

A Universal Human Machine Speech Interaction Language for Robust Speech Recognition Applications

Ebru Arisoy, Levent M. Arslan

Boğaziçi University, Electrical and Electronics Engineering Department,34342,
Bebek, Istanbul, Turkey

Abstract. Automatic speech recognition systems are prone to errors when there are confusable words in the dictionary. In this paper, a new approach to the solution of this problem is proposed. The idea is to create a human machine speech interaction language (HUMSIL) with acoustically orthogonal words. In order to minimize pronunciation variations among different nationalities, we selected a common subset of phonemes across world's major languages and generated a vocabulary set using the algorithm described in this paper. We performed two experiments to compare English, Turkish and HUMSIL in terms of digit recognition performance using microphone recordings from multi-national speakers. We found that in both of the experiments, the proposed vocabulary resulted in a significantly smaller error rate.

1 Introduction

With recent advances in technology, speech recognition is becoming more widely used in our daily lives. For example, instead of pressing the digits you can now speak the digits when entering your credit card number, or you can access voice-driven information in your automobile [1]. The ability of a recognition system to perform accurately under noisy conditions is critical to the success of a voice-enabled application. The confusable words that are close to each other in acoustic space are misunderstood by recognition systems. Even human beings sometimes make perceptual mistakes when they have to decide between words like "fix" and "six". Moreover, the system performance degrades for non-native speakers [2]. A solution to this problem may be to create a human machine speech interaction language having acoustically orthogonal words in its vocabulary. Using these new words, the performance of the recognition system will not be degraded under noisy conditions significantly. In order to make the system robust to speaker variations, the words of the new vocabulary should be easily pronounced by the majority of the people in the world. Therefore, the phonetic alphabet of the new language will include the common phonemes among the major languages spoken and the orthogonal words will be constructed from these phonemes.

Our primary objective in designing HUMSIL is to formulate an international and effective way of communication between "humans and machines" in speech

recognition applications. In that respect, it differs from the proposed international language Esperanto, which was primarily designed for international or interethnic "human to human" communication [3]. Considering human-machine interaction, the success of single stroke alphabet Graffiti, used for character entry in character recognition systems, becomes a standpoint to our study. Although the symbols in Graffiti are very similar to Roman letters, there are five symbols that do not match either uppercase or lower case letters (A,F,K,Q,T) and users must learn these new strokes in order to become proficient with Graffiti. Whereas a continuous stream of users consider Graffiti as an integral part of their daily interaction with PDA's when they feel easier to hand-write rather than type on a small size keyboard [4].

In this paper we limit the vocabulary of the new language to include only a few very essential words (i.e., digits). In such a case, the users may have the option of learning ten words in HUMSIL and get better service in return.

2 The Design of the New Language

Throughout the ages, philosophers and linguists have been discussing whether there are universal properties that hold for all human languages and are unique to them. The grammar is everything speakers know about their language, the sound system, called phonology; the system of meanings, called semantics; the rules of word formation, called morphology; and the rules of sentence formation, called syntax [5]. Therefore, in HUMSIL, the common phonemes of the world languages will construct the phonology; the proposed vocabulary will have the meaning of digits; the acoustical orthogonality will be the rule of word formation. We will not deal with sentence formation in this paper.

2.1 Fonetetic Alphabet

In this section, we will investigate the phoneme classes, search for the special cases and orthogonality principals and decide on the final version of the new alphabet. In the remainder of this paper, we will use single-letter ARPAbet Symbols to denote phonemes [6]. The study of the International Phonetic Association [7] is a very useful guide to search for the common phonemes among the languages. There are 29 natural languages examined extensively in this work and they found out that the most common phonemes (included in at least 70% of these languages) in descending order are the following: /m/, /n/, /k/, /t/, /l/, /b/, /d/, /p/, /s/, /g/, /f/, /y/, /z/ as consonants, /i/, /u/, /a/, /o/, /e/ as vowels. These common phonemes guide us for the decision process of the minimal alphabet.

If we investigate vowels, in a controlled sample of 317 languages, the vowels /i/, /u/, and /a/ all appeared in the phonemic contents of over 250 languages [8]. The vowels /a/, /e/, /i/, /o/, and /u/ are the common ones in most of the popular natural languages. When the frequency characteristics of the vowels are considered, the vowels /a/, /i/, and /u/ take place at the three corners of the

”vowel triangle”. Therefore, we conclude that these three vowels are distant from each other in acoustic space [9]. Also, in a vowel recognition experiment, it is seen that these three vowels have the least error rates in the confusion matrix [10]. However, the pronunciation of the phoneme /u/ may have variations in different languages and even in different words. Depending on these facts, we select the phonemes /a/, /i/ and /o/ as the vowels of our minimal alphabet.

In a perceptual study [11], it is found that the phoneme groups /ptk/ and /bdg/ have a very high rate of within confusions. Therefore, taking one or two phonemes from each group may result in a better recognition performance. In addition to that, about 83% of all languages have some kind of /s/ sound. Next most frequent is the voiced counter part of /s/, namely /z/ [8]. In another perceptual study, it is also found that the voiceless forms of the cognate pairs are heard more successfully than the voiced form (/s/ > /z/ and /f/ > /v/) [12]. Also, the bilabial nasal /m/ appeared in almost 300 languages [13]. The presence of /m/ in a language implies the presence of /n/ in 99.3% [8]. However, the confusion rate between /m/ and /n/ is the highest among other consonant pairs [14]. In light of these facts, the final version of our minimal alphabet will include the phonemes: /a/, /i/, /o/, /b/, /t/, /k/, /s/, /f/ and /n/.

2.2 Syllable structure

Most of the linguists suggest that there is a universal preference for open syllables of the CV (consonant-vowel) type [13]. Therefore, for the sake of universality, the syllables of the proposed words will obey the CV rule. The only exception will be the words ending with the nasal consonant /n/. Now that we have discussed the alphabet and the syllable structure, we can elaborate on the choice of words in HUMSIL.

3 Choice of the words for the new digit vocabulary

Acoustic orthogonality is the main consideration while designing the vocabulary of HUMSIL. In addition to acoustic orthogonality, the factors affecting human learning of new words in a second language such as number of syllables within a word and familiarity of the word to the speaker are other important criteria. Although one-syllable words are easier to learn, they result in higher error rates. Also using balanced number of one-syllable, two-syllable, three-syllable words in the vocabulary will improve speech recognition performance [15]. Therefore, using equal number of two and three syllable words seemed to be a compromising solution in our new language. For ease of learning HUMSIL, familiar words seem as an advantage, however that will cause some pronunciation variations because the speakers will try to pronounce the word as it is pronounced in their own language. Since multi-nationality is a more important criterion in this study, we prefer to use unfamiliar words.

In speech recognition, both acoustic and linguistic knowledge is used to decode a given utterance. Therefore, as well as acoustic orthogonality, the order

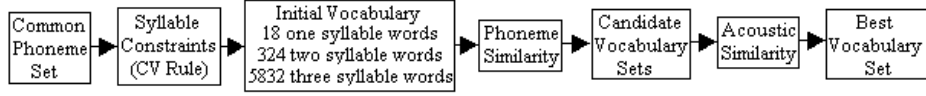


Fig. 1. Flowchart of the vocabulary design process of the HUMSIL.

of phonemes is also important. Now we will investigate these issues separately using the flowchart given in Fig. 1.

3.1 Phoneme String Distance

The phoneme string distance is some metric of how alike two strings are to each other [16]. For finding the phoneme string distance between two words we use the minimum edit distance algorithm, in which insertion and deletion operations have a cost of 1, and substitution has a cost of 2.

Operation List		intention
	delete i =>	ntention
	substitute n by e =>	etention
	substitute t by x =>	exention
	insert u =>	exenuhion
	substitute n by c =>	execution

Fig. 2. Operation list between the sequences "intention" and "execution", the phoneme string distance between these sequences is computed as 8 [16].

3.2 Acoustic Distance

Acoustic dissimilarity is related with the frequency characteristics of the phonemes. For finding the acoustic distances between the phonemes of our new alphabet, we used the Mahalanobis distance between the mel-cepstrum coefficients of each phoneme.

$$d_{ij} = \sum_{k=1}^p \frac{1}{\sigma_k^2} (MFCC_i(k) - MFCC_j(k))^2 \quad (1)$$

Here, d_{ij} stands for the acoustic distance between the i 'th and the j 'th phoneme. Then, for every substitution operation, the acoustic distance between the actual phoneme and the substituted phoneme is calculated using (1) and then they are summed to find the total acoustic distance between word pairs.

3.3 Development of The Algorithm

The digit vocabulary of the HUMSIL is selected from huge number of words. Using the phonemes of the new alphabet and the proposed syllable structure,

we generate all possible words having one, two and three syllables. Now we have to decide on which words will be selected for our new vocabulary. During the selection process, we use the string distance to determine the level of similarity between two words and acoustic distance to select the most orthogonal word pairs. The aim of our algorithm is to select the word pairs having larger string distances and at the same time to select the word pairs that are as distant as possible from each other in the acoustic space.

In our word selection algorithm, the first word of our new vocabulary is selected randomly from the generated two-syllable words. The second word is selected such that it has the highest string distance from the first word. The words from the third to the tenth are selected in a way that the minimum of the string distances between the new selected word and the previously selected words will be the highest. The selection algorithm for the fourth word is shown in Fig. 3.

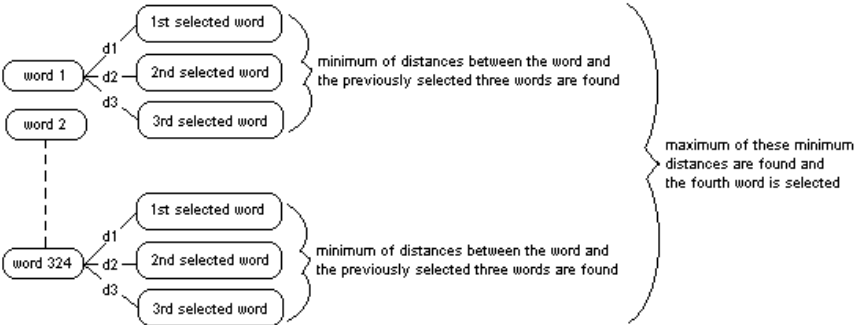


Fig. 3. Explanation of the algorithm for the selection process of the fourth word.

In Figure 3, d_1 , d_2 and d_3 stand for the phoneme string distances between one of the words in the generated word list and the firstly, the secondly and the thirdly selected words respectively. For the words from 1 to 324 the minimums of d_1 , d_2 and d_3 are found. Then the word with the maximum of these minimum distances is selected as the fourth word.

We select all possible vocabulary sets by using the algorithm described above. Then we consider the acoustic similarity criterion. For all the vocabulary sets, we add the effect of acoustic distance to the phoneme string distances, and then calculate the minimum of these total distances. The vocabulary set having the maximum of these minimum total distances is selected as our proposed vocabulary. Our proposed words with their counterpart digits are listed in Table 1.

Table 1. The proposed vocabulary

Proposed Digit Set in HUMSIL										
Digit	0	1	2	3	4	5	6	7	8	9
HUMSIL	/biko/	/nana/	/fofi/	/siti/	/toso/	/babisi/	/titaba/	/kobati/	/satabo/	/fibata/

4 Evaluations

We performed two recognition experiments using microphone recordings. Telephone speech database of GVZ Speech Technologies is used to train the Hidden Markov models for recognition. The training data does not contain the recordings of the new vocabulary. It only contains of Turkish utterances spoken by Turkish native speakers. For the first experiment, test recordings of English, Turkish, and HUMSIL digits are taken from 30 Turkish people, 15 females and 15 males, whose second language is English. Recordings are taken in a noisy office environment. A low quality microphone and a low sampling rate (8 kHz) was used in the recordings in order to simulate a difficult speech recognition scenario.

For English digits, the obtained error rate is 25.6%. 77 utterances out of 300 are misrecognized by the system. The most confused word pairs are evaluated as "six-eight" and "seven-zero". The results are not as we expected because these most confused word pairs are not the most correlated ones in English digit vocabulary. One reason of that unexpected result may be the training data used in HMM's for recognition experiments, because only Turkish utterances were used in these training data sets.

For Turkish digits, 14 utterances out of 300 are confused by the system. The obtained error rate is 4.6%. Most confusable words were "iki" and "sekiz". In Turkish recordings, there is a high increase in recognition performance. One reason of this may be the pronunciation variations in their second language.

In the recognition test of the HUMSIL, we used the recordings of the words given in Table 1. The number of misrecognized words for the new-digit vocabulary is 4 out of 300, so there is an error rate of 1.3%. If we compare these results with English digits (25.6% of error), we obtain an error rate reduction of 94.9%. There is also an error rate reduction of 71.7% for Turkish digits (See Table 2).

For testing the effect of pronunciation variations on the recognition performance of our proposed vocabulary, we performed a new experiment with the English digits and new-vocabulary recordings of a multinational group of subjects. We take test recordings from 30 speakers (15 females and 15 males). 10 of them can barely speak Turkish, and 10 of them were native English speakers. For this second experiment, error rate obtained from English digits is 37.0%. The error rate decreases to 4.0% for the new vocabulary set (See Table 2). There is an error rate reduction of 89.1% for HUMSIL digit set. Compared to the first experiment, pronunciation variation and training data set of the HMM's increase

Table 2. Error rates for the experiments

	Error Rate (%)		
	Turkish	English	HUMSIL
Experiment 1	4.6%	25.6%	1.3%
Experiment 2	-	37%	4.0%

the error rate in the recognition of HUMSIL. However, there is still a considerable increase in the recognition performance in HUMSIL rather than in English digits spoken by multinational subjects. In a further study, if the recordings of the new speech interaction language taken from a multinational group of subjects are used to train the Hidden Markov models, our new language may be even more robust to pronunciation variations.

5 Conclusion

In this paper we proposed a solution for the confusable word pair problem in speech recognition applications. Our solution is to create a new human-machine speech interaction language (HUMSIL) with orthogonal word pairs. To minimize pronunciation variations among nationalities, common phonemes from most of the world languages are selected. New words are constructed from these phonemes using the proposed syllable structure and the new digit vocabulary is chosen using the algorithm described in this paper. We proposed ten acoustically orthogonal words instead of the digit-set from zero to nine. Digits are chosen since digit recognition is a common task in many speech recognition applications.

We perform a recognition experiment with Turkish speakers in their mother tongue, second language and the new language. In HUMSIL, we observed an error rate reduction of 71.7% compared to Turkish and 94.9% compared to English. Also we performed the same experiment with people from different nationalities. The error rate reduction in these recordings is 89.1% compared to English. With the second experiment, we demonstrated the robustness of the new language against pronunciation variations.

The main disadvantage of our proposed idea is that people have to learn new words. We believe that when the vocabulary size is small, some part of the population might be convinced to learn these words for faster and better services. At first glance, this attempt to modify a language is not viewed as a welcome effort by many people. However, acoustically similar words in existing languages will always degrade performance of SR engines under noisy conditions and for speakers with heavy accents, therefore we think that the proposed idea provides a good alternative to the solution of this problem.

Acknowledgements

We would like to thank to Tuba Islam for her contribution to this research.

References

1. Hemphill, C.T., Agarwal, R., Muthusamy, Y.K., and Gong, Y.: Voice-Driven Information Access in the Automobile. IEEE Vehicular Technology Society News, August, 8-11 (2000)
2. Arslan, L.M., and Hansen, J.H.L.: Likelihood Decision Boundary Estimation between HMM Pairs in Speech Recognition. IEEE Trans. On Acoust. Speech, and Signal Processing,6(4) (1998) 410- 414
3. Schubert, K(ed.): Interlinguistics Aspects of the Science of Planned Languages, Trends in Linguistics, Studies and Monographs 42.(Mouton de Gruyter, Berlin and New York) (1989) 10
4. Mackenzie, I. S. and Zang, S.: The immediate usability of Graffiti. Proc. of Graphics Interface'97. (1997) 129-137
5. Fromkin, V. and Rodman, R.: An Introduction to Language. Holt, Rinehart and Winston, Inc.,Orlando. (1998)
6. Deller, J.R., Proakis, J.G. and Hansen J.H.L.: Discrete-Time Processing of Speech Signals. Macmillan Publishing Company. (1993)
7. IPA, Handbook of the International Phonetic Association, Cambridge University Press, (1999)
8. Maddieson, I.: Patterns of Sounds, Cambridge University Press. (1984)
9. Rabiner, L. R. and Schafer, W.: Digital Processing of Speech Signals, Prentice Hall, (1978)
10. Forgie, J. W. and Forgie, C. D.: Results Obtained from a Vowel Recognition Computer Program. The Journal of the Acoustical Society of America, 31(11). (1959) 1480-1489
11. Miller, G. A. and Nicely, P. E.: An Analysis of Perceptual Confusions Among Some English Consonants. The Journal of the Acoustical Society of America, 27(2), (1955) 338-352
12. House, A. S.; Williams, C. E.; Hecker, M. H. L. and Kryter, K. D.: Articulation-Testing Methods: Consonantal Differentiation with a Closed-Response Set, The Journal of the Acoustical Society of America, 37(1), (1965)
13. Odlin, T.: Cross-linguistic Influence in Language Learning, Cambridge University Press, (1989).
14. Roe, D. B. and Riley, M. D.: Prediction of Word Confusabilities for Speech Recognition, ICSLP, Yokohama, (1994), 227-230.
15. Arslan, L. M.: A New Universal Language for Speech Recognition Applications, IEEE Proc. ICASSP, Istanbul Turkey, (2000)
16. Jurafsky, D. and Martin J. H.: Speech and Language Processing, Prentice Hall, (2000)