Available online at www.sciencedirect.com

ScienceDirect

Computer Speech and Language 34 (2015) 83–86

COMPUTER SPEECH AND LANGUAGE

www.elsevier.com/locate/csl

Introduction for Speech and language for interactive robots

Abstract

This special issue includes research articles which apply spoken language processing to robots that interact with human users through speech, possibly combined with other modalities. Robots that can listen to human speech, understand it, interact according to the conveyed meaning, and respond represent major research and technological challenges. Their common aim is to equip robots with natural interaction abilities. However, robotics and spoken language processing are areas that are typically studied within their respective communities with limited communication across disciplinary boundaries. The articles in this special issue represent examples that address the need for an increased multidisciplinary exchange of ideas.

Keywords: Interactive robots; Speech localisation; Dialogue management; Multimodal interaction; Speech synthesis; Human–robot interaction

1. Aim and scope

Speech-based communication with robots faces important challenges in pushing current technology over the edge of usability in real world scenarios. In contrast to conventional interactive systems, a talking robot always needs to take its physical environment into account when communicating with users. Environments are typically unstructured, dynamic and noisy and therefore challenging for robots. The objective of this special issue is to highlight research that applies speech and language processing to robots that interact with people through speech as the main modality of interaction. For example, a robot may need to communicate with users via distant speech recognition and understand with constantly changing degrees of noise.

Alternatively, the robot may coordinate its verbal and non-verbal turn-taking behaviours as when generating speech and gestures at the same time. Speech and language technologies have huge potential to equip robots that interact naturally with their human users. However, the effectiveness of interactive robots needs to be demonstrated in real (or at least realistic) environments. This special issue presents some case studies.

2. Articles in the special issue

This special issue received a total of 20 submissions, 11 of which were originally accepted for publication. Each accepted article went through two or three rounds of reviewing, and each submission was assigned two or three reviewers. The contents of this special issue cover the following four broad topics, which are vital for interactive robots across domains:

http://dx.doi.org/10.1016/j.csl.2015.05.006
0885-2308/

2.1. Speech localisation

An important ability of interactive robots is to accurately estimate the direction of arrival of human speech, also referred to as 'sound source localisation'. This is required in order to analyse auditory scenes around robots and is an important pre-processing step followed by sound source separation and automatic speech recognition. This ability is also important to exhibit socially-interactive behaviours such as moving the robot platform and gaze to the speaker(s) in focus.

The article **A survey on sound source localisation in robotics: from binaural to array processing methods** of authors Argentieri et al. (2015) presents a survey of the state of the art in sound source localisation in robotics. It discusses topics such as embeddability, real-time, broadband environments, noise and reverberation—which are rarely taken into account simultaneously in the areas of acoustics or signal processing. The authors review binaural approaches as well as array processing techniques for localisation of human speech in robot audition.

The article **Subspace-based DOA with linear phase approximation and frequency bin selection preprocessing for interactive robots in noisy environments** of authors Lee et al. (2015) proposes a method for predicting the direction of arrival of human speech in noisy environments. The authors motivate their work in regards to the requirement of robust operation of interactive robots in the real world. The proposed method rectifies the speech signals from a microphone array affected by noise and reconstructs a representation of the received signals. The authors find that their proposed method yields improved results over conventional methods.

The article **Robust speaker localisation for real-world robots** of authors Athanasopoulos et al. (2015) describes a series of enhancements to existing acoustic localisation techniques. The authors propose novel preprocessing and time-delayed techniques for more robust localisation of human speech, which takes into account the imperfect frequency response of microphone arrays. Experimental results using a humanoid robot listening to multiple speakers report that the proposed and extended techniques improve the localisation performance in noisy and reverberant conditions.

2.2. Language understanding

Another important ability of interactive robots is to understand the meaning of human speech taking into account the entities in the spatial environment, e.g. understanding references to objects in motion. The article **Employing distance-based semantics to interpret spoken referring expressions** of authors Zukerman et al. (2015) proposes a method for interpreting spoken referring expressions. The approach considers multiple alternative recognition hypotheses at different stages including lexical, syntactic, semantics and pragmatics. At each stage, uncertainty scores are calculated and subsequently combined to reduce speech recognition errors and ambiguity. The proposed method considers the lexical similarity between the referring expression and the properties of potential referents using distance metrics. The authors report promising results from a comparison between the proposed method against humans doing the same task.

The article **Situated language understanding for a spoken dialogue system within vehicles** of authors Misu et al. (2015) studies situated language understanding in a mobile in-car system that can answer questions about the user's surroundings. The authors propose methods for understanding user queries taking into account changes in spatial relationships between the car and target buildings. They carry out an analysis of the timing in user utterances collected in a real driving setting. Language understanding takes into account the spatial relationships between car and targets, head pose of the user, and linguistic cues (expressions such as 'across the street' or 'colourful'). The analysis is then used to train probabilistic methods that can identify points of interest in referring expressions.

The article **The roles and recognition of haptic-ostensive actions in collaborative multimodal human–human dialogues** of authors Chen et al. (2015) investigates referential expressions for interactive robots for the elderly in home environments. They focus on a specific type of interaction based on multimodal actions referred to as 'Haptic-Ostensive', which not only manipulate objects but also perform a referring function in the spatial environment. The authors collect and analyse human–human dialogues in the home domain including haptic (force) actions, and train supervised models for reference resolution and dialogue act recognition. Additional experiments are reported on the recognition of actions from haptic signals measured through a sensory glove whose pressure sensors are relatively imprecise.

2.3. Dialogue management

A central component of all dialogue systems, including those for interactive robots, is the dialogue manager. It models the dialogue and makes decisions about the system's verbal actions. Although a dialogue manager can be programmed with handcrafted rules, a nowadays common approach is to train dialogue policies in presence of uncertainties that arise from, for example, speech or visual recognition errors.

The article **A hybrid approach to dialogue management based on probabilistic rules** of author Lison (2015) investigates the design of dialogue policies under uncertainty for human–robot interaction. The author is motivated by the large amount of training dialogues required by purely statistical approaches, which typically are not readily available for most dialogue domains. A middle ground between the traditional rule-based approach and a fully-trained framework is presented: the dialogue designer can define probabilistic rules and the parameters are then learned from Wizard-Of-Oz data. The framework is evaluated with a humanoid robot commanded by a human to pick up and deliver objects on a table.

The article **Reinforcement-learning based dialogue system for human–robot interactions with socially-inspired rewards** of authors Ferreira and Lefévre (2015) presents a method for efficiently inducing dialogue policies from social rewards, i.e. user appraisals gathered during the interaction. The authors argue that their method can be used at early stages of training from explicit positive and negative feedback. At later stages of training, it can be used to adapt to specific user profiles such as novice and expert. An evaluation in a virtual world reports that dialogue policies using social rewards lead to better performance than those without social rewards.

The article **Conversational system for information navigation based on POMDP with user focus tracking** of authors Yoshino and Kawahara (2015) proposes an information and navigation dialogue system that goes beyond task-oriented systems by being able to engage in small talk interactions. The authors formulate the dialogue management problem by optimising the selection of conversational modules such as greeting, silence, confirmation, story telling, proactive presentation, and question-answering. Since the performance function encourages long conversations, the longer the user interacts with the system, the better the system performance. Statistical classification is used to estimate the user's intention(s), which are passed on to a POMDP-based dialogue model for selecting the module to focus on. A human evaluation study reports that the trained dialogue policies outperform a rule-based baseline.

2.4. Speech synthesis

A further important ability of interactive robots is to generate human-like speech including emotional speech and even singing.

The article **Emotion transplantation through adaptation in HMM-based speech synthesis** of authors Lorenzo-Trueba et al. (2015) presents a method for learning speaker-independent emotional and identity features and in order to transfer them to an HMM-based speech synthesiser with neutral speech. The authors focus on transferring the following gender-independent emotions: anger, happiness, sadness and surprise. The proposed method was integrated into a robot laboratory assistant with support for facial expressions including mouth and eye movements. They find that adding emotions to neutral speech is preferred by humans in both perceptual evaluations of speech quality and human–robot interactions.

The article **HMM-based expressive singing voice synthesis with singing style control and robust pitch modelling** of authors Nose et al. (2015) proposes an HMM-based method for a singing robot with different singing styles. To control the singing style, the proposed technique requires users to provide a style vector that assigns different emphasis to different styles. The singing styles explored include 'child-like' and 'adult-like'. A listening experiment shows that the different styles and their intensities can gradually be combined using different style vectors, and that subjects perceive the different styles without experiencing the resulting speech as unnatural when compared to a baseline system.

3. Future directions

If robots are to communicate with humans in a natural way, using speech, then the topics above are essential. Although this special issue does not feature articles explicitly focused on the topics of speech recognition and language generation, these are also essential for recognising and generating text from spatially-aware dialogues.

While a multitude of future directions could be enumerated, there is one main research direction that needs to be explored: spoken language processing for robots in public spaces interacting with people with genuine needs. Efforts in this direction would be valuable for challenging speech perception (see Section 2.1), language understanding (see Section 2.2), dialogue management (see Section 2.3), and speech synthesis (see Section 2.4), among others. This direction has been motivated in previous publications (see Cuayáhuitl (2015); Mavridis (2015); Bordes et al. (2014); Cuayáhuitl et al. (2014)), and we hope that it will motivate future research towards the real application of conversational robots in the real world.

Acknowledgements

We thank our chief editor, Roger K. Moore, for encouraging this special issue, and the editorial office of this journal for their continuous support. Special thanks go to all authors for submitting their research work to this special issue. Finally, we thank the anonymous reviewers for their support in maintaining the high standard of this special issue.

References

Argentieri, S., Danées, P., Souéres, P., 2015. A survey on sound source localization in robotics: from binaural to array processing methods. Comput. Speech Lang. 34, 87–112.

Athanasopoulos, G., Verhelst, W., Sahli, H., 2015. Robust speaker localisation for real-world robots. Comput. Speech Lang. 34, 129–153.

Bordes, A., Bottou, L., Collobert, R., Roth, D., Weston, J., Zettlemoyer, L., 2014. Introduction to the special issue on learning semantics. Mach. Learn. 94 (2), 131, 127.

Chen, L., Javaid, M., Eugenio, B.D., Zefran, M., 2015. The roles and recognition of haptic-ostensive actions in collaborative multimodal human–human dialogues. Comput. Speech Lang. 34, 201–231.

Cuayáhuitl, H., March 2015. Robot learning from verbal interaction: a brief survey. In: Proceedings of the AISB Workshop on New Frontiers in Human–Robot Interaction, Canterbury, United Kingdom.

Cuayáhuitl, H., Frommberger, L., Dethlefs, N., Raux, A., Marge, M., Zender, H., 2014. Introduction to the special issue on machine learning for multiple modalities in interactive systems and robots. TiiS 4 (3).

Ferreira, E., Lefévre, F., 2015. Reinforcement-learning based dialogue system for human–robot interactions with socially-inspired rewards. Comput. Speech Lang. 34, 256–274.

Lee, S.-C., Chen, B.-W., Wang, J.-F., Liao, M.-J., Wen, J., 2015. Subspace-based DOA with linear phase approximation and frequency bin selection pre-processing for interactive robots in noisy environments. Comput. Speech Lang. 34, 113–128.

Lison, P., 2015. A hybrid approach to dialogue management based on probabilistic rules. Comput. Speech Lang. 34, 232–255.

Lorenzo-Trueba, J., Barra-Chicote, R., San-Segundo, R., Ferreiros, J., Yamagishi, J., Montero, J.M., 2015. Emotion transplantation through adaptation in HMM-based speech synthesis. Comput. Speech Lang. 34, 292–307.

Mavridis, N., 2015. A review of verbal and non-verbal human–robot interactive communication. Robot. Auton. Syst. 63 (22), 35.

Misu, T., Raux, A., Gupta, R., Lane, I., 2015. Situated language understanding for a spoken dialog system within vehicles. Comput. Speech Lang. 34, 186–200.

Nose, T., Kanemoto, M., Koriyama, T., Kobayashi, T., 2015. HMM-based expressive singing voice synthesis with singing style control and robust pitch modelling. Comput. Speech Lang. 34, 308–322.

Yoshino, K., Kawahara, T., 2015. Conversational system for information navigation based on POMDP with user focus tracking. Comput. Speech Lang. 34, 275–291.

Zukerman, I., Kim, S.N., Kleinbauer, T., Moshtaghi, M., 2015. Employing distance-based semantics to interpret spoken referring expressions. Comput. Speech Lang. 34, 154–185.

Heriberto Cuayáhuitl
Heriot-Watt University, Edinburgh, United Kingdom
Kazunori Komatani
Osaka University, Osaka, Japan
Gabriel Skantze
KTH Royal Institute of Technology, Stockholm, Sweden