You Just Do Not Understand Me! Speech Recognition in Human Robot Interaction

Omar Mubin [1], Joshua Henderson [1] and Christoph Bartneck [2]

***Abstract*— Speech Recognition has not fully permeated in our interaction with devices. Therefore we advocate a speech recognition friendly artificial language (ROILA) that initially was shown to outperform English, however under constraints. ROILA is intended to be used to talk to robots and therefore in this paper we present an experimental study where the recognition of ROILA is compared to English when speech is input using a robot's microphones and both when the robot's head is moving and stationary. Our results show that there was no significant difference between ROILA and English but that the type of microphone and robot's head movement had a significant effect. In conclusion we suggest implications for Human Robot (Speech) Interaction.**

I. INTRODUCTION

Social Robots are beginning to permeate in our society at a rapid and almost exponential pace [1]. This seamless integration can be witnessed in all facets of our daily life, such as homes, schools and hospitals [2]. Therefore researchers in Human Computer Interaction (HCI) and Human Robot Interaction (HRI) have been devoting significant efforts to provide a user-friendly interactive experience between users and robots [3]. The most natural, simplest and ideal modality to interact with robots is speech using natural language [4], in particular for robots that resemble humans or who are social in nature. Speech Interaction with non-robotic devices such as smart phones (Apple's Siri) and TV's (Samsung) [5] is now also being promoted and explored.

However, the usage of speech is compounded by the difficulties not only robots but machines have with recognizing natural language [6]. The difficulties with getting speech recognition to work to an acceptable accuracy are well documented [7]. The main problems with speech recognition are primarily due to how natural language has evolved, for e.g. words can sound the same but have different meaning (i.e. homophones) [8] and the semantics of language can be ambiguous. We have witnessed extensive research effort in the development of algorithms for the recognition of natural language but we are yet to see an as evident usage of speech interaction in our daily lives [9]. Problems with speech recognition are a cause for concern for most researchers in HRI [10]. Prasad et al [11] go as far as describing Speech Interaction with robots as the Holy Grail. Therefore in our research we carried out the design of a spoken speech recognition friendly artificial language (ROILA) that humans can use to talk to robots. Constraining natural language [12] or providing a novel artificial language [13] for improving speech recognition has been discussed, although in the former, it was command and control and in the latter it comprised of 10 solitary words only. Others have also addressed the unreliability of speech recognition by giving machines potentially easier to recognize non-natural linguistic input such as musical tones and whistles [14]. However, one may wonder how expressive such a modality would be.

[1]School of Computing, Engineering and Mathematics, University of Western Sydney, Australia o.mubin@uws.edu.au

[2]HIT Lab, University of Canterbury, New Zealand christoph.bartneck@canterbury.ac.nz

A. Background

A detailed description of ROILA is available in [15], here we summarise its main design principles. ROILA was designed on the basis of two key attributes, i.e. the proposed language should be easy for robots to recognize and at the same time easy for humans to learn. In order to provide ease in pronunciation, the phonetics of the language were built from phonemes comprised in the most widely spoken languages. In addition, word structure was based on what would be easiest to pronounce (only consonant-vowel units were included). The grammar rules were regular and inflections were not allowed (thereby reducing the number of rules speakers had to remember). To support ease of recognition by a robot, a genetic algorithm generated the vocabulary such that words would be acoustically unique from each other. We conducted an evaluation of ROILA with high school children [15], who learnt the language for 3 weeks and then took part in a controlled experiment where they used ROILA to interact with a LEGO Mindstorms robot. ROILA was shown to outperform English by 18.9% using the open source Sphinx-4 speech recognizer [16]. We used the North American Acoustic Model from Sphinx-4 for the recognition of ROILA because of a) All phonemes of ROILA exist in North American English and are meant to be pronounced in the same way and b) we do not have any native speakers of ROILA hence we cannot derive data to create our own acoustic model.

We acknowledged that the results of our initial evaluation were derived under certain constrained conditions. Firstly, the children's use of ROILA in the gaming scenarios was of typically command and control nature. One would expect that in an ideal HRI setup, users would wish to talk freely to robots. Secondly, LEGO Mindstorms robots do not have a native or in-built microphone, hence only an external (desktop) based microphone was used. In a typical HRI interactive scenario we would expect users to talk to the robot by directing speech towards the robot's speech sensor. We would also anticipate that the recognition accuracy of

Preprint submitted to The 23rd IEEE International Symposium on Robot and Human Interactive Communication. Received February 14, 2014.

the native microphone of the robot would be influenced to an extent by any motor movement within the vicinity of the sensor, a conundrum that has worried several HRI researchers [17]. Therefore in this paper we present an empirical study which set out to determine the extent of the recognition advantage of ROILA over English in more natural conditions (i.e. using the robot's microphone and for semantically richer sentences).

B. Motivation and Research Questions

We also wished to compare the recognition accuracy and consequently the viability of a robot's native microphone sensor against other microphones when the motors near the sensor were both operational and nonoperational. Prior work in Speech Recognition [18] has compared the recognition accuracy of natural language across different microphones, where a headset microphone was shown to outperform lapel and desktop microphones but only after training the system with a large corpus of data. Prior work [19] in HRI has compared the recognition accuracy of various microphones against those of a robot but this was only in the context of natural language, without considering ambient motor noise within the robot and with recordings from only 3 human participants. In summary, our experiment outlined the following main research questions: 1) *On a speech recognizer that is untrained for ROILA* does ROILA outperform English in terms of recognition accuracy regardless of the microphone employed and 2) is the sound sensor of a robot a viable microphone for speech recognition in comparison to other microphones given that it may be influenced by the robot's motor movements.

II. Method

We conducted an empirical study where the recognition accuracy of ROILA was compared to English across a range of conditions. Input was provided by participants in the form of audio recordings of pre-defined sentences which were then passed offline to Sphinx-4 [16]. Ideally, we would have liked to use real time speech recognition but this was not feasible because a) participants could not learn ROILA quickly and b) executing ROILA on the robot had a tendency to cause processing delays. Nevertheless because we would use similar speech recognition configurations for both online and offline input we expect our results from one setup to be transferable to the other. Our methodology of performing offline speech based recognition using the Sphinx-4 speech recognizer to understand and test recognition accuracy is adapted from prior work [20], [19], [21]. The experiment was setup as a 2 (Language Type: ROILA or English) X 3 (Microphone Type: Robot, Headset or Dekstop) X 2 (Robot Head Movement) within subjects design. Since the microphones were listening to data simultaneously, every participant took part in 4 data recording sessions. Appropriate ethics clearances were attained from the host institution prior to conducting the experiment (Reference Number: 13/006583-H10241).

A. Procedure

Every participant was invited to a recording setup in a university tutoring room where they were requested to record a set of N=26 sentences for each of the four sessions, resulting in a total of 4 X 3 (microphone) X 26 recordings. Besides the participant and facilitator, there was no one else in the room. Prior to beginning the recording sessions, the participants were explained that the purpose of the experiment was to evaluate ROILA. Afterwards, they signed consent forms and were guided on how to pronounce ROILA (none of them had prior experience with the language). A simple GUI interface was displayed on a computer screen that guided the participant on what sentence to record. Again as a guide, sample recordings were played out of ROILA sentences to the participants via a headset to assist them in their articulations. The sample recordings were from a North American Native English speaker and had achieved 100% recognition accuracy when passed through Sphinx-4 on a prior occassion. The participants were allowed to hear the sample recordings as many times as they wished via a simple button press. For the sessions in English, no sample recordings were provided. The presentation order of the sentences within each session was also randomized. The four sessions lasted in total for 30-45 minutes, where the order of exposure to the sessions was counterbalanced using a 4 X 4 Latin Square. The facilitator controlled the recording software which was a networked setup of the Nao and the two microphones.

B. Materials

The 26 ROILA sentences were extracted from the ROILA textbook [22] (for e.g. [fekef jutof wikute;they like fruit], [lobo buse tiwil;a robot is not a person], [mona bobuja;she ran], [luluno bamas pelake;taste this soup]). A deliberate effort was made to ensure that a wide phonetic spread was represented in the choosen words (w.r.t ROILA). The pool of ROILA sentences was based on two factors: 1) extracting phrases from different chapters of the book (each chapter addresses a different social setting in daily life) and 2) attempting to have a wide phonetic spread in the ROILA words. Therefore since the ROILA phrases were from various domains and the ROILA vocabulary being phonetically diverse we expect our sentences to have ecological validity. In addition the number of words in each sentence was between two and four (again w.r.t ROILA only, average length of ROILA sentences = 2.9 words). Longer sentences would have increased the articulation effort for participants who had no training in ROILA. The total number of unique ROILA words across all the sentences was 72. The English sentences were translations of the ROILA sentences (average length of English sentences = 3.9 words, total unique words = 107). Each sentence was recorded via three microphone sources, namely a headset, a desktop based microphone (brand: Blue snowball, set to conference mode) and the native microphone of the robot which is located inside the head of the robot (See Appendix for microphone configurations). The choice of microphones was similar to that of [19], with the headset

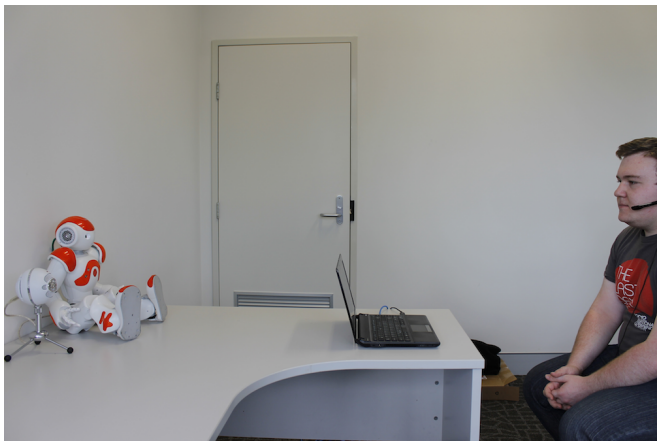

Fig. 1. Experiment Setup

condition acting as a "control condition". The robot that was employed for the experiment was the Nao humanoid robot from Aldeberan Robotics. The Nao robot is perhaps the single most popular humanoid robot that is used significantly in research on Human Robot Interaction (HRI). More than 2500 versions have been sold to researchers worldwide, however a thorough and empirical analysis of the speech recognition abilities of the Nao robot has been unexplored in prior work (as we mention). For two out of the four sessions the robot's head was continuously moving sideways while the recordings took place.

C. Setup

The distance between the participant and the robot was approximately 2 m. The desktop microphone was placed adjacent to the robot. The distance of 2 m was carefully chosen and informed by a) prior work in HRI where speech recognition accuracy was investigated using the Nao robot [19] and b) prior work in Speech Recognition Literature [23] where speech recognition algorithms were evaluated. The room where the experiment was conducted was in a generally quiet space/corridor with the door closely shut. The room was not sound proof though. Our intention was to replicate natural settings of a user talking to a robot. A picture of the setup is provided (see Figure 1).

D. Measurements

The dependent variable was word recognition accuracy, which is simply a word-level Levenshtein distance between what was said and what was recognized. This metric is commonly used in experiments where word recognition accuracy is computed [24]. All the recordings were transformed to a Sphinx-4 friendly format (16 bit, 16K Hz) and then processed offline through Sphinx-4. As mentioned earlier, a default North American English acoustic model (untrained for ROILA) and N-gram grammar was used to process the recordings, in line with our earlier research on the evaluation of ROILA [15]. For both ROILA and English the dictionary employed during speech recognition comprised of the words from the 26 sentences only.

E. Participants

15 university students were recruited for the experiment on a voluntary basis. Their participation was rewarded with a $20 gift card. All participants by requirement were born in Australia and spoke Australian English as their first language. This choice was an attempt to reduce any bias that varying dialects would have on the results. The data from one participant was excluded because of an error in the recordings.

III. Results

Given below is the table summarizing the recognition accuracy means (in %'s) across all conditions (see Table I). A repeated measures ANOVA revealed main effects for Robot Head Movement ($F(1, 13)=20.1$, $p<0.001$) and an even stronger effect for Microphone Type ($F(2, 26)=204.6$, $p<0.0005$). Language Type did not have an effect ($F(1, 13)=4.03$, $p=0.07$). We performed post-hoc tests to complete pairwise comparisons (Bonferroni) on the independent variable Microphone Type. As expected all comparisons were significant ($p<0.001$). There was a significant interaction effect between Language Type and Microphone Type ($F(2, 26)=6.08$, $p=0.007$) and between Microphone Type and Robot Head Movement ($F(2, 26)=10.68$, $p<0.001$).

IV. Discussion and Conclusion

Overall, the recognition accuracy in the headset condition across both languages was found to be 69.5%, which is what we would expect from Sphinx-4 on untrained test data in an ideal ambient environment [21]. We will now analyse our results on the basis of our independent variables, speculate on their implications to HRI and HCI and contemplate on the potential of ROILA as an interaction language between humans and machines.

A. Choice of Microphone

The results of our study shows that the single most important factor for improving the speech recognition rate has been the choice of microphone. Even a $20 off the shelf headset performed significantly better than the $100 semi-professional desktop microphone. To no surprise, the microphone built into the NAO robot performed at an unacceptable level. In order to circumvent the low recognition accuracy of robots, some researchers in HRI have employed the use of external microphones such as ceiling microphones; for example in the context of smart homes [25], however our results show that the desktop microphone achieves better recognition than the results in [19] using a ceiling microphone (where a 22% recognition accuracy is reported). In conclusion, we recommend that designers and researchers in human-robot interaction should use a headset if they intend to use an automatic speech recognition system. But even then our results show that the recognition accuracy does not exceed 70% and it cannot be expected of users to wear

R=ROILA E=English Acc=Accuracy	Head Movement					
	True			False		
	Headset	Desktop	Nao	Headset	Desktop	Nao
R Acc	62.5 (15.4)	23.0 (13.1)	5.8 (3.0)	63.7 (19.1)	30.7 (10.0)	7.6 (3.8)
E Acc	75.0 (7.8)	26.9 (16.1)	3.1 (2.6)	76.6 (6.4)	41.1 (23.9)	7.9 (5.9)

TABLE I

MEANS (STD'DEVS IN BRACKETS) TABLE FOR RECOGNITION ACCURACY ACROSS ALL CONDITIONS

headsets at all times to be able to communicate with robots [26] or even in general HCI scenarios [27], [28].

B. Robot Head Movement

We also have to conclude that the noise generated by Nao's head movement significantly influenced the recognition accuracy. But a robot that does not move its limbs can hardly be considered a (social) robot given that movement is an integral part of human robot communication [29]. In our study we only used the neck motor to move the head. If we had used the whole robot to walk or move around then the recognition accuracy is very likely to have been much worse. We recommend that the robot should pause its actions as soon as it realizes that it is being talked to.

C. Implications to Speech Interaction in Human Robot/Computer Interaction

Both of the afore-mentioned results w.r.t the choice of microphone and robot head movement might not have necessarily surprised speech technology experts ten years ago. As a by-product of our results (and not necessarily emergent from our research questions) we have also witnessed very low recognition accuracy; similar to results found years ago in Speech Recognition Literature [30]. We would have expected that the speech recognition software and hardware would have made more progress in the last decade. The continued difficulty in today's age to get speech recognition to reliably work is definitely a wake up call for researchers in HCI and HRI. One might argue that in certain circumstances commercial automatic speech recognition software have achieved better recognition accuracy but that comes at a cost of a large learning curve, hours of training and generally low usability [31], [32]. We express surprise that our study has revealed depressing results regarding recognition accuracy, coupled with the fact that our research was a unique, thorough and empirical attempt (with a decent sample size) on studying the accuracy of speech recognition in the domain of humanoid robots, we are of the opinion that our results are of practical value and ironical at the same time.

In summary, it is unfortunate, that a very expensive humanoid robot has an almost non-functional speech sensor, i.e. in its current settings we should not expect the Nao robot to understand anything that is said to it, regardless of whether its head is moving or not. The recognition accuracy when the headset was worn was over 70%. In our view, a 70% accuracy is below what is acceptable for a dialog between humans and robots, especially in critical scenarios involving assistive robots. Therefore we must conclude that because the speech understanding capabilities of social robots have not advanced to an extent that we would anticipate, hope and require; the only reliable speech recognition engine for HRI is another human being. HRI designers today are still constrained by the unreliability of speech recognition [33]. As far as robots are concerned, wizards are still required in the robotic land of Oz. The most natural interaction method between humans and robots remains still the most "unnatural" to implement since it does require *artificial* artificial intelligence: a wizard. Our assertion that for verbal interaction in HRI, a wizard of oz setup is perhaps the most optimal is also supported by findings in [34], where it was reported that for about 73% of HRI experiments which involve verbal interaction, a wizard of oz setup was used.

D. Recogntion Accuracy of ROILA

The performance of ROILA has not been significantly worse than English despite the fact that the participants whose native language was English had almost no training in ROILA. However this result should be considered in light of the configuration of the acoustic model of the Sphinx-4 speech recognizer that was employed. In very simple words, an acoustic model entails how phonemes and consequently words are pronounced and it is statistically trained and developed. Since there is no customised acoustic model for ROILA available yet, an English acoustic model was used instead. Therefore, from the outset ROILA was already at a disadvantage. In addition, participants did not have any training in ROILA, as the training of participants requires considerable logistic and practical effort. Consequently, explicitly training participants was not incorporated in the research described in this paper. We expect that given the results from prior research on ROILA [15], where for every 1 minute in training approximately 6 minutes of interaction time were required so that the learning effort would pay off; that the benefits of learning ROILA will lie in the long term. It is very much like learning to type with ten fingers on a QWERTY keyboard (i.e. initial investment, rewards in the long run [35]), is ROILA to languages what "QWERTY" is to keyboards? Imagine you would be asked to speak Finnish for the first time. It cannot be expected that your pronunciation would be good enough for an automatic speech recognition engine to work properly. Even native speakers of Finnish might struggle to understand you. Atleast ROILA has been designed to be easier to learn than most natural languages.

ROILA stands in the tradition of Esperanto and other artificial languages that succeeded in overcoming the drawbacks

of the irregularities in natural languages. And this tradition means that ROILA also faces the problem of accumulating a critical mass of speakers; one of the main reasons why some artificial languages do not prosper [36]. Even natural languages are challenged by this problem.Only if you have somebody to speak ROILA to, will this be a useful language. The advantage that ROILA has over other artificial languages is that it does have at least a chance to quickly gain many speakers. And these speakers will not be humans, but machines. With only update of popular operating systems, millions of computers can understand ROILA. If you can talk to all the machines around you in ROILA, then the utility of ROILA will increase dramatically, leading to a general adaptation of the language.

To reiterate, we believe that the benefits of ROILA are long term in nature and providing some form of training to participants is of paramount importance. As further evidence to this claim, we were able to achieve 100% recognition accuracy for the same 26 ROILA sentences that we employed in this experiment; for the North American English speaker who articulated the sample recordings. The speaker only required a 1-2 hour training session and was able to provide us with the pool of sample recordings in 2 iterations (recognition accuracy in first iteration was more than 75%). Moreover, our prior work on ROILA [20] shows that when participants were requested to pronounce solitary words *without* any training in ROILA the recognition accuracy of ROILA was significantly better than English on the first attempt. The necessity of training participants in ROILA (or any foreign language for that matter) seems to be essential especially when they requirement is to articulate semantically richer sentences. Speech Recognition literature [37] shows that automatic speech recognition (ASR) is easier when the vocabulary/dictionary of the ASR comprises of solitary words only as compared to complete phrases or sentences, as isolated words are easier to articulate (albeit unnatural) and in sentences or phrases word boundaries tend to interact with each other, causing variations in pronunciation from the speaker. We foresee that with basic training in ROILA (pertaining mainly to pronunciation), improvements in recognition accuracy can be attained for free-form speech. But that would only seem to be possible if a headset microphone or any microphone which is closer to the mouth is used. Obtaining similar results on the Nao robot would require at the very least hardware overhaul (better microphone and better sound localisation) and ultimately refinements to ASR software. A multimodal HRI platform might also be a mechanism to circumvent the problems robots have in recognising speech. For e.g. using computer vision to resolve ambiguous deictic references [38].

E. Setup Limitations

We would like to comment on our choice of proximal distance of 2m between user and robot. It was informed by a) HRI literature and b) ASR literature, with the goal to compare our results with those studies. The choice of 2m might seem high for desktop interactions but we expect that interaction with social robots in the future (especially in home environments) might be at larger distances. Optimal proximal distance will be determined by application context, user type, robot size and perhaps most importantly how anthropomorphic the robot is. Prior research in HRI [39] indicates that 1-3 m is an intermediate proximal distance in terms of comfort when interacting with a large humanoid robot. Moreover research in [39] showed that the participants would start to feel uncomfortable at proximal distances of less than 3 m. A second limitation is the issue of simultaneous recording. Would users have been influenced in their articulations to the robot because they were wearing a headset? A between subjects design would have been logistically difficult. Simultaneous recordings were therefore the most suitable choice for us. In an attempt to circumvent a bias, we instructed participants to speak as they would naturally to the robot (without any hyper-articulations).

F. Future Work

In our future work, we aim to repeat a similar study but by initially engaging participants in formal ROILA training and hopefully that will enable ROILA to outperform English in terms of recognition accuracy using the Nao robot. Our long term aim is to promote the adoption of ROILA as an efficient off the shelf platform for speech based interaction in HRI, because lets face it, it is a continuous struggle to solve the dilemma of speech recognition of natural language (within the domain of HRI and otherwise) [40], such that the robot/machine/system can interact autonomously. We have not even touched upon the issue of semantics and a machine's understanding of natural language. By design and nature, natural language is ambiguous [41]. Atleast ROILA attempts to reduce ambiguity and complexity by having a grammar/syntax which is simple and regular.

In summary, the main contributions of our research as described in this paper are:

1) Speech Recognition Accuracy in general is still not at an acceptable level that we may have expected it to be and the best choice of microphone is still a headset microphone.
2) The Nao robot has a weak speech sensor, despite being a state of the art humanoid robot with several autonomous capabilities. Therefore a wizard of oz setup for the Nao robot seems the most logical choice.

APPENDIX

The Nao robot has four omni-sound sensors in its head (each having its own channel), we choose the front one. Its sensitivity is 40+/-3dB and a frequency range of 20Hz-20kHz. The Blue snowball has three settings for various situations according to ambience. We choose Setting 3 which activates the omni capsule designed for conferences. It has a frequency range of 40Hz-18KHz. The headset was a Logitech H390, with a frequency range of 100 Hz - 10kHz.

REFERENCES

[1] S. D. IFR, "World robotics survey," Tech. Rep., 2012.

[2] M. A. Goodrich and A. C. Schultz, "Human-robot interaction: a survey," *Foundations and Trends in Human-Computer Interaction*, vol. 1, no. 3, pp. 203–275, 2007.

[3] D. Perzanowski, A. C. Schultz, W. Adams, E. Marsh, and M. Bugajska, "Building a multimodal human-robot interface," *Intelligent Systems, IEEE*, vol. 16, no. 1, pp. 16–21, 2001.

[4] A. A. Abdelhamid, W. H. Abdulla, and B. A. MacDonald, "Roboasr: a dynamic speech recognition system for service robots," in *Social Robotics*. Springer, 2012, pp. 485–495.

[5] C.-T. Liu, L.-G. Chen, C.-H. Hahm, A. Sung, and Y.-S. Chang, "Evolving technology integration for consumer electronics," in *Consumer Electronics (ISCE), 2013 IEEE 17th International Symposium on*. IEEE, 2013, pp. 11–17.

[6] B. Shneiderman, "The limits of speech recognition," *Communications of the ACM*, vol. 43, no. 9, pp. 63–65, 2000.

[7] M. Forsberg, "Why is speech recognition difficult," *Chalmers University of Technology*, 2003.

[8] H. Lieberman, A. Faaborg, W. Daher, and J. Espinosa, "How to wreck a nice beach you sing calm incense," in *Proceedings of the 10th international conference on Intelligent user interfaces*. ACM, 2005, pp. 278–280.

[9] J. Aron, "How innovative is apple's new voice assistant, siri?" *New Scientist*, vol. 212, no. 2836, p. 24, 2011.

[10] P. W. Schermerhorn, J. F. Kramer, C. Middendorff, and M. Scheutz, "Diarc: A testbed for natural human-robot interaction." in *AAAI*, 2006, pp. 1972–1973.

[11] R. Prasad, H. Saruwatari, and K. Shikano, "Robots that can hear, understand and talk," *Advanced Robotics*, vol. 18, no. 5, pp. 533–564, 2004.

[12] S. Tomko and R. Rosenfeld, "Speech graffiti vs. natural language: Assessing the user experience," in *Proceedings of HLT-NAACL 2004: Short Papers*. Association for Computational Linguistics, 2004, pp. 73–76.

[13] E. Arısoy and L. M. Arslan, "A universal human machine speech interaction language for robust speech recognition applications," in *Text, Speech and Dialogue*. Springer, 2004, pp. 261–267.

[14] U. Esnaola and T. Smithers, "Whistling to machines," in *Ambient Intelligence in Everyday Life*. Springer, 2006, pp. 198–226.

[15] O. Mubin, C. Bartneck, L. Feijs, H. Hooft van Huysduynen, J. Hu, and J. Muelver, "Improving speech recognition with the robot interaction language," *Disruptive Science and Technology*, vol. 1, no. 2, pp. 79–88, 2012.

[16] P. Lamere, P. Kwok, W. Walker, E. B. Gouvêa, R. Singh, B. Raj, and P. Wolf, "Design of the cmu sphinx-4 decoder." in *INTERSPEECH*, 2003.

[17] K. Nakadai, T. Lourens, H. G. Okuno, and H. Kitano, "Active audition for humanoid," in *AAAI/IAAI*, 2000, pp. 832–839.

[18] M. Lincoln, I. McCowan, J. Vepa, and H. K. Maganti, "The multichannel wall street journal audio visual corpus (mc-wsj-av): Specification and initial experiments," in *IEEE Workshop on Automatic Speech Recognition and Understanding*. IEEE, 2005, pp. 357–362.

[19] S. Heinrich and S. Wermter, "Towards robust speech recognition for human-robot interaction," in *Proceedings of the IEEE RSJ International Conference on Intelligent Robots and Systems*, 2011, pp. 468–473.

[20] O. Mubin, C. Bartneck, and L. Feijs, "Using word spotting to evaluate roila: a speech recognition friendly artificial language," in *CHI'10 Extended Abstracts on Human Factors in Computing Systems*. ACM, 2010, pp. 3289–3294.

[21] K. Samudravijaya and M. Barot, "A comparison of public-domain software tools for speech recognition," in *Workshop on Spoken Language Processing*, 2003.

[22] A. Stedeman, D. Sutherland, and C. Bartneck, *Learning ROILA*. Charleston: CreateSpace, 2011.

[23] D. Giuliani, M. Omologo, and P. Svaizer, "Experiments of speech recognition in a noisy and reverberant environment using a microphone array and hmm adaptation," in *Proceedings of Fourth International Conference on Spoken Language (ICSLP 96)*, vol. 3. IEEE, 1996, pp. 1329–1332.

[24] M. Boros, W. Eckert, F. Gallwitz, G. Gorz, G. Hanrieder, and H. Niemann, "Towards understanding spontaneous speech: Word accuracy vs. concept accuracy," in *Proceedings of Fourth International Conference on Spoken Language (ICSLP 96)*, vol. 2. IEEE, 1996, pp. 1009–1012.

[25] S. Kagami, S. Thompson, Y. Nishida, T. Enomoto, and T. Matsui, "Home robot service by ceiling ultrasonic locator and microphone array," in *Robotics and Automation, 2006. ICRA 2006. Proceedings 2006 IEEE International Conference on*. IEEE, 2006, pp. 3171–3176.

[26] H. Sakai, T. Cincarek, H. Kawanami, H. Saruwatari, K. Shikano, and A. Lee, "Voice activity detection applied to hands-free spoken dialogue robot based on decoding using acoustic and language model," in *Proceedings of the 1st international conference on Robot communication and coordination*. IEEE Press, 2007, pp. 180–187.

[27] C. M. Rebman Jr, M. W. Aiken, and C. G. Cegielski, "Speech recognition in the human–computer interface," *Information & Management*, vol. 40, no. 6, pp. 509–519, 2003.

[28] H. A. Mangold, "Realistic hands-free applications-the key for really user-friendly applications," in *International Workshop on Hands-Free Speech Communication*, 2001, pp. 35–38.

[29] T. Fong, I. Nourbakhsh, and K. Dautenhahn, "A survey of socially interactive robots," *Robotics and autonomous systems*, vol. 42, no. 3, pp. 143–166, 2003.

[30] D. R. Reddy, "Speech recognition by machine: A review," *Proceedings of the IEEE*, vol. 64, no. 4, pp. 501–531, 1976.

[31] R. M. Issenman and I. H. Jaffer, "Use of voice recognition software in an outpatient pediatric specialty practice," *Pediatrics*, vol. 114, no. 3, pp. e290–e293, 2004.

[32] H. Yang, C. Oehlke, and C. Meinel, "German speech recognition: A solution for the analysis and processing of lecture recordings," in *Computer and Information Science (ICIS), 2011 IEEE/ACIS 10th International Conference on*. IEEE, 2011, pp. 201–206.

[33] V. A. Kulyukin, "On natural language dialogue with assistive robots," in *Proceedings of the 1st ACM SIGCHI/SIGART conference on Human-robot interaction*. ACM, 2006, pp. 164–171.

[34] L. D. Riek, "Wizard of oz studies in hri: a systematic review and new reporting guidelines," *Journal of Human-Robot Interaction*, vol. 1, no. 1, 2012.

[35] P. Buzing, "Comparing different keyboard layouts: aspects of qwerty, dvorak and alphabetical keyboards," *Delft University of Technology Articles*, 2003.

[36] A. Okrent, *In the land of invented languages: Esperanto rock stars, Klingon poets, Loglan lovers, and the mad dreamers who tried to build a perfect language*. Random House LLC, 2009.

[37] H. Strik and C. Cucchiarini, "Modeling pronunciation variation for asr: A survey of the literature," *Speech Communication*, vol. 29, no. 2, pp. 225–246, 1999.

[38] Z. M. Hanafiah, C. Yamazaki, A. Nakamura, and Y. Kuno, "Human-robot speech interface understanding inexplicit utterances using vision," in *CHI'04 extended abstracts on Human factors in computing systems*. ACM, 2004, pp. 1321–1324.

[39] K. L. Koay, K. Dautenhahn, S. Woods, and M. L. Walters, "Empirical results from using a comfort level device in human-robot interaction studies," in *Proceedings of the 1st ACM SIGCHI/SIGART conference on Human-robot interaction*. ACM, 2006, pp. 194–201.

[40] M. Anusuya and S. K. Katti, "Speech recognition by machine, a review," *arXiv preprint arXiv:1001.2267*, 2010.

[41] K. Church and R. Patil, "Coping with syntactic ambiguity or how to put the block in the box on the table," *Computational Linguistics*, vol. 8, no. 3-4, pp. 139–149, 1982.