Using Word Spotting to Evaluate ROILA: A Speech Recognition Friendly Artificial Language

Omar Mubin
Department of Industrial Design,
Eindhoven University of
Technology (TU/e),
The Netherlands.
o.mubin@tue.nl

Christoph Bartneck
Department of Industrial Design,
Eindhoven University of
Technology (TU/e),
The Netherlands.
c.bartneck@tue.nl

Loe Feijs
Department of Industrial Design,
Eindhoven University of
Technology (TU/e),
The Netherlands.
l.m.g.feijs@tue.nl

CHI 2010, April 10–15, 2010, Atlanta, Georgia, USA.
ACM 978-1-60558-930-5/10/04.

Abstract
In our research we argue for the benefits that an artificial language could provide to improve the accuracy of speech recognition. We briefly present the design and implementation of a vocabulary of our intended artificial language (ROILA), the latter by means of a genetic algorithm that attempted to generate words which would have low likelihood of being confused by a speech recognizer. Lastly we discuss the methodology and results of two word spotting experiments that were carried out to evaluate if indeed the vocabulary of ROILA achieved better recognition than English. Our results reveal that our initial vocabulary was not significantly better than English but when the vocabulary was modified to include CV-type words only, the vocabulary nearly significantly outperformed English.

Keywords
Artificial Languages, Speech Recognition, Sphinx-4

ACM Classification Keywords
H5.2. User Interfaces.

General Terms
Design

Introduction

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 [1], constraining language is a plausible method of improving recognition accuracy. In [2] the user experience of an artificially constrained language 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 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 languages (for e.g. Esperanto, Interlingua) which attempt to make 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.

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, it should be learnable by a user, two possibly contradictory requirements. For e.g., speech recognizers prefer longer words whereas humans would prefer shorter words which are easier to learn. By designing an artificial language which is entirely new 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 it's designed such that it is speech recognition friendly, which might end up outweighing the price a user has to pay in learning the language. In this paper initially, we present the design and implementation of the vocabulary for our artificial language. Lastly, the vocabulary is evaluated by running recordings from users in a speech recognizer via a word spotting experiment. The language design is still an ongoing project, but this paper describes the principles and the experiments that will drive the development forward.

Design of ROILA

In order to obtain a selection of phonemes that could be used to generate the vocabulary of ROILA we conducted a phonological overview of natural languages [3]. Extending from our research 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 languages of the world. We used the UCLA Phonological Segment Inventory Database (UPSID) [4]. The database provides an inventory of the phonemes of 451 different languages of the world. We generated a list of phonemes that are found in 5 or more, major languages of the world, based on number of speakers. 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: {a, b, e, f, i, j, k, l, m, n, o, p, s, t, u, w}.

Word Type	Examples
CVCV (CV-type)	babo, wimo
VCVC	ujuk, amip
VCCV	obbe, uwjo
CVCVC (CV-type)	mejem, kutak
VCVCV	ofeko, ejana
CVCVCV (CV-type)	panowa, fukusa
VCCVCV	ukboma, emfale
VCVCCV	onabbe, emenwi

Table 1. ROILA Word Examples

Generating the Vocabulary of ROILA

As a starting point for the first version of the vocabulary of ROILA we choose the artificial language Toki Pona [5]. Toki Pona is designed on the basis of simplicity and caters for the expression of very simple concepts by just 115 short words. Therefore this number formed the initial size of the ROILA vocabulary. In order to define the exact and scalable representation of the words we utilized a genetic algorithm that would explore a population of words and converge to a solution, i.e. a vocabulary of words that would have the lowest confusion amongst them and in theory be ideal for speech recognition. In order to maintain a balance between ease of learnability for users and efficient speech recognition we set the word length to 4, 5 and 6 characters each word having 2, 2 or 3 and 3 syllables respectively. The following word types were deemed as possible constituents of the vocabulary: 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, again a design decision based on the assumption that such words would be easy to learn and pronounce.

The genetic algorithm randomly initialized a vocabulary of 115 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. Mutation was set to a standardized rate of 1%. For a given vocabulary 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 against each word. In every generation, 6% of the best fit (low confusion) parents were retained and new offspring was reproduced to complete the population. Cross-over was done by word selection (not within words). Parents were selected for breeding using the standard roulette wheel selection [6]. Note that in absolute terms low fitness or low confusion was preferred, so the selection had to be reversed. To choose the best vocabulary of words a fitness function was required which could somehow rank the populations based on the inter-confusion of its words. 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. The confusion between any two words within a vocabulary was determined by computing the probabilistic edit distance, as suggested in [8]. The edit distance was a slight modification of the conventional Levenshtein distance algorithm [9]. Insertion and deletion probabilities of each phoneme were also utilized from the same confusion matrix. The first ROILA vocabulary was generated by running the algorithm for P=G=200. Example words from this vocabulary are shown in table 1. The ROILA vocabulary had 54 six character words, 31 five character words and 30 four character words, which clearly exemplified the tendency of the genetic algorithm to favor longer words. The average length of the ROILA vocabulary was 5.2 characters per word.

In order to have a benchmark of English words to compare against in the subsequent speech recognition performance test we set the English vocabulary as the meanings of all the 115 Toki Pona words. The average length of the English vocabulary thus obtained was 4.5 characters per word.

Figure1. Recording Setup

Evaluating ROILA via Word Spotting

In order to adjudicate whether ROILA was indeed better than its counterpart English vocabulary we ran a word spotting test, where participants were asked to record samples of every word from both English and ROILA and the recordings were then passed offline through the Sphinx-4 [10] speech recognizer.

Participants

16 (6 female) voluntary users were recruited for the recordings. Participants had various native languages but all were university graduate or post graduate students and hence had reasonable command over English. The total set of Native Languages of the participants was 10 (American English-3, British English-1, Dutch-5, Spanish-1, Urdu-1, Greek-1, Persian-1, Turkish-1, Bengali-1, Indonesian-1).

Material

Recordings were carried out in a silent lab with little or no ambient sound using a high quality microphone (see Figure 1). A recording application was designed that would one by one display the words to be recorded. Participants would record all the words from a particular language before moving on to the next language. Recordings of every participant were then passed through the Sphinx-4 Speech recognizer. The choice of speech recognizer was carefully ascertained keeping in mind the requirement that the speech recognition engine should be open source and allow for the recognition of an artificial language. Moreover Sphinx has been quantitatively evaluated as acoustically superior to other open source speech recognition engines such as HTK [11]. Sphinx was tuned such that it was able to recognize ROILA by means of a phonetic dictionary; however the acoustic model that we used was that of English. Note that the ROILA words were generated from a confusion matrix that extracted its data from the basis of another speech recognizer [7] and not Sphinx; this might be a limitation but most speech recognizers operate on the same basic principles. In addition, we did not carry out any training on the acoustic model for ROILA.

Pilots and Procedure

In order to ascertain the recognition of ROILA within Sphinx-4, we carried out some pilot recording sessions. We noticed that for certain Native American English speakers, the recognition accuracy was relatively higher. Therefore we choose a Native American English speaker and conducted several recording iterations until we had a pool of sample recordings of that voice that rendered a recognition accuracy of 100%. These sample recordings of every word would be played out once before other participants recorded their own pronunciations of each ROILA word. The participants had a choice of listening to the sample recording again. This was done to ensure that the native language of participants would not affect their ROILA articulations. We instructed participants to follow the sample recordings as much as possible. Naturally no sample voice was played out in English.

Experiment Design and Measurements

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 number of errors in recognition by Sphinx. Words from both English and ROILA were randomly presented to counter any training or tiring effects. The order of recording English or ROILA first was also controlled between participants.

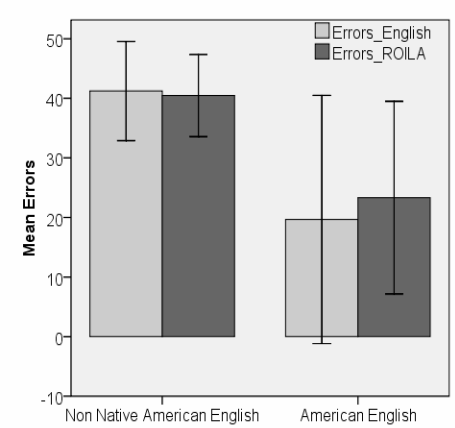

Figure 2. Bar chart comparing average recognition errors of Native Americans and Non Native Americans, significant difference even with only 3 Native American English participants

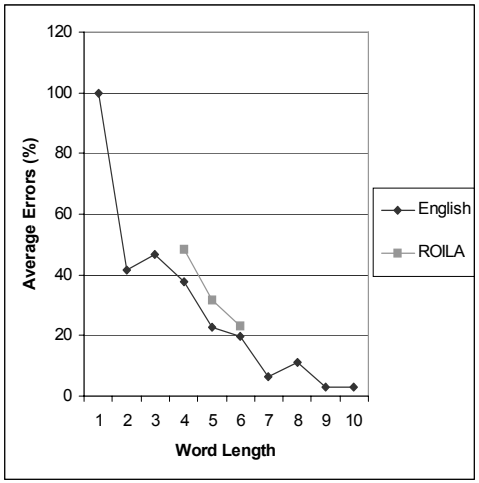

Figure 3. Graph illustrating the relation between word length and recognition accuracy

Results

We carried out a two factor repeated measures ANOVA with recording order (ROILA or English first), gender and whether participants were Native American as the three between subject factors. The REMANOVAs revealed that language type did not have a main effect $F(1,9) = 0.758$, $p = 0.41$ and there were no interaction effects either. Both ROILA and English performed equally in terms of accuracy (ROILA recognition accuracy=67.61% and English = 67.66%). Without any training data, such accuracy is expected from Sphinx on test data [11]. From the between subject factors, the factor of whether participants spoke Native American English had a significant effect $F(1,9) = 6.25$, $p = 0.034$ as they achieved higher recognition accuracy for both ROILA and English (see Figure 2). Recording order was not significant $F(1,9) = 0.019$, $p = 0.893$ and neither was Gender $F(1,9) = 1.07$, $p = 0.327$. We carried out a second REMANOVA with word length of ROILA words as the independent variable. It had 3 levels (4, 5 or 6 characters). The dependent variable was the recognition accuracy within each category of every participant, as each category had a different total number of words. The ANOVA analysis revealed that Word Length had a significant effect on the number of recognition errors $F(2,18) = 20.97$, $p < 0.0001$. Pair-wise comparisons (Bonferroni) between all three categories were significant ($p<0.05$). The average accuracy for 4, 5 and 6 character words was 52.6%, 69.33% and 77.7% respectively. Therefore longer words performed better in recognition, as is evident in the graph (see Figure 3). In order to understand if word structure of ROILA words 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 (CV or non-CV type, the former having three word types and the latter five-see Table 1). Our vocabulary had in total 73 non-CV type words and 42 CV type words. The dependent variable was the number of participants who got that type of word wrong. 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.

ROILA vocabulary – second iteration

In this iteration the vocabulary was set to include only the three CV word types. The genetic algorithm was run again with the same parameters G=P=200. The new vocabulary had an average word length of 5.1 characters (see Table 2). We asked 9 (3 female) from the earlier 16 participants to carry out recordings of the new vocabulary using the same setup and procedure. We did not have them record the English words again. The same Native American speaker as in the first setup was used as the sample voice, where the sample recordings of the new ROILA vocabulary had 100% recognition accuracy in Sphinx. The recordings from the 9 participants were run in Sphinx to evaluate the recognition accuracy of the new vocabulary. A repeated measures ANOVA revealed that the new ROILA vocabulary nearly significantly outperformed English $F(1,8) = 4.75$, $p = 0.06$. The average accuracy for the nine participants was English: 65.11%, ROILA 1: 66.22% and ROILA 2: 71.11%.

Conclusion and Future Work

Our results revealed some interesting insights. Firstly, we were able to achieve improved speech recognition

Word Type	Examples
CVCV	bama, pito
CVCVC	fenob, topik
CVCVCV	simoti, banafu

Table 2. ROILA-2 Word Examples

accuracy as compared to English even for a larger vocabulary. Similar endeavors have only been carried out for a vocabulary size of 10 [12]. Secondly, we confirmed the result that longer words perform better in speech recognition [13]. Thirdly, we quantitatively illustrated that CV type words perform better in recognition. This has only been discussed [12] but was not quantitatively proven. Lastly, we showed that Native American English speakers significantly outperformed other speakers, probably due to the acoustic model of Sphinx, which is trained using Native American speakers. 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. Secondly, the acoustic model of Sphinx was primarily designed for English, yet our ROILA accuracy results in the first iteration were just as good and in the second significantly better. As the next immediate step we aim to add grammar to our vocabulary and evaluate its effect on improving recognition accuracy. Once the grammar is in place, we aim to train participants in ROILA and evaluate the language 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. Exploiting peculiarities of the application domain would be a further boost for whatever recognition gain we achieve. We aim to explore applications for children, medical tasks, machine maintenance tasks or care robots.

Acknowledgements

We would like to thank Vanessa Vakili for assisting in the sample ROILA recordings. We would also like to thank Andy Lovitt for providing the confusion matrix.

References

[1] Rosenfeld, R., Olsen, D. and Rudnicky, A. Universal speech interfaces. *Interactions 8*, 6 (2001), 34-44.
[2] Tomko, S. and Rosenfeld, R. Speech Graffiti vs. Natural Language: Assessing the User Experience. In *Proc. HLT/NAACL*. (2004).
[3] Mubin, O., Bartneck, C. and Feijs, L. Designing an Artificial Robotic Interaction Language. In *Proc. INTERACT 2009: Part II*. Springer (2009), 851-854.
[4] UPSID Info. http://web.phonetik.uni-frankfurt.de/upsid_info.html.
[5] Toki Pona - the language of good. http://www.tokipona.org/.
[6] Selection - Genetic Algorithm. http://wikipedia.org/wiki/Selection_(genetic_algorithm).
[7] Lovitt, A., Pinto, J. and Hermansky, H. On Confusions in a Phoneme Recognizer. *IDIAP Research Report, IDIAP-RR-07-10*, (2007).
[8] Amir, A., Efrat, A. and Srinivasan, S. Advances in phonetic word spotting. ACM NY, USA (2001), 580-582.
[9] Levenshtein Distance. http://wikipedia.org/wiki/Levenshtein_distance.
[10] Sphinx-4. http://cmusphinx.sourceforge.net/sphinx4/.
[11] Samudravijaya, K. and Barot, M. A Comparison of Public-Domain Software Tools for Speech Recognition. In *Proc. Workshop on Spoken Language Processing*. ISCA (2003), 125-131.
[12] Arsoy, E. and Arslan, L. A Universal Human Machine Speech Interaction Language for Robust Speech Recognition Applications. In *Proc. Intl Conference on Text, Speech and Dialogue*. (2004), 261–267.
[13] Hämäläinen, A., Boves, L. and De Veth, J. Syllable-length acoustic units in large-vocabulary continuous speech recognition. In *Proc. SPECOM 2005*. (2005), 499-502.