Towards the Design and Evaluation of ROILA: A Speech Recognition Friendly Artificial Language

Omar Mubin, Christoph Bartneck, and Loe Feijs

Department of Industrial Design
Eindhoven University of Technology
Den Dolech 2, 5612 AZ Eindhoven, The Netherlands
{o.mubin,c.bartneck,l.m.g.feijs}@tue.nl

Abstract. In our research we argue for the benefits that an artificially designed language that we call ROILA could provide to improve the accuracy of speech recognition given that it is constructed on speech recognition friendly principles. We also contemplate the trade off effect of users investing some effort in learning such a language. Initially we present the design and evaluation of the vocabulary of ROILA and subsequently we describe the ROILA grammar and the method by which we rationally chose grammar rules. Our evaluation results indicated that the vocabulary of ROILA significantly outperformed English whereas we could not yet replicate similar trends while evaluating the grammar.

Keywords: Artificial Languages, Automatic Speech Recognition, Sphinx-4.

1 Introduction

Recent research in speech recognition is gradually progressing towards altering the medium of communication in a bid to improve the quality of speech interaction. As stated in [12], constraining language is a plausible method of improving recognition accuracy. In [15] the user experience of an artificially constrained language was evaluated within a movie-information dialog interface and it was concluded that 74% of the users found the constrained language interface to be more satisfactory than natural language interface. The limitations prevailing in current automatic speech recognition technology for natural languages is an obstacle behind the unanimous acceptance of Speech Interaction. Generally in speech interfaces the focus is on using natural language; it may be time to explore a different balance in the form of a new language. 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. Using the same analogy we aim to design a "Speech Recognition Friendly Artificial Language" (ROILA) where an artificial language as defined by the Oxford Encyclopedia is a language deliberately invented or constructed. In linguistics, there are numerous artificial

H. Loftsson, E. Rögnvaldsson, S. Helgadóttir (Eds.): IceTAL 2010, LNAI 6233, pp. 250–256, 2010.

languages (for e.g. Esperanto, Interlingua) whose goal is easier communication amongst users; however there has been little or no attempt to optimize a spoken artificial language for speech recognition. In summary, our research is constructed on the basis of two main goals. Firstly the artificial language should be optimized for efficient automatic speech recognition and secondly, there should be an attempt to make it learnable for a user, two possibly contradictory requirements, for e.g. users would prefer shorter words but shorter words would be harder to recognize.

2 Vocabulary Design

In order to obtain a group of phonemes that could be used to generate the vocabulary of ROILA we conducted a phonological overview of natural languages [10]. Extending from our goal of designing a language that is easy to learn for humans, we extracted a set of the most common phonemes present in the major 13 natural languages of the world based on number of speakers. We used the UCLA Phonological Segment Inventory Database (UPSID) [11]. The database provides an inventory of the phonemes of 451 languages of the world. We generated a list of phonemes that are found in 5 or more, major languages. This resulted in a total of 23 phonemes. Certain other constraints were employed to reduce this list further; diphthongs were excluded; and phonemes that had ambiguous behavior across languages were ignored. Therefore the final set of 16 phonemes that we wished to use for our artificial language was: (in ArpaBet notation) {AE, B, EH, F, IH, JH, K, L, M, N, AA, P, S, T, AH, W}.

As a starting point for the first version of the vocabulary of ROILA we choose the artificial language Toki Pona [6] which caters for the expression of very simple concepts by just 115 words. Therefore this number formed the size of the ROILA vocabulary. In order to maintain a balance between our two research goals we set the word length to 4, 5 and 6 characters, with each word having 2 or 3 syllables rendering the following word types: CVCV, VCVC, VCCV, CVCVC, VCVCV, CVCVCV, VCCVCV, VCVCCV, where V refers to a vowel and C to a consonant from our pool of 16 phonemes. The 8 word types were simple extensions of words existing in Toki Pona based on the assumption that such words would be easy to learn and pronounce. To define the scalable representation of the words we utilized a genetic algorithm that would converge to a vocabulary of words that would have the lowest confusion amongst them and in theory be ideal for speech recognition. The genetic algorithm randomly initialized a vocabulary of N words, for P vocabularies, where each word was any one of the 8 afore-mentioned word types. The algorithm was then run for G generations with mutation and cross-over being the two primary offspring generating techniques. The fitness function was determined from data available in the form of a confusion matrix (from [7]), where the matrix provided the conditional probability of recognizing a phoneme p_j by a speech recognizer when phoneme p_i was said instead. Therefore, the confusion between any two words was determined by computing the probabilistic string edit distance, as suggested in [1]. The first ROILA vocabulary was generated by running the algorithm for P=G=200. In order to have a benchmark of English words to compare against we set the English vocabulary as the meanings of the 115 Toki Pona words.

2.1 Vocabulary Evaluation

In order to evaluate ROILA 16 (6 female) voluntary participants were asked to record samples of every word from both English and ROILA. The recordings were then passed offline through the Sphinx-4 [4] speech recognizer. Participants had various native languages but all were graduate students and hence had reasonable command over English. Recordings were carried out using a high quality microphone. Sphinx was tuned such that it was able to recognize ROILA by means of a phonetic dictionary; however an acoustic model for English was used. In addition, we did not carry out any training on the acoustic model for ROILA. One of the researchers conducted rounds of ROILA recordings until we had a pool of recordings that rendered a recognition accuracy of 100%. These sample recordings of every word would be played out before participants recorded each ROILA word. This was done to ensure that the native language of participants would not affect their ROILA articulations. The experiment was carried out as a 2 condition within subject design, where the language type (English, ROILA) was the main independent variable. The dependent variable was the total number of errors in recognition. Words from both English and ROILA were randomly presented and the order of recording English or ROILA first was also controlled between participants. We carried out a repeated measure ANOVA which revealed that language type did not have an effect F (1,9) = 0.758, p = 0.41. Both ROILA and English performed equally in terms of accuracy (67.61% and 67.66% respectively). Without any training data, such accuracy is expected from Sphinx on test data [14]. To judge if ROILA word structure had an effect on recognition accuracy, we executed an analysis in which the type of word was the independent variable. This factor had 2 levels namely CV or non-CV type, where CV-type words were CVCV, CVCVC and CVCVCV. The ANOVA analysis revealed a nearly significant trend F (1, 113) = 3.6, p = 0.06. CV-type words performed better on recognition (on average 4.19 participants got such words wrong, as compared to non CV type words, where 5.75 participants got them wrong). Therefore for our second iteration of the evaluation we generated a new vocabulary that comprised of CV type words only. The genetic algorithm was run with the parameters G=P=200. We had 11 (4 female) from the earlier 16 participants carry out recordings of the new vocabulary using the same setup and procedure. We did not have them record the English words again. Participants would once again hear sample pronunciations. The REMANOVA revealed that the new ROILA vocabulary significantly outperformed English F (1, 10) = 4.86, p = 0.05 (see Figure 1). The accuracy for the 11 participants was English: 65.11%, and ROILA_CV: 71.11%. This vocabulary was hence declared as the first ROILA vocabulary.

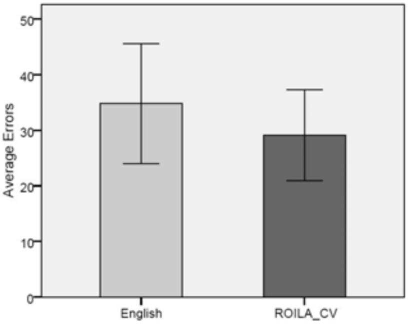

Fig. 1. Average Errors Bar Chart for English and ROILA_CV

3 Grammar Design

In conjunction with conducting a phonological overview of artificial languages we also carried out a morphological overview of artificial languages individually and also in contrast to major natural languages of the world [10]. This aided us in identifying grammar features which were popular in both sets of languages. We determined several grammatical categories based on properties defined in various linguistic encyclopedias [5]. Gender, numbering, tense and aspect are some examples. However within each category there were a number of options that we could choose from, for e.g. should we have gender? How many tenses should we have? In order to make our choice we carried out a rationale decision making process by utilizing the Questions, Options and Criteria (QOC) technique [8]. For this purpose we defined the following important criteria for every grammatical property: Learnability; defines whether the grammatical marking in question would be easy to learn or not, Expected recognition accuracy; defines the effect the grammatical marking would have on the anticipated word error rate given that the more constrained a grammar (lower perplexity) is the better it would be for recognition [9], Vocabulary size; describes the effect the grammatical marking would have on increasing or decreasing the vocabulary size, Expressive Ability of the language; defines whether using the grammatical marking in question would actually enable speakers to express more concepts then they would have been unable to do so otherwise, Efficiency; simply relates the grammatical marking to how many words would be required to communicate any solitary meaning, Acknowledgement within Natural and Artificial Languages; states the popularity of the particular grammatical marking amongst each type of languages. Appropriate weights were assigned to the criteria based on importance, for e.g. learnability and expected recognition accuracy were assigned higher weights with recognition accuracy being given twice as much weight as learnability. The total sum of the weights was 1.

Table 1. ROILA Example Sentences

ROILA Sentence	English Translation	Literal Meaning
pito fosit bubas.	I am walking to the house.	I walk house
pito fosit jifi bubas.	I walked to the house.	I walk < past tense marker > house
pito fosit jifo bubas.	I will walk to the house.	I walk < future tense marker > house

All possibilities of each grammatical category were listed and every category was then ranked across the criteria by giving a number between 1 and 3 with 3 being the best fit. The category which yielded the highest output was then chosen to be as the grammar category of choice. After filling in a matrix we concluded firstly that the ROILA grammar would be of isolating type. Affixes would not be added as this might alter the word structure hereby reducing their efficiency for speech recognition. Therefore grammatical categories in ROILA would be represented by word markers (see Table 1). At the end we arrived at the following properties: Gender (male, female) on the level of pronouns only and not nouns, Numbering (singular, plural) on the level of nouns, Person references (first, second, third) on the level of pronouns, Tense (past, present, future) and word order would be SVO.

3.1 Grammar Evaluation

In order to evaluate the grammar in terms of recognition we formulated some sample sentences (N=30) based on a hypothetical interaction scenario for a dialog system. These sentences were evaluated against their English semantic counterpart. Sphinx-4 Language Models were created using the Sphinx Knowledge Base tool [13]. An identical setup was followed as done in the evaluation of the vocabulary except that participants would now record sentences and not isolated words. Participants would once again hear a sample voice as a guide of how to pronounce sentences. The dependent variable was word accuracy, a common metric to evaluate continuous speech recognition [3] with the independent variable yet again language type. In the initial evaluation we conducted recording sessions with 8 participants. However we were unable to achieve significant results in favor of ROILA; as indicated by the REMANOVA results F (1, 7) = 1.97, p = 0.21.

4 Discussion and Conclusion

Our results revealed some interesting insights. Firstly, we were able to achieve improved speech recognition accuracy as compared to English for a relatively larger vocabulary. Similar endeavors have only been carried out for a vocabulary size of 10 [2]. Secondly, we quantitatively illustrated that CV type words perform better in recognition; co-articulation of CV syllables could be one explanation for that. We must keep in mind several implications to our results. Firstly, participants recorded words without any training in ROILA, whereas they were already acquainted with English. Potentially, by training participants in ROILA the accuracy could be further improved. This effect was observed to be more pronounced when participants had to speak ROILA sentences, which could explain the insignificant difference between ROILA and English in terms of the accuracy of speech recognition. The acoustic models of Sphinx are trained with dictation training data and from what we observed the ROILA sentence articulations of participants did not fall within the domain of dictation speech. There were pauses between words and pronunciations were not smooth, which could have been caused by the inexperience of the participants in ROILA. In the future, we aim to conduct more evaluation sessions of ROILA sentences after carrying out training with additional participants. It may also be observed that the acoustic model of Sphinx was primarily designed for English, yet our ROILA accuracy in the second vocabulary iteration were significantly better as compared to English; a promising result indeed. What we would also like to determine in the future is the magnitude of the difference between ROILA and English. This could be accomplished by using the same English acoustic model for another natural language and comparing the differences in recognition accuracy between the three languages (English, ROILA and the second natural language).

We acknowledge the trade-off factor of humans having to invest some energy in learning a new language like ROILA, even though in various steps of the design process we have tried to accommodate the aspect of human learnability and introduce language features which were conducive to learnability. In summary, by designing an artificial language we are faced with the effort a user has to put in learning the language. Nevertheless, we wish to explore the benefits that an artificial language could provide if its designed such that it is speech recognition friendly. This factor might end up outweighing the price a user has to pay in learning the language and would ultimately motivate and encourage them to learn it. Another criticism that might be levied on ROILA is that many artificial languages were created already but not many people ended up speaking them. Where our approach is different is that we aim to deploy and implement our artificial language in machines and once certain machines can speak the new language it could encourage humans to speak it as well. In the future, we aim to train participants in ROILA and evaluate it by deploying it in an interaction context. We acknowledge that a meaningful societal application of our language would provide an extra gain in addition to recognition performance. We aim to explore applications for children, medical tasks, or care robots[1].

Acknowledgements. We would like to thank the reviewers for their helpful comments and feedback to revise the paper.

References

1. Amir, A., Efrat, A., Srinivasan, S.: Advances in phonetic word spotting. In: The Tenth International Conference on Information and Knowledge Management, pp. 580–582. ACM Press, New York (2001)

[1] To know more about ROILA and its latest developments please visit: http://roila.org

2. Arsoy, E., Arslan, L.: A universal human machine speech interaction language for robust speech recognition applications. In: Sojka, P., Kopeček, I., Pala, K. (eds.) TSD 2004. LNCS (LNAI), vol. 3206, pp. 261–267. Springer, Heidelberg (2004)
3. Boros, M., Eckert, W., Gallwitz, F., Grz, G., Hanrieder, G., Niemann, H.: Towards understanding spontaneous speech: Word accuracy vs. concept accuracy. In: CORR (1996)
4. Carnegie-Mellon-University: Sphinx-4 (2008), http://cmusphinx.sourceforge.net/sphinx4/
5. David, C.: The cambridge encyclopedia of language (1997)
6. Kisa, S.E.: Toki pona - the language of good (2008), http://www.tokipona.org/
7. Lovitt, A., Pinto, J., Hermansky, H.: On confusions in a phoneme recognizer. IDIAP Research Report, IDIAP-RR-07-10 (2007)
8. MacLean, A., Young, R., Bellotti, V., Moran, T.: Questions, options, and criteria: Elements of design space analysis. Human-Computer Interaction 6(3), 201–250 (1991)
9. Makhoul, J., Schwartz, R.: State of the art in continuous speech recognition. Proceedings of the National Academy of Sciences 92(22), 9956–9963 (1995)
10. Mubin, O., Bartneck, C., Feijs, L.: Designing an artificial robotic interaction language. In: Gross, T., Gulliksen, J., Kotzé, P., Oestreicher, L., Palanque, P., Prates, R.O., Winckler, M. (eds.) INTERACT 2009. LNCS, vol. 5727, pp. 851–854. Springer, Heidelberg (2009)
11. Reetz, H.: Upsid-info (2008), http://web.phonetik.uni-frankfurt.de/upsid_info.html
12. Rosenfeld, R., Olsen, D., Rudnicky, A.: Universal speech interfaces. Interactions 8(6), 34–44 (2001)
13. Rudnicky, A.: Sphinx knowledge base tool (2010), http://www.speech.cs.cmu.edu/tools/lmtool-new.html
14. Samudravijaya, K., Barot, M.: A comparison of public-domain software tools for speech recognition. In: Workshop on Spoken Language Processing, pp. 125–131. ISCA (2003)
15. Tomko, S., Rosenfeld, R.: Speech graffiti vs. natural language: Assessing the user experience. In: Proceedings of HLT/NAACL (2004)