Preface

The International Workshop on Spoken Dialog Systems (IWSDS) series provides an international forum for the presentation of research and applications and for lively discussions among researchers as well as industrialists, with a special interest to the practical implementation of spoken dialog systems in everyday applications.

Following the success of IWSDS'09 (Irsee, Germany), IWSDS'10 (Gotemba Kogen Resort, Japan), and IWSDS'11 (Granada, Spain), the Fourth IWSDS'12 took place at the castle of Ermenonville, near Paris (France), on November 28–30, 2012.

This book consists of the revised versions of a selection of the papers that were presented at the IWSDS'12 conference.

Spoken dialog has been a matter of research investigations for many years. The first spoken language processing systems aimed at providing such an interaction between humans and machines. It slowly appeared that the problem was much more difficult than it was initially thought, as it involves many different components: speech recognition and understanding, prosody analysis, indirect speech acts, dialog handling, maintenance of the communication with verbal or nonverbal events such as backchannels, speech generation and synthesis, multimodal fusion and fission. Social interaction among humans is characterized by a continuous and dynamic exchange of information carrying signals. Producing and understanding these signals allow humans to communicate simultaneously on multiple levels. The ability to understand this information, and for that matter adapt generation to the goal of the communication and the characteristics of particular interlocutors, constitutes a significant aspect of natural interaction. It shows that it is actually very complex to develop simple, natural interaction means.

Even if the research investigations kept on being conducted, it induced a shift of interest to easier tasks, such as voice command, voice dictation, or speech transcription. However, scientific achievements in language processing now result in the development of successful applications such as IBM Watson, the Evi, Apple Siri, Google Voice Action, Microsoft Bing Voice Search, Nuance Dragon Go!, or Vlingo for access to knowledge and interaction with smartphones, while the coming of domestic robots advocates for the development of powerful communication means with their human users and fellow robots.

We put this year workshop under the theme "Towards a Natural Interaction with Robots, Knowbots and Smartphones," which covers:

- Dialog for robot interaction (including ethics)
- Dialog for open-domain knowledge access
- Dialog for interacting with smartphones
- Mediated dialog (including multilingual dialog involving speech translation)
- Dialog quality evaluation

We enjoyed the invited Keynote Talks of Jérôme Bellegarda (Apple, USA), Alex Waibel (Karlsruhe Institute of Technology (Germany) and Carnegie Mellon University (USA)), Axel Buendia (SpirOps) and Laurence Devillers (LIMSI-CNRS and University Paris-Sorbonne, France) and Marilyn Walker (UCSC, USA) on those topics. We also had an invited talk on the conclusions of the SemDial workshop on the semantics and pragmatics of dialog, which took place in Paris in September 2012, by its organizer, Jonathan Ginzburg (University Paris Diderot). We warmly thank all of them.

We also encouraged the presentation and discussion of common issues of theories, applications, evaluation, limitations, general tools, and techniques. We particularly welcomed papers that were illustrated by a demonstration.

This book first includes several parts on the implementation of spoken dialog systems for various areas of application and especially those related to the main topics of the conference: smartphones, robots, and knowbots. It then has a part on spoken dialog systems components and a final one on spoken dialog management.

The first part deals with spoken dialog systems in everyday applications. First, Jérôme Bellegarda from Apple Inc. presents the Siri experience, which has had a tremendous impact in the actual use of spoken interaction on personal assistants. He introduces the two major semantic interpretation frameworks, statistical and rule-based, discusses the choices made in Siri, and speculates on how the current implementation might evolve in the near future. Hansjörg Hofmann and colleagues from Daimler AG depict the development of speech-based in-car human-machine interaction for information exchange. The permanent use of smartphones impacts the automotive environment, necessitating an intuitive interface in order to reduce driver distraction. They investigate two different dialog strategies, command-based or conversational speech dialog, and different graphical user interfaces, one including an avatar. Those prototypes are evaluated regarding usability and driving performance. Alan Black and Maxine Eskenazi address the problem of developing spoken dialog systems with controlled users, who may not act as real users, in a study related to a task of providing bus information hosted at Carnegie Mellon University. They report on several lessons learned from the experience and provide recommendations on various approaches, including crowdsourcing. Daniel Sonntag and Christian Schulz from DFKI describe the use of a multimodal multi-device infrastructure for collaborative decision-making in the medical area: the Radspeech industrial prototype. In their study, two radiologists use two different mobile speech devices (Apple iPhone and iPad) and collaborate via a connected large screen installation, jointly using pointing and spoken interaction.

The second part presents five examples of spoken dialog prototypes and products in different domains such as crosslingual communication, city exploration and services, or ambient intelligence environments.

First, Feiyu Xu and colleagues from Yocoy and DFKI LT Lab (Germany) describe Yochina, a mobile multimedia and multimodal crosslingual dialog system. The mobile application combines language technologies such as speech synthesis, template-based translation, and dialog to offer language and travel guide without depending on an Internet connection. A novel strategy of linking provided knowledge with covered communication situations is explained. Yochina is available for two language pairs: English to Chinese and German to Chinese. Johan Boye and colleagues from KTH and Liquid Media (Sweden) address the challenging problem of giving navigation instructions to pedestrians through a spoken dialog approach rather than a map-based approach. It means interpreting and generating utterances within a rapidly changing spatial context even though the pedestrian's position, speed, and direction are uncertain due to possible GPS errors. They present the results of a user experiment conducted in Stockholm. The paper by Nieves Ábalos and colleagues from the Department of LSI, University of Granada, and from Systems Laboratory, University Rey Juan Carlos (Spain), deals with a multimodal dialog system to enable user control of home appliances in an Ambient Intelligence environment (lights, TV, etc.). It describes the interaction of Mayordomo, a multimodal dialog system which uses either spontaneous speech or a traditional GUI, with Octopus, a system which enables AmI applications through a file-based service access. Sunao Hara and colleagues from the Graduate School of Information Science at the Nara Institute of Science and Technology (Japan) depict a toolkit for multi-agent server-client spoken dialog systems: *tankred on rails (ToR)*. iTakemaru is the client software for mobile phones. It provides a speech-guidance service handling one main agent and multiple subagents. It allows the client to obtain more information thanks to the communication between the main agent and the subagents based on a server-to-server communication. The last paper of this part describes a voice portal based on the VoiceXML standard to provide the citizens with municipal information (city council, city services, etc.). The authors, David Griol and colleagues from the Computer Science Department, Carlos III University of Madrid, and the Department of Languages and Computer Systems, University of Granada (Spain), give the results of both a subjective evaluation, through quality assessments, and an objective evaluation (successful dialogs, average number of turns per dialog, confirmation rate, etc.).

The third part (Multi-domain, Crosslingual Spoken Dialog Systems) deals with model adaptation when facing changes of languages or domains.

Teruhisa Misu and colleagues, from the National Institute of Information and Communication Technology, address a very actual issue of cross-domain/cross-language portability of dialog systems. They present an approach for extending a language model designed for one task in a given language to another task by using resources in other languages or tasks using statistical machine translation systems. They propose a selection mechanism to automatically extract relevant parts in those resources, based on a spoken language understanding module corresponding to the

source language and task. Pierre Lison, from the University of Oslo, addresses the problem of online learning of dialog policy. The proposed approach relies on probabilistic rules (in order to simplify the inference) and on a Monte Carlo sampling method to determine the best action to perform. Injae Lee and colleagues, from the Pohang University of Science and Technology (Korea) and the Institute for Infocomm Research (Singapore), address the problem of the domain selection for a multiple-domain dialog system. The proposed approach includes a domain preselection, which provides, for each user utterance, a list of possible domains associated with scores. Then a content-based filtering method is performed on the domain candidate list to select the final domain. The experimental results show an improvement in terms of accuracy and processing time compared to more standard approaches.

The fourth part deals with dialog for robot interaction, including ethics.

First, Alex Buendia from the French SpirOps SME and Laurence Devillers from LIMSI-CNRS and University Paris-Sorbonne address the challenges for going from informative cooperative dialogs to long-term social relationship with a robot. They aim at exploring the ability of a robot to create and maintain a long-term social relationship through more advanced dialog techniques. They expose the social, psychological, and neural theories used to accomplish such complex social interactions. From these theories, they build a consistent, computationally efficient model to create a robot that can understand the concept of lying and have compassion: a robotic social companion. Taichi Nakashima, Kazunori Komatani, and Satoshi Sato from the Graduate School of Engineering at Nagoya University in Japan propose the integration of multiple sound source localization results for speaker identification in a multiparty dialog system. They present a method of identifying who is speaking more accurately by integrating the multiple sound source localization results obtained from two robots. The experimental evaluation revealed that using two robots improved speaker identification compared with using only one robot.

Ina Wechsung, Patrick Ehrenbrink, Robert Schleicher, and Sebastian Möller from the Quality and Usability Lab of the Berlin Telekom Innovation Laboratories at the Technical University of Berlin investigate the social facilitation effect in human-robot interaction. The current study indicates that a higher degree of human likeness results in a social inhibition effect. In this experiment, the reported differences were caused by the appearance of the robot, whereas its synthetic voice was kept constant. After the social inhibition as well as the uncanny valley effect could be confirmed for this setup, it would be interesting to study whether the same effect can also be observed for voices with different degrees of anthropomorphism. Emer Gilmartin and Nick Campbell from the Speech Communications Lab, Trinity College Dublin, present how to build a chatty robot. Their work describes the design and implementation of a robot platform for the extraction of data and acquisition of knowledge related to spoken interaction, by capturing natural language and multimodal/multisensorial interactions using voice-activated and movement-sensitive sensors in conjunction with a speech synthesizer.

Takaaki Sugiyama, Kazunori Komatani, and Satoshi Sato from the Graduate School of Engineering at Nagoya University tackle the novel problem of predicting when a user is likely to begin speaking to a humanoid robot. Clément Chastagnol, Céline Clavel, Matthieu Courgeon, and Laurence Devillers from LIMSI-CNRS show how to design an emotion detection system for a socially intelligent human-robot interaction. This work is part of the French ANR ARMEN project that aims at designing and building a prototype for a robotic companion (RC) for the elderly and disabled people. In their paper, Kristiina Jokinen and Graham Wilcock from the University of Helsinki present ongoing work on multimodal interaction with the Nao robot, including speech, gaze, and gesturing. It also describes the interaction with the Nao robot from the point of view of constructive dialog modeling and demonstrates how the framework can be applied to the WikiTalk open-domain interaction. Finally, Ridong Jiang, Yeow Kee Tan, Dilip Kumar Limbu, Tran Anh Dung, and Haizhou Li from the Institute for Infocomm Research in Singapore describe a component pluggable dialog framework, which is domain-independent, cross-platform, and multilingual, and its application to the interface with social robots, showing a shorter development cycle while improving the system robustness, reliability, and maintainability.

The last two parts of this book are about the development of specific aspects of dialog systems. The fifth part (Spoken Dialog Systems components) is about specific components while the sixth specifically concerns the dialog management module.

In the fifth part, Martin Heckmann, from the Honda Research Institute Europe, investigates the use of acoustic and visual cues to detect prominent (e.g., corrected) words in an utterance. The experiment shows that when using only the fundamental frequency as an acoustic feature, the improvement of the classification is interesting when combining to this acoustic feature the visual features but that when all possible acoustic features are used, the combination with visual features allows for a less important gain. Bart Ons and colleagues, from ESAT-PSI (KU Leuven), address the problem of robustness of a direct mapping between an acoustic signal and a command in the context of a learning system. The proposed approach is based on a supervised nonnegative matrix factorization. The results show that this learning approach is robust to label noise. Rafael Torres and colleagues, from the Nara Institute of Science and Technology and from the Institute of Statistical Mathematics in Tokyo, present a work on topic classification of spoken user utterances received by a guidance system. They specifically study a semi-supervised approach, using a transductive support vector machine and the impact of the inclusion of unlabeled examples during the training process of the classifier. Experimental results show that this approach can be useful for taking advantage of unlabeled samples, which are simpler to obtain than labeled ones.

Yoo Rhee Oh and colleagues, from the Spoken Language Processing Team, Electronics and Telecommunications Research Institute (ETRI, Korea), address the problem of the decoding of nonnative speech. Most automatic speech recognition systems have to face one important problem: speakers can be nonnative and then the performance of the system decreases. The proposed decoding strategy consists in decoding speech with both native and nonnative speakers models and selecting,

based on the likelihood scores, which model to use for each frame to decode. The experimental results show a reduction of the word error rate. Marcela Charfuelan and Geert-Jan Kruijff, from DFKI GmbH, are interested in analyzing speech under stress. They address the problem of acoustical analysis of stress in a USAR database and examine a range of acoustical cues which are annotated by two annotators into the categories of neutral, medium, or high stress. Analysis results show that traditional prosody and acoustic features are robust enough to discriminate among the different types of stress and neutral data.

In the sixth part, Marilyn Walker and colleagues, from the University of California at Santa Cruz, address the problem of adapting the answers of dialog agents to a particular user, either within the context of a single interaction or over time. A general spoken language generation framework is presented along with dynamic generation for task-oriented dialog systems and most importantly expressive generation. Stefan Ultes and colleagues, from the Institute of Communications Technology (University of Ulm), address the problem of an interaction quality estimator in spoken dialog systems. They describe how conditioned hidden Markov models (CHMM) can be used to estimate the interaction quality of a spoken dialog system, developed for the "Let's Go Bus Information System." Unfortunately using CHMM does not allow for improvements in the results compared to standard approaches such as HMM or SVM. Fabrizio Morbini and colleagues, from the Institute for Creative Technologies (University of Southern California), present a dialog manager based on the information-state update approach that performs forward inference and exploits local dialog structures. This approach is related to plan-based approaches of dialog management with the addition of rewards attributed to specific states. Two examples of implementation are described. Zoraida Callejas and colleagues, from the University of Granada, Carlos III University of Madrid, and the Quality and Usability Lab (Deutsche Telekom Laboratories), are interested in using user profiles to implement intelligent dialog systems. They proposed an approach to cluster user profiles using interaction parameters and overall quality prediction. They provide experimental results related to young and senior user groups and to users with high vs. low technical skills. The general conclusion is that a better grouping of users should distinguish between three groups and not four: young users with high technical affinity, senior users with low technical affinity, and a third group considering the remaining users.

Etsuo Mizukami and Hideki Kashioka, from the National Institute of Information and Communications Technology (NICT), introduce an extension to the dialog mechanism of grounding, called the extended grounding networks. They implemented this extended grounding network using the concept of contribution topics, in the context of touristic information systems. The contribution topics are units of achievement corresponding to discourse segments. Senthilkumar Chandramohan and colleagues, from Supelec, CNRS-Georgia Tech and University of Avignon/LIA-CERI, present a work developed in the context of stochastic-based dialog management. They describe a coadaptation framework and a method to learn optimal dialog policies by taking into account the adaptation of users to systems over time. Experimental results show that this coadaptation framework is

a robust approach for facilitating dialog evolution. Lasguido and colleagues, from the Nara Institute of Science and Technology and the Faculty of Computer Science (Universitas Indonesia), are interested in non-goal-oriented dialog systems. In this framework, they present a method, based on the example-based dialog management approach, for developing a dialog manager by generalizing from examples from drama television (the Friends TV show) in order to achieve more natural dialog interaction. The main problem in such an approach is to select the useful examples. They propose a tri-turn unit for dialog extraction and semantic similarity analysis techniques to ensure that the content extracted from drama script files forms an appropriate dialog example.

Klaus-Peter Engelbrecht, from the Quality and Usability Lab, Telekom Innovation Laboratories (TU Berlin), presents a causal user model for user simulation as it is used for spoken dialog systems development. The approach is based on connectionist models of human behavior. The objective of this work is to generate user simulators which are more meaningful and portable across tasks. The presented approach relies on parameters of the model that are related to the characteristics of the users and the task, and the model is useful to explain why a specific behavior is observed. Finally, Sanat Sarda and colleagues, from Nanyang Technological University, are interested in providing real-time feedback about an ongoing conversation to speakers. The system extracts various kinds of information such as speaking time, speaker turns, and duration. This information is then displayed in real time. This is somehow a monitoring system on ongoing conversations. The extracted information is then displayed in different ways to the speakers using icons, animation, etc. Haruka Majima and colleagues, from the Graduate School of Information Science at Nara Institute of Science and Technology, the Graduate School of Natural Science and Technology at Okayama University, and the Department of Statistical Modeling at the Institute of Statistical Mathematics (Japan), present a method for detecting invalid inputs for a spoken dialog system. Invalid inputs include background voices, which are not directly uttered to the system, and nonsense utterances. The main idea is to feed the decision method with different features like signal-noise ratio, utterance duration, and bag of words (BOW) when available. They compare two different methods, one based on SVM and the other on maximum entropy. The SVM-based methods reached an F-measure of 0.870 while the ME-based one obtained a $F = 0.837$. This has to be compared to the baseline method (GMM-based) which reached $F = 0.817$.

Finally, we wish to thank the IWSDS Steering Committee, the members of the IWSDS 2012 Organizing Committee and Scientific Committee, the participating and supporting organizations, and our sponsors: ELSNET (the European Language and Speech Network), ELRA (the European Language Resources Association), and the QUAERO project.

Orsay, France

Joseph Mariani
Sophie Rosset
Martine Garnier-Rizet
Laurence Devillers

http://www.springer.com/978-1-4614-8279-6

Natural Interaction with Robots, Knowbots and Smartphones
Putting Spoken Dialog Systems into Practice
Mariani, J.; Rosset, S.; Garnier-Rizet, M.; Devillers, L. (Eds.)
2014, XVI, 397 p. 132 illus., 73 illus. in color., Hardcover
ISBN: 978-1-4614-8279-6