected context information or they are tightly coupled within their context-aware systems. Since sensing techniques are well developed, existing sensors utilize these techniques through instrumentation or polling mechanisms, and extend their capability by acquiring context information from existing systems.

As described in [24], we propose the use of a Facilitator and Positioning Systems to acquire and process external contextual information. The Positioning System communicates with the ARUBA positioning system to extract and transmit positioning information to other agents in the system

The Facilitator System is implemented using the Appear IQ commercial platform (AIQ, `www.appearnetworks.com`). The platform consists of two main modules: the Appear Context Engine (ACE) and the Appear Client (AC). The ACE is installed in a server, while the ACs are included in the users' devices.

The ACE implements a rules engine, where the domain-specific rules that are defined determine what should be available to whom, and where and when it should be available. These rules are fired by a context-awareness runtime environment, which gathers all known context information about a device and produces a context profile for that device (e.g., physical location, date/time, device type, network IP address, and user language).

The ACE is divided into three modules that collaborate to implement a dynamic management system that allows the administrator to control the capability of each device once they are connected to the wireless network. The Device Management Module provides management tools to deploy control and maintain the set of mobile devices. The Synchronization Module manages the exchange of files between corporate systems and mobile hand-held devices. Finally, the Device Management is continuously provided with updated versions of the configuration files. Figure 3 shows the integration of the Positioning and Facilitator systems in the proposed framework for developing multimodal dialog systems.

2.4 Fusion of input modalities and dialog management

As previously described, the objective of fusion in multimodal dialog systems is to process the input information and assign a semantic representation which is eventually sent to the dialog manager. Two main levels of fusion are often used: feature-level fusion, semantic-level fusion. The first one is a method for fusing low-level feature information from parallel input signals within a multimodal architecture. The second one is a method for integrating semantic information derived from parallel input modes in a multimodal architecture.

Semantic-level fusion is usually involved in the dialog manager and needs to consult the knowledge source from the dialog history and data repositories. Three

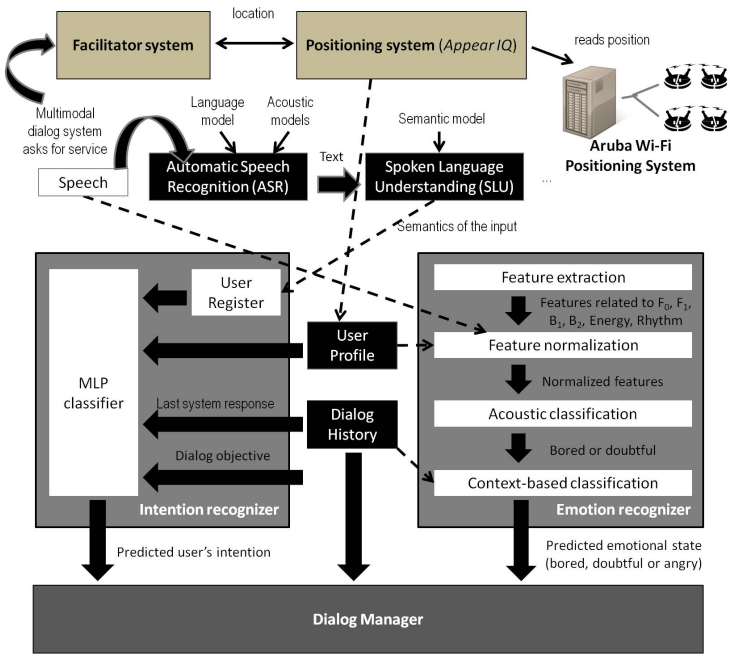

Fig. 3. Schema for the acquisition and processing of external and internal contextual information

popular semantic fusion techniques are used. Frame-based fusion is a method for integrating semantic information derived from parallel input modes [25].

Unification-based fusion is a logic-based method for integrating partial meaning fragments derived from two input modes into a common meaning representation during multimodal language processing. Compared with frame-based fusion, unification-based fusion derives from logic programming, and has been more precisely analyzed and widely adopted within computational linguistics (e.g. [26]).

Hybrid symbolic/statistical fusion is an approach to combine statistical processing techniques with a symbolic unification-based approach (e.g. Members-Teams-Committee (MTC) hierarchical recognition fusion [27]). Another related work on low-level fusion is sensor fusion, which is the combining of sensory data from disparate sources such that the resulting information is in some sense better than would be possible when these sources were used individually

To deal with the input information sources and transmit this information to the dialog manager, we propose the use of EMMA (Extensible MultiModal Annotation markup language, `www.w3.org/TR/emma/`), developed by the W3C Multimodal Interaction Framework (`www.w3.org/TR/mmi-framework/`) and intended for use by systems that provide semantic interpretations for a variety of inputs, including speech recognition, handwriting recognizers, natural language understanding engines, and other input media interpreters (e.g. DTMF, pointing,

keyboard), as well that multimodal integration component and the interaction manager.

EMMA is focused on annotating single inputs from users, which may be either from a single mode or a composite input combining information from multiple modes, as opposed to information that might have been collected over multiple turns of a dialog. The language provides a set of elements and attributes that are focused on enabling annotations on user inputs and interpretations of those inputs. The attribute *emma* : *hook* can be used to mark the elements in the application semantics within an *emma* : *interpretation*, which are expected to be integrated with content from input in another mode to yield a complete interpretation. Figure 4 shows an example EMMA code in which this attribute is used to integrate a spoken and a visual input.

```
<emma:emma version="1.0"
    xmlns:emma="http://www.w3.org/2003/04/emma"
    xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
    xsi:schemaLocation="http://www.w3.org/2003/04/emma
     http://www.w3.org/TR/2009/REC-emma-20090210/emma.xsd"
    xmlns="http://www.example.com/example">
  <emma:interpretation id="voice2"
      emma:medium="acoustic"
      emma:mode="voice"
      emma:function="dialog"
      emma:confidence="0.4"
      emma:tokens="I want to go there"
      emma:start="1087995961500"
      emma:end="1087995963542">
    <command>
      <action>send</action>
        <arg1>
          <object emma:hook="ink">
            <type>file</type>
            <number>1</number>
          </object>
        </arg1>
       <arg2>
         <object emma:hook="ink">
           <number>1</number>
         </object>
       </arg2>
    </command>
  </emma:interpretation>
</emma:emma>
```

Fig. 4. Example of EMMA document dealing with several input modalities

The methodology that we propose for the multimodal data fusion and dialog management processes considers the set of input information sources (spoken interaction, visual interaction, user intention modeling, and user emotional state) by means of a machine-learning technique. The dialog manager receives EMMA files containing the results processed by the modules that deal with each input modality. As in our previous work on user modeling and dialog management [21, 11], we propose the definition of a data structure to store the values for the different concepts and attributes provided by means of the different input modalities along the dialog history.

The information stored in this data structure, that we called Interaction Register (IR), is coded in terms of three values, $\{0, 1, 2\}$, for each field according to the following criteria:

- **0**: The value of the specific position of the IR has not been provided by means of any of the input modalities or sources defined as interaction context.

- **1**: The value of the specific position of the IR has been provided with a confidence score that is higher than a given threshold. Confidence scores are provided by different modules that process the information acquired for each input modality (e.g., the ASR and SLU modules for the spoken utterances).
- **2**: The value of the specific position of the IR has been provided with a confidence score that is lower than the given threshold.

The information contained in the IR at each time i has been generated considering the values extracted from the EMMA files along the dialog history. Each slot in the IR can be usually completed by means of more tan one input modality. If just one value has been received for a specific dialog act, then it is stored at the corresponding slot in the IR using the described codification. Confidences scores provided by the modules processing each input modality are used in case of conflict among the values provided by several modalities for the same slot. Thus, a single input is generated for the dialog manager to consider the next system response.

As in our previous work on dialog management [11], we propose the use of a classification process to determine the next system response given the single input that is provided by the interaction register after the fusion of the input modalities and also considering the previous system response. This way, the current state of the dialog is represented by the term (IR_i, A_{i-1}), where A_{i-1} represents the last system response. The values of the output of the classifier can be viewed as the a posteriori probability of selecting the different system responses given the current situation of the dialog.

3 Conclusions and future work

In this paper we have described a framework to develop multimodal systems that considers information provided by means of spoken, visual and tactile input modalities. We carry out an additional step towards the adaptation of these systems by also modeling the context of the interaction in terms of external and internal context, which in our case is related to the detection of the user's intention and emotional state.

Several modules have been incorporated in the classical architecture of a spoken dialog system to achieve the integration of the additional input modalities and contextual information sources. These modules respectively allow to predict the next user response for the conversational agent and carry out the fusion of visual and spoken information. The proposed multimodal fusion and dialog management technique allows considering these heterogeneous information sources to select the next system action by means of a classification process.

Although the different methodologies proposed to develop the described modules integrated in the multimodal dialog system have been evaluated in previous works [21, 23, 24, 11], as a future work we propose the application of the described framework to develop and evaluate a practical system in a real environment.

Acknowledgements

This work was supported in part by Projects MINECO TEC2012-37832-C02-01, CICYT TEC2011-28626-C02-02, CAM CONTEXTS (S2009/TIC-1485).

References

1. Filipe, P., Mamede, N. In: Ambient Intelligence Interaction via Dialogue Systems. Intech (2010) 109–124
2. López-Cózar, R., Callejas, Z. In: Multimodal Dialogue for Ambient Intelligence and Smart Environments. Springer (2010) 559–579
3. Jaimes, A., Sebe, N.: Multimodal human-computer interaction: A survey. Computer Vision and Image Understanding **108** (2007) 116–134
4. Turk, M.: Multimodal interaction: A review. Pattern Recognition Letters **36** (2014) 189–195
5. López-Cózar, R., Araki, M.: Spoken, Multilingual and Multimodal Dialogue Systems. John Wiley & Sons Publishers (2005)
6. Pieraccini, R.: The Voice in the Machine: Building Computers that Understand Speech. The MIT Press (2012)
7. Wahlster, W.: SmartKom: Foundations of Multimodal Dialog Systems. Springer (2006)
8. Dumas, B.: Frameworks, description languages and fusion engines for multimodal interactive systems. Master's thesis, University of Fribourg, Fribourg (Switzerland) (2010)
9. Traum, D., Larsson, S. In: The Information State Approach to Dialogue Management. Kluwer (2003) 325–353
10. Williams, J., Young, S.: Partially Observable Markov Decision Processes for Spoken Dialog Systems. In: Computer Speech and Language 21(2). (2007) 393–422
11. Griol, D., Hurtado, L., Segarra, E., Sanchis, E.: A Statistical Approach to Spoken Dialog Systems Design and Evaluation. Speech Communication **50**(8-9) (2008) 666–682
12. Ruiz, N., Chen, F., Oviatt, S. In: Multimodal input. Elsevier (2010) 211–277
13. Dai, X., Khorram, S.: Data fusion using artificial neural networks: a case study on multitemporal change analysis. Computers, Environment and Urban Systems **23**(1) (1999) 19–31
14. Tsilfidis, A., Mporas, I., Mourjopoulos, J., Fakotakis, N.: Automatic speech recognition performance in different room acoustic environments with and without dereverberation preprocessing. Computer Speech & Language **27**(1) (2013) 380–395
15. Wu, W.L., Lu, R.Z., Duan, J.Y., Liu, H., Gao, F., Chen, Y.Q.: Spoken language understanding using weakly supervised learning. Computer Speech & Language **24**(2) (2010) 358–382
16. Minker, W.: Design considerations for knowledge source representations of a stochastically-based natural language understanding component. Speech Communication **28**(2) (1999) 141–154
17. Traum, D., Larsson, S.: The Information State Approach to Dialogue Management. Current and New Directions in Discourse and Dialogue. Kluwer Academic Publishers (2003)

18. Möller, S., Englert, R., Engelbrecht, K., Hafner, V., Jameson, A., Oulasvirta, A., Raake, A., Reithinger, N.: MeMo: towards automatic usability evaluation of spoken dialogue services by user error simulations. In: Proc. Interspeech'06. (2006) 1786–1789
19. Chung, G.: Developing a flexible spoken dialog system using simulation. In: Proc. ACL'04. (2004) 63–70
20. Schatzmann, J., Weilhammer, K., Stuttle, M., Young, S.: A Survey of Statistical User Simulation Techniques for Reinforcement-Learning of Dialogue Management Strategies. Knowledge Engineering Review **21**(2) (2006) 97–126
21. Griol, D., Carbó, J., Molina, J.: A statistical simulation technique to develop and evaluate conversational agents. AI Communication **26**(4) (2013) 355–371
22. Schuller, B., Batliner, A., Steidl, S., Seppi, D.: Recognising realistic emotions and affect in speech: state of the art and lessons learnt from the first challenge. Speech Communication **53**(9-10) (2011) 1062–1087
23. Callejas, Z., López-Cózar, R.: Influence of contextual information in emotion annotation for spoken dialogue systems. Speech Communication **50**(5) (2008) 416–433
24. Griol, D., Carbó, J., Molina, J.: Bringing context-aware access to the web through spoken interaction. Applied Intelligence **38**(4) (2013) 620–640
25. Vo, M., Wood, C.: Building an application framework for speech and pen input integration in multimodal learning interfaces. In: Proc. of ICASSP'96. (1996) 3545–3548
26. Johnston, M.: Unification-based multimodal parsing. In: Proc. of ACL'96. (1996) 624–630
27. Wu, L., Oviatt, S., Cohen, P.: From members to teams to committee - a robust approach to gestural and multimodal recognition. IEEE Transactions on Neural Networks **13**(4) (2002) 972–982