Review

Robots that can hear, understand and talk

RAJKISHORE PRASAD*, HIROSHI SARUWATARI
and KIYOHIRO SHIKANO

Graduate School of Information Science, Nara Institute of Science and Technology,
8916-5 Takayama-cho, Ikoma-shi, Nara 630-0101, Japan

Received 4 July 2003; accepted 28 September 2003

Abstract—In this survey paper we analytically examine the state of the art in speech and natural
language processing technologies, and one of their most promising applications in the robotics world
as a user interface to facilitate human–robot interaction/communication and robot control by spoken
natural language. Theoretical aspects of spoken language technology and the main bottlenecks in
developing a conversational interface for a robot have been presented in depth with results found
while searching the literature related to the major breakthroughs made in this field. In this study,
we present a brief technical introduction to talk-active robots, and to discuss related future technical
challenges and technical approaches used. Efforts have been made to highlight the limitations and
missing directions of the research and development in the spoken language technology which are
creating hurdles in the development of voice-active robots for real-world applications.

Keywords: Spoken language processing; voice-active robot; conversational interface; human–robot
corner; social robots.

1. INTRODUCTION

Since antiquity, it has been one of the dreams of humans to build an intelligent
machine that can listen and understand their spoken natural language, and can
talk back in the same way. Continuous research and development in various
disciplines has paved the way for developing such machines in general, and robots
in particular. Over the last two decades, various such machines with different
levels of intelligence and voice activation capabilities have been developed, among
which robots are the most important. However, the language capabilities of many
such robots have been very poor. Robots were primarily designed to perform very
specialized tasks, especially for industrial use, e.g. in part assembling, hazardous
situations, painting, welding, picking and placing, etc., but now the scenario

*To whom correspondence should be addressed. E-mail: kishor-p@is.aist-nara.ac.jp
is drastically changed. The application areas of robotics technology have gone beyond the initial boundaries. Robots are now not only employed to automate a factory or manufacturing process or in hazardous situations, but also in various other applications such as entertainment, domestic work, health care and hospital management, for the extension of the human body and remote embodiment, as a remote brain, as an artificial social agent, and other related applications. Robots are being developed as pets and much more as social entities to meet the emotional and mental needs of the people. The most interesting feature in robotic pets is that they can communicate in human understandable spoken natural language. Human beings have been trying patiently to develop such a capability in biological pets since antiquity, but have never succeeded. The concept of robotic technology has been adequately modified and areas of applications are increasing rapidly, which in turn has led to the need to redefine the terms robot and robotics [1]. Such new application areas also include the development of robot assistants capable of social interaction with people and learning from them to develop cooperation and collaboration. Thus, this technological field provides a very good platform for scientific study of human behaviors, their ways of performing a task and implementation in machines or robots [2]. Over the last two decades robots have become increasingly important in the lives of humans. Fantastically, one may hope that in future robots will be playmates of our children at home, servants for adults, leisure managers for the elderly, nurses in the hospital, etc: They will perform dangerous military and space tasks for us. They will clean pollution and save us from hazards [3]. Thus, autonomous service robots are expected to do much more complicated tasks than industrial robots. A result of a pilot study on the domestic acceptability of service robots shows the highest affirmation and aspiration of the users [4]. Conventional autonomous robots are capable of performing their tasks almost without any visual and audio information aids. They can explore their work area using dedicated sensors and task plans. However, service robots have to share the workspace with the other robots and humans, and have to establish collaboration, like a life-like character, based on different principles such as stigmergy [5, 6], inference and communication to perform collective tasks. Such robots need capabilities to communicate with trained and untrained people.

For a robot as an assistant in public places there is need for robots that can learn and behave in the correct social manner. Humans learn social manners in society from skilled individuals through direct conversation, observational conditioning, goal emulation and several other methods, and socially competent robots are supposed to do the same. Learning of such rules and behaviors by a social robot is essential because interaction and co-working with others may lead to resource-sharing conflicts which can be resolved by social rules. Therefore, it is also important for service robots to know, learn and obey the social rules [7]. These complications in behaviors also enhance the technical sophistication in the robots and disqualify their use by unskilled people. It poses a challenging task for making a simpler system of robotics engineering since the success lies in this area. If
it is not possible to make the robotic systems simpler from the inside, then they must seem simpler to the users. Such extrinsic simplicity can help a general user by overlooking and bypassing the hidden intrinsic complexities of the robots. A user should not need to be a robot engineer or scientist to use robot or to develop new applications using robots. Thus, socially competent robots need sensors and a human-friendly user interface with easier functionalities through which an average citizen can have access, even to the most sophisticated robots, in an intuitive, natural, efficient and enjoyable manner. With this concept, the research area of human–robot interaction is very important and holds the key to success of the emerging robotic technology.

Research is underway to make it possible for an ordinary user to task, train, control, and supervise robots through natural communication channels [8]. The two main channels of human–human communication are speech and visual interpretation of gesture, mimics, lip-readings, etc. Use of human speech for communication between human and robots [4, 9] is one of the most natural, intuitive and highly user-acceptable means, and can play a very effective role in solving the general problems of human–robot interaction, e.g. acceptability, communication and safety. Vocal means of control and command to the robots will be the most wanted amenity where a user’s hands and eyes are not free or are beyond the reach of the robot control terminal. It will be also one of the best possible options for physically handicapped people, one of the most needy candidates for robotic assistance, to control and use robots. In recent years many robots have been developed and demonstrated with a natural spoken language-based interface in a limited and technically restricted framework. Since a number of robots may work simultaneously in the same workplace with a human partner, the communication between robot and robot should also be realized in the natural spoken language. Robot–robot communication has been internalized through cable or wireless, but results in information loss for others working in the same environment and such robots do not qualify to live with humans. With the social robotic architecture, if one wants to achieve collaborative goals, the robot–robot interaction should also be externalized in spoken natural language or in other natural ways of communication, because such an intelligent agent uses both reactive and deliberative reasoning based on theories of intentionality to show collaborative behavior, and it is essential to know each others objectives, intentions and actions for close collaboration and safety [10].

Human society is highly collaborative. Hearing and speaking are the two separate collaborative behaviors, and are one of the most obvious capabilities of life. Equipping a robot with the capability of talking in natural language is a highly knowledge-intensive process and its implementation requires a good understanding of different aspects of the speech communication process, such as acoustics, semantics, pragmatics, etc. Spoken communication among human beings is not simply an exchange of linguistic elements, but is accompanied by several other elements like paralinguistic elements, non-verbal communicative elements, e.g. natural and symbolic gestures (body actions, facial expressions), non-speech
sounds and different prosodic cues such as variation in pitch and speaking rate. All these paralinguistic and non-verbal actions convey some clues for the listener to decode and understand the message, even more than what textually appears in the said spoken words. Thus, for a human being, speech conversation is much more than mere integration of the speech recognition, understanding and synthesizer modules. Such an assembly of functional modules can give only a unimodal talk-active module, just like a blind person’s ear. The auditory process in humans is multimodal. In humans there is a ‘hidden ear in the eyes and an eye inside the ears’ [11]. Humans can perceive the emotional content of the speech and the psychological aspect of the speaker from his/her spoken words. This all makes implementation of a human-like or human-compatible and -comparable auditory system in a machine or robot much more complicated and cumbersome. Palpably, very efficient implementation of the auditory system in a robot is multidisciplinary and calls for a proper merger of supporting research and development from diverse disciplines such as acoustics, signal processing, pattern recognition, phonetics, image processing, digivision, graphics, machine learning, linguistics, psychology, neuroscience, artificial intelligence and computer science, revealing the secrets of our natural spoken communication skills.

Spoken communication among humans is bidirectional and happens as a dialogue or multilogue depending upon the situation and the number of participants. Therefore, it is not only important to make a robot capable of receiving, recognizing and understanding human speech, but also robots have to be compatible and well skilled in producing vocal responses with exact, not optimized, naturalness so that a user or other robot can hear, interpret and respond to them easily and usually as he/she would do with other humans. If service robots lack such qualities, their popularity, acceptance by society and applications will suffer because human society is not going to bear the extra burden of learning how to recognize, interpret and understand the robotic utterances, clues, gestures, postures and other modalities that accompany vocal communication. Under such circumstances, robots might have to be confined within laboratories, or in an expert environment as most ‘complicated toys’, and it may take very long time like computers that despite tremendous computational power have acquired the present status in more than 60 years, to enjoy familiarity with the general population. One of the most significant causes behind the delay in the domestication of computers was the introduction of computers with several unnatural and uncommon means of access or interfaces such as the mouse, keyboard and artificial languages. Such accessibilities compel people to learn many new things (imagine the terms computer literacy, computer programming, etc.) before using computers.

As the working area of symbiotic robots will be shared with human beings, the matter of safety from the robot and of the robot is a crucial and challenging task while ensuring close cooperation among them. In service robots safety matters are highly concerned with the robot’s behaviors. If robots behave with a communication gap it will be formidable and a great danger in the status of a social partner.
Safety management in robotic systems based on only physical augmentation is not enough. In addition, strong safety strategies based on the intelligent augmentation of robotic behaviors are required to stop the chances of undesirable contact between the robot and human. In such an approach, robots need to share spatial, temporal and modal information about their events or tasks with human beings or cooperating partners [12, 13]. Naturally, for this purpose too, vocal communication will be one of the more conducive channels.

In this paper, our aim is to outline the state of the art of talk-active robots and to procure first-hand knowledge of this technology. We also describe the problems faced in designing such an interface for service robots for real-world application and give an overview of the different techniques adopted to mitigate the problems. The rest of this paper is organized as follows. Section 2 presents a short technical review of talk-active robots. Section 3 deals with the general architecture of voice-active robots. Section 4 and its subsections present the basic research issues and technical problems related to voice-active robot development. This is followed by a conclusion with a look at future possibilities.

2. MOTIVATING TALK-ACTIVE ROBOTS

Keeping in view the influential role of the personality and physical body of the robot in the psychological aspects of interaction, symbiotic robots developed so far can be categorized in the following four groups: (i) humanoids, (ii) familiar animal pets, (iii) non-familiar animals and (iv) new characters [14]. Social robots developed in every category are still in the laboratory. However, some robots like Sony’s AIBO, NEC’s PaPeRo, etc., have entered the commercial market. The list of such experimental robots is too long to include here. However, we will mention some of them for reference. The starting work for voice-active robots can be traced back to the development of SHAKEY-II in 1960. This robot was capable of stacking wooden blocks in compliance with given verbal instructions [15]. Other early examples of voice-controlled robots can be found in Ref. [16]. Robots described in that were capable of responding to some vocal commands. There is also a description of a voice-controlled hand in Ref. [17] developed by Palo Alto Veteran’s Administration Center to perform manipulation tasks.

The preliminary work related to voice-controlled robots started with the voice control of graphic robots with very limited capabilities such as to command the movement of a simulated human or other graphic character with speech segments [18, 19]. Other earlier work reports on voice-active robots can be found in the CUBRICON project in which multimodals interfaces to an air-mission planning system were developed [20–22]. Almost all early research on voice-active robots has been centered on instructing robots with a limited number of verbal commands, which in turn activated the programmed procedures. However, talk-active robots in public places need to have human-comparable learning ability from on-going conversation [22], which has gained much research attention in recent years. In recently
developed voice-active robots, attention has been paid to this point, e.g. the humanoid robot HERMEM (Humanoid Experimental Robot for Mobile Manipulation and Exploration Services) is a behavior-based voice-active robot [23]. Its speech interface is based on a speaker-independent speech recognition system, and uses a context-dependent grammar rule and a domain-dependent word list. A behavior-based architecture has been used to integrate the key technologies. It is anthropomorphic in shape, but lacks hardware-based speech production. The user can interact with it using close talking or head-mounted microphones. Hands-free communication is not available. Noise robustness is very poor and thus it can fail to respond to vocal commands in a noisy environment. The Human–Robot Interaction system presented in Ref. [24] provides limited conversation with the robot, including commands to the robot, and queries about the robot’s abilities and internal states. ASKA is anthropomorphic receptionist she-robot and is under development at NAIST, Japan under the joint venture of Speech and Acoustics Laboratory and Robotics Laboratory for reception services [25, 26]. It can provide information about NAIST and weather reports today’s news, etc., by vocal communication. Its speech interface is based on the real-time Large Vocabulary Continuous Speech Recognition (LVCSR) system JULIUS [27] for Japanese language. ASKA’s present generation uses closetalking microphones and has template, fixed format text and slot-filling text based dialogue management. The response templates are selected by keyword matching and N-best recognition candidates. Its noise robustness is poor and lip movement is toyish, and cannot handle vocal requests from multiple speakers speaking simultaneously. Jijo-2 is another nomad-based mobile talking robot without a manipulator. This system combines speech recognition, task-dependent dialog management and statistical learning procedures to communicate with a human and know its environment [28]. This robot shows some generic type of intelligence to adapt using the new environment. The system builds a probabilistic map of its environment by acquiring missing information from a nearby human being through dialogue [29]. It uses a circular array of microphones for speech pick-up based on the beam-forming technique. This system has improved the robustness for noise than others using systems that use a single microphone. However, it has a task-dependent dialogue module and a very informal technique for semantic interpretation of the heard utterance. Therefore, the system fails to cover a wide variety of tasks. The other communication robot comparable to Jijo is GODOT [30], which uses its reasoning ability to mitigate uncertainty in navigation by dialogue with the user, but the system’s performance is poor in an uncontrolled acoustic environment. A spoken dialog system for the robot described in Ref. [31] is based on semantic frames and can be easily extended to different tasks by describing new tasks in different slots of the semantic frame.

One of the most sophisticated systems in this framework can be found in Ref. [32]. It is a communicative humanoid robot with a hand and graphic face, which appears on a small monitor in front of the user. It is capable in real-time face-to-face dialog with a human user with various hand gesture, facial expressions, body language
and meaningful utterances. Its underlying architecture is a model of human psychological dialog skills that includes implementation of user behavior with respect to the current state of the dialogue. However, graphics robots are perceived by vision and audition only. Such bimodal robots are less suitable for subjective evaluation in social robotics and evaluation results of such robots cannot be made valid for symbiotic robots with a complete physical body. Physical existence of the robot is an important factor for symbiotic robots. Reference [33] describes a speech interface for the KHEPERA mobile robot with a speaker-independent speech recognition module based on a Multirate Neural Network [34] to improve the robustness in adverse acoustic environments. However, the speech recognizer used is an isolated word recognizer, so the robot responds to a small number of selected command words. In Ref. [35], the speech interface with CARL (Communication, Action, Reasoning and Learning) has been described. A description of some other voice-active robots can be found in Refs [36–38]. The other important development of the voice-activated robotic assistant AESOP 3000 can be found in Ref. [39], which is a voice-controlled surgical robot used in heart surgery.

A lot of work has been done in this area. However, no such voice-active robots has reached the performance and acceptance level of the industrial robots. There are several causes for this. Industrial robots operate in modified and conducive natural environments but service robots have to accomplish tasks in pure natural environments inhabited by human beings, which put many technical requirements on the robots that are not possible to implement with the available technology or can be implemented for very restricted use. Among the many technological bottlenecks such as the robot’s sensory perception capabilities, mobility and dexterity, task planning, reasoning and decision making capabilities, etc., lack of human-friendly interfaces capable of facilitating human–robot communication in a natural way is also prominent. There are many candidates for natural communication with a machine and no one channel alone can fill perfection in the system. What is required is the meaningful integration and proper fusion of all the channels of information to create human-like communication [40, 41]. However, in this paper, we address only the audio channel, and have tried to trace out progress made so far and some of the existing limitations in this direction.

3. GENERAL ANATOMY OF VOICE-ACTIVE ROBOTS

The concept of speech communication with a robot has been inspired by the speech communication mechanism in humans. However, the state of art in this technology does not model and implement the auditory system in a robot exactly as is in a human. Basically, in the human’s speech communication system, the brain of the speaker sends a command signal (which is the message to be conveyed) to the speech production system that consists of the lungs, trachea, glottis, velum, nasal cavity and oral cavity (including tongue, jaw and lips). Accordingly, the speech production system converts the received command signal into speech. The
listener’s ears collect the transmitted speech. The ear represents these signals by some specific patterns and transmit them to the brain. The brain does the decoding and understanding by performing some type of pattern matching, and accordingly gives a control signal to the speech production system to continue or end the conversation. However, in doing so, besides using various (known and unknown) acoustic features [42], the brain also uses visual information. The general architecture of the audition module of talkative robots is based on the integration of such analogous functioning modules. The human ear is replaced by speech acquisition and processing algorithm modules; the human brain’s function by Automatic Speech Recognition (ASR) and understanding algorithms; the control signal of the brain corresponds to the text and the speech production system corresponds to Text to Speech Synthesizer (TTS). Based on such an analogy, the generally used block diagram of the robotic auditory system is shown in Fig. 1. However, in the future robots with hardware-based speech synthesis organs like tongue, lungs, lips and a vocal tract may be developed which are not shown here. In the block diagram in Fig. 2, the classification of the voice in different types of command for further action by robots is shown. Even with the explicit architectural analogy with the human auditory system, the functioning is not same. Speech recognition research has its genesis in engineering departments where the aim has been to develop practical applications rather than modeling the real cognitive process of a human hearing. The difference can be better described by the metaphor that plane and birds fly, but a plane does not have wings to flap. The human auditory system is a learning system, but the robotic auditory system is a programmed one. The state of the art ASR systems are based on phonetic modeling of the speech signal. However, the human auditory system may not be based on phonetic
modeling, as a human can learn and pronounce other spoken languages by hearing without having any knowledge of the phonemes. Human beings never keep a record of the phonemes of the language never heard before, but can speak it just the same. This indicates the possibility of some generic model, not yet discovered, that equips the human brain with distinguished subtleties for communication such as recognition, language understanding, learning from some patterns observed in the sentences of the unknown language, and formulating and expressing responses. It has been found that Plathe, the language-handling area of the human brain, is organized differently from non-humans [43], which may be responsible for the distinguished language abilities in humans.

4. RESEARCH CHALLENGES AND APPROACHES USED

Communication between a human and robot through speech seems easy and enjoyable, but the mathematical modeling and physical implementations of the underlying myths have proven to be one of the major challenges of modern computing. A lots of basic underlying processes of speech communication are still not crystal clear or are undiscovered [42]. The main technologies envisaging the speech interface to a robot are speech recognition, speech understanding and dialog management, and speech synthesis [44]. The problems standing in the way of facilitating human–robot vocal communication can be best decribed by broadly looking into the challenges and limitations of the state of the art in these technologies. Speech recognition in the spoken interface design is in the central role. However, this Holy Grail is still too hostile to facilitate easy and flexible ways for recognition. During recent decades, statistical signal processing-based
approaches have been prevalent in speech recognition. Hidden Markov Models (HMM) and Neural Networks (NN) or their hybrid approaches have been widely used [45]. The mechanism of the state of the art HMM-based speech recognizer is shown in Fig. 3 [46]. The HMM-based speech recognizer uses a statistical pattern recognition technique in forward chaining. The recognizer estimates the most probable word (as recognized) sequence $w'$ by maximizing the a posteriori probability $p(w|x)$, which can be expressed as follows using Bayes rule:

$$w' = \arg \max_{w \in T} p(w|x) = \arg \max_{w \in T} p(x|w)p(w),$$

where $T$ is the set of all possible word sequence, $p(w)$ is the probability that word sequence $w$ is spoken by the speaker and is calculated by the language model and $p(x|w)$ is the output probability of the given speech $x$ for word sequence $w$ given by the acoustic model. Such speech recognition systems have been developed for different languages such as English, Japanese, French, Italian, Hindi, etc., using the same type of the acoustic feature vectors, stochastic acoustic models, lexicon size and statistical language model, and surprisingly almost the same recognition results have been reported irrespective of language-specific characteristics. LVCSR performance has been improved up to the mark under restrictions in certain domains such as for dictation and planned or read speech, and has entered the commercial market. However, the system performance degrades drastically in real acoustic environment. This is one of the biggest hurdles in developing a speech interface for a robot or any machine for real-world applications. At present, commercially
available speech recognition software for spoken English are IBM Via Voice, Dragon Naturally Speaking, Voice Xpress and FreeSpeech 2000 from Philips. All of these products boast an accuracy of 90% or more, but getting to that optimal recognition is a tricky and painful process. In recent years, improvement in performance of HMM-based ASR has been brought about by combining an automatic speech reading technique, which advocates fusion of visual features of speech, extracted from the speaker's lips or mouth region, with audio features analogous to a human's sensory fusion process of the information, as demonstrated by McGurk effect [47].

The performance of a HMM-based ASR under various noise conditions is shown in Fig. 4 [48]. This presents the performance on the IBM ViaVoice audio-visual database which consists of the full face frontal video and continuous read speech. The audio data used has been degraded by the artificially generated non-stationary, wideband speech babble noise. In the graph, audio-visual (3) and audio-visual (4) represent performance for the two different audio-visual feature fusion techniques, i.e., feature concatenation and Hierarchical Linear Discriminate Analysis (HiLDA), respectively. It is evident from Fig. 4 that the audio-visual mode outperforms the audio-only or visual-only mode, but in every mode the recognition performance degrades drastically as the speech signal quality degrades from clean to noisy. It is important to note that this performance is on read speech, not on general conversational speech data, which is highly spontaneous. The human auditory system is very robust under such conditions. Comparative studies on the
Figure 5. Relative recognition performance of a human and a state of the art ASR. The x-axis shows the SNR of the input speech and the y-axis shows the Word Error Rate (WER), a measure of the recognition efficiency for ASR.

performance of ASR and humans has been performed by many researchers [48, 49]. Figure 5 [48] shows the result of one of such study. It can be inferred from Fig. 5 that the state of the art speech recognizers are very much less robust in comparison to a human’s recognition performance. Thus, such ASRs are not optimal to make a conversational interface for real-world robotic application. There are many causes of deterioration in recognition performance and their eradication has emerged as one of the biggest challenges in developing a robust speech recognizer.

4.1. Real-world speech variation and signal pick-up

Commercially available speech recognition software is based on statistical signal processing techniques in which each fundamental unit of speech is modeled by a HMM, NN or hybrid approach. The models used are trained using clean speech and thus become suitable to handle the training-like situations, which are hard to meet in a real environment. In the real-world, the speech signals reaching the recognizer suffer distortion and modifications, as shown in Fig. 6, which produce a serious difference between training and testing conditions. These contaminations lead to degradation in the performance of the recognizer. Under such perturbed conditions, a speech recognizer puts several constraints on the speaker such as use of a close-talking microphone, no movement in the speaker, etc., to minimize the difference between the training and testing conditions so that laboratory-level recognition
performance can be achieved. However, in real-world applications speakers are not happy to accept such restrictions. Speakers like to talk with their own styles, from their own position (distant or near, moving or static), in a natural manner free from such restrictions. An artificial recognizer responds to such variations in the speaker differently and shows a wide variation in performance. Factors creating a mismatch between the training and testing conditions are considered to be contamination with noise (additive, convolutional, reverberational) arising from the acoustic channel, speaking style (Lombard effect, speaking rate), inter-speaker variations (voice quality, pitch, gender, dialect) and task/context (dialogue, dictation, conversation). Different approaches have been developed to cope with these problems and can be grouped in following three categories: (i) speech signal enhancement, (ii) using robust features and (iii) model adaptation [50]. Based on these problems, the state of the art speech recognizers find it very difficult to overcome environmental problems and there is need to create environmental robustness in the ASRs so that the existing systems can be used in a variety of environments. Most of the techniques used to overcome environmental problems assume that the additive noise source is stationary, and convolution distortion is time-invariant and uses some a priori knowledge of environmental interference. However, in practice these assumptions breakdown and call for a method that can handle a signal corrupted by stationary or non-stationary additive noise and time-invariant or time-variant convolutional distortion with the adaptability for the new environments [51]. The spoken words in real-world conversations are different from the dictation, planned or read speech. Conversational speech is spontaneous in nature and contains random variations in language characteristics such as speaking rate, speaking style, accents, speech disfluencies (hesitations, incomplete words or fragments, repeated words, restarts), syntax and grammatical
errors [52]. Speech disfluencies are one of the most important factors responsible for the very poor performance of the LVCSR, showing best recognition efficiency for dictation, planned or read speech. The other important episode in speech recognition is the affective content of the speech. As mentioned previously, speech contains prosodic information which plays a vital role in speech recognition and perception in humans [53]. Different prosodic cues modify the literal content of the linguistic message. Prosody packs information such as the speaker’s gender, age and physical condition, and the speaker’s opinions, emotions, and attitudes towards the concerned topic inside the speech. The attitude of the speaker depends on his/her emotional state. Under different emotional states, the physics and chemistry of the human body are changed, and corresponding reflections occur as changes in the different activities/symptoms of the body, such as a change in skin conductivity, blood pressure, and the face and voice of the speaker [54, 55]. Also, the different emotive intents of the speaker are not independent; many of them occur simultaneously, and mutually induce and transform one another. The effects of the emotive state of the speaker on speech have been broadly presented in [56]. The prosody of the utterance has a modification effect on the emotive content of the speech, which further complicates recovery of the affective message and representative acoustic features from the speech. However, even for a human being it is difficult to perceive and distinguish emotional states that have almost similar acoustic features. A service robot in an office or house is like a public device which needs to handle multiple speakers/users. It has been scientifically confirmed that emotion plays a vital role in human life. Humans are affective beings, motivated to action by a complex system of emotions, drives, needs and environmental conditioning in addition to cognitive factors. The recognition of the affective contents of utterance as well as the utterance of the human speech is essential for smooth and enjoyable verbal communication between humans and robots. The robotic system that fails to or is incapable of recognizing and understanding the emotive content of the speech cannot only frustrate, irritate and induce irrational behavior in the interactors, but can also fail to learn from social interaction even in the presence of prohibitory or encouraging feedback from the instructor. The main problem in this direction is that almost state of the art ASRs do not consider prosodic features for the cepstral features in the context of the Gaussian mixture models. It is a speaker-specific feature and recognizers do not care. However, in some works [57–59], the importance of prosody in recognition has been realized and used.

Human society will digest spoken activities of the talk-active robot without concentration if the rules of speech communication between the human and robot are almost same as that of among humans. The mismatch in the style and personality of the robot can badly affect its quality of vocal communication with humans. The problems do not end with the recognition and perception of different acoustic emotive information packed in the speech — the robots need to respond vocally with an emotive quality analogous to that of humans to please and influence
interacting humans [60, 61]. Obviously, the talk-active social robot needs to have elementary skills for recognizing, understanding, handling and manipulating the emotional content of the interlocutor’s speech as well as of itself. It is important to note that there may be differences in the experience of the feelings of the same emotion in the robot and human due to their structural and functional differences. However, to make a robot feel the emotions of others as he/she feels may be a very powerful tool in deciding the emotive behaviors of the robot, but the chances of such possibilities are far beyond the state of the art technologies. Despite various complexities, efforts are underway in affective computing to recognize the emotive content of speech. Based on the variation in prosodic features such as variation in pitch, rhythm, loudness, etc., many vocal emotion recognition systems have been developed [62, 63]. Some other researchers have developed an emotion recognition system that can recognize speaker approval versus speaker disapproval from child-directed speech [64].

It has been found that ASR performance is reasonably good for close-talking microphones, but begins to deteriorate significantly as the talkers moves away. However, while talking with the robot and in many other applications a user does not like to be imposed by such constraints and may prefer hands-free communication. Also, especially in robotic applications, hands-free communication is necessary to command a robot from a distance, as the physical situation or safety aspects may not allow the user to use other means. As the separation between the microphone and real speaker widens, the speech signal becomes increasingly susceptible to background noise and the reverberation effect of the acoustic environment, which affects the Mel Frequency Cepstral Coefficients (MFCC) parameters used to represent speech in the speech recognizers and ultimately results in a reduction in recognition efficiency [65]. Speech signal pick-up of a distant speaker with a good signal-to-noise ratio (SNR) in a real environment is a challenging area and is important for further processing [51]. One of the brilliant solutions to this problem has been offered by the microphone array, which picks up the speech signal simultaneously over a number of spatially separated channels. Further, many array signal processing algorithms such as fixed beam-forming, adaptive beam-forming, dereverberation techniques, Blind Signal Separation (BSS), auditory model-based array processing techniques, etc. are used to obtain the signal with improved SNR [66]. Use of a particular technique depends upon the application and computational requirements. The beam-forming technique is widely used and is robust to adverse situations. Dereverberation techniques are used to reduce the negative effect of the reverberation. Most research in this area has been based on using the impulse response between the sound source and microphones a priori. These techniques require that the room transfer function between the source and each sensor in the array is static; however, in practice, impulse response changes over time and its measurement is unrealistic [67].

The other important and challenging area is speech signal pick-up of a particular speaker in the hotchpotch of sounds — a very common situation in public places.
This is equivalent to focusing hearing attention of the robot on a particular speaker. This has been well documented as the Cocktail party problem [47]. The BSS techniques have been accepted with great promise over the Computational Auditory Scene Analysis (CASA) to solve the Cocktail party problem, but this computationally intensive technique is in its infancy [68]. BSS techniques have been successful in separating non-convolutive mixtures of non-real world signals. In real-world speech the signal picked up by microphones is a convolutive mixture of source signals and noise. The assumptions made behind the BSS algorithm are hard to meet in real-world applications [69]. In recent years research on real-world acoustic data has gain impetus, but to date there exist hardly any real-time BSS algorithms that can be used to perform real-world speech signal separation online on the fastest available processor. In the Cocktail party problem, humans are capable of steering and switching their hearing attention on a particular source, but BSS algorithms do separation computation for all sources and thus with increasing number of sources, the separation performance deteriorates. Most of the successful BSS algorithms for real environment audio source separation work in the frequency domain with independent component analysis. These techniques suffer from the problem of permutation, which makes the order of source separation undeterministic. Although there are several ways to solve it after the estimation of the permuted separation matrix, there is no way to compel the separation algorithm to extract sources in a particular order. Still, it is an open question how to sieve out a speech signal of interest, while rejecting others from the mixture [66]. Even deflationary algorithms for the separation of mixtures of convoluted speech fail to extract a particular source in a fixed order [70, 71]. The other technical handicap with the BSS algorithm is that it is hard to cope with acoustic dynamics, e.g. BSS, fails when all the simultaneously speaking speakers except one become silent, which is a very common event in a multi-speaker environment.

4.2. Spoken Language Understanding (SLU) module

Understanding what has been said in the spoken words is the task of the SLU module in the voice-active robots. The perfection in the response of the speech interface-based system largely depends on this module. The SLU system is different from the textual Natural Language Understanding (NLU) system in many respects. In the written message, humans produce grammatically correct sentences with a richer lexical density than in spoken language production. Thus, language understanding techniques based on the statistical model of text fail to understand the spoken input due to mismatch. However, research done in the traditional NLU systems is useful, but not satisfactory and sufficient, and needs due modification to handle the spoken input. An NLU system exploits syntactic and semantic knowledge of the language concerned to understand the hidden meanings. Humans use such knowledge not only in language understanding, but also in speech recognition. Traditionally, in the computational framework NLU and speech recognition have not been linked with such functionality for several reasons. Speech recognition and lan-
guage understanding technologies have their genesis in different faculties. Speech recognition is largely an engineering approach and thus research emphasis has been to produce the effect of the art of human speech recognition. Language understanding is motivated by a psycho-linguistic background so that research priorities have been to model the real cognitive process of language understanding in humans. Originally, research in natural language understanding was based on the syntactic analysis of grammatically complete sentences, but failures of such systems [72–74] for spoken input have shifted research in the area of semantic-driven approach, e.g. meaning extraction based on keyword spotting [75], selected phrase parsing, etc. However, such approaches cannot do well in the understanding of complex linguistic constructs. Since the importance of semantic and syntactic knowledge in speech recognition and prosodic information of utterances in language understanding has been realized, a major breakthrough is needed in the area of interaction between both modules. How can semantic and syntactic information from NLU be used in speech recognition and prosodic information of utterance from speech recognizer into the NLU? The mere concatenation of two modules can give suboptimal performance [75]. There are a lot of problems in establishing a tightly coupled integration between these two. Several mechanisms for this integration have been proposed, among which the $N$-best interface [76] and the word graph method [77] have been popular. In the $N$-best approach, the speech recognizer passes only the top $N$-best sentence hypotheses one by one to the NLU module to do grammatical knowledge analysis to determine the best scoring hypotheses. Since many of the top $N$ sentences hypotheses may differ minimally in a bad acoustic situation, the $N$-best interface eliminates them from the full semantic and syntactic analysis.

Other important features of a human’s cognitive process of natural language understanding is that words and utterances are understood in the light of the context acquired from the speaker’s world or conversation background. Ultimately, humans understand utterances and words through sensory motor experience [78], not in terms of other words. In this way, this establishes a common multisensor ground for every word and concept leading to the development of complex linguistic subtleties. Once such ground becomes strong, people become able to learn and understand words of other languages largely by word substitution. Thus, the language understanding process in humans is grounded understanding. In contrast to this, most of the spoken language processing systems are trained using recorded utterances with manual transcription and semantic labelueling. Such semantic representation is abstract and ungrounded. Obviously, humans and machines do not share the common ground for the same utterance. Such representations may be suitable for human–computer interaction because computers are not coming from the electronic domain, but for robotic applications the same schemes are problematic. A robot has a physical body and needs knowledge of objects, location, action, etc., with respect to the physical situation to perform proper manipulation in the physical world. Understanding and concepts not relevant to the context of the physical world will be impractical for the robot. For example, if somebody asks
a service robot 'Take the second one and put it in the red bag', the robot needs to decide lots of things in the context of dialogue history and the speaker's attention before accomplishing the said task. The robot needs to know what is the second one? For what is the second one used? If known, from which side should counting be done to decide the second position? Where is the red bag (this information may be in the attention of the user)? If such information is not available the robot has to query as well as interpretate with respect to the physical situation. However, the robot may fail to select such information if it does not share common perception ground with the instructing human.

A service robot's utterances and word understanding capabilities should also be in context with the real-world. Very little research in this direction has been reported. Work reported in Ref. [74] deals with such a robotic system that learns to understand and generate a spoken utterance grounded in camera-derived visual semantics. Major research challenges in this direction are how to obtain non-linguistic knowledge information from a speaker's world, and integrate them with linguistic information to implement human-like grounded spoken utterance learning, understanding and production. The biggest problem in this direction is that many aspects of human cognition and language understanding are still not clear or undiscovered, so less can be copied in the artificial counterpart. However, such approaches are very natural and have a good resemblance to linguistic capability development in the human right from infancy [80].

4.3. Dialogue management

In symbiotic robotics, dialogue management has proved to be a very powerful tool in implementing collaborative control in a robot [81]. Dialogue modeling and management are crucial tasks in establishing spoken communication between humans and robots. Human–robot dialogue management is different from human–computer management mainly due to the manipulative power of the robot in the real-world. The robot has a physical body and can operate in the physical world during dialogue or as a result of dialogue. However, research on human–computer dialogue management is a good foundation for human–robot dialogue management [82]. In the dialog between human and robot, the effect of physical constraints on the meaning of utterances becomes important because the robot has a physical body. It is like human–human dialogue where physical constraints mediated by body movement modify the meaning of the utterance [83]. For example, when humans face each other or walk together there is chance relationships emergencing. In particular, the meanings of demonstrative pronouns used by the speaker are altered by the viewpoint, a physical constraint, of the speaker [84]: Thus, if the physical constraint-sensitive words are to be used in human–robot conversation, the robot must be able to detect or select appropriate viewpoints. There may be a transition in the viewpoint of the speaker if the speaker shifts from a static to mobile status [85]. Thus, for robotic applications, different dialogue strategies are required for spoken reciprocation with the same user in mobile and static positions.
The fundamental tasks in spoken dialogue management are concerned with the translation of the user’s request in robot understandable language and vice versa, management of disambiguation, conversational fillers, ellipsis, turn-taking, indirectness, anaphoric references, etc. Different dialogue modeling approaches have been used to model human–robot dialogue. However, a dialogue pattern based on that of human–human dialogue will be the best suited for human–robot dialogue [86]. Out of system initiative, user initiative and mixed initiative or goal-directed computational approaches of dialog modeling, the last one has maximum resemblance with human–human dialogue management. However, even the goal-driven approach fails to handle all the variability due to spontaneity in human–human dialogue and still it is not clear whether the modeling of human–robot dialogue on the basis of human–human dialogue is essential and appropriate [87]. In most of the state of the art spoken dialog systems, dialogue strategy is hand crafted by the system developer, depending upon his/her design institution of a particular dialogue flow. Such systems give unsatisfactory performance in the public or domestic places where different users have different thoughts and institutions. There is a need for the development of an adaptive dialogue strategy that can learn and decide dialog strategy automatically from little user data to adjust to different speakers. There has been research in this direction too, but very rare [88] and there is need to go further in this direction. The state of the art in dialogue management is restricted to a particular domain and language; however, in the future it will be necessary to develop methods and tools that can be easily used in other languages and domains with less effort even by a non-expert. The task of developing such an automatic or semi-automatic dialogue module that can fit in new applications is challenging and less well studied. However, such modules are essentially helpful to develop systems for new applications at less efforts.

Service robots in public or domestic places are supposed to be able to do unrestricted dialogue. The state of the art in dialogue management technology is not so highly developed and at present there exist no such systems that can do unrestricted dialogue. It may not be possible to develop such a perfect system in the programming paradigm as real-world dialogue contains a wide variability and dynamism. Designing a dialogue system such that it can expose the capabilities and activities of a robot to the user is a better compromise in the practical system design. Disclosure of capabilities and activities of the robot will let the user know about the system and help in ingraining the idea about range of commands or verbal instructions and type of task specification that can be given to the robot. In another approach, the robot may direct the user to stay within its capabilities by issuing some warning or help message. However, to manage issuance of such system initiative is difficult as it depends on the context of the discourse and dialogue state. Also, the user will notice the handicap of the system, as humans are used to having unrestricted dialogue.

Another challenging area in dialogue management is how to cope with the problematic or incomplete spoken input and the arising misunderstanding or commu-
nication deviation. Such a problem may be due to various phenomena such as acoustics, out of vocabulary word, speaking style, recognition error, etc. The dialogue management module is input with logical expression developed by SLU to decide further action like message generation, robot control, etc. As mentioned previously, the state of the art ASRs suffer from misunderstanding, mishearing and are far away from maturity, so it makes the SLU produce incomplete input for the dialogue manager. Another serious cause of problematic input to the dialogue manager is the spontaneous nature of human dialogue [89]. Large proportions of spontaneous speech are ungrammatical and have very poor lexical density [90]. For such input the worst performance of the recognizer is well known and gives extremely problematic input for the dialogue management module. Therefore, the dialogue management module should be able to mitigate such difficulties and recover dialogue. For such a technicality, dialog tracking is needed to interpret information as a correction versus new information, and to know the end of dialog and any change of context. The technology of dialog tracking, still in its infancy, needs research in various areas such as speaker tracking, recognition of speech acts, context or topic tracking, etc. [57, 91].

Another equally important challenging area in dialog management is the modeling and representation of shared attention, behavior and knowledge. Attention and behavior sharing are very primitive functions in human–human vocal communication in the shared world. Human beings use shared attention as a tool to spotlight things and events in the shared world, and to monitor behavior of the attending and communicating partner. Human–robot communication can be made more effective using some cognitive modeling of attention and behavior sharing [92], however, research in this direction is slow and little reported. In human–human dialogue, three types of knowledge, i.e. perceptual, linguistic and cultural, are shared between the speakers and listeners. These different types of shared knowledge influence the language and structure of the dialogue [93]. A human partner uses a number of heuristics and cognitive skills to understand the meaning of the dialogue. However, if the same is expected from a robot partner, it is not possible with the state of the art technologies. We do not have perfect tools to model knowledge and there is no alternative for human intelligence. Modeling and representation of the first two shared knowledges i.e. perceptual and linguistic, are relatively easy and appreciable research has been done in this area. On the other hand, cultural knowledge is community knowledge — people learn it from the community where they dwell in — and it is cumbersome to model and represent. However, it is important to note that if robots come close to society, they will be one of the best keepers of community knowledge as they can survive from generation to generation.

The other challenging area in dialogue management requires establishing good coordination with the other interacting modules as shown in Fig. 7 [94]. For example, while generating a response, discourse history must be taken into consideration. The response of the system must be related to what has been said in the current session and to an individual user. Discourse history influences the form of response, new in-
formation being passed and how it is related with information already passed. What information is available from a previous discourse? The dialog manager should help with such information for the spoken language generation module so that an appropriate response is generated. A lot of research has been done in this direction for human–computer interaction. Last, but not least, the dialogue manager should establish proper coordination with other channels of communication such as graphics, animation as well as written language. For example, if the robot is giving information about the location it may display a map, and it should be able to decide what to speak and what to place on the graphic display. However, robots having only a speech modality dialog system should be able to provide alternative explanations if the first response is not understood through follow-up questions [95]. These issues have surfaced recently and require in-depth exploration.

4.4. Speech generation

Talking robots must be able to pass information in natural sounding sentences. For this purpose, the Natural Language Generation (NLG) module generates a textual response and then TTS converts the same into spoken message. The verbal response passed by the robot should have enough conveyance, comprehensibility and be best suited for the given context because this is the ultimate representation of the robot’s understanding as well as concern about the situation, and is one of the factors responsible for personifying the robot to the user’s satisfaction. The NLG module maps internal computer representation of the information in human language. Many tasks of the output processing are endowed with this module. Since service robots are expected to be used for domestic purposes, while generating response for the user the NLG module of the robot must be able to take account of the user’s status such as age (children, adult, etc.), and experience or level of comprehensibility. However, such flexible and self-adapting NLG do not exist at present. There has been much research on the NLG on coherent paragraph generation [96], but paragraph generation is not so useful in real conversation. However, underlying
technologies will impact on the further development of the real conversational system for human–robot conversation. In real conversations, the NLG needs to generate a response in one or two sentences in tune with the dialog management, not a well-planned paragraph. Responses generated on identical patterns are boring to users and such static NLG modules are useless as the user needs consistent, natural and non-monotonic feedback from the system.

The response generated by the NLG module is converted into a spoken message by a software program called TTS. Much research in this area has been reported in recent years. TTS software is commercially available that can produce sound with an acceptable level of intelligibility and comprehensibility. The state-of-the-art TTS are mostly based on formant and concatenation synthesis or their combination [97], since these methods provide the best possible quality of sound with less computational load. In contrast, the articulation-based methods are complex and have a high computational load. Their scope and potency are only partially explored; however, this fundamental technique is expected to give better results. As mentioned earlier, it is the output-spoken message that leaves an impression of the robot on the user so that naturalness in sound quality is of prime importance, especially for service robots. If the quality of sound is unnatural it may have a bad effect on the linguistic development of a child if interacting with the system or listening to the response of the system regularly for long time. So far as the naturalness in the sound quality of TTS is concerned, this has a long way to go. Natural speech has many dynamic changes, so perfect naturalness may be impossible to achieve. Even for the production of natural-like sound, much of research is still to be done in the prosodic, text prepossessing and pronunciation field. The TTS needs to be based on the concept of speech synthesis using some generic model that can completely describe speech not merely as a text-to-speech synthesizer, as written text does not contain much information such as explicit emotions, proper pronunciation and the mental state of the speaker. A TTS cannot add these flavors in speech simply as a text reader. This calls for advances in high-level speech synthesis and tight coupling between the NLG and TTS to control voice quality and prosody, which provide crucial clues as to intended meanings. So far, in high-level speech synthesis processes, e.g. text preprocessing [98], creation of linguistic data for correct pronunciation, analysis of prosodic features for correct intonation [99], stress, and duration, etc., perhaps the least progress has been made and many problems are still outstanding.

Obviously, technological reality is far away from practice and it may take many years or decades to develop fully acceptable a text to speech synthesizer. Even if a full-fledged TTS is developed there will still arise problems in the application with
service robots. Using TTS with the appropriate transducer can generate a spoken response of the robot; however, if we examine such spoken responses from the psychological point of view of human–human or human–robot conversation, such conversation is not satisfactory. As we mentioned earlier, the physical embodiment of robotic actions in human-like functionality and modality is influential in the effective conversation — in this respect, a TTS generated signal converted into speech has no visual clues. However, most talk-active robots have either toyish mechanical mouth movement or a graphical face [100] with the facial animation properly aligned with words being spoken to show the lip movement, facial gesture, gaze, etc., but still lack in real appearance of mouth dynamics. Human beings are accustomed to conversation with real talking faces, so while listening to a graphical face they notice the lack in sensing modalities and feel uneasy about such technological handicaps. The face acts as an interface during the vocal communication. In the view of most psychologists, spoken language and the appearance of the human face has a special status in the mind. Seeing a speaker’s face and mouth seems to support the auditory perception process [101]. Thus, there is a need to develop real talking head model functionality for service robots. Research to develop such electro-mechanical voice synthesis machines or real talking heads started much earlier [102] with less attention and achievements are still far away. There exists no such talking head to date. There is need for intensive research to model the control mechanism of vocal movement, which produces acoustic signals. In recent years, serious initiatives have been taken by the leading research organizations like JST (Japan Science Technology) to develop and analyze human speech dynamics. Such an anthropomorphic talking head to produce human vocal movement can be found in Ref. [103]. This head is capable of pronouncing vowel and consonants sound, but the naturalness of the sound is lost. Natural sound contains fluctuations of amplitude and pitch of the sound waves, which is an important factor for the naturalness in the voice. The implemented control mechanism in this system fails to achieve such an effect. However, it is undecided what to use for the robot — TTS with graphical lip animation or mechanical voice synthesis machines that model vocal movement.

The other important issue for a voice-active robot is utterance imitation. TTS uses speaker-dependent fundamental sound units stored in the database for speech generation. Thus, it is not easy to produce speech of different speakers using the same speech database. It is found in humans that when they narrate a message of others or quote part of a spoken message received in dialog with others, they try to deliver the message with the same tone and prosody so that third listener can understand the intention of the original speaker. However, this capability for a robot seems not to be so important; however, in the advanced stage when a robot will be domestic partner, it will be desired. For example, say a voice-active robot is taking care of children, if children ask the robot how the sound of a cow or crow is spoken (a childish question), then it will be impressive for the childminder robot to try to produce a matching sound (as humans can produce a matching sound).
However, it is not out of place to mention that even professional mimics cannot imitate fully the speech of others due to physiological differences, but can imitate speaking styles to a good degree. Will the programmed TTS help the robot for such purpose? Of course there have been engineering efforts in this direction in the field of voice conversion (not imitation) research (but not for non-human speech as stated in the above example) where ASR adaptation techniques have been used for speech synthesis voice conversion [104–106]. This technique transforms the initial source model to the target speaker model using a small amount of speech data from the target speaker. However, such techniques are not a complete solution, as they may not capture the entire speaker-specific oratory characteristics such as speed of articulation, preference for certain words, dynamics, idiosyncratic articulation and/or intonation patterns, and characteristic fillers. Research in parametric speech synthesis, suitable for a wide variety of speech synthesis, may open some new avenues in this direction, but is limited by the high costs of rule development [107].

In reality, the field of spoken language generation is in its infancy with very little research that deals with all aspects, like determining what to say, how to say it and how to pronounce it. Research thrusts are needed in deciding the syntactic structure and word use for the spoken message from the psycholinguistic perspective so that the spoken output becomes appropriate to the spoken situation. Language generation often follows discourse features and intonations. Thus, it is required to refine the interaction and integration of the speech synthesizer with language generation so that the parameters controlling intonations should come in accordance with the discourse features. There is also a need to improve the assessment methods of the speech generation system. There has been very little research to evaluate the spoken response generated in human–robot communication.

4.5. Implementation issues

This issue is of prime importance. Even with all the modules of the voice-active robots we have, they will give malfunction if not implemented properly. As the capability of vocal communication adds a life-like feature to the robot, the most important thing required is a real-time response. If the response delay is appreciable, it invites unnaturalness in the conversation even with all the ideal modules. From the software perspective, the state of the art spoken dialog systems provide appreciable delay even on the fastest available computers. Thus, to make system in real-time, not only trimmings in software side required, but also faster processors are needed from the hardware side. The other important issue is the placement of microphones and speakers (TTS based) on the robot. As mentioned in the previous sections, noise corrupts the speech signal reaching the pick-up device and the robot itself is a source of noise (vibrations, motoractuator noise, hammering, etc.). As prevention is better than cure, it is very crucial to prevent a contaminated speech signal reaching the microphones. Contamination of the speech signal can be reduced by placing microphones at a suitable place (inside or outside). Placement of the microphone also depends on the physical embodiment of the robot. Also, the relative placement
of system speakers and microphones effects the barge-in-free [108] performance of the system. Safety of microphones from the manipulating part of the robot body is also important, e.g. if the robot moves its hand while making conversation and if the hand movement control is not precise it may strike its head or other parts (possibly where the microphone is placed) and damage the sensors. Coordination of hardware and software components also plays a major role in gestureposture coordination with the spoken words by voice-active robots. If proper coordination between the hardware and software functioning of the robot is not good, a difference in speech and action will be observable, and will be highly unacceptable.

There are other practical issues associated with the universality of acoustic feature used in the recognition. The ASR trained by adult speech data shows very poor performance for child speech recognition [109, 110]. The lipmovement is also important. For short-time conversation, synchronized lip movement with the spoken word may not be important; however, for long-term conversation, it is important. At present most of experiments with the speaking robot have been done for short-time (from few words to a week) conversation. However, if the robots come to live as a social partner it will be as a long-term partner (even inter-generaction) and under such cases lip movement matters. The other important issue with long-term conversation is that the robot has to take care of the individual or distinct existence of each interacting user, which is not cared for in short-term conversation [111]. As the activity of the voice-active robots will be like a human being in society, they will learn different things through interaction, observation, etc. Such robots need to organize acquired knowledge. How should the memory be organized for the service robot? This is important from both the hardware and software side. How should the robot organize its whole knowledge so that during conversation it can make the best use of the stored knowledge and experience for the response generation in real-time? The other practical issue is related to speech–non-speech detection. It is important, but very difficult, for the system to differentiate between natural pauses in the dialog and the end of the dialog. If the utterance boundaries are not determined then it is problematic for the parser to produce result in real-time [112]. Not only this, if symbiotic robots deviate from naturalness in conversation, they need to keep estimates of such deviations because they may also need to understand such perturbations from other interacting robots with similar shortcomings. Lastly, robustness in the conversational interface should be very high because misrecognition and misunderstanding of commands in a noisy environment may harm the user or robot.

5. CONCLUSIONS

In this paper we have attempted to look into the development of voice-active assistant robots. We have attempted to trace out achievements and limitations of the state of the art in the spoken language processing technologies and requirements for the developing talk-active robots. We have tried to spotlight some key research
issues needed to develop voice-active robots. The state of the art in technology
is not mature enough to develop fully-fledged voice-active social robots for real-
world applications. Unless the speech recognition, synthesis and understanding
technologies matured, the possibility of doing unrestricted conversation with a robot
is reduced. Research in almost related technologies is underway to find suitable and
feasible avenues. However, one cannot make the exact prediction when research
breakthroughs in spoken language technologies will be advanced enough to develop
voice-active robotic characters with the spoken language subtleties of humans.

Acknowledgements

The first author thanks Manbusho, Government of Japan, for providing a Doctoral
fellowship. Valuable discussions with Professor Dr Hiromichi Kawanami and Dr
Ryuichi Nishimura, NAIST, Japan are acknowledged. We also thank the anonymous
reviewers whose valuable comments were highly helpful in rectifying this paper.

REFERENCES

1. J. P. Trevelyan, Redefining robotics for the new millennium, Int. J. Robotics Res. 18 (12),
2. G. C. Atkeson, J. G. Hale, F. Polick, M. Riley, S. Kotosaka, S. Schaul, T. Shibata, G. Teviata,
A. Ude, S. Vijaykumar, E. Kawato and M. Kawato, Using humanoid robots to study human
behavior, IEEE Intelligent System, Special Issue on Humanoid Robotics (July/August), 46–56
(2000).
4. Z. Khan, Attitudes towards intelligent service robots, IPL Technical Report No. TRITA-NA-
5. R. Beckers, O. E. Holland and J. L. Denebourg, From local actions to global tasks: stigmergy
and collective robotics, in: Proc. Artificial Life IV, Fourth International Workshop on the
(1996).
dynamic of collective sorting robot-like ants and ant-like robots, in: Proc. Int. Conf. on
7. Y. Nakauchi and R. Simmons, A social robot that stands in line, Autonomous Robots 12 (3),
8. L. Oestreicher, H. Hüttnerauch and K. S. Ekklund, Where are you going little robot? Prospects
of human–robot interaction, in: Position paper for the CHI 99 Basic Research Symposium,


47. E. C. Cherry, Some experiments on recognition of speech, with one and with two ears, J. Acoust. Soc. Am. 25, 975–979 (1953).


ABOUT THE AUTHORS

Rajkishore Prasad was born in Bihar, India on 23 January 1970. He received BS, MS (1995) and PhD (2000) degrees in Electronics from BRA Bihar University, Muzaffarpur, India. He received the JRF and eligibility for Lectureship from CSIR, Government of India in 1995 and joined the University Department of Electronics BRA Bihar University as a Lecturer in 1996. He received a Japan Government fellowship 2000. Currently he is second year doctoral student at Nara Institute of Science and Technology, Japan. His research interests include expert system development, voice-active robots and blind source separation. He is a life-member of IETE, New Delhi, India.

Hiroshi Saruwatari was born in Nagoya, Japan, on July 27, 1967. He received the BE, ME and PhD degrees in Electrical Engineering from Nagoya University, Nagoya, Japan, in 1991, 1993 and 2000, respectively. He joined Intelligent Systems Laboratory, SECOM Co., Ltd., Mitaka, Tokyo, Japan, in 1993, where he engaged in research and development on the ultrasonic array system for the acoustic imaging. He is currently an Associate Professor of the Graduate School of Information Science, Nara Institute of Science and Technology. His research interests include speech processing, array signal processing, non-linear signal processing, blind source separation, blind deconvolution and sound field reproduction. He received the Paper Award from IEICE in 2000. He is a member of the IEEE, IEICE and Acoustical Society of Japan.
Kiyohiro Shikano received the BS, MS and PhD degrees in Electrical Engineering from Nagoya University in 1970, 1972 and 1980, respectively. He is currently a Professor of Nara Institute of Science and Technology (NAIST), where he is directing the speech and acoustics laboratory. His major research areas are speech recognition, multi-modal dialog systems, speech enhancement, adaptive microphone arrays, and acoustic field reproduction. From 1972, he had been working at NTT Laboratories, where he had been engaged in speech recognition research. During 1990–1993, he was the Executive Research Scientist at NTT Human Interface Laboratories, where he supervised the research of speech recognition and speech coding. During 1986–1990, he was the Head of Speech Processing Department at ATR Interpreting Telephony Research Laboratories, where he was directing speech recognition and speech synthesis research. During 1984–1986, he was a Visiting Scientist at Carnegie Mellon University, where he was working on distance measures, speaker adaptation and statistical language modeling. He received the Yonezawa Prize from IEICE in 1975, the Signal Processing Society 1990 Senior Award from IEEE in 1991, the Technical Development Award from ASJ in 1994, IPSJ Yamashita SIG Research Award in 2000 and Paper Award from the Virtual Reality Society of Japan in 2001. He is a member of the Institute of Electronics, Information and Communication Engineers of Japan, Information Processing Society of Japan, Acoustical Society of Japan, Japan VR Society, IEEE and International Speech Communication Society.