What you say is not what you get: Arguing for Artificial Languages Instead of Natural Languages in Human Robot Speech Interaction

Omar Mubin, Christoph Bartneck, Loe Feijs
Department of Industrial Design, Eindhoven University of Technology (TU/e), the Netherlands

Abstract— **The project described hereunder focuses on the design and implementation of a "Artificial Robotic Interaction Language", where the research goal is to find a balance between the effort necessary from the user to learn a new language and the resulting benefit of optimized automatic speech recognition for a robot or a machine. We also discuss the rationale of creating our artificial language and highlight the possibility of improving speech recognition by virtue of an artificial language. In conclusion we present the methodology by which we have designed an initial vocabulary of our artificial language.**

I. Introduction

Robots are becoming a part and parcel of our life and research has already been contemplating in the domain of social robotics [1]. Numerous studies have investigated various controversial issues related to the acceptance of Robots in our society. We are already at a juncture, where importance must now be levied onto how can we as researchers of Human Robot Interaction (HRI) provide humans with smooth and effortless interaction with robots. Organizational studies have shown that the use of robots is gradually growing in large numbers [2] and that they are deployed in diverse domains such as Entertainment, Education, Assistive Technologies, Search and Rescue Acts, and Military and Space Exploration [3]. Given their increasing commercial value it is not very surprising that the emphasis in HRI research has recently been on enhancing the user experience of humans who are directly and indirectly affected by robots. Speech is one of the primary modalities utilized in Human Robot Interaction and is a vital and natural means of information exchange [3]. Therefore, improving the status of speech interaction in HRI could consequently lead to more efficient and more pleasant user-robot-interaction.

Manuscript received September 8, 2009.

O. Mubin is a PhD researcher at the Department of Industrial Design, Eindhoven University of Technology (TU/e), Netherlands, (phone: +31402473842; e-mail: o.mubin@tue.nl).

C. Bartneck, is an Assistant Professor at the Department of Industrial Design, Eindhoven University of Technology (TU/e), Netherlands, (e-mail: c.bartneck@tue.nl).

L.M.G. Feijs is a Full Professor at the Department of Industrial Design, Eindhoven University of Technology (TU/e), Netherlands, (e-mail: l.m.g.feijs@tue.nl).

II. Speech in human robot interaction

Some researchers in HRI have concentrated on designing interaction which can provide or at least to some extent, imitate a social dialogue between humans and a robot. Reviewing various state of the art dialogue management systems unearthed several hindrances behind the adoption of natural language for robotic and general systems alike, which are described next.

A. Speech Recognition

The limitations prevailing in current speech recognition technology for natural language is a major obstacle behind the unanimous acceptance of Speech Interfaces for robots. Existing speech recognition is at times not good enough for it to be deployed in natural environments, where the ambience influences its performance. Recent attempts to improve the quality of automatic speech recognition of natural language for machines have not advanced sufficiently [4].

B. Difficulties in mapping dialogue

Dialogue Management and Mapping is one of the popular techniques used to model the interaction between a user and a machine or a robot [5]. However the inherent irregularity in natural dialogue is one of the main obstacles against deploying Dialogue Management systems accurately [6]. A conversation in natural language involves several ambiguities that cause breakdown or errors. These include issues such as turn taking, missing structure, filler utterances, indirect references, etc. There has been attempt to solve such ambiguities by utilizing non verbal means of communication. As reported in [7], a robot tracks the gaze of the user in the case when the object or the verb of a sentence in a dialogue may be undefined or ambiguous. A second argument related to the difficulties in mapping dialogue is which approach to adopt when building a dialogue management system. Several exist, such as state based, frame based and plan or probabilistic based, with an increasing level of complexity. A state based approach is one in which, the user input is predefined and so the dialogue is fixed. Consequently there is limited flexibility in a state based approach. On the other end of the scale are probabilistic approaches that allow dynamic variations in dialogue [8]. It has been argued by [9] that for most applications of Robotics, a simple state based or frame based approach would be sufficient. However a conflict

arises when it is important to support an interaction which affords a natural experience. In [10] it is stated that a mixed initiative dialogue, that is more natural than a master slave configuration, can only be sustained by adopting a probabilistic approach, which is as stated before, more complex. The hardest dialogue to model is one in which the initiative can be taken at any point by any one.

C. Technological Limitations

The hardware platform of the robot and the speech recognition engine can be out of sync, causing uncertainty to the user [11]. This has been precisely the reason why some HRI researchers have concentrated on using speech more as an output modality instead of as a form of input. As a direct after effect of un-synchronization, both speech recognition and generation are far from optimal and is also one of the reasons why speech technology has not grown as anticipated earlier [12].

D. An after effect: Miscommunication

As a consequence of the prior discussed problems miscommunication occurs between the user and robot. The mismatch between humans' expectations and the abilities of interactive robots often results in frustration. Users are disappointed if the robot cannot understand them properly even though the robot can speak with its mechanical voice. To prevent disappointment, it is important to match the communication skills of a robot with its perception and cognitive abilities. Generally in speech interfaces for robots or otherwise the focus is on using natural language and given their unpopularity, inapplicability and unsuitability for automatic speech recognition, it is perhaps time to find a different balance in the form of a new language.

III. A NEW BALANCE: ARTIFICIAL LANGUAGES

Recent research in speech recognition is already moving in the direction of trying to alter the medium of communication in a bid to improve the quality of speech interaction. As stated in [13], constraining language is a plausible method of improving recognition accuracy. In [14] the user experience of an artificially constrained language ("Speech Graffiti") was evaluated and it was concluded that 74% of the users found it more satisfactory than natural language and also more efficient in terms of time. The field of handwriting recognition has followed a similar road map. The first recognition systems for handheld devices, such as Apple's Newton were nearly unusable. Palm solved the problem by inventing a simplified alphabet called Graffiti which was easy to learn for users and easy to recognize for the device. Therefore, using the same analogy we aim to construct an "Artificial Interaction Language" where an artificial language as defined by the Oxford Encyclopedia is *a language deliberately invented or constructed, especially as a means of communication in computing*. Numerous artificial languages have been designed to improve communication between humans and it remains to be seen if they can improve communication between a human and a machine. As stated earlier, constrained languages can have better performance in terms of recognition and efficiency as compared to natural languages; therefore we aim to determine if artificial languages can exhibit similar results. Our research is constructed on the basis of two main goals. Firstly the artificial interaction language should be learnable by the user and secondly, it should be optimized for efficient automatic speech recognition. There have been attempts to design such a language [15], but the emphasis was only on improving speech recognition and the seemingly conflicting aspect of learnability of a language for humans was ignored. In linguistics, there are numerous artificial languages which address a user perspective by making communication between humans easier and/or universal; however there has been little or no attempt to optimize a spoken artificial language for automatic speech recognition.

IV. TYPES OF ARTIFICIAL LANGUAGES

As a first step in our research, we have analyzed various artificial languages and extending from [16] the following language continuum was designed (see Figure 1). A particular language can be placed in any of the eight categories. Constrained languages were determined to have two main categories which differed by the manner in which the vocabulary was altered. For e.g. in Basic English the vocabulary is just reduced in size but other techniques could be to change the words within the vocabulary as in the Kalle and Astrid approach [17]. Artificial Languages were observed to have four basic types. As described in [16], an artificial language can have naturalistic derivations or be completely artificial in nature [18]. Artificial Languages have been developed for various reasons. The primary one being universal communication i.e. to provide humans with a common platform to communicate, other reasons include, reducing inflections and irregularity from speech and introducing ease of learnability.

V. DESIGNING AN ARTIFICIAL INTERACTION LANGUAGE

The overview of artificial languages was further extended across other dimensions to ascertain what we could learn from existing Artificial Languages, especially in reference to what could be easier to learn for humans. The overview was carried out across two aspects, namely morphology or grammar and phonology. Various encyclopedias such as [19] define the major properties of a language of which morphology and phonology are two key aspects.

In summary it was revealed from the overview that artificial languages created prior were based primarily on Germanic languages, at least phonetically.

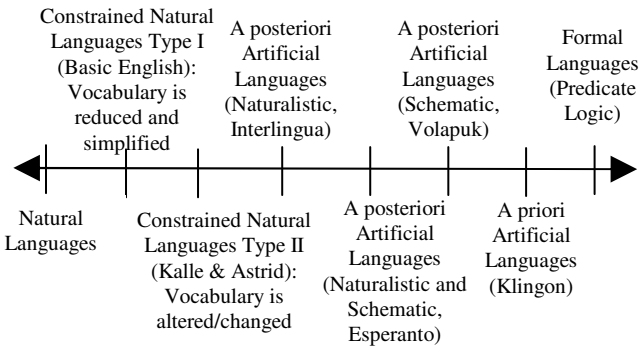

Fig. 1. Language Type Continuum

We presented a morphological overview of artificial languages where, two primary grammar types were derived, of varying grammatical complexity, one involving more inflections than the other. In the future, we aim to evaluate which of the mentioned grammar types will be easier to learn for our intended artificial language and which will be less ambiguous, using methods as advocated in [20]. Moreover, our phonological overview revealed a set of phonemes that might be desirable to include in our artificial language to render it conducive for human learnability. However for both aspects of morphology and phonology what also needs to be determined is how both could contribute to improve speech recognition. For example unique phonemes that have less confusion amongst them would be easier to recognize [15]. Similarly, selecting a particular grammar type could also influence the quality of speech recognition, and determining this effect is something that we aim to address in the future. The afore-mentioned aspects are also important to how speech recognition functions. Typically the grammar of a language is built into the language model of a recognizer and the phonological information is placed in the acoustic model [21]. It has also been shown that longer length units be at word, syllable or phoneme level are more favourable to continuous speech recognition [22]. Therefore, we aim to incorporate and focus on longer words as one of the design principles of our intended artificial language. The size of the vocabulary could also play some role in the design of the language. Users would tend to want as few words as possible to remember but that could be at the cost of an increase in ambiguity for the speaker as the semantic span of the language will be smaller.

Another factor that could influence the speech recognition of the artificial language could be the mother tongue of speakers, as how words of a new language are pronounced would tend to vary from speaker to speaker.

VI. Designing The Vocabulary Of The Artificial Language

As a first step in the design process we aimed to inherit the vocabulary set or word concepts of the simple artificial language Toki Pona [23]. It has 118 word concepts and sufficiently caters for the needs of a simple language. Moreover the pronunciations of the words of Toki Pona were adapted based on the requirements of word length and phonetic information. For example, given that Toki Pona is a simple language it has some words which are very short; of course to be easier to learn for humans. However to assist speech recognition, some of its words will need to be elongated based on a specific methodology, which will also attempt to improve the phonetic discernability of words hereby aiding recognition and would also be scalable and allow for the generation of new words. In order to define the exact representation of the words we utilized a genetic algorithm that would explore a population of words and converge to a solution, i.e. a group or dictionary of words that would have the lowest confusion amongst them and in theory be ideal for speech recognition.

Extending from the phonological overview we utilized a common phoneme list which gave a set of phonemes found in major natural languages of the world. Certain other constraints were employed to reduce this list further, such as diphthongs were excluded; and phonemes that had ambiguous behaviour across languages were ignored. Therefore the final set of phonemes that we wished to use for our artificial language was: {a, b, e, f, i, j, k, l, m, n, o, p, s, t, u, w} or in the Arpabet [24] notation {AE, B, EH, F, IH, JH, K, L, M, N, AA, P, S, T, AH, W}. Extending the word and syllable structure of Toki Pona we designed our own word types. We started off with 8 word types and attempted to maintain a balance of learnability and appropriate word length. In the first iteration of our design cycle we have restricted the maximum word length to 6 characters and/or 3 syllables. Word types were (VCCVCV, VCVCV, VCVCCV, CVCVC, CVCVCV, VCCV, VCVC, CVCV). Minimum word length was 4 characters. The manner in which the words would be constructed would need to be carefully implemented as to render the vocabulary to be speech recognition friendly. Moreover, the method would need to be scalable as well to allow the generation of as many words as required.

The genetic algorithm was randomly initialized for a population of N dictionaries/plausible solutions each having W words or genes, where each word was any one of the afore-mentioned 8 word types. The algorithm was then run for G generations with mutation and cross over being the two primary infant generating techniques. Mutation was set to a standardized rate of 1%. For a given dictionary its confusion was defined as the average confusion of its all constituent words or genes, i.e. pair wise confusions were computed for each word. In every generation, 6% of the best fit (elite) parents were retained and infants were reproduced to complete the population. Parents were selected for breeding using the standard roulette wheel selection [25]. Note that in absolute terms low fitness or low confusion was preferred, so the selection had to be reversed.

The fitness function was determined from data available in the form of a confusion matrix from source [26], where the matrix provided the conditional probability of recognizing a phoneme p_i when phoneme p_j was said instead. The confusion matrix was generated via a phoneme recognizer using the TIMIT corpus for English words [26]. The confusion between any two words within a dictionary was determined by computing the probabilistic edit distance, as suggested in [27]. The edit distance was a slight modification of the conventional Levenshtein distance algorithm [28]. Insertion and deletion probabilities of each and every phoneme were also utilized from [26].

Shown in the table (see Table I) is a sample vocabulary containing 25 words generated over 200 generations. The vocabulary shown is the dictionary that had the least confusion across the N solutions, where N = 200.

VII. FUTURE WORK

In the evaluation of the language and its suitability for speech recognition we aim to compare its performance with a natural language such as English for both conditions: with and without grammar. Firstly, we aim to compare on a word level only and will later add grammar as part of the evaluation. The next obvious steps will be to add grammar as part of the vocabulary and also to identify explicitly how every word will be pronounced.

We also aim to measure the subjective satisfaction of such a language and also evaluate its learnability for human users using various techniques such as the SASSI approach [18] or objective measures such as in [20]. It should be stressed here that we have presented a design idea and the rationale behind it. We can only claim that our concept works until we perform a successful evaluation and this should be noted as one of the limitations of our research accomplished so far.

Our intention is to carry out future research in the form of two or three iterative cycles as a spiral model. Each cycle typically would have four phases: requirements, design, implementation and evaluation. We intend to deploy our interaction language within the domain of robotics, however our proposed interaction language does not necessarily have to be restricted to robots only, but it could be applied to any behavioral product that employs speech interaction.

VIII. SUMMARY

In summary we believe that our idea is novel and might seem controversial, provocative and untraditional at first sight. The first criticism that might be drawn is that for any artificial language it would need to be learnt by users. However, we wish to explore the benefits that an artificial language could provide if it's designed in such a way that it is speech recognition friendly. This benefit might end up outweighing the price a user has to pay in learning a new language. A second criticism that might be levied on our idea is that many artificial languages were created already but nobody ended up speaking them. Where our approach is different is that we aim to deploy and implement an artificial language in a robot and once several robots can speak a certain language it might lead and encourage humans to speak it as well. Through this workshop we hope to be provided with an opportunity to present our proposal to experts in the field of speech interaction in HRI who would be able to provide constructive feedback and valuable insights.

Table. 1. Sample Vocabulary of 25 words

Sample Words	
bifaf	bosib
tesif	aktubi
fasin	iptabe
pomi	tasime
kekof	itoka
ojsuko	otajki
wemupu	nobelu
miwawu	awejja
amop	famifo
ajamo	lipum
lowu	ajji
alalso	tomite
obonu	

ACKNOWLEDGMENT

We would like to acknowledge Andrew Lovitt for providing access to the data of the confusion matrix and also for suggesting insights related to our research.

REFERENCES

[1] T. Fong, I. Nourbakhsh, and K. Dautenhahn, "A survey of socially interactive robots," *Robotics and Autonomous Systems,* vol. 42, pp. 143-166, 2003.

[2] I. S. Department, "World Robotics Survey " 2008.

[3] M. A. Goodrich and A. Schultz, "Human Robot Interaction: A Survery," *Foundations and Trends in Human–Computer Interaction,* vol. 1, pp. 203-275, 2007.

[4] B. Shneiderman, "The limits of speech recognition," *Commun. ACM,* vol. 43, pp. 63-65, 2000.

[5] J. Fry, H. Asoh, and T. Matsui, "Natural dialogue with the Jijo-2 office robot," in *In Proceedings of the 1998 IEEE/RSJ International Conference on Intelligent Robots and Systems* Victoria, B.C., Canada, 1998, pp. 1278-1283 vol.2.

[6] G. E. Churcher, E. S. Atwell, and C. Souter, "Dialogue management systems: a survey and overview," University of Leeds, Leeds, UK1997.

[7] 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* Vienna, Austria: ACM, 2004, pp. 1321-1324.

[8] T. H. Bui, "Multimodal Dialogue Management - State of the art," University of Twente, Enschede, The Netherlands, Technical ReportJanuary 3, 2006 2006.

[9] D. Spiliotopoulos, I. Androutsopoulos, and C. D. Spyropoulos, "Human-robot interaction based on spoken natural language dialogue," in *Proceedings of the European Workshop on Service and Humanoid Robots (ServiceRob 2001)* Bari, Italy, 2001.

[10] L. S. Lopes and A. Teixeira, "Human-robot interaction through spoken language dialogue," in *In Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems, 2000. (IROS 2000)*, Takamatsu, Japan, 2000, pp. 528-534 vol.1.

[11] V. A. Kulyukin, "On natural language dialogue with assistive robots," in *ACM conference on Human-robot interaction* Salt Lake City, Utah, USA: ACM, 2006, pp. 164-171.

[12] M. Liberman, "Computer speech synthesis: its status and prospects," *Proceedings of the National Academy of Sciences of the United States of America,* vol. 92, pp. 9928-9931, 1995.

[13] R. Rosenfeld, D. Olsen, and A. Rudnicky, "Universal speech interfaces," *Interactions,* vol. 8, pp. 34-44, 2001.

[14] S. Tomko and R. Rosenfeld, "Speech Graffiti vs. Natural Language: Assessing the User Experience," in *Proceedings of HLT/NAACL*, 2004.

[15] S. Hinde and G. Belrose, "Computer Pidgin Language: A new language to talk to your computer?," Hewlett-Packard Laboratories2001.

[16] P. Janton, *Esperanto: language, literature, and community*: State University of New York Press, 1993.

[17] Wikipedia, "Rövarspråket." , 2009.

[18] K. Hone and R. Graham, "Towards a tool for the subjective assessment of speech system interfaces (SASSI)," *Natural Language Engineering,* vol. 6, pp. 287-303, 2001.

[19] C. David, "The Cambridge encyclopedia of language," in *The Cambridge encyclopedia of language*: Cambridge University Press, 1997.

[20] O. Mubin, S. Shahid, C. Bartneck, E. Krahmer, M. Swerts, and L. Feijs, "Using Language Tests and Emotional Expressions to Determine the Learnability of Artificial Languages," in *CHI 2009 Conference on Human Factors in Computing Systems*: ACM, 2009, pp. 4075-4080.

[21] C. Lee, F. Soong, and K. Paliwal, *Automatic speech and speaker recognition: advanced topics*: Kluwer Academic Pub, 1996.

[22] A. Hämäläinen, L. Boves, and J. De Veth, "Syllable-length acoustic units in large-vocabulary continuous speech recognition," 2005, pp. 499-502.

[23] S. E. Kisa, "Toki Pona - the language of good." 2008.

[24] Wikipedia, "Arpabet," 2009.

[25] Wikipedia, "Selection - Genetic Algorithm," 2008.

[26] A. Lovitt, J. Pinto, and H. Hermansky, "On Confusions in a Phoneme Recognizer," *IDIAP Research Report, IDIAP-RR-07-10,* 2007.

[27] A. Amir, A. Efrat, and S. Srinivasan, "Advances in phonetic word spotting," ACM New York, NY, USA, 2001, pp. 580-582.

[28] Wikipedia, "Levenshtein Distance," 2009.