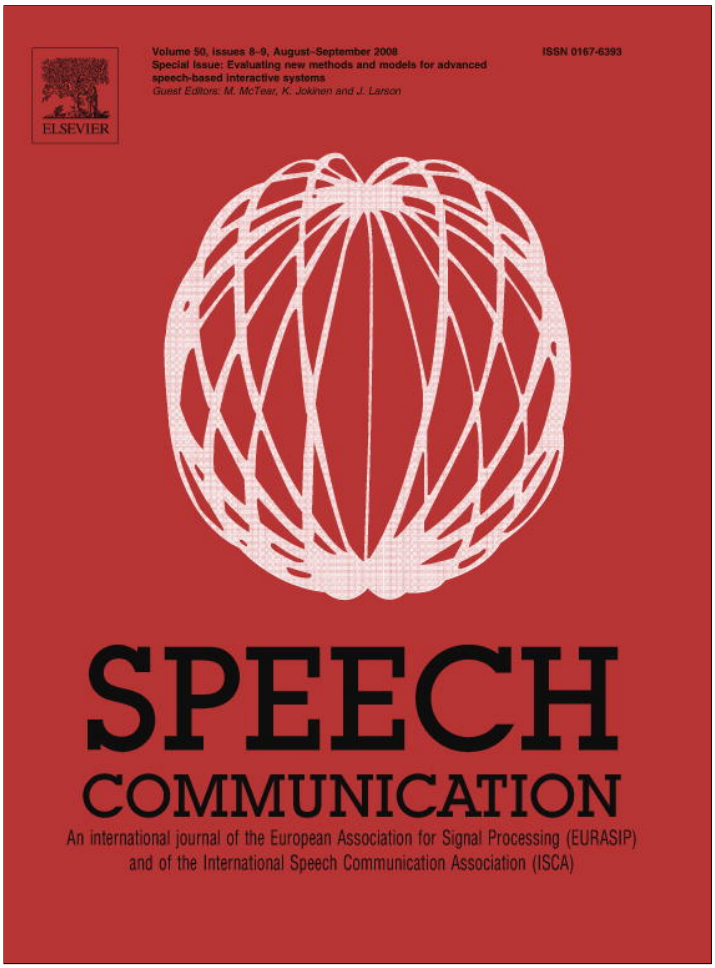

Available online at www.sciencedirect.com
ScienceDirect
Speech Communication 50 (2008) 627–629
SPEECH COMMUNICATION
www.elsevier.com/locate/specom

Editorial

Special Issue on "Evaluating new methods and models for advanced speech-based interactive systems"

Introduction

In recent years there has been a growth of research focusing on communication models for speech-based interactive systems that go beyond those currently used in commercial spoken dialogue systems. One aim of this research is to increase the system's communicative competence by incorporating more advanced characteristics of spoken interaction including speech segmentation, disfluencies, turn-taking, emotions, and adaptation. Another research direction involves dialogue strategies: the development of more advanced models of dialogue that enable more natural, more robust, and more efficient communication. To support these additional functionalities various advanced architectures have been developed that incorporate sophisticated components and techniques not present in current commercial systems. At the same time, researchers have been applying new methods for the automatic design and optimization of spoken dialogue systems, such as example-based learning and reinforcement learning.

One of the motivations for this research is to provide interaction capabilities that will ultimately lead to more natural and more usable practical applications. However, relatively little work has so far been devoted to defining the criteria for evaluating speech interaction systems. Typical criteria for evaluating interactive dialogue systems include subjective and objective measures. Subjective measures involve user surveys and feedback aggregated and reduced into statistics measuring the users' opinion about evaluation criteria such as ease of use, ease of learning, and problem coverage. Objective criteria include measurable aspects of the interaction and system properties such as the length of interaction, processing time, number of errors per time unit, and amount of data entered per time unit. Together these criteria are used to indicate user satisfaction and usability of the system. They do not necessarily demonstrate that the model/technique is useful or even promising for practical purposes, and subjective measures can also be difficult as the collection of user data requires time and resources that may not always be possible or easily accessible. Often it would also be useful to run quick experiments with the system without elaborated user studies in order to achieve understanding of the strengths and weaknesses of various modelling techniques, algorithms, and architectures used in the implementation and internal functioning of the system. Thus there is also a need for simulations and automated tools and techniques that would allow efficient evaluation of the system's performance and of the suitability of new algorithms for particular tasks and typical interactions.

In terms of evaluation it is also important to distinguish the user's expectations from their real experience with the system. The user's expectations guide their evaluation of the system, and their experience with the system may be positive or negative depending on how well the system addresses their expectations. Also the frequency of use of the system influences the user's evaluation of the system. Users adapt to particular strategies and standard techniques, and their views may change accordingly. In general, only by running experiments using two or more methods/models for speech-based interactive systems can researchers determine how much better one advanced system is over the other with respect to well-defined criteria.

These issues were the motivation for a workshop 'Dialogue on Dialogues' that was organized as a satellite event at the InterSpeech 2006 conference in Pittsburgh, Pennsylvania on September 17, 2006. The workshop involved more than 40 participants from Europe, the United States, Australia, Japan, and South Korea. For a summary of the workshop, see Jokinen et al. (2007).

Following the workshop it was decided to invite contributions for a Special Issue on the topics of the workshop. The aim of this Special Issue is to explore new evaluation techniques and strategies as applied to advanced dialogue systems, including new models and methods. Submissions were invited on the following topics:

- What characteristics of spoken language interaction can and should be incorporated into advanced spoken dialogue systems?

0167-6393/$ - see front matter
doi:10.1016/j.specom.2008.05.001

- What are the best methods for designing such systems? To what extent are automatic design methods appropriate or possible?
- What criteria can be defined for the evaluation of the performance of advanced spoken dialogue systems?
- Under what circumstances can and should these criteria be used?
- How effective are these criteria in isolating problems with a dialogue strategy and in measuring how addressing and correcting the problems leads to improvements in the dialogue?
- How can these criteria be used to compare and evaluate alternative dialogue strategies and methods for the design and implementation of dialogue systems?
- How can the evaluation process be streamlined so that it can be frequently and effectively applied to the improvement and comparison of dialogue strategies?

Each submission was reviewed by three reviewers, and eight papers were selected to be included in this issue. The papers can be grouped broadly into two main headings: those describing new models and methods for advanced spoken dialogue systems and those focusing primarily on new evaluation methods. Several of the papers cover both topics in that they introduce a new methodology and also techniques for its evaluation.

In the first paper *Edlund*, *Gustafson*, *Heldner*, *and Hjalmarsson* address the issue of more human-life spoken dialogue systems, beginning with an exploration of the nature of human and interface metaphors as well as their implications and desirability for design. The paper then reviews methods for the elicitation and analysis of user responses to human-like spoken dialogue systems such as variations on Wizard-of-Oz techniques that provide a set of tools to be used for evaluating human-like spoken dialogue systems. The issue of more natural spoken dialogue systems is also addressed by *López-Cózar and Callejas*, but more specifically in terms of a technique for post-correction of speech recognition errors that takes into account the context of the error to be corrected. Correction is carried out at syntactic–semantic and lexical levels which employ statistical models of the words, and at a contextual level which is used to decide whether a recognition result is correct or whether it may be changed. Experimental results indicated improvements in word accuracy, speech understanding, implicit recovery and task completion rates through the use of this technique.

The next three papers are concerned with new methods for the development of dialogue models using statistical methods. *Griol*, *Hurtado*, *Segarra*, *and Sanchis* describe an approach which is used to determine the system's dialogue strategy based on models learned from a training corpus. A user simulator was employed to overcome the issue of data sparseness. The paper addresses issues such as the efficient representation of the dialogue history and the classification of user responses that are absent in the corpus. *Tetreault and Litman* discuss the use of reinforcement learning for the development of a dialogue manager to be used in a spoken dialogue tutoring system. They focus specifically on metrics for the evaluation of the best features to be used in the state space representation of a model that is used to learn an optimal dialogue strategy. The choice of appropriate features is motivated by the need to train dialogue policies more quickly and to avoid issues of data sparseness. The authors present a methodology of confidence intervals that support a better assessment of the reliability of the proposed metrics, and they empirically evaluate the effects of these features in terms of the system's strategies. *Jung*, *Lee*, *Kim and Lee* describe a data-driven spoken dialogue system workbench to support the development and management of data-driven dialogue systems. The tool helps the designer to manage otherwise burdensome and time-consuming tasks such as corpus preparation and annotation, testing, and component integration. The usability of the tool was evaluated by developing dialogue systems in different domains using different dialogue management methods.

The final three papers focus primarily on new methods for evaluation. *Paek and Pieraccini* address the issue of the deployment in industry of new methods and models developed in academic research. In particular, they evaluate the strengths and weaknesses of new methods that use machine-learning techniques such as reinforcement learning. They examine current approaches to the automation of spoken dialogue management in industrial applications, and how these new methods and models might be deployed in practical applications. The paper by *Möller, Engelbrecht and Schleicher* is concerned with different approaches for predicting the quality and usability of spoken dialogue systems. Different methods and models are compared in terms of their predictive power, and their usefulness for taking decisions concerning component optimisation is assessed. *Callejas and López-Cózar* discuss the relationship between instrumentally based (objective) and quality based (subjective) measures for evaluation They report empirical results from statistical studies which involve real users interacting with spoken dialogue systems. The value of these results is that they can provide guidelines for the refinement of systems as well as showing how aspects of quantitative evaluation may affect more subjective user perceptions of the systems.

We are confident that the papers give a good overview of the issues related to the evaluation of advanced spoken dialogue systems, and they also point to important challenges that will be studied and further discussed in the future.

The editors would like to thank the following people for their help in reviewing for this Special Issue.

Masahiro Araki, Kyoto Institute of Technology, Japan.
Mats Blomberg, KTH, Sweden.
Dan Bohus, Microsoft Research, USA.
Elizabeth Owen Bratt, CSLI, Stanford University, USA.
Morena Danieli, Loquendo, Italy.
Giuseppe Di Fabbrizio, AT&T, USA.

Kallirroi Georgila, University of Edinburgh, UK.
Juan Gilbert, Auburn University, USA.
Dilek Zeynep Hakkani-Tur, ICSI, Berkeley, USA.
Jaakko Hakulinen, University of Tampere, Finland.
Stefan Hamerich, University of Ulm, Germany.
Junling Hu, Bosch, USA.
Yasuhiro Katagiri. ATR, Japan.
Jörn Kreutel, SemanticEdge, Germany.
Gary Geunbae Lee, POSTECH, Republic of Korea.
David A. van Leeuwen, TNO Human Factors, Soesterberg, The Netherlands.
Esther Levin, SAC Capital Advisers, LLC, USA.
Gina-Anne Levow, University of Chicago, USA.
Ramon Lopez-Cozar, University of Granada, Spain.
Dominic Massaro, UCSC, USA.
Colin Matheson, University of Edinburgh, UK.
Helen Meng, The Chinese University of Hong Kong, China.
Susan McRoy, University of Wisconsin-Milwaukee, USA.
Wolfgang Minker, University of Ulm, Germany.
Sebastian Möller, Deutsche Telekom, Germany.
Jan Nouza, Technical University of Liberec, Czech Republic.
Matthew Purver, CSLI, Stanford University, USA.
Gabriel Skantze, KTH, Sweden.
Ronnie W. Smith, East Carolina University, USA.
Thora Tenbrink, University of Bremen, Germany.
Gokhan Tur, SRI International, USA.
Markku Turunen, University of Tampere, Finland.
Marilyn Walker, University of Sheffield, UK.
Fuliang Weng, Bosch, USA.
Steve Young, University of Cambridge, UK.

Reference

Jokinen, K., McTear, M., Larson, J.A., 2007. Dialogue on dialogues – multidisciplinary evaluation of advanced speech-based interactive systems; a report on the Interspeech 2006 satellite event. AI Mag. 28 (2), 133–136.

Michael McTear
University of Ulster,
School of Computing and Mathematics,
Shore Road, Newtownabbey,
Co. Antrim BT37 0QB,
UK
Tel.: +44 01232 368 166; fax: +44 01232 366 803
E-mail address: mf.mctear@ulster.ac.uk

Kristiina Jokinen
University of Helsinki, Finland

James Larson
Portland State University, Oregon, USA