Towards Speech Acquisition in Natural Interaction on ASIMO

Tobias Rodemann, Martin Heckmann, Claudius Gläser, Frank Joublin, Christian Goerick

2010

Towards Speech Acquisition in Natural Interaction on ASIMO

Tobias Rodemann, Martin Heckmann, Claudius Gläser, Frank Joublin, and Christian Goerick

***Index Terms*—Speech Acquisition, Robot Audition, Online-Learning, Auditory Scene Analysis**

I. Introduction

The standard approach for teaching robots to communicate via speech is by providing the structure, statistics, and semantics of speech through a supervised, offline learning process. This process imposes constraints like a high degree of specialization to certain, predefined tasks. The resulting system is very rigid and lacks the ability to acquire new skills (e.g. words and their semantics). In contrast to this, children acquire language through observation of adults' speech and, more importantly, in interaction with them. As a result their speech capabilities are very flexible and can adapt to new situations. Our research target is therefore to build a system that can learn to acquire speech in interaction with humans. The interaction aspect requires a hardware platform that can engage in a natural communication with humans in real-world environments. For this purpose we employ our humanoid robot ASIMO (see Fig. 1). To provide the robot with human-like speech communication abilities we are working on several aspects of sound processing, scene representation, and learning that will be outlined in more detail in the next sections.

Fig. 1. HONDA's humanoid robot ASIMO modified to have human-inspired outer ears to improve sound processing.

II. From Sound Localization to Scene Representation

In order for a robot to interact with a human partner, it is necessary to provide the robot with a basic ability to localize sound sources. This facilitates an orientation of the robot (and its cameras) toward the human and improves sound source separation and auditory scene analysis. In contrast to most state-of-the-art approaches employing microphone arrays (see e.g. [1]), we are using a biologically-inspired binaural localization approach that was shown to be effective even under noisy conditions [2]. We also demonstrated that the approach can be extended to both azimuth and elevation estimation [3] (results shown in Fig. 2) and we are currently investigating the potential of our system for source distance measurement. In order to adapt to changing environmental conditions, we are studying how sound localization systems could recalibrate online - that is during normal operation of the robot. These initial steps towards life-long learning and adaptation would reduce the maintenance effort for the robot substantially. First results have been presented in [4].

Inspired by recent work on auditory and visual attention [5] we have introduced the concept of Audio Proto Objects as an interface between audio processing and behavior control [6]. We have shown that in addition to providing a substantial benefit for localization, the framework can also be used to perform a rough but flexible and reliable categorization of sounds. This ability could be used e.g. to ignore background sounds and to build up a representation of the auditory scene (a step towards auditory scene analysis, [7]) in terms of sound source positions and their characteristics.

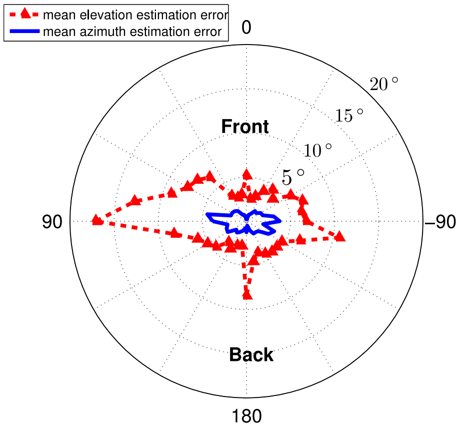

Fig. 2. Localization performance for azimuth (inner solid line) and elevation (dashed outer line) direction estimation. The plot shows the mean localization error (in degrees) of our binaural approach for different horizontal positions around the robot.

III. Noise Suppression and Auditory Attention

Since we are using a humanoid robot as a platform, we have to deal with the robot's ego-noise. In general, humanoid robots, due to their many degrees of freedom and considerable onboard computing power, often produce substantial noise,

The authors are with the Honda Research Institute Europe, Carl-Legien-Strasse 30, 63073 Offenbach, Germany, phone: +49-69-89011-732, fax: +49-69-89011-749, e-mail: {firstname.lastname}@honda-ri.de, WWW: www.honda-ri.de

especially when moving. This poses a serious challenge for sound processing. We are using several approaches to deal with this problem. Stationary background noise, as for example from fans, can be taken care of by spectral subtraction methods. The prediction and suppression of dynamic noise (e.g. from motors) is currently investigated. For an outline and first results, see [8].

To enable the robot to selectively concentrate on one aspect of the environment while ignoring other sensory events, we investigate different attention mechanisms [9]. An essential top-down information is the current interaction status of ASIMO which we determine based on feedback from the vision system. When ASIMO neither sees an object in its peri-personal space or a human in its inter-personal space it assumes that nobody is interacting with it and ignores all sounds around it.

IV. Robust Speech Features

Robust speech recognition can be supported by the development of speech features that are more tolerant to noise and better adapted to the characteristics of speech than state-of-the-art methods. Towards this goal we have investigated how features can be learned data-driven. Based on inspirations from neurobiology, i.e. the receptive fields in the mammalian primary auditory cortex [10], and visual object recognition models, we developed an acoustic feature extraction framework. The approach uses two hierarchical layers and on each layer we perform an unsupervised learning of spectro-temporal receptive fields. When using this framework we see significant improvements for the recognition in noisy conditions [11] and could also show that an adaptation of the features to different environments is beneficial [12]. We have also worked on the robust extraction of pitch and formants under noise conditions [13], [14]. Fig. 3 depicts the extracted formant tracks for a speech utterance example. Furthermore, Table I shows improvements of our formant tracking approach relative to the different state-of-the-art methods [15], [16], [17]. The results are averaged for speech disrupted by white, babble, and car noise at signal-noise-ratios of -3 to 15 dB.

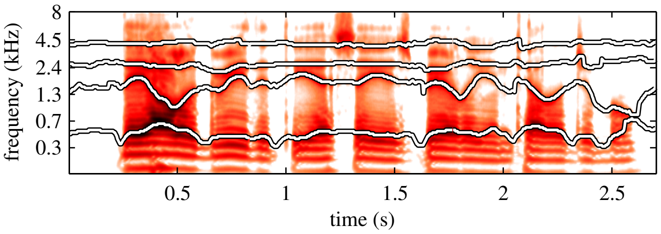

Fig. 3. The extracted formants overlaid to the original spectrogram of the utterance ”They all agree that the essay is barely intelligible.” are shown.

The performance of the pitch and formant extraction has also been demonstrated in a parrot-like imitation scenario [18]. The specific setting is shown in Fig. 4.

TABLE I
Relative improvements (and 95 % confidence intervals) in % of our formant extraction method compared to [15], [16], [17].

Formant	Praat [15]	Snack [16]	Mustafa [17]
F1	71.6 (+1.3;-1.4)	39.6 (+2.2;-2.3)	53.3 (+2.3;-2.4)
F2	55.4 (+2.7;-2.9)	29.1 (+3.2;-3.2)	16.2 (+2.7;-2.8)
F3	50.4 (+2.8;-3.0)	35.0 (+3.2;-3.3)	33.3 (+4.0;-4.3)

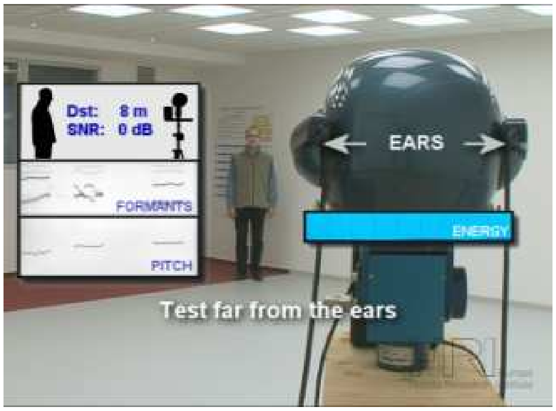

Fig. 4. Imitation in the parrot scenario. The system records human speech from several meters distance, extracts pitch and formants, and finally resynthesizes the speech using the extracted parameters.

V. Audio Visual Association Learning in Interaction

To demonstrate the potential of online speech acquisition in interaction scenarios we have implemented a demonstration system on our robot ASIMO, that is capable of acquiring new words and their meaning in interaction with a human tutor [9].

The interactive learning system is designed to bootstrap multimodal representations with minimal initial knowledge and to enable a continuous development by learning in interaction. Our system can learn associations like the one between a cluster in relative visual positions, in which an object is presented, and an arbitrary speech label for this position.

Initially the system has only very little knowledge. The visual clusters and the speech labels are fully learned in interaction. During a learning session samples in the different perceptual modalities are accumulated. Within a session an object with the property to be labeled is presented, and matching speech labels are uttered several times. If no speech models have been learned yet a new model is initialized with the best matching phone sequence learned. In later learning steps the current utterance is compared to the best matching speech label and the best matching phone sequence. The interaction with the tutor (from ASIMO’s perspective) is shown in Fig. 5.

VI. Speech Imitation

In order to allow a bi-directional speech communication we investigate how words can be imitated using previously acquired internal representations of speech sounds. To avoid that the system simply replicates the tutor’s voice and to give the robot its own distinct characteristics, we artificially impose

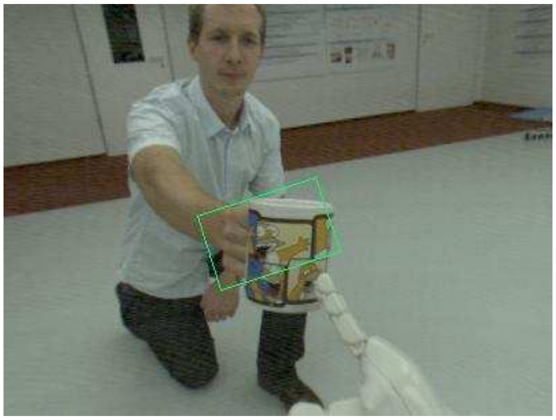

Fig. 5. Internal view of ASIMO when learning the relation between a speech label and the position of an object.

constraints on the speech production that mimic those of the vocal tract in humans. These constraints are chosen in a way that the resulting voice has the characteristics of a child. This requires learning a mapping between the tutor's voice and the system's voice. For the imitation this mapping is learned in interaction with the tutor. In contrast to the work of e.g. [19], we are not using explicit models of the human vocal tract but rather produce speech at the signal level directly using a source/filter model similar to [20]. For the synthesis we employ vocal motor primitives (formant trajectories) that are morphed for continuous speech output. Such formant configurations can be also extracted in interaction from the tutor's voice [14], [18]. Although the approach is currently limited to the imitation of vowel sequences (filling in consonants with the closest vowel), the system can imitate the tutor's utterance using its own voice [21].

VII. Conclusion

The acquisition of speech — including the semantics of an utterance — in an interaction scenario with a humanoid robot requires a broad range of auditory capabilities. In this article, we have described our previous and current efforts for developing those functionalities with a special emphasis on adaptability and robustness to noise. We often take inspiration from the development of speech capabilities in children and brain-like information processing in general.

Although our system already exhibits a substantial array of capabilities in different areas, more progress in the described research fields and other related areas will be necessary before robots can start to match the ability of children to acquire speech.

References

[1] K. Nakadai, S. Yamamoto, H. G. Okuno, H. Nakajima, Y. Hasegawa, and H. Tsujino, "A robot referee for rock-paper-scissors sound games," in *ICRA*, 2008, pp. 3469–3474.

[2] T. Rodemann, M. Heckmann, B. Schölling, F. Joublin, and C. Goerick, "Real-time sound localization with a binaural head-system using a biologically-inspired cue-triple mapping," in *Proceedings of the IEEE-RSJ International Conference on Robots and Intelligent Systems (IROS)*. Beijing: IEEE Press, 2006, pp. 368–373.

[3] T. Rodemann, G. Ince, F. Joublin, and C. Goerick, "Using binaural and spectral cues for azimuth and elevation localization," in *Proceedings of the IEEE-RSJ International Conference on Intelligent Robot and Systems (IROS)*. IEEE Press, 2008, pp. 2185–2190.

[4] T. Rodemann, K. Karova, F. Joublin, and C. Goerick, "Purely auditory online-adaptation of auditory-motor maps," in *Proceedings of the IEEE-RSJ International Conference on Intelligent Robot and Systems (IROS)*. IEEE, 2007, pp. 2015–2020.

[5] B. G. Shinn-Cunningham, "Object-based auditory and visual attention," *Trends in Cognitive Sciences*, vol. 12, no. 5, pp. 182–186, 2008.

[6] T. Rodemann, F. Joublin, and C. Goerick, "Audio proto objects for improved sound localization," in *Proceedings of the IEEE-RSJ International Conference on Intelligent Robot and Systems (IROS)*. IEEE Press, 2009.

[7] D. Wang and G. J. Brown, Eds., *Computational Auditory Scene Analysis*. IEEE Press, 2006.

[8] G. Ince, K. Nakadai, T. Rodemann, Y. Hasegawa, H. Tsujino, and J. Imura, "Ego noise suppression of a robot using template subtraction," in *Proceedings of the IEEE-RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE Press, 2009.

[9] M. Heckmann, H. Brandl, J. Schmüdderich, X. Domont, B. Bolder, I. Mikhailova, H. Janssen, M. Gienger, A. Bendig, T. Rodemann, M. Dunn, F. Joublin, and C. Goerick, "Teaching a humanoid robot: Headset-free speech interaction for audio-visual association learning," in *Proceedings of the 18th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. Toyama, Japan: IEEE, 2009, pp. 2185–2190.

[10] S. Shamma, "On the role of space and time in auditory processing." *Trends in Cognitive Sciences*, vol. 5, no. 8, pp. 340–348, 2001.

[11] X. Domont, M. Heckmann, F. Joublin, and C. Goerick, "Hierarchical spectro-temporal features for robust speech recognition," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2008)*. Las Vegas, Nevada: IEEE, 2008, pp. 4417–4420.

[12] X. Domont, M. Heckmann, H. Wersing, F. Joublin, and C. Goerick, "A hierarchical model for syllable recognition," in *European Symposium on Artificial Neural Networks (ESANN)*. Bruges, Belgium: d-side publications, 2007, pp. 573–578.

[13] M. Heckmann, F. Joublin, and C. Goerick, "Combining rate and place information for robust pitch extraction," in *Proceedings of the INTERSPEECH conference*. Antwerp, Belgium: ISCA, 2007, pp. 2765–2768.

[14] C. Gläser, M. Heckmann, F. Joublin, and C. Goerick, "Combining auditory preprocessing and bayesian estimation for robust formant tracking," *to appear in IEEE Transactions Audio, Speech, Language Processing*, 2009.

[15] P. Boersma, "Praat, a system for doing phonetics by computer," *Glot. Int.*, vol. 5, no. 9/10, pp. 341–345, 2001.

[16] K. Sjölander and J. Beskow, "Wavesurfer - an open source speech tool," in *Proc. ICSLP*, vol. 4, 2000, pp. 464–467.

[17] K. Mustafa and I. Bruce, "Robust formant tracking for continuous speech with speaker variability," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 2, pp. 435–444, 2006.

[18] M. Heckmann, C. Gläser, M. Vaz, T. Rodemann, F. Joublin, and C. Goerick, "Listen to the parrot: Demonstrating the quality of online pitch and formant extraction via feature-based resynthesis," in *Proceedings of the IEEE-RSJ International Conference on Intelligent Robots and Systems (IROS)*. Nice: IEEE-RSJ, 2008, pp. 1699–1704.

[19] H. Kanda, T. Ogata, T. Takahashi, K. Komatani, and H. G. Okuno, "Phoneme acquisition model based on vowel imitation using recurrent neural network," in *Proceedings of the IEEE/RSJ International Conference on Intelligent RObots and Systems*, 2009.

[20] H. Zen, T. Nose, J. Yamagishi, S. Sako, T. Masuko, A. W. Black, and K. Tokuda, "The HMM-based speech synthesis system (HTS) version 2.0," in *Proceedings of the 6th ISCA Workshop on Speech Synthesis*, Bonn, Germany, 2007, pp. 294–299.

[21] M. Vaz, H. Brandl, F. Joublin, and C. Goerick, "Learning from a tutor: embodied speech acquisition and imitation learning," in *Proceedings of International Conference on Development and Learning*. Shanghai: ICDL, 2009, pp. 1–6.

Tobias Rodemann studied physics and neuroinformatics at the Ruhr Universität Bochum, Germany, and received his Dipl.-Phys. degree from the Universität Bochum in 1998 and a Ph.D. degree from the Technische Universität Bielefeld, Germany in 2003. Since 1998 he is working at the Honda Research Institute Europe, Offenbach, Germany. Previous research fields were evolutionary algorithms, biologically-inspired vision systems, information processing with spiking neurons and learning of sensory-motor maps. Since 2003 he is working as a senior scientist on sound localization, auditory scene analysis and audio-visual interaction.

Martin Heckmann studied electrical engineering at the Universität Karlsruhe, Karlsruhe, Germany, and the Institut National des Sciences Appliquées, Lyon, France. He received a Dipl.Ing. degree and a Ph.D. degree in electrical engineering from Universität Karlsruhe in 1997 and 2003. While working as a Research Assistant at Universität Karlsruhe from 1998 to 2003 he was engaged in the Collaborative Research Center Humanoid Robots at Universität Karlsruhe, and a visiting researcher at gipsa-lab Grenoble, France. Since 2004 he is a Senior Scientist at Honda Research Institute Europe, Offenbach, Germany. His research interests include speech and image processing, auditory scene analysis, robust speech recognition, and language acquisition.

Claudius Gläser was born in Gera, Germany, in 1982. He received a Dipl.-Inf. degree (with honors) in computer science from the Technische Universität Ilmenau, Ilmenau, Germany, in 2006. Since 2006 he is a Scientist at the Honda Research Institute Europe, Offenbach, Germany. Since 2007 he is also a Ph.D. student at the Honda Research Institute Europe focusing on the acquisition of sensorimotor competencies and their use for grounding language during social interaction. His research interests include speech analysis, language acquisition, neural information processing, and cognitive behavior control.

Frank Joublin received the Dipl.Ing. degree in electrical engineering from the Ecole Universitaire D'Ingénieurs de Lille (EUDIL), Lille, France, in 1987 and the Ph.D. degree in neuroscience from the University of Rouen, France, in 1993. From 1994 to 1998 he was a Postdoctoral Research Fellow at the Institute for Neural Computation, Ruhr-Universität Bochum, Germany. From 1998 to 2001 he was a Customer Project Manager at Philips Speech Processing, Aachen, Germany Since 2001, he is a Principal Scientist at the Honda Research Institute Europe, Offenbach, Germany, and since 2008 a Board Member of the CoR-Lab Graduate School of the Bielefeld University, Germany. His research interests include auditory scene analysis, language and semantic acquisition, and developmental robotics.

Christian Goerick studied electrical engineering at the Ruhr-Universität Bochum, Bochum, Germany, and at the Purdue University, West Lafayette, IN, USA. He received a Dipl.Ing. degree in electrical engineering and a Ph.D. degree in electrical engineering and information processing from the Ruhr-Universität Bochum in 1993 and 1998. From 1993 to 2000 he was with the Institute for Neural Computation, Ruhr-Universität Bochum. Since 2000 he is a Chief Scientist at the Honda Research Institute Europe, Offenbach, Germany. His special interest is in embodied brain-like intelligence covering research on behavior based vision, audition, behavior generation, cognitive robotics, car assistant systems, systems architecture as well as hard- and software environments.