FAR-FIELD AUDIO-VISUAL SCENE PERCEPTION OF MULTI-PARTY HUMAN-ROBOT INTERACTION FOR CHILDREN AND ADULTS

Antigoni Tsiami [1,3], *Panagiotis Paraskevas Filntisis* [1,3], *Niki Efthymiou* [1,3], *Petros Koutras* [1,3], *Gerasimos Potamianos* [2,3], *Petros Maragos* [1,3]

[1] School of E.C.E., National Technical University of Athens, Greece
[2] E.C.E. Department, University of Thessaly, Volos, Greece
[3] Athena Research and Innovation Center, Maroussi, Greece

{antsiami,pkoutras,maragos}@cs.ntua.gr,{filby,nefthymiou}@central.ntua.gr,gpotam@ieee.org

ABSTRACT

Human-robot interaction (HRI) is a research area of growing interest with a multitude of applications for both children and adult user groups, as, for example, in edutainment and social robotics. Crucial, however, to its wider adoption remains the robust perception of HRI scenes in natural, untethered, and multi-party interaction scenarios, across user groups. Towards this goal, we investigate three focal HRI perception modules operating on data from multiple audio-visual sensors that observe the HRI scene from the far-field, thus bypassing limitations and platform-dependency of contemporary robotic sensing. In particular, the developed modules fuse intra- and/or inter-modality data streams to perform: (i) audio-visual speaker localization; (ii) distant speech recognition; and (iii) visual recognition of hand-gestures. Emphasis is also placed on ensuring high speech and gesture recognition rates for both children and adults. Development and objective evaluation of the three modules is conducted on a corpus of both user groups, collected by our far-field multi-sensory setup, for an interaction scenario of a question-answering "guess-the-object" collaborative HRI game with a "Furhat" robot. In addition, evaluation of the game incorporating the three developed modules is reported. Our results demonstrate robust far-field audio-visual perception of the multi-party HRI scene.

Index Terms— Human-robot interaction, speaker localization, distant speech recognition, gesture recognition, adaptation, fusion

1. INTRODUCTION

HRI systems have been gaining increasing popularity, following advances in interaction technologies and robotic platforms [2], with a wide range of applications developed for edutainment [3–5] and assisted living [6, 7], among others. In such systems, it is highly desirable that the interaction mimics typical human-to-human communication involving the exchange of audio-visual information, most critically via speech and hand gestures [7, 8]. For this purpose, on the perception side of HRI systems, three crucial components can be readily identified: automatic speech recognition, recognition of hand gestures, and speaker localization. The latter is necessary to scene diarization in multi-party interaction scenarios, allowing for example to guide robotic attention towards the active speaker [9, 10].

The aforementioned perception components should be capable of supporting natural HRI scenarios, involving interaction with multiple users, without restricting their movement or requiring them tethered to the robot. Further, performance should remain robust to audio-visual noise due to the environment and the interaction scenario complexity, which can, for example, imply acoustic reverberation to speech, or visual occlusion and pose variation of user gestures. Albeit recent progress [11–13], successfully achieving such goals by robot-based sensing alone remains challenging.

A suitable alternative for indoors HRI is to employ robot-external sensing instead, based on multiple audio-visual sensors located in the far-field, thus providing a "smart space" where the interaction is unobtrusively observed. Such approach allows the fusion of multiple data streams within the same modality (audio or visual) and/or across modalities (audio-visual), improving robustness to audio-visual noise, while bypassing limitations of current robotic sensing and providing perception solutions to HRI in a robot-independent fashion. Not surprisingly, the external sensing paradigm has been considered in recent HRI works [14–16], with limited however exploration of multi-sensory and multi-modal fusion, thus failing to fully exploit relevant research on perception technologies inside smart spaces, for example [17, 18].

In this work, robot-external sensing is adopted based on Kinect sensors [19] that have become popular in HRI systems [15, 16, 20, 21]. Specifically, four Kinects are employed providing a multitude of data streams (see Fig. 1), leading to the design of novel perception components for: (i) multi-sensory audio-visual speaker localization, (ii) multi-microphone distant speech recognition, and (iii) multi-view gesture recognition, as discussed in detail in this paper.

Further to the above, a major HRI challenge involves robustness to variations in user group characteristics [22]. Of particular interest is the case of different age groups, i.e., children vs. adults, especially since child-robot interaction has been the focus of intense research efforts [1–5, 14–16], while, with few exceptions, perception component development in the literature has primarily focused on adults. The two age groups differ both in interaction behavior, as well as

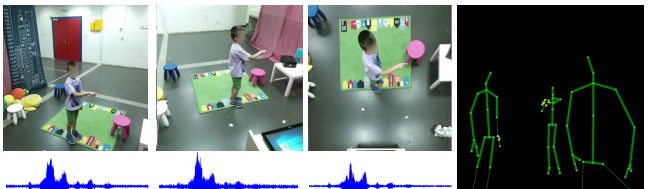


Fig. 1: Examples of the data streams recorded by the four Kinects of the proposed multi-sensory setup. Three of the Kinects provide RGB video and beamformed audio, while the fourth user skeletons.

This work was supported by EU Horizon 2020 project BabyRobot [1], under grant agreement no. 687831.

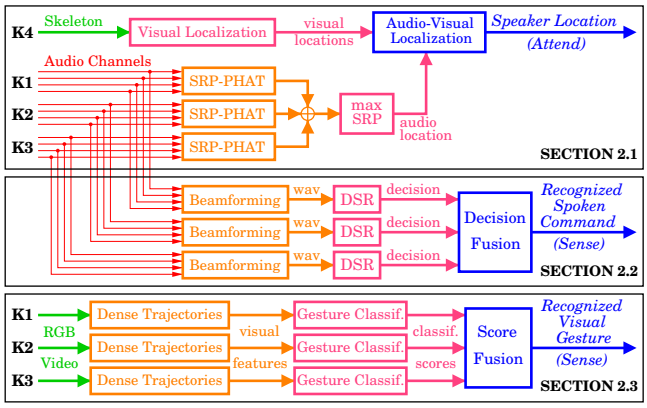


Fig. 2: Schematic of the three multi-sensory perception modules.

in "articulatory" characteristics (vocal tract, arm lengths), deeming robust component performance across them challenging. Motivated by the above, a second contribution of this paper constitutes the investigation of the developed recognition modules for both children and adults, with suitable adaptation and training schemes proposed within the adopted multi-sensory far-field approach.

Additional contributions involve the integration and evaluation of the developed modules. In particular, the perception components are integrated under an architecture that controls dialog flow and robot actions, providing "intelligent" HRI that exploits the audio-visual scene perception results. For this purpose, an interaction scenario of a question-answering "guess-the-object" collaborative HRI game with a "Furhat" robot [20] is presented. The latter choice is primarily driven by the 3D photorealistic appearance of the Furhat robotic head, its abilities to speak and turn towards a desired direction "engaging" the user, and accompanying software integration environment within the IrisTK dialog framework [23]. For the development and evaluation of the three perception components, a corpus of both children and adults has been collected supporting this interaction scenario. In addition, evaluation of the HRI game incorporating the three developed modules is reported. The results demonstrate robust far-field audio-visual perception of the multi-party HRI scene.

2. THE AUDIO-VISUAL PERCEPTION SYSTEM

The developed audio-visual perception system is depicted in Fig. 2. Details of its three components are provided next.

2.1. Audio-Visual Speaker Localization Module

Localization may prove useful in cases where the auditory scene consists of multiple speakers and there is need for speaker tracking and diarization, being for example essential to audio denoising and robot's attention-guiding in natural HRI. Although visual-only localization may be more accurate than audio-based, it does not suffice when a speaker has to be tracked in multi-party scenarios.

There exist various techniques for audio speaker localization and diarization [24], some of them adapted specifically to HRI setups for microphones mounted on robots [11, 13]. In our case, microphones are static, and we seek a real-time algorithm. For this purpose, we have developed a real-time 3D audio localization system that is robust to noise and errors, based on the steered response power – phase transform (SRP-PHAT) algorithm [25, 26]. Regarding audio-visual speaker localization, several methods exist [27–30], most of them employing Bayesian filtering techniques or fusion between audio and video features, primarily developed for conventional RGB cameras. In our setup where Kinects are used, visual tracking is accomplished exploiting the skeleton data stream provided by the sensor (see also Fig. 1).

In more detail, for audio-visual speaker localization the skeletons of all persons present in the scene are first retrieved, as returned by one of the Kinects (K4) of the adopted sensory setup. In parallel, SRP-PHAT based audio speaker localization is performed separately for the microphone array of each of the three other Kinects (K1, K2, K3), based on a "global" pre-defined 3D grid. The three computed SRP-PHAT energies are subsequently added, and maximization over the entire 3D grid yields the possible sound source location. The single-modality results are then fused by simply computing distances between the audio- and visual-only locations, and selecting the visual location with the smallest distance from the audio one. The speaker position is then used for turning the robot's head towards the active speaker.

2.2. Distant Speech Recognition Module

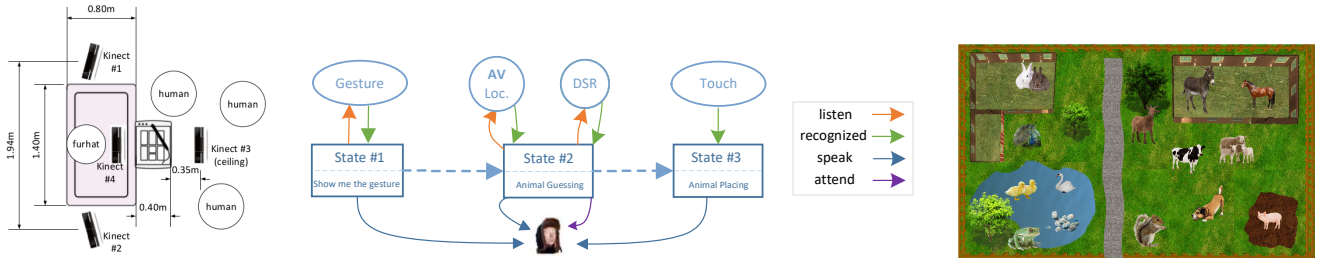
Several factors mentioned earlier, such as noise, reverberation, and speaker-robot distance [31], render speech recognition a challenging task. We employ distant speech recognition (DSR) [32–34], based on three Kinect microphone arrays distributed in space (K1, K2, K3). Extending our earlier work [34], the DSR module is always-listening, being able to detect and recognize user utterances at any time, among other speech and non-speech events, possibly degraded by environmental noise and reverberation. Further, it is grammar-based, so the speaker communicates with the robot via a set of utterances suitable for the HRI use-case of interest. The system can recognize both English and Greek. For the use-case and the evaluation that will be described later, we have employed the Greek language.

Regarding acoustic modeling, we employed GMM-HMMs and trained 3-state, cross-word triphones (about 8k) with 16 Gaussians per state on standard MFCC-plus-derivatives features. To detect one of the target utterances, we use a 2.5 sec sliding window with a 0.6 sec shift. To improve recognition in noisy and reverberant environments, we employ delay-and-sum beamforming using the available 4 microphones from each Kinect. To reduce mismatch with the acoustic conditions in the target environments we have trained models on artificially distorted data. Data contamination has been performed on the available clean training data of the Logotypografia database [35]. The distortion process involves convolution of all utterances with room impulse responses (RIRs) and addition of white Gaussian noise [36]. The employed RIRs were measured in real environments using the exponential sine sweep technique [37, 38]. However, mismatch between training and test conditions necessitates model adaptation, thus maximum likelihood linear regression (MLLR) has been employed. Each microphone array outputs an individual DSR result, and we subsequently fuse them via an appropriate majority voting scheme.

2.3. Gesture Recognition Module

Our multi-sensor gesture front-end is an extension of our previous work on single-view gesture recognition [39] and employs state-of-the-art dense trajectory features [40] along with the bag-of-visual-words (BoVW) framework. First, we sample feature points from each RGB frame and track them over time based on optical flow. Following the trajectory extraction, different descriptors can be computed within space-time volumes along each trajectory. More specifically, motion boundary histogram features [41], describing the motion along each trajectory, are computed on the gradient of the horizontal/vertical optical flow components.

The extracted features are encoded using visual codebooks, constructed by clustering a subset of selected training features. The cen-



(a) (b) (c)

Fig. 3: (a) Spatial arrangement of sensors (four Kinects) relative to the table hosting the "Furhat" robot, the touch-screen, and the overall human interaction area in the HRI system; (b) Dialog flow in the multi-party HRI system; (c) A snapshot of the HRI game as shown to the system users on the touch-screen (the farm with animals in their correct position is depicted).

troid of each cluster can be considered as a visual word, and each trajectory is assigned to its closest visual word using the Euclidean distance. We use BoVW encoding that yields a sparse video representation, which is essentially the histogram of visual word occurrence frequencies over the space-time volume. Videos are classified based on their BoVW representation, using non-linear support vector machines (SVMs) with the χ^2 kernel [41]. Since we face multi-class classification problems, we follow the one-against-all approach and select the class with the highest score. For a given Kinect (K1, K2, K3), we have trained a different SVM for all gesture classes and obtain the probabilities as described in [42]. We apply a soft-max normalization to each sensor's probabilities, and then take their average. Finally, we select the class with the highest fused probability.

3. MULTI-PARTY HRI EXPERIMENTAL SETUP

3.1. Use-Case Scenario

A multi-party game for multiple humans and a robot has been designed, aiming to entertain, educate, but also establish a natural interaction between all parties. The game, mostly designed for children but also enjoyed by adults, is called "Form a Farm", and is a guess-the-object game involving two roles, the picker and the guesser. These can be equally played and interchanged between the humans and the robot. The picker chooses an animal and utters characteristics of this animal. The guesser has to guess the picked animal.

More specifically, humans play the "Show me the gesture" game (State #1 in Fig. 3b), so as to decide who plays first. If the robot recognizes the gesture correctly, it becomes the picker, otherwise it becomes the guesser. In the first case, the human players take turns guessing the chosen animal (State #2). After every wrong guess, the robot reveals another characteristic of the animal. After the identification of the animal, the robot asks the humans to properly place the animal in a farm with some distinct segmented areas that appear in a touch-screen in front of them (State #3). This has also an educational purpose for children, because they are prompted to learn animal characteristics. In the second case the roles are reversed: Humans consult and choose an animal, revealing one characteristic. Then the robot tries to guess the picked animal. Subsequently, the humans take turns revealing more animal characteristics, until the robot guesses correctly. The number of animals is 19, and their characteristics belong in 5 different classes: color, size, species, number of legs, and a distinctive property. A snapshot of the farm with animals correctly placed is depicted in Fig. 3c.

3.2. System Interconnections and Dialog Management

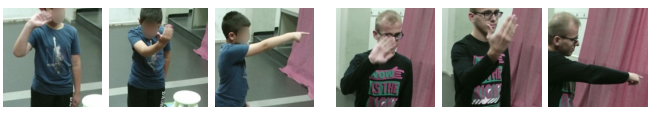
Our multi-party HRI system adopts a modular architecture and follows the IrisTK dialog framework [23]. Communication between the modules is event-driven, with events being divided in three different types: *action* events signaling what the system should do, *sense* events that report what the system perceives from its surroundings, and *monitor* events that report feedback about actions executed by the system. The dialog is managed by a module that translates the information provided by the perception components to actions according to the dialog state. The dialog follows a variation of the Harel statechart [23, 43].

An example of the dialog flow and the system interconnections can be seen in Fig. 3b. In the first state (#1), a "listen" *action* event (depicted with an orange arrow) is sent to the gesture recognition module. The recognized gesture is then sent back to the dialog through a "recognized" *sense* event (depicted with a green arrow), and the appropriate utterance is sent to the robot text-to-speech synthesis (TTS) system along with a "speak" *action* event (depicted with a blue arrow). The dialog then moves to the second state (#2), where speech input is required by the humans, so "listen" events are sent to the audio-visual speaker localization and DSR components. According to the result that is fed back (with the "recognized" events), the appropriate utterance is sent to the robot TTS system, and the position recognized by the audio-visual localization component is sent to the robot to attend (with an "attend" *action* event depicted with a purple arrow). In the final third (#3) state, the dialog awaits input on the touch-screen, and sends again the appropriate speak event.

As already discussed, our setup involves the use of multiple Kinect sensors distributed in space at about 2.5 m average distance from the users. The Furhat robot head [20], created by Furhat Robotics, which is an animated face back-projected on a 3D mask, has been employed as the robotic agent. Among other things, Furhat is capable of speech and head movement with 2 degrees of freedom. We have also employed a Greek TTS engine [44] to enable speech in Greek. The spatial arrangement of the four Kinect sensors along with the Furhat robot can be seen in Fig. 3a.

4. EVALUATION

We employ two evaluation strategies: We perform an objective evaluation of the core perception technologies (speaker localization, gesture and speech recognition) for both adults and children, as well as a higher-level evaluation of the multi-party HRI system for the "Form a Farm" game. Regarding objective evaluation of the perception system, we have collected audio-visual data from 20 adults and



greeting "come closer" pointing greeting "come closer" pointing

Fig. 4: Examples of 3 gesture types by a child (left) and adult (right).

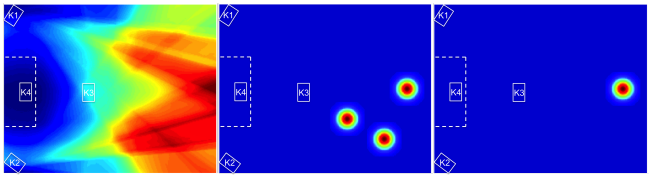


Fig. 5: An example of the audio-visual speaker localization. Left: Audio-only (the SRP output is shown with high values in red); Middle: Visual-only; Right: Audio-visual. Positions of the table and the four Kinects are also shown (see also Fig. 3a).

28 children for model adaptation, training, and testing. Each subject has uttered about 130 utterances out of 300 that constitute the speech recognition grammar from several pre-defined positions (and among other non-speaking people present), and performed 7 gestures related to various HRI scenarios: "agreement", "come closer", "circle", "point", "stop", "sit down", and "greeting" (see also Fig. 4). Background data with random movements have also been collected. Regarding evaluation of the high-level performance of the HRI online system, we invited 12 pairs of adults and 14 pairs of children to interact with Furhat in the "Form a Farm" scenario.

4.1. Objective Evaluation of Core Technologies

An example of audio-visual speaker localization can be seen in Fig. 5. For audio-only speaker localization, the employed metrics are PCOR (percentage correct), which is the percentage of correct estimations (deviation from ground truth less than 0.5 m) over all estimations, and RMSE (root mean square error) between the estimation and the ground truth. For audio-visual speaker localization, since person locations are computed by the Kinect skeleton, the problem is essentially transformed into a speaker diarization task. Thus, evaluation is performed in terms of correct speaker estimation, where PCOR is used. Audio-only localization does not perform sufficiently well, yielding a PCOR of 45%, but the average RMSE is 60 cm, meaning that the average source localization error is 60 cm, which is not very large. If both audio and visual information are used, then speaker localization performance is boosted to a PCOR of 86%.

For distant speech and gesture recognition evaluation, we have experimented with training/adapting adult, children, and mixed models, and testing them with both sets of data. Results for DSR are presented in Table 1 in terms of sentence accuracy, denoted by SCOR, for the two different age groups as test set and with two different decision strategies: "Average" refers to the average result over all three Kinect arrays, while "Fusion" is the result of the three Kinect arrays decision fusion. We present results for unadapted models ("no-adapt") and MLLR-adaptated models on adult data, children data, and both (denoted by "mixed"). Adaptation and testing has been 4-fold cross-validated. Speech recognition achieves satisfactory performance for adults, even without adaptation. Adaptation improves performance in all cases, even when it is performed on a different age group than testing. Table 1 indicates that the best results are obtained when adapting and testing on the same group, which was expected. Decision fusion further boosts performance in all cases, except for children testing using the unadapted and adult-adapted models, where SCORs for independent decisions are relatively low. The best achieved results are 99.82% for adults and 98.97% for children.

In a similar fashion, to evaluate the gesture recognition system, we have trained a separate model for each Kinect sensor using as training sets: a) the children gesture data, b) the adults gesture data, and c) both. Testing was carried out for both children and adult data separately, using leave-one-out cross-validation. Table 1 presents the average accuracy (Acc.) results (%) for the 7 gesture classes and the background model. Results indicate that fusion of the three Kinect sensors improves performance significantly compared to the best single sensor result. Also, when training and test data come from the same age group, the recognition accuracy is high. We can see that accuracy on adult data is enhanced when the model is trained on mixed group age data, since the diversity with which children perform their gestures accommodates the generalization of the model. On the other hand, using the mixed training set deteriorates performance slightly on children gesture recognition, since the range of adult gestures is significantly larger than children's.

		DSR-Adaptation scheme				Gesture Rec.-Training scheme		
		No-adapt	Adults	Children	Mixed	Adults	Children	Mixed
Test		SCOR	SCOR	SCOR	SCOR	Acc.	Acc.	Acc.
Adults	K1	91.76	98.95	94.52	98.69	84.79	60.21	87.81
	K2	90.60	98.70	90.99	97.85	89.27	53.13	92.19
	K3	91.39	98.95	94.11	98.75	85.42	55.63	82.08
	Avg	91.25	**98.87**	93.20	98.43	86.49	56.32	87.36
	Fuse	92.41	**99.82**	94.42	99.77	92.19	62.08	**95.10**
Children	K1	70.53	72.31	95.95	82.95	60.42	76.85	77.31
	K2	72.48	73.85	95.95	82.52	46.99	67.82	68.75
	K3	66.83	67.63	94.60	80.70	42.36	68.29	70.83
	Avg	69.95	71.20	**95.50**	82.06	49.92	70.99	72.30
	Fuse	64.17	66.02	**98.97**	95.51	56.25	**83.80**	80.09

Table 1: Evaluation of the DSR and gesture recognition modules.

Online Evaluation Statistics						
#Trials	Human	3.13	# Corr. guesses(%)		Human	96.70
	Furhat	2.7			Furhat	86.35
Subjective Evaluation Results						
	Dis.	Mostly Dis.	Neutral	Mostly Agr.	Agree	MOS
It was easy to play with the robot	0	3.57	0	28.57	67.86	4.61
The robot behaves like humans	0	7.14	25.00	42.86	25.00	3.86

Table 2: High-level statistics and subjective evaluation for the online HRI system. "MOS" is the mean opinion score (in a $1-5$ scale).

4.2. High-level Evaluation of the HRI System

Table 2 presents two high-level measures relevant to the interaction success: the average number of trials required by each party in order to recognize an animal, as well as the percentage of successfully guessed animals. Results demonstrate that the interaction is effective: almost all rounds ended with correct identification, with each party needing approximately 3 tries. A subjective evaluation of the system was also carried out by asking the children to grade two statements regarding their interaction, using a 5-point ordinal scale from disagree to agree. Table 2 presents these results: The significant majority finds it easy to play with the robot, while a large number of children sees a strong resemblance of the robot behavior to humans, a result that is consistent with the fact that children tend to anthropomorphize robotic agents. These evaluation results confirm the effectiveness of the whole system regarding both its core perception technologies and their integration into a unified HRI system.

5. CONCLUSIONS

In this work, we have proposed and developed an audio-visual perception system, including audio-visual speaker localization, distant speech and gesture recognition, for natural multi-party HRI using multiple distributed sensors. We have also integrated all the components with a social robot and designed a game for multi-party interaction. After the evaluation of the core technologies with both adult and children data, we conducted an evaluation of the online system by humans according to the proposed scenario. The obtained results confirmed the success of the proposed system, highlighting the need for adapting and training perception systems especially for children.

6. REFERENCES

[1] "BabyRobot project," [Online] Available: http://babyrobot.eu.

[2] M.A. Goodrich and A. Schultz, "Human-robot interaction: a survey," *Foundat. Trends HCI*, 1(3): 203–275, 2007.

[3] M. Fridin and M. Belokopytov, "Acceptance of socially assistive humanoid robot by preschool and elementary school teachers," *Comp. Human Behavior*, 33: 23–31, 2014.

[4] R. Ros and Y. Demiris, "Creative dance: An approach for social interaction between robots and children," in *Human Behavior Understanding*, LNCS vol. 8212, pp. 40–51, 2013.

[5] J. Kennedy, P. Baxter, E. Senft, and T. Belpaeme, "Higher nonverbal immediacy leads to greater learning gains in child-robot tutoring interactions," in *Proc. ICSR*, 2015.

[6] A. Zlatintsi, *et al.*, "Social human-robot interaction for the elderly: two real-life use cases," in *Proc. HRI*, 2017.

[7] G. Canal, S. Escalera, and C. Angulo, "A real-time human-robot interaction system based on gestures for assistive scenarios," *Comp. Vision Image Under.*, 149: 65–77, 2016.

[8] R. Stiefelhagen, C. Fugen, R. Gieselmann, H. Holzapfel, K. Nickel, and A. Waibel, "Natural human-robot interaction using speech, head pose and gestures," in *Proc. IROS*, 2004.

[9] C. Oertel, J. Gustafson, and A.W. Black, "Towards building an attentive artificial listener: On the perception of attentiveness in feedback utterances," in *Proc. Interspeech*, 2016.

[10] J. Hemminghaus and S. Kopp, "Towards adaptive social behavior generation for assistive robots using reinforcement learning," in *Proc. HRI*, 2017.

[11] J. Cech, *et al.*, "Active-speaker detection and localization with microphones and cameras embedded into a robotic head," in *Proc. Humanoid Robots*, 2013.

[12] H.W. Löllmann, H. Barfuss, A. Deleforge, S. Meier, and W. Kellermann, "Challenges in acoustic signal enhancement for human-robot communication," in *Proc. SPECOM*, 2014.

[13] C. Evers, Y. Dorfan, S. Gannot, and P.A. Naylor, "Source tracking using moving microphone arrays for robot audition," in *Proc. ICASSP*, 2017.

[14] A. Baird *et al.*, "Automatic classification of autistic child vocalisations: A novel database and results," in *Proc. Interspeech*, 2017, pp. 849–853.

[15] J.C. Pulido, J.C. González, C. Suárez-Mejías, A. Bandera, P. Bustos, and F. Fernández, "Evaluating the child-robot interaction of the NAOTherapist Platform in pediatric rehabilitation," *Int. J. Social Robotics*, 9(3): 343–358, 2017.

[16] P.G. Esteban, *et al.*, "How to build a supervised autonomous system for robot-enhanced therapy for children with autism spectrum disorder," *Paladyn, J. Behav. Rob.*, 8(1):18–38, 2017.

[17] A. Waibel and R. Stiefelhagen, Eds., *Computers in the Human Interaction Loop*, Springer, London, 2009.

[18] "AMI," [Online] Available: http://www.amiproject.org.

[19] I. Tashev, "Kinect development kit: A toolkit for gesture- and speech-based human-machine interaction [Best of the Web]," *IEEE Sig. Proc. Mag.*, 30(5): 129–131, 2013.

[20] A.S. Moubayed, J. Beskow, G. Skantze, and B. Granström, "Furhat: a back-projected human-like robot head for multiparty human-machine interaction," in *Cognitive Behavioural Sys.*, LNCS vol. 7403, pp. 114–130, 2012.

[21] N. Magnenat Thalmann, L. Tian, and F. Yao, "Nadine: A social robot that can localize objects and grasp them in a human way," in *Front. Electronic Techn.*, LNEE vol. 433, pp. 1–23, 2013.

[22] G. Skantze, "Predicting and regulating participation equality in human-robot conversations: Effects of age and gender," in *Proc. HRI*, 2017.

[23] G. Skantze and S. Al Moubayed, "IrisTK: a statechart-based toolkit for multi-party face-to-face interaction," in *ICMI*, 2012.

[24] X. Anguera, S. Bozonnet, N. Evans, C. Fredouille, G. Friedland, and O. Vinyals, "Speaker diarization: A review of recent research," *IEEE TASLP*, 20(2): 356–370, 2012.

[25] H. Do, H.F. Silverman, and Y. Yu, "A real-time SRP-PHAT source location implementation using stochastic region contraction (SRC) on a large-aperture microphone array," in *Proc. ICASSP*, 2007.

[26] A. Brutti, M. Omologo, P. Svaizer, and C. Zieger, "Classification of acoustic maps to determine speaker position and orientation from a distributed microphone network," in *Proc. ICASSP*, 2007.

[27] D. Gatica-Perez, G. Lathoud, J.-M. Odobez, and I. McCowan, "Audiovisual probabilistic tracking of multiple speakers in meetings," *IEEE TASLP*, 15(2): 601–616, 2007.

[28] V.P. Minotto, C.R. Jung, and B. Lee, "Multimodal multi-channel on-line speaker diarization using sensor fusion through SVM," *IEEE Trans. MM*, 17(10): 1694–1705, 2015.

[29] V. Kiliç, M. Barnard, W. Wang, and J. Kittler, "Audio assisted robust visual tracking with adaptive particle filtering," *IEEE Trans. MM*, 17(2): 186–200, 2015.

[30] I.D. Gebru, C. Evers, P.A. Naylor, and R. Horaud, "Audio-visual tracking by density approximation in a sequential Bayesian filtering framework," in *Proc. HSCMA*, 2017.

[31] C.T. Ishi, *et al.*, "A robust speech recognition system for communication robots in noisy environments," *IEEE Trans. Robotics*, 24(3): 759–763, 2008.

[32] M. Wölfel and J. McDonough, *Distant Speech Recognition*, John Wiley & Sons, 2009.

[33] A. Katsamanis, I. Rodomagoulakis, G. Potamianos, P. Maragos, and A. Tsiami, "Robust far-field spoken command recognition for home automation combining adaptation and multichannel processing," in *Proc. ICASSP*, 2014.

[34] I. Rodomagoulakis, A. Katsamanis, G. Potamianos, P. Giannoulis, A. Tsiami, and P. Maragos, "Room-localized spoken command recognition in multi-room, multi-microphone environments," *Comp. Speech Lang.*, 46: 419–443, 2017.

[35] V. Digalakis, *et al.*, "Large vocabulary continuous speech recognition in Greek: Corpus and an automatic dictation system," in *Proc. Interspeech*, 2003.

[36] M. Matassoni, M. Omologo, D. Giuliani, and P. Svaizer, "Hidden Markov model training with contaminated speech material for distant-talking speech recognition," *Comp. Speech Lang.*, 16(2): 205–223, 2002.

[37] A. Farina, "Simultaneous measurement of impulse response and distortion with a swept-sine technique," in *Audio Engin. Soc. Conv. 108*, 2000.

[38] M. Ravanelli, A. Sosi, P. Svaizer, and M. Omologo, "Impulse response estimation for robust speech recognition in a reverberant environment," in *Proc. EUSIPCO*, 2012.

[39] N. Kardaris, I. Rodomagoulakis, V. Pitsikalis, A. Arvanitakis, and P. Maragos, "A platform for building new human-computer interface systems that support online automatic recognition of audio-gestural commands," in *Proc. ACMMM*, 2016.

[40] H. Wang, A. Klaser, C. Schmid, and C.L. Liu, "Action recognition by dense trajectories," in *Proc. CVPR*, 2011.

[41] H. Wang, M.M. Ullah, A. Klaser, I. Laptev, and C. Schmid, "Evaluation of local spatio-temporal features for action recognition," in *Proc. BMVC*, 2009.

[42] C.C. Chang and C.J. Lin, "LIBSVM: a library for support vector machines," *ACM Trans. IST*, 2(3): 1–27, 2011.

[43] D. Harel, "Statecharts: A visual formalism for complex systems," *Sc. Comp. Prog.*, 8(3): 231–274, 1987.

[44] A. Chalamandaris, P. Tsiakoulis, S. Karabetsos, and S. Raptis, "The ILSP/INNOETICS text-to-speech system for the Blizzard Challenge 2013," in *Proc. Blizzard Chall. Works.*, 2013.