A Hybrid Statistical Model to Generate Pronunciation Variants of Words

Bahram Vazirnezhad[1], Farshad Almasganj[2], Mahmood Bijankhan[3]
[1, 2] Biomedical Engineering Faculty, Amirkabir University of Technology, Tehran
[1, 2, 3] Research Center of Intelligent Signal Processing, Tehran
[3] Letters and Humanities Faculty, Tehran University, Tehran
Fax: +9821-66495655, Email: {bvazirnezhad, almas}@aut.ac.ir, mbjkhan@ut.ac.ir

***Abstract*-Generating pronunciation variants of words is an important applicable subject in speech researches and is used extensively in automatic speech segmentation and recognition systems. In this way, Decision trees are extremely used to model pronunciation variants of words and sub-word unites. In the case of word unites and very large vocabulary, to train necessary decision trees we will need a huge amount of speech utterances which contains all of the needed words with a sufficient number of each one. This approach besides demanding very large data, for new words will need some new extra corpus. To solve these problems we have used generalized decision trees, that each tree is trained for a group of words with similar phonemic structure instead of a single word. These trees can predict regions of the words in which substitution, deletion and insertion of phonemes would occur. Next to this step, appropriate statistical contextual rules, which are extracted from a large speech corpus, will be applied to these regions in order to generate words variants. This new hybrid d-tree/c-rule approach takes into account word phonological structures, stress, and phone context information simultaneously and an ordinary size speech corpus will be sufficient to train its models. By using the word variants obtained by this method in the lexicon of "SHENAVA", a Persian ACSR, a relative WER% reduction of as high as 6% was obtained.**

I. INTRODUCTION

Pronunciation Variation is a well known phenomenon which is a result of co-articulation, assimilation, reduction, deletion and insertion of phones. The degree to which these phenomena occur will vary depending on factors such as rate of speech, speaking styles, speaker specifications and other factors and mechanisms. Some of these mechanisms are categorized as interspeaker variations while some of them are intraspeaker variations. Pronouncing words in different ways makes speech recognition a difficult task [1]. So pronunciation models are necessary in order to overcome this difficulty. Almost any of the current ASR systems needs a lexicon containing pronunciation variants of words to describe how the entries can be pronounced, or more precisely, how they can be realized as a sequence of phones. This method is referred in the literature as explicit pronunciation modeling. In the last few years, several researchers have put efforts in introduction of pronunciation models comprising pronunciation variants [2, 3, 4, 5, 6]. Experiments have shown that introducing appropriate pronunciation variants would improve performance of ASR systems. It is important to choose the source from which information on pronunciation variation will be retrieved. In this regard a distinction can be drawn between data-driven vs. knowledge-based methods. In data-driven methods, the formalizations are derived from the data. In general this is done in the following manner. The phonetic transcription of an utterance is aligned with its corresponding phonemic transcription obtained by concatenating the transcriptions of individual words. Alignment is done by means of a dynamic programming algorithm. The resulting DP-alignments can then be used to derive rewrite rules using statistical approaches, decision trees, artificial neural networks, etc. Here, we will have a review on data-driven approaches considering that our proposed new hybrid d-tree/c-rule technique is in the same category.

Some researchers extracted some contextual rules to model word phonemes variations using data-driven approaches, in order to generate pronunciation variants of the words from their phonemic transcriptions [2, 6]. Mostly, application likelihoods are assigned for such rules to estimate variant's probability conditioned on occurrence of the word. This is an applicable approach, besides a great weakness. Contextual rules take no notice of word level information, because they have only limited contextual condition, for being applied. Decision trees are another way to represent information on pronunciation variation achieved from data. Various types of features, such as phoneme context, speaking rate, speaker specifications, etc. have been used to train these models [7]. It is shown in [3] that the mapping of canonical phones to surface phones has a dynamic nature. Dynamic pronunciation models based on decision trees has been also designed [8]. It is shown that auxiliary factors of words like stress, syllabification, syntactic role and prosody parameters may affect pronunciation variants. Artificial neural networks are also used to model pronunciation variation. Phoneme context is again used to predict pronunciation variants of word segments [4]. It is shown that pitch accent can improve the prediction of pronunciation variation [5]. In addition of the mentioned approaches there are other approaches, like [9] in which finite-state transducer is used as a representation for pronunciation variation, developed to face this problem. This is important that the majority of these works focused on finding phonetic deviations of the phonemic segments of the words as the main way to find the whole word variants.

In this paper, we introduce a method for automatic generating of pronunciation variants of words which takes into account whole word information such as word's phonological structure and stress besides its phones context information simultaneously. We have designed a hybrid

0-7803-9361-9/05/$20.00 ©2005 IEEE

statistical model. This model is composed of decision trees and contextual rules. First, Trees predict regions in the word which are susceptible of change, by asking some questions about phoneme categories and position of stressed syllable in the input word. Consequently, appropriate contextual rules are applied to permissible regions and not other regions, to generate the pronunciation variants of the input word. Decision trees which we used are similar to those trees which are used as triphone models in [10]. We should emphasize that in our method each decision tree is not trained for a word as it is done in [3, 8], but for group of words with similar phonological structure; so we have chosen the term of generalized decision trees for them. By using such generalized decision trees we will overcome the practical difficulty of insufficient speech data for each word. It should be considered that contextual rules take no notice of word level information, because they have only limited context phones as condition to being applied. We have solved this limitation by combining models of generalized decision trees with contextual rules. In other words some word level constraints are introduced by generalized decision trees, and contextual rules can be applied only on permitted regions. Results have shown that hybrid d-tree/c-rule model can generate pronunciation variants that are closer to real pronunciation variants of words in comparison with variants generated by using only contextual rules.

This paper is organized as follows. The materials and methods, procedure of training generalized decision trees and contextual rules are detailed in section II. Two ways of performance measurement of the proposed approach and the corresponding results are described in section III. And conclusions are considered in section IV.

II. MATERIALS AND METHODS

A. *How models can generate pronunciation variants*

In this part the procedure of generating pronunciation variants will be explained as it is shown in Fig.1. Hybrid statistical model generates pronunciation variants of words in two steps. First, generalized decision tree corresponding to the phonemic structure of the input word, predicts which phonemes in the word can be substituted, deleted or where an insertion can take place. Choosing the tree corresponding to the input word is based on phonological structure or arrangements of consonants (represented by C) and vowels (represented by V) of the word. An example is shown in the Fig.1. to clarify the procedure. The input word is a Persian word with Phonemic transcription of /ket/b/, it means "book" in English. It is a disyllabic word and its consonants and vowels are arranged as "CVCVC" pattern. So the corresponding generalized decision tree is "CVCVC" tree, which is trained for words with similar structure to predict how consonants and vowels can be substituted, deleted or inserted. Generalized decision tree asks for specifications of each consonant and vowel and location of stressed syllable that is attached to word as a label, to predict variation patterns.

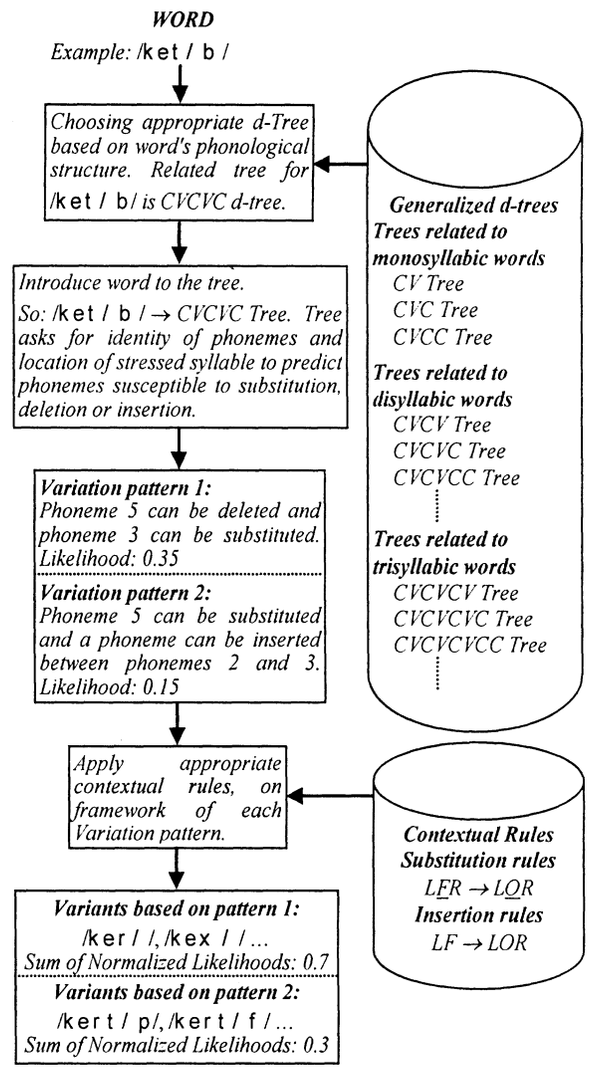

Fig.1. The procedure of generating pronunciation variants.

Specifications of consonants and vowels are defined by their membership to different categories. And categorization of phonemes is based on their linguistic similarity or phone confusion matrix. Variation patterns means here the ways that pronunciation variations can take place. Each variation pattern defines which phonemes can be substituted, deleted or where an insertion can be occurred. Each variation pattern will have a probability, which is determined by the tree. We have introduced a cut-off threshold of probability to limit accepting variation patterns. In the second step we would apply contextual rules to phonemes that are candidate to be altered. We will define such context dependent rewrite rules in section D. Substitution rules will be applied to phonemes that are chosen by tree for substitution. And insertion rules will be used to apply to regions that insertion is permitted. So the variants will be generated after applying contextual rules. The likelihoods of variation patterns will be normalized to have a sum of one. Variants which have produced in framework of same variation pattern have same likelihood which is equal to normalized likelihood divided by number of variants.

Consequently, such variants of words were introduced to the lexicon of the continuous Persian speech recognizer to improve recognition accuracy.

B. *Database and transcriptions used in training models*

The database, used in this research was "large FARSDAT". This is a Persian speech database, created by research center of intelligent signal processing (RCISP). It includes 100 speakers selected with regard to age, gender, educational level and belonging to one of ten frequent dialects of Persian in Iran. These dialects are Tehrani, Torki, Isfahani, Jonubi, Shomali, Khorasani, Baluchi, Kordi, Lori, and Yazdi. Each speaker uttered about 4000 words of various texts of newspapers in an office room. The material covers a variety of fields such as politics, economics, culture, sports and .... Phonemic labeling of the sentences is produced by IPA characters with similar format to FARSDAT [11]. In order to train the pronunciation models; first, we have used the phone recognizer of SHENAVA ACSR system to decode speech utterances as realized phone strings. The phone recognizer of SHENAVA has a highly acceptable performance with minimal errors, as it is reported in [12]. In the next step realized phone transcriptions (transcribed automatically by phone recognizer) are automatically aligned with the baseline phonemic transcriptions. By using realized phone transcriptions instead of phonetic transcriptions we can model both variations due to pronunciation and phone recognizer errors simultaneously. Although different factors may affect each of these variation sources, but we believe that our statistical approach models both of these variations sources. Automatic alignment is done by a dynamic programming algorithm which minimizes an alignment distance. Substitution cost is dependent to the distinctive features that differ between baseline phoneme and realized phone. An example of the alignment is given below. For example, the alignment is shown for phonemic and realized transcriptions of a Persian word with Phonemic transcription /ket/bx/neh/, it means "library" in English.

/k e t/ b x/ n e h/
/p e t/ # f/ n e #/

The first row in this example is the phonemic transcription and the second row the realized phonetic transcription. Notice to unvoiced plosive substitution (k→p), deletion of voiced bilabial plosive (b→#), consonant substitution (x→r) and unvoiced glottal fricative deletion (h→#). Symbol # is used to represent deletions and insertions.

C. *Generalized decision trees and input features*

A generalized decision tree is trained for each group of words. We categorized words based on their number and phonemic structure of syllables. Persian has 3 syllable structures. These are "CV", "CVC" and "CVCC". "C" represents consonants and "V" is a representation for vowels. Thus in Persian there is 3 kinds of monosyllabic words, 9 kinds of disyllabic words and 27 kinds of trisyllabic words. The term generalized tree which we have used, gets its name from the basic idea of our approach, as we didn't train, a tree for each word as it is done in [3, 8], but for a group of words with similar phonological structure. For example, the word /ket/b/ and /med/d/ (means pencil in English) are used to train the same "CVCVC" tree. The input features that are used to train trees, are based on membership of phonemes to various categories and the place of stressed syllable in the input word. Categorization of consonants was based on their distinctive specifications and their linguistic differences and similarities. Table I show, 7 categories of consonants and their linguistic descriptions. Vowels were categorized by taking notice of phone confusion matrix based on vowel mispronunciation or misrecognition. Categories of vowels are chosen as {/, a, o}, {e, i}, {u}.

The training algorithm tries to gain lowest sum of entropy in all nodes. The large variety of variation patterns and limited occurrence of each variation pattern is a challenge in training generalized decision trees. So we used a vector quantization (VQ) technique using radial basis functions (RBF) to quantize code vector related to variation pattern of each datum to reduce the large variety of variation patterns. Quantization process finds the nearest RBF center for each variation pattern code vector by exhaustively computing its distance to each of the RBF center code vectors. Then, the index of the nearest neighbor code vector is used to encode the variation pattern. The proposed algorithm reduces the number of various variation patterns' indexes attached to each datum (phonemic and realized transcription). Code vectors are in dimension of 4N+1 (where N is the phonemic length of words) and are composed of zeros and ones which demonstrates phonemes matching, substitutions or deletions and insertions. Each phoneme has 3 corresponding cells in the vector. The first cell is related to match (pronounced as its phonemic format), second cell is related to substitution and third cell is related to deletion. Insertions of phonemes are defined by setting to one, cells between the above mentioned 3 cells or at the beginning or at the end of the word. Assume a word with structure "CVCVC" is realized as "C'VCV#" (means that the first consonant is substituted and the last consonant is deleted). The corresponding variation pattern code vector for this utterance is the following vector which is in dimension of $4\times5+1$.

0 0 1 0 0 1 0 0 0 1 0 0 0 1 0 0 0 1 0 0 0 1 0 0 0 1 0
i m s d i m s d i m s d i m s d i m s d i

"i" represents insertion cells, "m" illustrates match cells,

TABLE I
CATEGORIES OF CONSONANTS AND LINGUISTIC DESCRIPTIONS

IPA Symbol	Linguistic Description
b, d, q, ., g	Voiced, Plosive
p, t, k, '	Unvoiced, Plosive
s, ., x, f	Unvoiced, Fricative
l, m, n, r	Sonorant
z, [, v	Voiced, Fricative
], h	Glottal Consonant
y	Palatal glide

"s" shows substitution cells and "d" is representation for deletion cells. The cells are set to 1 or 0 dependent on whether realized phone is matched, substituted, or it is deleted. The

cells related to insertions are set to 1 if an insertion is occurred in corresponding position. Fig.2. shows a part of generalized decision tree with CVCVC format, trained after vector quantization of variation patterns.

We can't train generalized decision trees for words with more than three syllables as they are so rare. So we apply only contextual rules to such multi-syllabic words in order to generate their variants. Our focus will be on a solution for this problem in the future. We state here as a linguistic justification why the word clustering idea and using a separate decision tree as a pronunciation model works. As it is said before each d-tree is trained for words with same phonemic structure; as the training procedure goes on the training algorithm asks for phone identities and location of stressed syllable of the words to split the data. So words in the terminal nodes, have similar phonemes at each place in the linguistic view. We believe that same word structures and phone similarities will cause same variation patterns to occur in each terminal node of the tree; and the results show the same thing. It is shown in the Fig.1. that first predicted variation pattern says that phoneme 5 can be deleted and phoneme 3 can be substituted. These two changes are predicted by getting relevant information about identities of adjacent phonemes to phoneme 3 and 5.

D. *Contextual rules*

We define more closely what is meant here by the term "Contextual Rule". We consider contextual rules to be context dependent rewrite rules of the form $L\underline{F}R \rightarrow O$ with L, F, R and O representing phoneme sequences. We emphasize that L, F, R and O do not represent sequences of phoneme classes, nor do they contain symbols representing such classes. The interpretation of such a rule is as follows. If in a reference pronunciation the focus F is found surrounded by a left context L and a right context R, then it may actually be pronounced as O. The combination of $L\underline{F}R$ is called the condition part of the rule. As it is the condition for applying

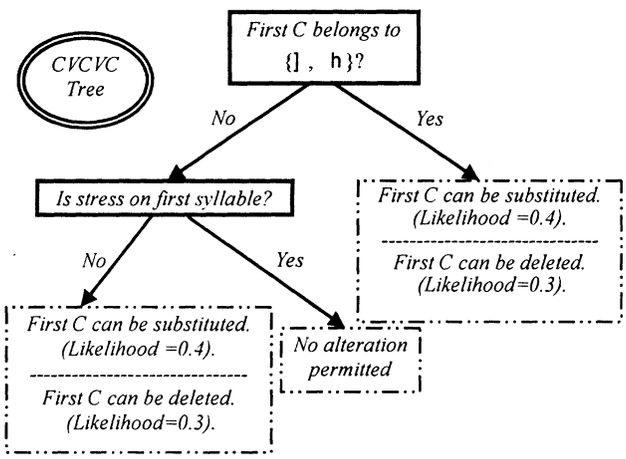

Fig.2. An example of CVCVC generalized decision tree

the rule. When we apply a rule in word boundaries a label should be attached to the variant so only compatible variants could be recognized. The procedure of learning rules and applying them to permitted regions (defined by generalized d-trees.) of words is the same as [2]. Only some implementation differences are between our work and the referred work. These rules involve only limited contextual phones as a condition for being applied. It means that they don't matter for word level information such as phonological structure of word or stress.

III. EXPERIMENTS

Two methods are used to measure the performance of our approach: First, generated variants of each word are aligned with the realized phonetic transcriptions of that word. The alignment is done by using standard DP-algorithm of NIST. Distances, between aligned pairs are normalized to the lengths of aligned sequences. Then, the average of normalized distances over all aligned pairs is computed. As the average of normalized distances comes down, it means the model variants are closer to realized ones, and the model performance is higher. We use a part of the large FARSDAT as the test database which contains about 200000 words. Hybrid model variants, variants generated by contextual rules and reference phonemic transcriptions are compared with each other in this way. The results are scheduled in table II.

Second, we put model outputs which are word phonetic variants into the lexicon of SHENAVA system which is a Persian ACSR system [12], and see if its word recognition accuracy is improved. SHENAVA is developed in RCISP research center as the output of a primary phase of a big project which aims a professional Persian ACSR system. Its vocabulary contains about 1200 words. Phoneme recognition is performed by a hybrid structure of some neural networks and rule-based engines. After the lexicon search and applying a semi viterbi algorithm to find the best 100 recognized phrases, a N-best rescoring block which benefits HMM models of Persian phones finds the best phrase. The SHENAVA version used in conducting our experiments doesn't exploit any language model to decrease its output word error rate. And the baseline absolute WER was 54%.

TABLE II

AVERAGE OF NORMALIZED DISTANCES BETWEEN ALIGNED MODEL VARIANTS WITH REALIZED VARIANTS FOR ALL WORDS IN THE LEXICON. THE DP-ALGORITHM IS THE STANDARD OF NIST.

Models used in generating variants of words	Average of normalized distances between aligned model variants with realized variants
Hybrid Statistical Model	0.25
Contextual Rules	0.27
Reference Transcription	0.34

TABLE III

WER% REDUCTION OF THE ACSR SYSTEM WHEN USING WORD VARIANTS IN ITS LEXICON AGAINST USING ONLY PURE PHONEMIC VERSION

Models used in generating variants of words in lexicon	WER% reduction relative to lexicon containing only reference forms.
Hybrid Statistical Model	6%
Contextual Rules	5.15%

In order to generate pronunciation variants of words, we used a training set of about 25 hours, to train the pronunciation models. There was no overlap between the train and test set. And on average about 3 variants for each word were used to construct the lexicon. Here the variants generated by our hybrid d-tree/c-rule model are compared with variants generated by only applying contextual rules. Improvements on word error rate in the both cases are reported relative to the word error rate of the system while lexicon contains only phonemic transcriptions. The results are scheduled in table III. Although improvements which are reported in table III are highly dependent on the base ACSR system performance and language but a comparison of results with those which are reported in other works confirms the efficiency of our approach.

IV. CONCLUSIONS

In this paper, we introduced a new hybrid method, to produce word pronunciation variants. The idea of generalized decision trees effectively handles blurred phone identities commonly found in conversational speech, instead of making a hard decision in the lexicon. Results of experiments showed that pronunciation variants which are produced by this hybrid technique are closer to realized variants than only using statistical contextual rules. Also, introducing variants produced by such hybrid model improves word recognition accuracy of the automatic continuous speech recognizer as shown in table III. Both parts of the experimental results showed that the hybrid proposed method has a better performance relative to finding out pronunciation variants of words only by using statistical contextual rules. We believe that it is due to the capacity of generalized decision trees in using word level information to model pronunciation variation based on whole structure of word. Generalized decision trees as explained earlier, are a solution for insufficient data problem when we want to consider the whole structure of words while producing its pronunciation variants. In spite of the significant results which we have obtained until now, which proves the benefit of this approach to model word pronunciation variation, we believe that this work is only a pilot study in this way, in which the feasibility of the idea of seeing whole structure of words while producing their variants is shown, but needs more research works to be completed, for example using extra features of the word structure in the modeling process or using better classification of realized pattern of words.

REFERENCES

[1] H. Strik, and C. Cucchiarini, "Modeling Pronunciation Variation for ASR: A survey of the literature," *Speech Communication,* Vol. 29, pp:225-246, 1999.

[2] N. Ceremelie, and J. P. Martens, "In search of better pronunciation models for speech recognition," *Speech Communication*, Vol. 29, pp:115-136, 1999.

[3] E. Fosler-Lussier, "Contextual word and syllable pronunciation models," *In Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding, Keystone, Colorado,* 1999.

[4] T. Fukada, T. Yoshimura, and Y. Sagisaka, "Automatic Generation of Multiple Pronunciations based on Neural Networks," *Speech Communication,* Vol. 27, pp.63-73, 1999.

[5] K. Chen, and M. Hasegawa-Johnson, "Modeling pronunciation variation using artificial neural networks for English spontaneous speech," *In Proceedings of the ICSLP, Jeju Island,* Oct. 2004, pp.1461-1464.

[6] B. Vazirnezhad, F. Almasganj, and M. Bijankhan, "Modeling pronunciation variants of words by contextual rules in Persian continuous speech recognition system," *In Proceedings of the First International Conference on Information & Knowledge Technology, Tehran,* 2003, pp.137-145.

[7] P. A. Jande, "Pronunciation Variation Modeling using Decision Tree Induction from Multiple Linguistic Parameters," *In Proceedings of the Fonetik, Stockholm,* 2004, pp.12-15.

[8] E. Fosler-Lussier, "Dynamic pronunciation models for automatic speech recognition," *PhD Dissertation*, university of California, Berkeley, 1999.

[9] T. Hazen, L. Hetherington, L. Shu, and K Livescu, "Pronunciation Modeling using a Finite-State Transducer Representation", *ISCA Tutorial and Research Workshop, Colorado,* Sep. 2002.

[10] H. Yu, and T. Schultz, "Enhanced Tree Clustering with Single Pronunciation Dictionary for Conversational Speech Recognition", *In Proceedings of the EUROSPEECH, Geneva,* 2003.

[11] M. Bijankhan, and M. J. Sheikhzadegan, "FARSDAT- the Farsi Spoken Language Database", *In Proceedings of the Fifth International Conference on Speech Sciences and Technology, Perth,* 1994, Vol. 2, pp.826-829.

[12] F. Almasganj, et al., "SHENAVA-1; A Persian Spontaneous speech recognizer", *In Proceedings of the Tenth International Conference on Electrical Engineering, Tehran,* 2001, pp.101-106.