Available online at www.sciencedirect.com

ScienceDirect

Computer Speech and Language 23 (2009) 1–24

COMPUTER SPEECH AND LANGUAGE

www.elsevier.com/locate/csl

Hybrid statistical pronunciation models designed to be trained by a medium-size corpus

Bahram Vazirnezhad [a], Farshad Almasganj [a,*], Seyed Mohammad Ahadi [b]

[a] *Biomedical Engineering Department, Amirkabir University of Technology (Tehran Polytechnic), Hafez Avenue, P.O. Box 15875-4413, Tehran, Iran*

[b] *Electrical Engineering Department, Amirkabir University of Technology (Tehran Polytechnic), Hafez Avenue, P.O. Box 15875-4413, Tehran, Iran*

Received 2 May 2007; received in revised form 3 February 2008; accepted 7 February 2008
Available online 15 February 2008

Abstract

Generating pronunciation variants of words is an important subject in speech research and is used extensively in automatic speech recognition and segmentation systems. Decision trees are well known tools in modeling pronunciation over words or sub-word units. In the case of word units and very large vocabulary, in order to train necessary decision trees, a huge amount of speech utterances are required. This training data must contain all of the needed words in the vocabulary with a sufficient number of repetitions for each one. Additionally, an extra corpus is needed for every word which is not included in the original training corpus and may be added to the vocabulary in the future. To overcome these drawbacks, we have designed generalized decision trees, which can be trained using a medium-size corpus over groups of similar words to share information on pronunciation, instead of training a separate tree for every single word. Generalized decision trees predict places in the word where substitution, deletion and insertion of phonemes may occur. After this step, appropriate statistical contextual rules are applied to the permitted places, in order to specifically determine word variants. The hybrids of generalized decision trees and contextual rules are designed in static and dynamic versions. The hybrid static pronunciation models take into account word phonological structures, unigram probabilities, stress and phone context information simultaneously, while the hybrid dynamic models consider an extra feature, speaking rate, to generate pronunciation variants of words. Using the word variants, generated by static and dynamic models, in the lexicon of the SHENAVA Persian continuous speech recognizer, relative word error rate reductions as high as 8.1% and 11.6% are obtained, respectively.

Keywords: Pronunciation models; Continuous speech recognition; Lexicon

1. 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

* Corresponding author. Tel.: +98 21 6454 2372; fax: +98 21 6649 5655.
E-mail addresses: bvazirnezhad@aut.ac.ir (B. Vazirnezhad), almas@aut.ac.ir (F. Almasganj), sma@aut.ac.ir (S.M. Ahadi).

0885-2308/$ - see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.csl.2008.02.001

on various factors such as rate of speech (ROS), speaking styles, speaker specifications and other factors and mechanisms. Some of these mechanisms are categorized as inter-speaker variations while others are intra-speaker variations. Therefore, pronouncing words in different ways makes speech recognition a difficult task (Strik and Cucchiarini, 1999).

If words were always pronounced in the same way, automatic continuous speech recognition would have been relatively easy. Meanwhile, for various reasons, words are normally pronounced in different manners. In fact, the words are strung together in continuous speech and this alters their pronunciations from their isolated form. In addition, in continuous speech, all sorts of interactions may take place between words, resulting in various phonological processes. However, for isolated speech the situation is different. The speakers have to pause between words, which results in the reduction of the degree of interaction between word pronunciations. Moreover, in this case, speakers also have the tendency to articulate more carefully. Although using isolated words makes the task of an ASR system easier, it certainly does not do the same for the speaker, since pausing between words is highly unnatural. While many current applications still make use of isolated word recognition, the emphasis is now on spontaneous speech. Nowadays, we face a range of ASR system applications from carefully read speech to ordinary conversational speech; and it is clear that pronunciation variation happens to a greater extent in natural speech in comparison to isolated-word speech. Since this phenomenon is believed to be the cause of some severe errors in ASR systems, modeling pronunciation variation is seen as an effective way to improve the performance of the current systems. During the last decade, a lot of research has been conducted on this topic.

ASR systems usually need a lexicon containing pronunciation variants of words to describe how the entries can be pronounced. In other words, how they can be recognized as a sequence of phones. In the literature, this method is referred to as explicit pronunciation modeling. In the last few years, several researchers have made efforts to develop pronunciation models comprising pronunciation variants. Experiments have shown that introducing appropriate pronunciation variants would improve performance of ASR systems (Cremelie and Martens, 1999; Fosler-Lussier, 1999a; Fukada et al., 1999). It is important to choose the source from which the information on pronunciation variation will be retrieved. In this regard, a distinction can be made between data-driven versus knowledge-based methods. In data-driven methods, the formalizations are derived directly from the data. In general, this is done in the following manner. The hand-labeled phonetic or sometimes recognized 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 model technique which is a combination of decision trees and contextual rules is in the same category.

Generating sophisticated pronunciation dictionaries are considered as an effective way to improve the system performance in large vocabulary continuous speech recognition (LVCSR) tasks since the past decade. Since manual construction of the pronunciation dictionary must be carried out by expert phoneticians and is quite expensive and time consuming, research efforts have been directed at automatic construction of pronunciation dictionaries. In the early 1990s, the emergence of honetically transcribed (hand-labeled) medium-size databases (e.g. TIMIT and Resource Management) encouraged a number of researchers to explore pronunciation modeling (Wooters and Stolcke, 1994; Riley, 1991; Randolph, 1990). Although all of these approaches were able to automatically generate pronunciation rules, hand-labeled transcriptions by expert phoneticians were still required. To solve this problem, automatic phone transcriptions generated by a phone recognizer, which enables one to cope with a large amount of training data, started to be used in pronunciation modeling (Humphries, 1997; Imai et al., 1995; Sloboda, 1995; Schmid et al., 1993). More recently, LVCSR systems have started to deal with spontaneous and conversational speech, such as the Switchboard corpus, and consequently, pronunciation modeling has become a more important issue as word pronunciation variations happen more often here in comparison to read speech (Saraclar and Khudanpur, 2004; Fosler-Lussier, 1999a).

Some researchers extracted contextual rules to model word-level phone variations using data-driven approaches, in order to generate pronunciation variants of the words from their phonemic transcriptions (Vazirnezhad et al., 2005a; Cremelie and Martens, 1999). Mostly, application likelihoods are assigned for such rules to estimate variant's probability conditioned on occurrence of the word. Contextual rules are easy to apply and since these models show variations in phone level, one can produce variants of every word, regardless of the number of its occurrences or even its existence in the training corpus. Meanwhile, this approach suffers from a major weakness; contextual rules do not take into account word level information such as phonological structure, stress, *n*-gram, etc. They only consider limited surrounding phones of the region in the word that its pronunciation supposed to be predicted. To overcome this shortcoming, researchers made use of decision trees as a way to represent information on pronunciation variation of words extracted from data, while taking into account word level information such as "phonemic structure". Different types of features, such as phoneme context, speaking rate, speaker specifications, etc. have been used to train the decision trees (Jande, 2008; Vazirnezhad et al., 2005b; Fosler-Lussier, 1999a). In Fosler-Lussier (1999a), it was shown that the mapping of canonical phones to surface phones has a dynamic nature. Dynamic pronunciation models based on decision trees have also been designed by Fosler-Lussier (1999a). These models use speaking rate and *n*-gram information to generate pronunciation variants in a dynamic framework. It was shown that auxiliary factors of words such as stress, syllabification, syntactic role and prosody parameters may affect pronunciation variants. Artificial neural networks have also been used to model pronunciation variation. Here, the phoneme context was used to predict pronunciation variants of word segments (Fukada et al., 1999). Moreover, it was shown that the pitch accent can improve the prediction of pronunciation variation (Chen and Hasegawa-Johnson, 2004). But artificial neural networks need a large amount of data to be trained as pronunciation models. Also they suffer from the difficulty of over-fitting to training data and even not converging to a global minimum. In addition to the mentioned approaches, there exist other approaches, like the approach proposed by Hazen et al. (2005), in which a finite-state transducer was used to represent pronunciation variation. It is important to note that the majority of these works have focused on finding phonetic deviations of the phonemic segments of the words as the main approach to find the word variants. It can be difficult to define these variants in a consistent way. It can also be difficult to extract generalized grapheme-to-phoneme rule sets from a lexicon containing variants. In Davel and Barnard (2006) both these issues are addressed by creating "pseudo-phonemes" associated with sets of "generation restriction rules" to model those pronunciations that are consistently realized as two or more variants.

In this paper, a hybrid statistical structure for automatic generation of pronunciation variants of words is introduced. The hybrid models are composed of decision trees and contextual rules. Decision trees predict regions in the word that are susceptible to change. Consequently, appropriate contextual rules are applied to permissible regions, and not to other regions, to generate the pronunciation variants of the input words. The proposed models are designed in static and dynamic types. Dynamic and static terms come from the fact that decision trees in the dynamic models consider rate of speech in generating pronunciation variants while the static models do not. Both dynamic and static models take into account the syllabic structure of the input word and ask questions about phone identities, unigram statistics, and the position of the stressed syllable simultaneously. In the final step, phonemic context information is considered by contextual rules. The decision trees used in this architecture are similar to those exploited to model triphone pronunciations in Yu and Schultz (2003). It should emphasized that in our proposed method, each decision tree is not assigned for just one word as done in Fosler-Lussier (1999a), but for a group of words with similar phonological structure; so we chose the term "generalized decision trees" for them to describe them better. However, while describing our approach, we have occasionally used the short terms "decision tree" or "tree" instead of "generalized decision tree" just for simplicity in several part of this paper. Using such generalized decision trees, we do not need speech data prepared distinctively to train a model for each of the investigated words. Experimental results show that both the static and dynamic hybrid decision trees/contextual rules (d-tree/c-rule) models can generate pronunciation variants closer to real pronunciation variants of words, in comparison to variants generated by using only contextual rules.

The remainder of this paper is organized as follows: an overview of the hybrid pronunciation models is provided in Section 2. The speech corpus "Large-FARSDAT" which is used for training and test of the models, and "SHENAVA" Persian ASR system which is used as our experimental workbench are introduced in Section 3. The effects of factors on phonetic pronunciation variation, which are considered in our models, are studied in Section 4. Comprehensive details and training procedure for generalized decision trees and contextual rules are detailed in Sections 5 and 6. A method to prune generated pronunciation variants is presented in

Section 7. Two assessment methods for the proposed approaches and the corresponding results are described in Section 8 and discussion and conclusions are considered in Sections 9 and 10, respectively.

2. Overall framework of the hybrid statistical pronunciation models

This section gives an overview of our approach to generate pronunciation variants. Technical details of various portions of the proposed model will be provided in later sections. The proposed pronunciation models are hybrids of generalized decision trees and statistical contextual pronunciation rules. Decision tree portions are implemented in static or dynamic forms. The static models which utilize static trees use phonological structure, word unigram statistics and the location of the stressed syllable in the word in addition to phonemic context information, to generate pronunciation variants. The dynamic models utilize dynamic trees which take into account the extra feature of speech rate which varies in time. In Section 8, the improvements obtained in the task of continuous speech recognition, due to the usage of both static and dynamic models, are compared.

Some previous works (Fosler-Lussier, 1999a) employed decision trees to find word pronunciation variants, but they used an individual decision tree for each word, in order to only capture pronunciation variation information of that word. For instance, the word “book” has a specific decision tree which is able to predict appropriate pronunciation variants of it by taking into account the specifications of surrounding words, the location of the stressed syllable, rate of speech and predictability. Our approach is different in that instead of designing and training an individual tree for each word, which imposes practical difficulties in providing enough training data for all words, a generalized decision tree is designed for all words with the same phonological structure. For further clarification, for example, the pronunciation variants of all words that have the same arrangement of consonants (represented by C) and vowels (represented by V), like the string of “CVCVC”, are used to train one decision tree. In this approach, the problem of sparse training data is solved, and the training database does not need to contain all words supposed to be in the lexicon. In addition, no new extra corpus is needed for new words. In other words, by sharing information of pronunciation variation in structurally similar words, the size of necessary training data is decreased and the problem of facing new words is solved. We chose the name of “generalized decision tree” instead of “decision tree”, to emphasize that in our approach each tree is trained for a group of words with similar phonological structure, and highlight the major differences between our approach and previous works. Comprehensive details of generalized decision trees and training algorithm are provided in Section 5.

Hybrid statistical models generate pronunciation variants of words in two main steps. First, the generalized decision tree corresponding to the syllabic structure of the input word, predicts which phonemes in the word can be substituted or 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 and vowels of the word. An example is shown in Fig. 1. The input word is a Persian word with Phonemic transcription of */ketɒb/*, which means “book”. It is a disyllabic word with a “CVCVC” phonemic pattern. Hence, the corresponding generalized decision tree is “CVCVC” tree, which is trained by pronunciation patterns of the words with same syllabic structure. The generalized decision tree asks for the specifications of consonants and vowels which are defined by their membership to different phonetic categories, the logarithm of unigram probability of the word and location of stressed syllable in the input word. The dynamic tree also uses speaking rate as a continuous feature. The influence of these factors on phonetic variation is discussed in Section 4. The outputs of the generalized decision trees are the predicted pronunciation patterns, which means the ways that pronunciation variations can take place. Each pronunciation variation pattern defines which phonemes can be substituted, deleted or where in the word an insertion may occur. Each of these variation patterns are tagged with a probability, which is determined by the tree. We set a cut-off threshold of probability to limit the number of accepted variation patterns.

In the second step, contextual rules are applied to the phonemes that are candidates to be altered. The figure shows two types of contextual rules which are *substitution* rules and *insertion* rules. In substitution rules, $L\underline{F}R \rightarrow LOR$, where phonemic string F can be recognized as string O due to pronunciation or phone recognizer errors, when it is surrounded by a left context L and a right context R. In insertion rules, $LR \rightarrow LOR$, string F is empty and string O can be inserted in a LR context. The substitution rules will be applied to the regions that are chosen by the tree for the substitution, and the insertion rules can be applied to regions that

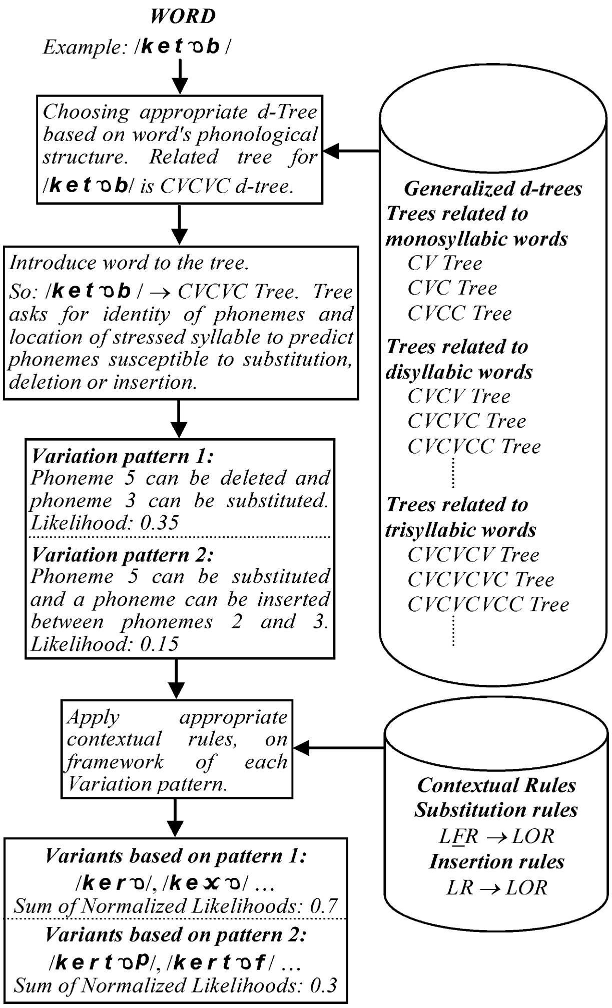
WORD
Example: /ketɒb/
Choosing appropriate d-Tree based on word's phonological structure. Related tree for /ketɒb/ is CVCVC d-tree.
Introduce word to the tree.
So: /ketɒb/ → CVCVC Tree. Tree asks for identity of phonemes and location of stressed syllable to predict phonemes susceptible to substitution, deletion or insertion.
Variation pattern 1:
Phoneme 5 can be deleted and phoneme 3 can be substituted. Likelihood: 0.35
Variation pattern 2:
Phoneme 5 can be substituted and a phoneme can be inserted between phonemes 2 and 3. Likelihood: 0.15
Apply appropriate contextual rules, on framework of each Variation pattern.
Variants based on pattern 1:
/kerɒ/, /kexɒ/ ...
Sum of Normalized Likelihoods: 0.7
Variants based on pattern 2:
/kertɒp/, /kertɒf/ ...
Sum of Normalized Likelihoods: 0.3
Generalized d-trees
Trees related to monosyllabic words
CV Tree
CVC Tree
CVCC Tree
Trees related to disyllabic words
CVCV Tree
CVCVC Tree
CVCVCC Tree
Trees related to trisyllabic words
CVCVCV Tree
CVCVCVC Tree
CVCVCVCC Tree
Contextual Rules
Substitution rules
LFR → LOR
Insertion rules
LR → LOR

Fig. 1. Block diagram of hybrid structure to generate pronunciation variants of words.

insertion is permitted. More details about context-dependent rules are provided in Section 6. Finally, the sum of probabilities of variants is normalized to unity. The resultant pronunciation variants of the words may be introduced to the lexicon of the ASR system to improve its recognition accuracy.

3. Speech corpus and ASR experimental workbench

3.1. Large-FARSDAT – a Persian speech corpus

Persian (Farsi) is a member of the Iranian branch of the Indo-Iranian languages, a subfamily of the Indo-European languages. It is the language of Iran, Afghanistan, Tajikistan and the Pamirs mountain region. Persian, as a language spoken by Iranians, has many varieties due to cultural interactions of different nationalities within the Persian land, now Iran. One grouping, also used in this research, consists of 10 major accent groups namely Tehrani, Azari, Isfahani, Shomali, Yazdi, Jonubi, Khorasani, Kurdish, Lori, and Baluchi. Persian speech is known to include around 29 phonemes consisting of 23 consonants and 6 vowels. Table 1 lists IPA symbols of Persian phonemes along with their orthographic symbols and linguistic descriptions (Bijankhan et al., 2003). Although Table 1 shows orthographic symbols for short vowels, which are usually not represented in written Persian, there is not a strict grapheme-to-phoneme correspondence in Persian writing system.

Table 1
IPA symbols and the equivalent alphabets of Persian phonemes

IPA symbol	Orthographic symbols	Phonetic description
i	یـ ، ی	High front vowel
e	ـــِـ ، ا، ـه	Mid front vowel
a	ـــَـ ، ا	Low Stint vowel
u	و	High back vowel
o	ـــُـ ، و	Mid back vowel
ɒ	ا ، آ	Low back vowel
b	بـ ، ب	Voiced bilabial plosive
p	پـ ، پ	Unvoiced bilabial plosive
d	د	Voiced dental plosive
t	تـ ، ت ، ط	Unvoiced dental plosive
g	گـ ، گ / گاری	Voiced velar plosive
k	کـ ، ک / کار	Unvoiced velar plosive
G	قـ ، ق ، غ ، غـ ، ـغـ	Voiced uvular plosive
ʔ	ا ، عـ ، ع ، ئـ ، ء ، ـعـ	Glottal stop
ʤ	جـ ، ج	Voiced alveopalatal affricate
ʧ	چـ ، چ	Unvoiced alveopalatal affricate
v	و	Voiced labiodental fricative
f	فـ ، ف	Unvoiced labiodental fricative
Z	ذ ، ز ، ضـ ، ض ، ظ	Voiced alveolar fricative
s	ثـ ، ث ، سـ ، س ، صـ ، ص	Unvoiced alveolar fricative
ʒ	ژ	Voiced alveopalatal fricative
ʃ	شـ ، ش	Unvoiced alveopalatal fricative
x	خـ ، خ	Unvoiced uvular fricative
h	حـ ، ح ، هـ ، ه، ـه	Unvoiced glottal fricative
m	مـ ، م	Bilabial nasal
n	نـ ، ن	Alveolar nasal
r	ر	Alveolar trill
l	لـ ، ل	Alveolar lateral
j	یـ ، ی	Palatal glide

Large-FARSDAT is a Persian speech database, created by "Research center of intelligent signal processing" (RCISP) (Bijankhan and Sheikhzadegan, 1994). Large-FARSDAT includes 100 speakers selected with respect to age, gender and educational level, from one of the 10 frequent dialects of Persian in Iran. Each speaker has read thousands of words from various kinds of newspapers in an office room. The material covers a variety of fields such as politics, economics, culture, sports, etc. Since one of the main contributions of our research work is that the proposed pronunciation models can be trained with a much smaller amount of data in comparison with similar works, we used the minimum required data, i.e. half of Large-FARSDAT data, for training hybrid pronunciation models. This portion of Large-FARSDAT includes speech material from 50 out of 100 speakers in the corpus which is equivalent to a total of 25 h of speech material in duration. These 50 speakers are selected in a way to have maximum diversity of speakers regarding their age, gender and dialect. Large-FARSDAT contains a single phonemic transcription for each uttered word; the corpus does not contain any phonetic transcriptions for the speech utterances. In order to train the pronunciation models, first, we used the phone recognizer module of the SHENAVA ASR system to decode speech utterances as recognized phone strings. This phone recognizer has been found to operate with an acceptable level of performance on similar tasks (Almasganj et al., 2001). In the next step, recognized phone transcriptions of the words (transcribed automatically by phone recognizer) were aligned with the phonemic transcriptions (originated from the corpus). In this way, we modeled variations due to both pronunciation and phone recognizer errors, simultaneously. Although the factors that affect these two sources of variation are different, we believe our approach can face both these variation sources. Alignment of the phonemic transcriptions of the words to their recognized phonetic transcription is carried out by a dynamic programming algorithm (DTW) that minimizes the alignment distance. Substitution cost is dependent on simple distinctive features that differ from the baseline phoneme to the recognized phone. The distinctive features differentiate vowels from consonants and voiced from unvoiced phonemes. An example of the alignment is given below. The alignment is shown for phonemic and recognized phonetic transcriptions of a Persian word with phonemic transcription */ketɒbxɒneh/*, which means "library".

/k e t ɒ b x ɒ n e h/

/p e t ɒ # f ɒ n e #/

The first row includes the phonemic transcription and the second row the recognized phonetic string. Symbol "#" stands for deletion/insertion. Unvoiced plosive substitution (*/k/* with */p/*), deletion of voiced bilabial plosive (*/b/* to "#"), substitution of unvoiced fricative consonant (*/x/* to */f/*) and unvoiced glottal fricative deletion (*/h/* to "#") occurred in this example. Here, it is possible to consider the effects of two variation sources separately. It is reasonable to assume that the substitution of */k/* with */p/* is a phone recognizer fault, while deletion of */h/* is a real pronunciation variation which is produced by the speaker.

3.2. SHENAVA – Persian ASR system

SHENAVA is a Persian ASR system, developed in RCISP research center as the output of a primary phase of a big project which aims to develop a professional Persian ASR system. We used a version of SHENAVA with a 1200-word vocabulary as our experimental workbench in this work. Phoneme recognition is performed by a hybrid structure of two MLP neural networks which phonetically classify frames of an input speech signal and a rule-based engine to extract phones due to the classified frames. After the lexicon search and applying a semi-viterbi algorithm to find the best 100 recognized phrases, an N-best rescoring block, which uses HMM models of Persian phones, finds the best output phrase (Almasganj et al., 2001).

As mentioned, Large-FARSDAT only contains phonemic transcriptions of utterances; and speech material is not labeled with phonetic transcriptions. Hence we used the SHENAVA phone recognizer to have access to recognized phone strings of utterances instead of phonetic transcriptions. By comparing aligned pairs of phonemic and recognized transcriptions of speech utterances, the training algorithm captures differences due to speakers' pronunciations and errors of the SHENAVA phone recognizer. Considering errors of the phone recognizer in pronunciation models improves accuracy rate of the SHENAVA ASR system while using model-generated variants in its lexicon, since the variants are generated by considering phone recognizer errors. We used a version of the SHENAVA ASR system as our experimental workbench in this paper. The version of SHENAVA used in our experiments did not employ any language model to decrease its output word error rate.

4. Factors affecting pronunciation variations of words

Here the influences of factors which are considered in the hybrid pronunciation models are studied. The influence of the syllabic structure of words, the location of the stressed syllable, phonetic context and the rate of speech on pronunciation variation is well known in Persian and many other languages and is emphasized in literature (Ladefoged, 2006; Adda Decker et al., 2005; Milanian and Hosseini, 2002; Haghshenas, 1995). Many researchers exploited such features to model pronunciation variations (Chen and Hasegawa-Johnson, 2004; Cremelie and Martens, 1999; Fosler-Lussier, 1999b; Fukada et al., 1999). A phoneme may be deleted or substituted in pronunciation if it appears in a certain context. Stressed syllables are pronounced with higher energy and lower rate of speech. Therefore, phoneme deletions and variations happen less in them. Researchers have considered word predictability as another important factor affecting pronunciation variation (Jurafsky et al., 2001). In the following subsections, the effects of rate of speech, word predictability, location of syllable in the word and word stress on pronunciation variations will be discussed quantitatively. The database used in the following studies is the same subset of Large-FARSDAT which is used for training pronunciation models and contains speech material from 50 speakers with 25-h duration.

4.1. Rate of speech

"Rate of speech" (ROS) is usually measured in terms of the number of linguistic units pronounced in 1 s. In this paper, ROS is considered as the number of syllables pronounced in 1 s. The corpus has time labels for all word units; it means that both the start and the end time of a word are accessible from the label files. In addition the phoneme recognizer determines time boundaries of each recognized vowel and also each recognized syllable. We calculated the rate of each word by considering its duration plus the durations of its two neighboring words from each side, ROS is calculated by dividing the overall duration by the total number of syllables counted in these five words. This value is considered as an estimate of the rate of speech for the middle word. This technique will effectively normalize the effect of phonetic syllable length on rate measurement with a smoothing effect. Fig. 2 shows a histogram of ROS, in syllables per second, for all words in Large-FARSDAT. The histogram is fairly similar to a Gaussian probability distribution and can thus be approximated by this function with a relatively small error. The estimated parameters of this function are: $\mu = 5.34$ (syl/s) and $\sigma = 1.59$ (syl/s). To analyze the effect of ROS on pronunciation variation, we divided our database into three portions according to the rates of speech production: low-rate, medium-rate and high-rate. This is done in accordance with the mean and standard deviation of ROS. The low-rate portion consists of utterances which are uttered in a rate lower than the average minus one standard deviation; the medium-rate portion consists of those with a rate between average minus one standard deviation to average plus one standard deviation and the high rate portion consists of the remaining utterances.

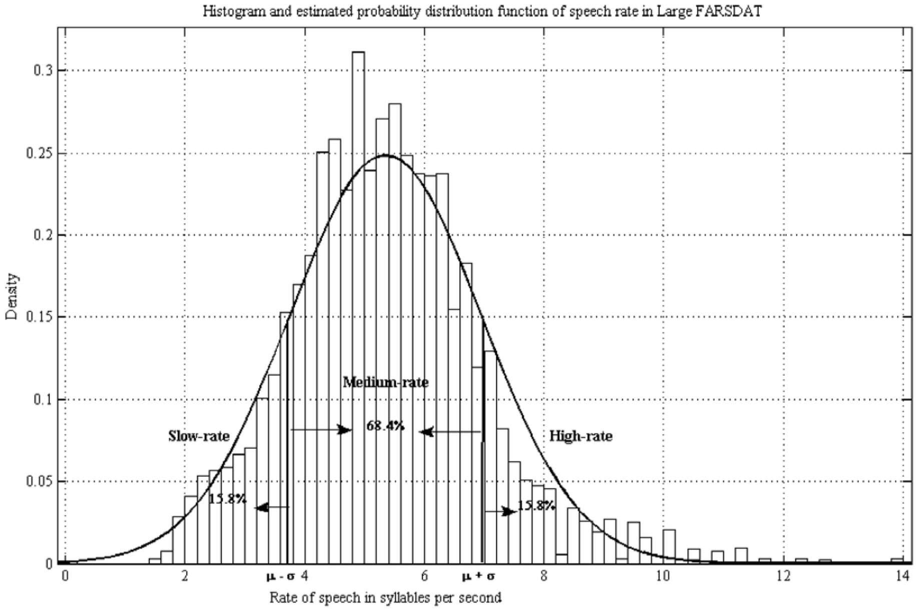

Fig. 2. Histogram of ROS in syl/s and the estimated Gaussian pdf.

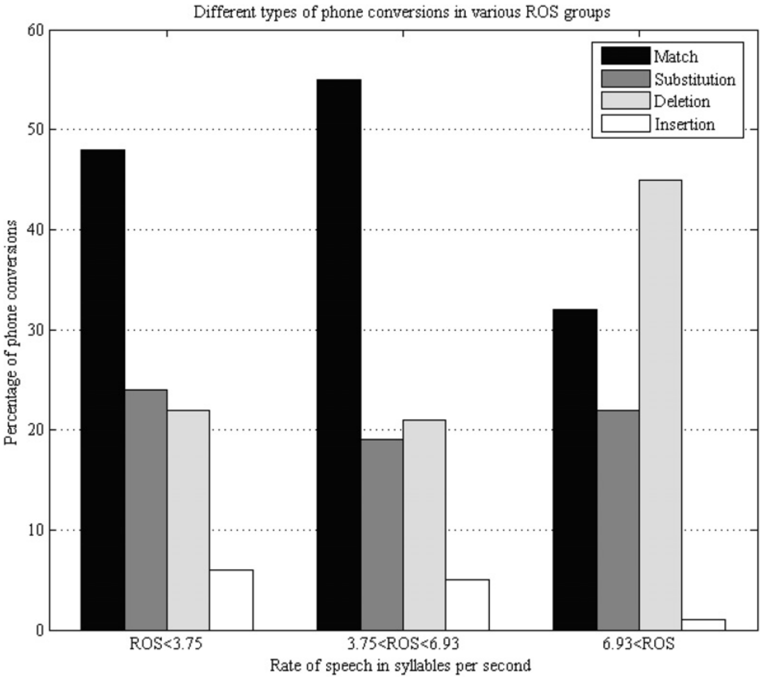

Fig. 3. The percentages of various kinds of phone conversions in different groups of ROS, comparing standard phonemic and recognized phone strings of words.

Large-FARSDAT utterances are produced in the form of read speech. Thus, speakers usually speak with a medium rate. However, it seems that there are a noticeable number of speakers who inherently utter faster or slower than the normal cases. It could be seen in Fig. 2, that 68.4% of the whole database is in medium rate of speech, in this partitioning. After alignment of phonemic and recognized phone strings of words and comparing the aligned strings, Fig. 3 depicts statistics of various kinds of phone conversions in each of the three portions of the speech corpus. The difference between aligned transcriptions can be a result of pronunciation or a fault in the phone recognition engine. In this figure, the percentages of phone matching, substitution, deletion and insertion, in each portion of the database, are shown. Black bars represent the percentages of phones that are the same in both aligned phonemic and recognized phone strings or percentages of identically aligned phones; from now on phone match. The dark grey bars represent the percentages of phone substitutions, the light grey bars represent the percentages of phone deletions and the white bars show the phone insertion percentages.

According to Fig. 3, the highest similarity or maximum match between the aligned phonemic and recognized phone strings is obtained for the medium ROS (between 3.75 and 6.93 syl/s). Furthermore, this ROS shows the lowest deletion rate. For the ROS higher than 6.93 syl/s, deletion is highly increased. In high ROS, speakers use shorter pronunciation variants of words. The performance of the phone recognizer is also decreased in high ROS, which results in an increase in the phone deletion rate. In the literature, it is shown that the word error rate will increase due to an increase in ROS (Fosler-Lussier and Morgan, 1999).

A detailed analysis of Fig. 3 reveals the importance of the ROS parameter in generating a dynamic lexicon for ASR systems. In a dynamic lexicon, variants of the words will be modified when the rate of the input speech is changed. In such a lexicon, for high rates of speech, shorter variants will generally experience higher probabilities than longer ones. In medium rates of speech, the number of word variants will decrease, because words are generally produced and recognized closer to their phonemic versions. Here, the probability of the reference forms will increase. Modeling and applying the effects of speech rate, on phonetic versions of words included in the lexicon of an ASR system, seems to decrease the word error rate.

Phone pronunciation entropy, sometimes called phone variation entropy, could also be used as another metric to investigate the effect of speech rate on phonetic pronunciation variation (Fosler-Lussier, 1999a). This metric measures the degree of variation in pronunciation. By comparing aligned phonemic and recognized strings, the level of conversions or substitutions of phones are considered to calculate phones' pronunciation entropy. If phone α can be potentially pronounced or recognized as one of the phones in set X, phone pronunciation entropy will be defined as

$$H(\alpha) = \sum_{x \in X} p(\alpha \to x) \log p(\alpha \to x), \tag{1}$$

in which X represents the set of language phonemes, x a member of this set and α the inspected phoneme. $p(\alpha \to x)$ is the probability of pronouncing α as x. The defined Entropy can be used as a measure for the level of phonetic pronunciation variation. This measurement will increase when the inspected linguistic unit (here phoneme) is pronounced or recognized in more different ways. Therefore, the pronunciation entropy of a phone can be interpreted in this way: phone pronunciation entropy is minimum when it has only one pronunciation variant. In this case, entropy will be zero. As an example, if a phoneme \u\ can be pronounced as phones \u\, \a\, \e\ and \o\ with probabilities of $p(u \to u)$, $p(u \to a)$, $p(u \to e)$ and $p(u \to o)$, respectively, its pronunciation entropy will be:

$$H(u) = \sum_{x \in \{u,a,e,o\}} p(u \to x) \log p(u \to x). \tag{2}$$

$p(u \to x)$ is the probability of pronouncing \u\as \x\ and can be calculated by the following equation:

$$p(u \to x) = \frac{N(u \to x)}{N(u)}, \tag{3}$$

where $N(u \to x)$ is the frequency of pronouncing \u\as \x\, and $N(u)$ is the overall frequency of \u\in the phonemic transcription of the processed utterances. Table 2 shows the averages of phone variation entropies over different groups of Persian phones, derived for three ROS ranges. Apparently, the measured average entropies vary by changing the ROS in different manners for various groups of phones. The lowest overall entropy when considering all of the phones are obtained for the rates of speech between 3.75 and 6.93 syl/s, thus it can be concluded that phones have lowest pronunciation variability in medium rate when considering all of them at the same time in the study. This could not necessarily be concluded when considering each phone group separately as shown in Table 2. According to Table 2, for the vowels, entropy increases as a result of an increase in ROS. Entropies in plosives are less affected by ROS and fricatives are almost not affected by ROS. The Liquids, Nasals and glottal group show lowest entropy in the medium ROS which means lowest pronunciation variation when considering these kinds of phones in the study.

The results reported in this table could be a subject of further study in phonetics. However, here, we only follow the effectiveness of ROS on the level of pronunciation variation in different phone groups. We conducted the previous experiments by dividing the used speech corpus into three main regions of ROS just to prove that the ROS is an important factor which effectively varies pronunciations of words. But the generalized decision trees in dynamic versions of hybrid pronunciation models, which are introduced in this paper, consider ROS as a continuous feature without any partitioning or quantization.

4.2. Word unigrams

It is already shown that the word predictability influences speaker pronunciations (Fosler-Lussier, 1999a; Jurafsky et al., 2001). Words with high n-gram probabilities are easier for listeners to recognize, because their

Table 2
Average entropies of various groups of Persian phones for different ROS ranges

Phone group	ROS < 3.75	3.75 < ROS < 6.93	6.93 < ROS
All phones	1.58	1.54	1.66
Vowels	0.51	0.80	1.24
Plosives	2.21	2.13	2.09
Fricatives	1.40	1.42	1.44
Liquids, Nasals	1.78	1.42	1.43
Glottals	2.03	1.88	2.42

predictabilities are high. To pronounce these words, speakers naturally use shorter pronunciation variants and pronounce them faster. In contrast, specific words with low predictabilities usually are pronounced slower than normal and phonetically closer to their reference or phonemic forms. This compensates for the unpredictability of this group of words and helps listeners to recognize them better. This speech phenomenon is an intrinsic and language independent reality. As a different view, when the meaning in a sentence is carried mainly by a word or a few specific words, they will be carefully pronounced in a lower rate of speech, and speakers use word variants near to their reference forms to transform the information carefully. In contrast, to produce unimportant words, speakers use short variants and high rates of speech. Here, to analyze the influence of words' predictabilities on their pronunciations, word unigram statistics are employed as a measure of word predictability. To handle this analysis we divided the database into frequent and infrequent parts. These parts consisted of words with more than and less than 100 samples in the used database, respectively. This threshold is chosen heuristically and results in almost identical numbers of occurrences for frequent words and infrequent words in the database. Using this threshold, 78,205 occurrences of frequent words and 70,002 occurrences of infrequent words are observed. This threshold is chosen also by considering the size of the database. Moreover, to control and normalize the effects of ROS in our experiments, we carried out our experiments individually on the parts of the database which are selected by their ROS levels. Fig. 4 shows the word phonetic conversions for the parts of the database which are selected by the ROS and unigram statistics, found by comparing aligned phonemic and recognized phone strings of words. The percentages of phonetic mappings are shown in this figure for the frequent and infrequent words in the medium and high rates of speech.

It can be seen that the percentage of phone match for the infrequent words is higher than the same percentage for the frequent words in both rates of speech. Furthermore, the percentages of phone deletions are relatively high for the frequent words, which support the claim that by increasing unigram statistics of words, speakers use shorter length pronunciation variants. It can also be concluded that the probability of pronouncing words in their standard phonemic forms is lower for frequent words, in contrast to infrequent words. The results reported here are obtained where the phone recognizer performance is not directly affected by unigram statistics of words.

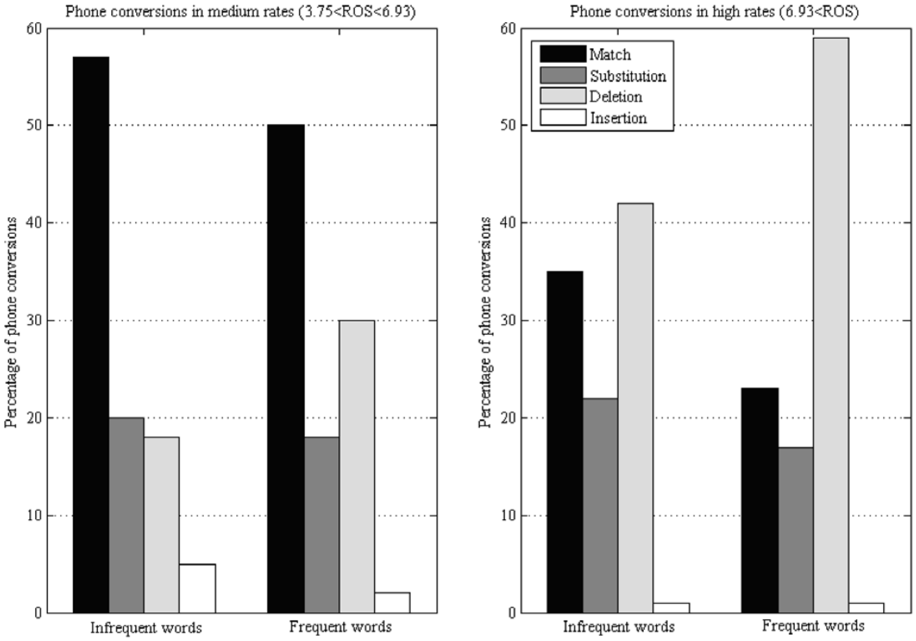

Fig. 4. Percentages of different types of phone conversions for the frequent and infrequent words in the medium and high ROS, while comparing standard phonemic and recognized phone strings of words.

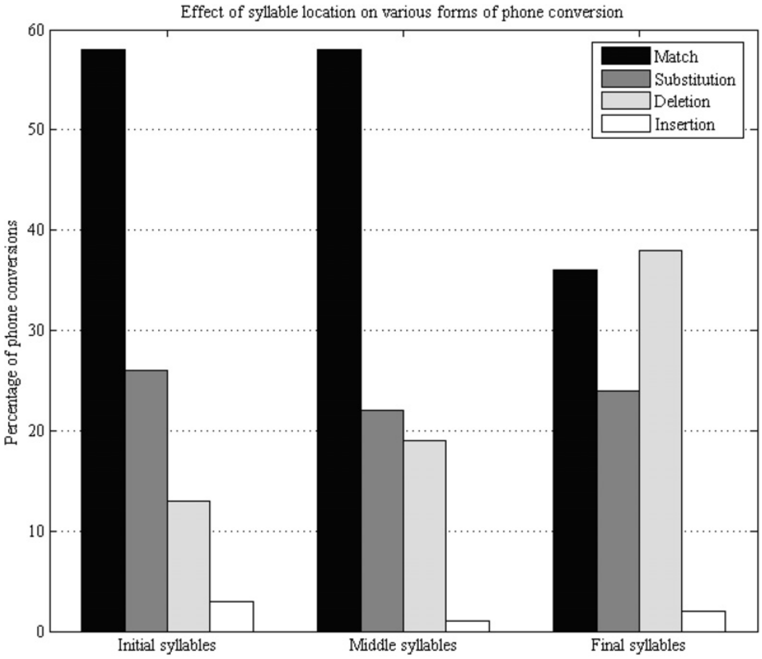

Fig. 5. Effect of “syllable location” on different type of phone conversions, while comparing standard phonemic and recognized phone strings of words.

Although in this analysis we defined frequent and infrequent words based on their numbers of occurrence, the generalized decision trees in hybrid pronunciation models use the logarithm of unigram probabilities as a continuous attribute of each word without any primary partitioning of data into frequent and infrequent words.

4.3. Syllable locations

The next issue to address is the influence of syllable location in the word on phonetic pronunciation variation. In Fig. 5, the percentages of phone match, substitution, deletion and insertion in the initial syllables, the middle syllables and the final syllables of the words available in the database are displayed. Our speech database consists of 36,680 monosyllabic words, 36,159 disyllabic words and 75,368 words with three or more syllables. Inflected forms of words and foreign names are included in these counts and treated as word units in this analysis. We did not use monosyllabic words just in this study because they can be treated as an initial, middle or final syllable. Therefore, only the words with two or more syllables were analyzed. For disyllabic words, first syllables were treated as initial syllables and second syllables as finals. The percentage of various phone conversions are extracted by comparing phonemic and recognized phone strings of words.

Fig. 5 reveals that the phone deletion rate is highest among the final syllables of words while phone match between phonemic and recognized phone strings of words is highest among the initial and lowest among the final syllables. These results indicate the possible usefulness of the syllable location in the pronunciation modeling of words.

4.4. Word stress

Word stress is another factor which affects pronunciation and performance of phone recognizers. Higher energy and duration are featured in stressed syllables, which result in better performance of phone recognizer engines. To study the effect of stress on phonetic pronunciation variations, we, once again used percentages of phone match, substitution, deletion and insertions by comparing the aligned standard phonemic and recognized phone strings of words. Just in this analysis and not throughout the whole paper, in order to cancel

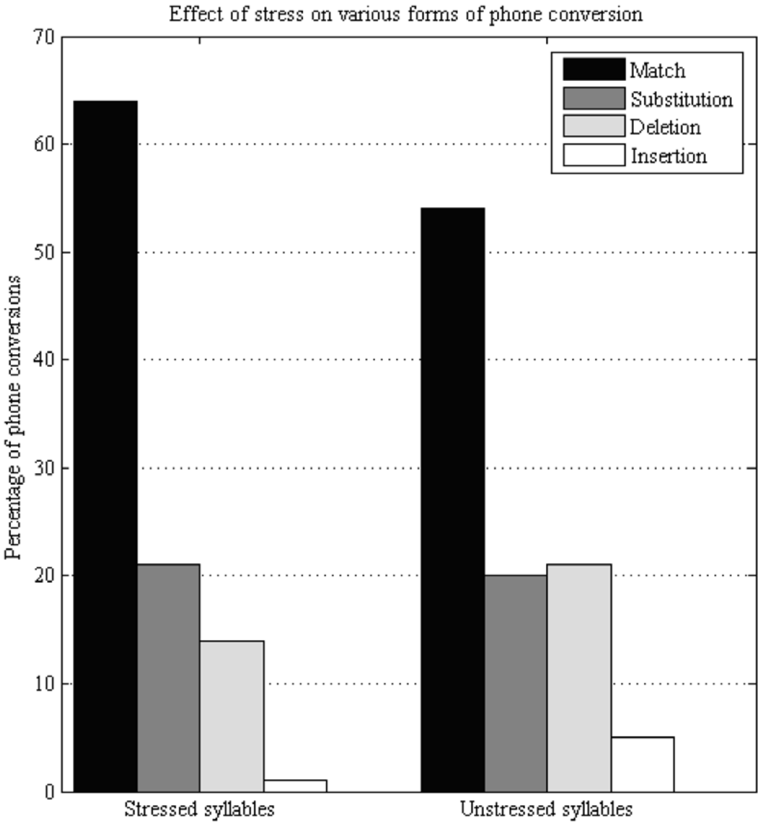

Fig. 6. Percentages of phone conversions in stressed syllable and un-stressed syllables.

the effect of syllable location factor, we used only middle syllables. Middle syllables used in this study originate from 75,368 words with three or more syllables in the database. Fig. 6 displays the results obtained. As can be seen, phone matching rate is higher and phone deletion rate is lower for stressed syllables in comparison to unstressed ones.

According to these results, we concluded that the word stress has an overall effect on phonetic pronunciation variation and also the performance of the phone recognizer. The generalized decision trees of our proposed hybrid pronunciation models utilize the word stress in predicting the pronunciation variants of words.

5. Generalized decision trees and training algorithm

In this section, generalized decision trees as the main part of hybrid statistical pronunciation models are discussed in detail. In our approach, words in the training database are partitioned into different groups based on the numbers and structures of their syllables. Syllable structure in Persian is limited to only three patterns, "CV", "CVC" and "CVCC". As a result, 3 groups of monosyllabic words, 9 groups of disyllabic words and 27 groups of trisyllabic words exist in Persian. A separate decision tree is trained for each of these groups. For example, the words */ketˈɒb/* and */medˈɒd/* (means "book" and "pencil", respectively) are used to train the same specialized "CVCVC" tree. In the following sub-sections, we will introduce and explain the features of words which are used by generalized decision trees, the encoding process of uttered words' pronunciation pattern as vectors to be used in training, a vector quantization method to quantize word's pronunciation vectors in order to reduce the variations in data before training, and finally the training algorithm of generalized decision trees and splitting criteria.

Table 3
Phonetic categories of vowels and consonants, their IPA symbol and assigned unordered category numbers

Category	IPA symbol	Phonetic description
1	ɒ, *a*, *o*	Low back, low front, mid back vowel
2	*e*, *i*	Mid front, high from vowel
3	*u*	High back vowel
4	*b*, *d*, ***G***, ʤ, *g*	Voiced plosive
5	*p*, *t*, *k*, ʧ	Unvoiced plosive
6	*s*, ʃ, *x*, *f*	Unvoiced fricative
7	*Z*, ʒ, *v*	Voiced fricative
8	*m*, *n*, *l*, *r*	Nasals, Liquids
9	ʔ, *h*	Glottals
10	*j*	Palatal glide

5.1. Features of words considered by generalized decision trees

As the first module of hybrid pronunciation models, generalized decision trees are trained and then used to predict pronunciation patterns based on several category-based, discrete and continuous features of words. The effects of some of these features on pronunciations are discussed in Section 4. The category-based features are defined as the membership of each of a word's phonemes in one of the 10 phonetic categories. The location of the stressed syllable in the word is used as a discrete feature. The logarithm of a word's unigram probability is the common continuous feature between static and dynamic trees. Dynamic trees, moreover, take into account the rate of speech as another continuous feature.

Phonetic categories of phonemes in the word are considered as input features in the trees during training and testing. In this work, 10 phonetic categories of Persian phonemes are introduced, as shown in Table 3. This categorization tends to put phones with similar conversion behavior into the same group, i.e. phonemes which have similar manners and tendencies in substitution, deletion and insertion. We considered both linguistic characteristics and the confusion matrix extracted after phone recognition process in this categorization simultaneously. The confusion matrix is derived by comparing aligned phonemic and recognized strings. The phoneme recognizer, for this purpose, is the same phoneme recognizer used in conducting experiments and preparing training data, i.e. the SHENAVA phoneme recognizer module. Therefore, the selected categories of vowels with similar conversion behaviors are: {*/*, *a*, *o*}, {*e*, *i*} and {*u*}. The assigned unordered category numbers for these groups are 1, 2 and 3, respectively, as shown in Table 3. Phonetic categories of Persian consonants are defined mainly based on their phonetic similarities and differences. We have also paid attention to their behavior in the recognition process. For example, it is observed that consonant /*y*/ has a tendency to be recognized as vowel /*i*/. Due to this special behavior it is inserted in a separate category. /*]*/ and /*h*/ are both glottal; they are inserted in the same category because they are very likely to be deleted in Persian continuous speech. Table 3 also shows seven categories of the consonants and their phonetic descriptions along with assigned unordered category numbers, which are 7–10.

For an input word, the corresponding generalized decision tree considers phone arrangement in the word, by asking about memberships of the consonants and the vowels, placed in various locations of the word, to different categories in order to predict the pronunciation pattern. As an example, the feature vector of the word /*ket/b*/ with stress on its second syllable, which is uttered in a rate of 6.3 syl/s and its logarithm unigram probability is −3.45, will be [5 2 5 1 4 2 6.3 −3.45], where the first five attributes define membership of each phone to different phonetic categories and are treated as unordered category-based features during training and prediction by trees, the 6th feature shows the location of stressed syllable and 7th and 8th features are continuous features of rate and logarithm of unigram probability, respectively. It is clear that the lengths of feature vectors differ for various words with different phonemic lengths. This may raise the question of how these vectors with various lengths could be used to train trees? As described earlier, in our approach, a specific tree is trained for each of the word groups with similar number and structure of syllables (which have also same phonemic lengths). Hence the feature vectors which are used to train each tree have the same lengths.

5.2. Encoding process of pronunciation patterns

Pronunciation pattern Code vector, assigned to each uttered word, represents mapping between aligned phonemic and recognized strings of the uttered word and shows occurrences of substitution, deletion and insertion in the utterance. The vectors have a length of $4N + 1$, where N is the phonemic length of the words. They are composed of zeros and ones that indicate identically aligned phones, substitutions, deletions and insertions. Each phoneme of a word has three corresponding cells in the code vector. The first cell will be set to one, if the phone is identically aligned and otherwise to zero; the second bit represents substitution and the third bit deletion. Phone insertions are defined by extra cells accommodated between the sets of three cells mentioned above. As an example, consider word */ket/b/* which is recognized as */]et/#/* (the first consonant is substituted and the last consonant is deleted) in the utterance. The assigned code vector for this uttered word will have a length of $4 \times 5 + 1$ and will be defined as:

0	0	1	0	0	1	0	0	0	1	0	0	0	1	0	0	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*

where *i*, *m*, *s* and *d* specify cells in vector which are related to phone match, substitution, deletion and insertion, respectively.

5.3. Vector quantization of pronunciation pattern: a way to deal with data insufficiency

The large variety of pronunciation patterns and limited occurrence of each one in the database is a challenge in training trees. To solve this problem and keep one of the main contributions of our approach, which is training models by a medium-size corpus, we employed vector quantization (VQ) to reduce the diversity of pronunciation patterns. Quantization vectors are then used in training and prediction.

As described earlier, data is partitioned according to various word groups, based on the number and structure of the syllables. The K-means algorithm is applied to each of these partitions to quantize various pronunciation patterns' code vectors. In this way, for each partition of data, we could quantize pronunciations and collect them in groups containing ones which are similar, based on their least squared distance error. Unique labels are assigned to each of the patterns in a cluster. These labels are just unique names for each quantization vector and can be chosen in any way, but there should be a one to one correspondence between labels and quantization vectors. Labels are attached to corresponding uttered words and are used as prediction outputs in the training phase. As a result, generalized decision trees predict labels of likely pronunciation patterns based on the features of input words. The pronunciation pattern related to the predicted label will be decoded and will be used as a framework for generating pronunciation variants by applying appropriate contextual rules in the final step of the procedure.

Here, the vector quantization process is further described by giving an example. For words with CVCVC structure, 11,356 uttered samples are available in the database. This is one of the most common structures among Persian words. Words with CVCVC structure are pronounced in 380 pronunciation patterns. Hence on average, there exist almost 30 uttered words for each of these pronunciation patterns. We quantized these 380 pronunciation vectors into 114 quantization vectors to have an average of almost 100 samples for each of the quantization vectors. The number of bins for all other disyllabic and trisyllabic groups of words is set in the same way to have almost 100 samples (uttered words) around each quantization vector. For monosyllabic words there are enough samples available for each of the observed pronunciation patterns. As a result, for monosyllabic words, no vector quantization is carried out. Fig. 7 is an example of vector quantization for the CVCVC group. The left side of the figure shows pronunciation patterns in groups containing similar patterns, and the number of the uttered words. Different cells in the vectors of each group are shown in dark grey. The corresponding quantization levels are shown on the right side. Unique labels related to these quantization vectors are assigned to corresponding uttered words to be used as prediction outputs in the training phase while the inputs in training will be the features derived from uttered words described in Section 5.1.

To design an M-level codebook for a group of words with similar syllabic structures, it is necessary to partition $4N + 1$ dimensional space (where N is phonemic length in this group as discussed in Section 5.2) into M

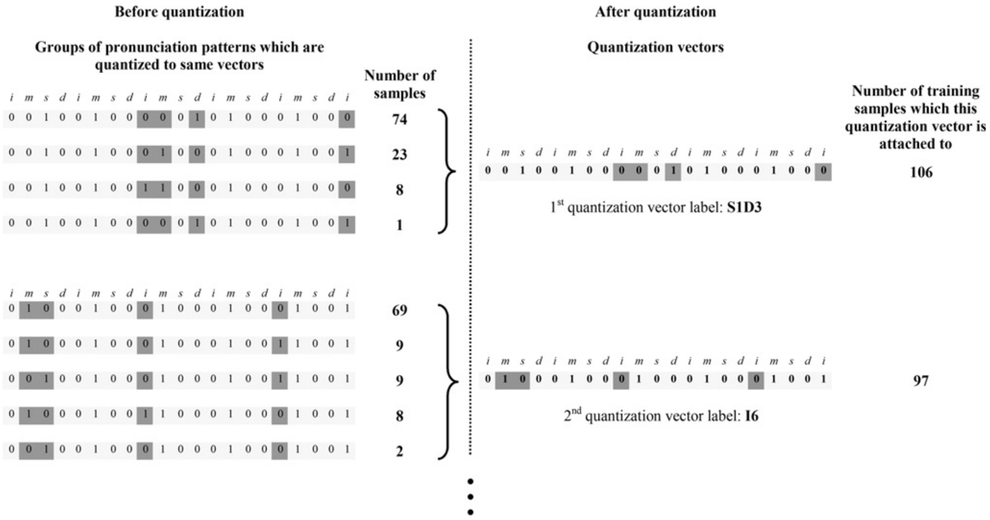

Fig. 7. An example of vector quantization for words with CVCVC syllabic structure.

cells and associate a quantized vector with each cell. The criterion for optimization of the vector quantizer is to minimize the overall average distortion over all M-levels of the VQ. The overall average distortion can be defined by

$$D = E[d(x,z)] = \sum_{i=1}^{M} p(x \in C_i) E[d(x,z_i)|x \in C_i] = \sum_{i=1}^{M} p(x \in C_i) \int_{x \in C_i} d(x,z_i) p(x|x \in C_i) \mathrm{d}x = \sum_{i=1}^{M} D_i, \tag{4}$$

where the integral is taken over all components of vector x(x is a variation pattern vector), $p(x \in C_i)$ denotes the prior probability of codeword z_i, $p(x|x \in C_i)$ denotes the multidimensional probability density function of x in cell C_i and D_i is the average distortion in cell C_i. An iterative algorithm which guarantees a minimum of the average distortion measure exists and works well in practice (Kovesi et al., 2001). By applying this VQ technique to the set of pronunciation patterns of a specific group of words (which are used to train a tree) and finding their centroids, each of the patterns will be evaluated by exhaustively computing its distance from each of the centroids. Then, the label of the nearest neighbor quantization vector is used to encode that pattern. In this way, we solve the difficulty of training trees due to diversity of pronunciation patterns and insufficiency of samples for each pattern by using quantization patterns instead. As a result, a medium-size database could be used to train trees in hybrid pronunciation models accepting a minimal loss of information on pronunciation variation due to the nature of quantization in information reduction. We accept this minimal loss of information in a trade off to gain the advantage of being able to train the models using a medium-size corpus which contains a very limited number of occurrences for some actual patterns.

5.4. Generalized decision tree training algorithm: splitting criteria and prediction process

We exploited a training algorithm which tries to reduce the sum of the entropies of terminal nodes of decision trees while training. The total entropy should not be very low, to avoid over-fitting the models to the training data. A splitting criterion is used for this purpose, i.e. when the number of samples or input/output training pairs (feature vectors and labels of patterns' pairs of uttered words) in the node is greater than a certain threshold, the node can be further split, and on the contrary, when the number of samples is less than the threshold, the node becomes a terminal node. The threshold is experimentally set to 200, in order to have a meaningful decision in each terminal node. Questions in the decision tree training framework are chosen automatically by the training algorithm based on the set of input features which are introduced in Section 5.1 to partition data samples of a node. The training task is to find the best question for any node split which is

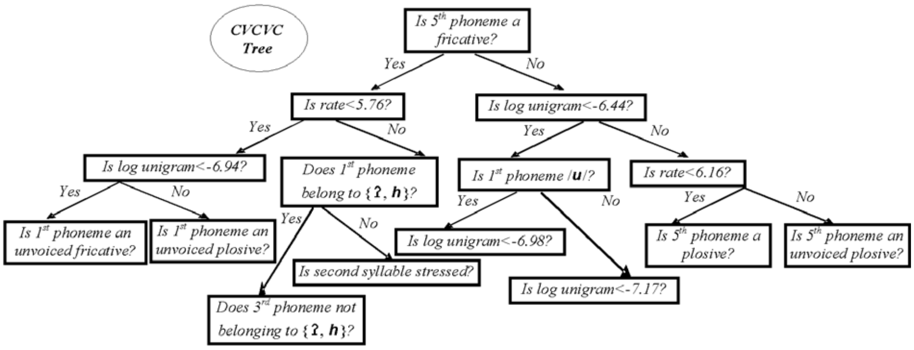
CVCVC Tree
Is 5th phoneme a fricative?
Yes
No
Is rate<5.76?
Is log unigram<-6.44?
Yes
No
Yes
No
Is log unigram<-6.94?
Does 1st phoneme belong to {ʔ, h}?
Is 1st phoneme /u/?
Is rate<6.16?
Yes
No
Yes
No
Yes
No
Yes
No
Is 1st phoneme an unvoiced fricative?
Is 1st phoneme an unvoiced plosive?
Is log unigram<-6.98?
Is 5th phoneme a plosive?
Is 5th phoneme an unvoiced plosive?
Is second syllable stressed?
Is log unigram<-7.17?
Does 3rd phoneme not belonging to {ʔ, h}?

Fig. 8. First four levels of CVCVC dynamic generalized decision tree used in the first stage of the hybrid model.

equivalent to splitting the node in a way to obtain leaf nodes that are as pure as possible, in terms of class distribution (classes here are labels of quantization pronunciation patterns). The training algorithm will be discussed shortly. Let Y be the random variable of classification decision. We could define the weighted entropy for any node t as follows

$$\overline{H}_t(Y) = H_t(Y)p(t) = -\sum_i p(i|t)\log p(i|t)p(t), \tag{5}$$

where $p(i|t)$ is the probability of data samples in node t belonging to class i and $p(t)$ is the prior probability of visiting node t (equivalent to the ratio of number of data samples in node t and the total number of training data samples). With this weighted entropy definition, the splitting criterion is equivalent to finding the question which gives the greatest entropy reduction, where the entropy reduction for a question q to split a node t into leaves l and r can be defined as

$$\Delta\overline{H}_t(q) = \overline{H}_t(Y) - (\overline{H}_l(Y) + \overline{H}_r(Y)) = \overline{H}_t(Y) - \overline{H}_t(Y|q) \tag{6}$$

The reduction in entropy is also the mutual information between Y (classification or identifying the quantization pronunciation pattern) and question q (which is based on input features). The task then becomes evaluating the entropy reduction $\Delta\overline{H}_q$ for each potential question, based on input features, and picking the question with the greatest entropy reduction, that is,

$$q^* = \arg\max_q(\Delta\overline{H}_t(q)). \tag{7}$$

Fig. 8 shows the first four levels of dynamic generalized decision tree trained for words with CVCVC structure. Showing more branches is avoided for the sake of saving space. Fig. 8 presents a real example –, numbers in the figure are actual values that appeared in the trained CVCVC tree.

As the prediction procedure runs, the decision tree asks for phone identities, the location of the stressed syllable, word unigrams and the rate of speech (in dynamic trees). Finally, the most probable labels of pronunciation patterns are defined in the terminal nodes as the outputs of the prediction process. Then, the most likely pronunciation pattern will be known as a framework to generate pronunciation variants. Each pronunciation pattern defines which phonemes can be substituted, deleted or where an insertion can occur. We discussed the effects of factors such as rate of speech, unigram, syllable location and stress on pronunciation variation in Section 4. Therefore, it should be clear why the idea of predicting pronunciation patterns using generalized decision trees works. We have used these features of spoken words besides information about phone categories, in every location of the word, to predict pronunciation patterns. The mutual information between the input features and the output, which are labels of pronunciation patterns, is discussed. A word clustering idea and a vector quantization technique is used to face the problem of the medium-size of the training data in the framework of generalized decision trees. We cannot train generalized decision trees for words with more than three syllables because in Persian, words with the same structures and more than three syllables are too rare. Hence, we applied only contextual rules to such multi-syllabic words in order to generate their variants. As the words with more than three syllables play a small role in Persian language, treating these cases in a conventional way has a minor effect on the overall performance.

6. Contextual rules

The contextual rules constitute another important building block in the suggested hybrid pronunciation models. After defining appropriate pronunciation patterns as a framework for generating pronunciation variants of the word by generalized decision trees, contextual rules can be applied to the permitted regions of substitution and insertion by taking into account the context of these regions. They are a set of context-dependent rewrite rules which are defined in the form of $L\underline{F}R \to LOR$, in which L and R represent left and right single phone contexts of the focus string F. F represents the phonemic string and can be recognized as O due to pronunciation or phone recognizer error. F and O can be more than a single phone, and F also can be an empty string in which case the rule becomes an insertion rule in the form of $LR \to LOR$. The combination of $L\underline{F}R$ is called the condition of the rule because it involves the contextual condition of applying the rule i.e. existence of the string $L\underline{F}R$ in the phonemic transcription of the word.

The main idea of learning the rules and applying them to dictated regions (marked by generalized decision trees) is the same as the method described in Cremelie and Martens (1999). However, the main idea and the differences with our implementation will be discussed shortly. The same 25 h medium-size training database which was described earlier is used for the extraction of contextual rules. The learning algorithm compares aligned phonemic string and the recognized string of phones. When a difference between aligned phone strings is detected, a rule will be derived, i.e. the context and focus F will be extracted from phonemic transcription and the output O from the recognized string. The overall number of times that the condition of a specific rule has occurred in the phonemic transcriptions of the training database is counted; this number is called the coverage of the rule which means how many times a specific rule could occur in the whole database. We can then calculate the application likelihood of each rule p_{AL} which is used in pruning statistically unimportant rules. The application likelihood of a rule is defined as the probability of pronouncing or recognizing the string F as O when its condition is satisfied, i.e. $L\underline{F}R$ string exists. When the likelihood of a rule is very low, it means that the rule is statistically unimportant. The likelihood is calculated by dividing the number of times a specific rule has been derived from the training database by the number of times it could have occurred, regardless of whether it has occurred or not (i.e. the coverage of that rule). The definition of application likelihood for ith rule r_i can be expressed and calculated by the following equations, where $\mathrm{count}_2(r_i)$ is the number of times r_i has occurred in the training database, and $\mathrm{count}_1(r_i)$ is the coverage of r_i:

$$p_{AL}(r_i) = p(O|LFR, r_i) = \frac{\mathrm{count}_2(r_i)}{\mathrm{count}_1(r_i)}. \tag{8}$$

A large set of 47,273 different rules are derived from the rule generation stage. The large size of this set is due to the fact that for each detected unique pronunciation variation (irrespective of whether it occurred frequently or not) a rule is derived. Hence, we used a rule pruning phase to remove rules which are not sensible from a statistical point of view and to select the rules that represent relevant pronunciation variation mechanisms. The pruning stage is accomplished by considering the application likelihood of each rule i.e. rules with less than 0.1 likelihood are pruned to achieve a number of 98,325 statistically meaningful rules; the used threshold is chosen experimentally. The approach is completely detailed in Vazirnezhad et al. (2005a).

A difference of our work in rule learning algorithm in comparison with Cremelie and Martens (1999) will be shortly summarized here. In the rule learning approach introduced in Vazirnezhad et al. (2005a), and of course in this work, we modified the algorithm to match with the nature of Persian language. We defined another rule format in our work which makes it possible to capture some frequent types of pronunciation variations in Persian language which cannot be captured by the simple $L\underline{F}R \to LOR$ format. Upon these new rule formats, for example, if there is a triple-phone set, including one of the group of {/m/, /n/, /l/, /r/} consonants, surrounded by two similar vowels, this group of phones will be most probably pronounced as just one vowel.

The rules described so far, are derived automatically from the training database. As a direct result of database size, it is sensible that the database does not have coverage for all possible phone arrangements and the final set of data derived rules may be deficient due to the fact that all sorts of Persian phone variations are not seen in the database and are not captured by the rules' learning algorithm. We later added some knowledge-based rules to solve this issue. Knowledge-based rules are a small proportion in comparison with data derived rules and have a minor effect on the whole performance of the system. However, they are useful and we keep them in the system. Knowledge-based rules are designed to give a generality to some data derived rules by taking into account Persian linguistic facts. They solve a small part of the problem described above due to data insufficiency matter. As an example, in Persian, when a voiced plosive precedes by any consonant, it is very likely that the vowel */e/* be inserted between them (Haghshenas, 1995). We add this knowledge-based general rule to the set of specific rules, from the same framework, which were derived from data such as insertion rules: *tg* → *teg* or *n*G → *ne*G. As described above, regarding the database size, all possible arrangements of a voiced plosive preceding by a consonant are not seen in the training database and as a consequence all rules in this framework are not extracted. To solve this issue we add the above mentioned general knowledge-based rule (i.e. vowel */e/* can be inserted between a consonant and a voiced plosive) to solve the issue of lack of coverage due to data insufficiency.

Utilizing only contextual rules in order to generate pronunciation variants have some drawbacks. The main disadvantage of the contextual rules as a sole tool for pronunciation variation modeling is that these rules involve only contextual information and do not consider word-level information, such as words' syllabic structure, stress pattern, unigram and rate of speech. However, the hybrid of generalized decision tree and contextual rules benefits from the advantages of contextual rules without suffering from its weaknesses.

7. A measure of confusability to discard confusable lexicon entries

Adding all variants of the words to the lexicon of an ASR system does not always decrease the word error rate (WER). Which variants must be included in the lexicon and which ones must be excluded is an important problem that remains to be solved. Some variants improve the system performance and others cause degradation. It is difficult to determine which one will decrease the WER. However, it has been shown that confusable variants will degrade the performance of an ASR system. Confusable variants are those which are homophones to other words' variants or only differ in just confusable phonemes. In fact, in designing an ASR lexicon, it is difficult to judge beforehand which variants contribute to describing the variances of the corpus at the level of pronunciation and also are not confusable, and is often postponed until all of the variants are generated. After generating all variants, the confusable ones will be rejected through a lexicon pruning phase (Wester, 2003).

Sloboda and Waibel (1996) chose to reduce confusability by eliminating learned pronunciations that exactly match entries for other words. We propose a softer measure, which introduces a metric to judge the confusability of individual word variants. We introduced a technique in calculating the confusability scores at the word level in order to be able to discard highly confusable entries from the lexicon. We name this measurement "Confusability Count" in this study. The confusability count of a pronunciation variant is defined as the number of times it is more similar to variants of other words in comparison with its phonemic transcription, i.e. the number of times its alignment distances with other words' entries is smaller in comparison with the alignment distance with its phonemic form itself. For example, the word */ketɒb/* has a model generated variant */kebɒb/* whose distance to its phonemic form is 1, the calculated confusability count for this variant is 2 because its alignment distances with other words' entries in two cases is smaller than 1. These two cases are related to variants related to word */kabɒb/* (which means roast meat).

A variant of a word is discarded from the lexicon when its confusability count is more than a threshold T_{prune}. In this way, we can reject variants that are almost homophones to already existing lexicon entries. We conducted a set of experiments using different values of T_{prune} to find out the effect of the pruning threshold on the size of the lexicon, the number of survived variants per word and consequently on the performance of ASR system. The results are detailed in the following section.

8. Experiments

Two sets of experiments were conducted to evaluate the performance of our approach. First, we compared the performance of various pronunciation models, by measuring the overall similarity between generated variants of each model to actual variants that occurred in the test set. The test set in this experiment consists of 1 h speech material from the held out speech utterances of Large-FARSDAT. This test set is completely separate from the training set. The performance of each model is measured by first generating variants of each word, which is available in the test set and consequently aligning generated variants of the word with corresponding recognized strings in the test set. The process is done for all existing words in the test set and distances between aligned pairs were normalized to the phonetic lengths of the aligned strings. The alignment was carried out using the standard DP algorithm of NIST (Black, 1999). Eq. (9) shows how the distances between aligned strings were calculated. Eq. (10) gives a way of normalizing the distance to the length of two aligned strings:

$$\text{dist} = N_{\text{sub}} + N_{\text{del}} + N_{\text{ins}}, \tag{9}$$

$$\text{normalized dist} = \frac{\text{dist}}{N_{\text{corr}} + N_{\text{sub}} + N_{\text{del}} + N_{\text{ins}}}, \tag{10}$$

N_{corr}, N_{sub}, N_{del}, N_{ins} are the numbers of phone match, substitution, deletion and insertion in aligned pairs. In this set of experiments, the lower normalized distances introduce the closer generated variants to actual recognized ones, and in turn, show the better performances for the models. Averaged normalized distances for hybrid models, contextual rules and phonemic transcriptions are calculated by conducting separate experiments. The results are summarized in Table 4. Results in the contextual rules row in Table 4 relate to the experiment which uses only contextual rules in generating variants to calculate average normalized alignment distance between generated variants and actual variants of words. The phonemic transcription pronunciation

Table 4
Average normalized distance between aligned model variants and recognized variants for all words in the lexicon

Pronunciation model	Average normalized distance
Dynamic hybrid statistical	0.43
Static hybrid statistical	0.45
Contextual rules	0.47
Phonemic transcription	0.52

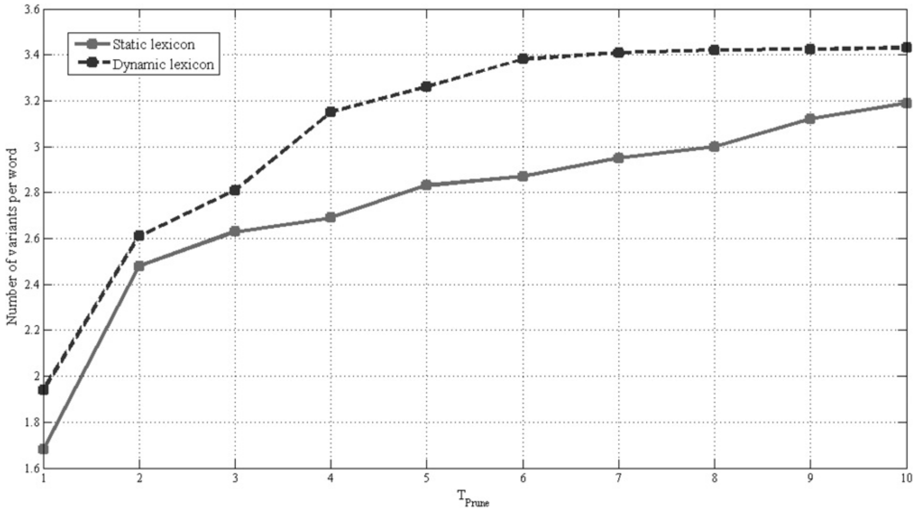

Fig. 9. Average number of pronunciation variants per word for different values of threshold T_{prune}.

model in Table 4 refers to the experiment in which we have used the phonemic transcriptions to align with actual variants. Table 4 reveals that variants generated by hybrids of decision trees and contextual rules are closer to the actual variants, over the test set. Among hybrid models, dynamic ones have better performance by considering the rate of speech while generating pronunciation variants.

In the second set of experiments, we construct lexicons using hybrid static and dynamic pronunciation models. Lexicons are then pruned in the same way described in Section 7, using different values of T_{prune}. This, in turn, produces lexicons with different numbers of variants per word. Fig. 9 shows how the average number of pronunciation variants per word evolves as a function of threshold T_{prune} for dynamic and static lexicons. The dashed line represents dynamic lexicons and the solid line static lexicons.

After generating lexicons with different numbers of variants per word using different T_{prune} values, we use them in a series of recognition experiments on the test set. The experiments evaluate how the word recognition accuracy is improved utilizing hybrid static and dynamic model generated lexicons with different numbers of variants per word. We employed generated lexicons in the SHENAVA Persian ASR system. The baseline SHENAVA dictionary contains only the phonemic transcription for every word. Therefore, the baseline dictionary contains 1200 phonemic entries for the 1200 words it contains. The version of SHENAVA used in our experiments did not employ any language model to decrease its output word error rate. We did not use a language model because we wanted to study the effectiveness of each of the plugged in lexicons without having an interacting system, here the language model. Language models compensate for the errors from their preceding modules. However, using a language model makes it difficult to understand the level of effectiveness of each lexicon. The reason is that the compensatory effect of the language model is highly nonlinear and it shows

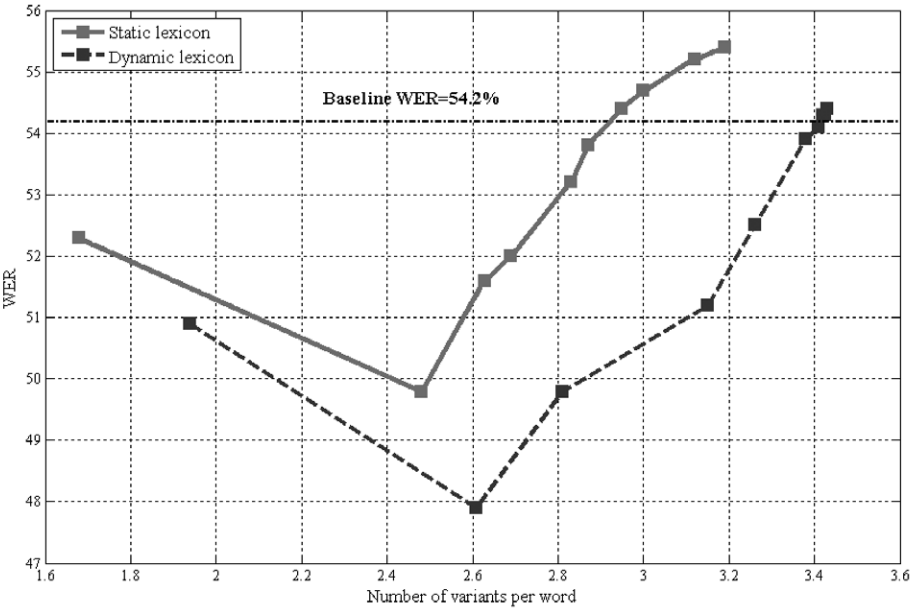

Fig. 10. Word error rates on the test set for different values of threshold T_{prune}. The baseline performance on test set is 54.2%.

Table 5
Reduction in the WER of the ASR system due to the employment of modified lexicon against original lexicon containing only reference forms of words

Pronunciation model	Absolute WERR (%)	Relative WERR (%)
Dynamic hybrid statistical	6.3	11.6
Sialic hybrid statistical	4.4	8.1
Contextual rules	2.9	5.4

different levels of error compensation while working with different lexicons. These experiments were conducted on a set of 100 utterances, taken from the remaining held out part of Large-FARSDAT which is completely separate from the training set. The absolute WER for the baseline system, which utilizes the lexicon with only phonemic version of the words, is 54.2%. The word error rates obtained via plugging dynamic and static lexicons into our recognizer are shown in Fig. 10. Each point in Fig. 10 corresponds to a T_{prune} value which can be read from Fig. 9.

Fig. 10 shows the minimum WERs obtained by using lexicons containing 2.48 and 2.61 words per variant for static and dynamic cases, respectively. In both cases, the pruned lexicons are obtained by $T_{\text{prune}} = 2$. Table 5 summarizes the best results obtained by hybrid static and dynamic models. WER reductions are reported in terms of absolute word error rate reduction (WERR) and relative WERR (against the baseline system). The best result obtained from our previous work (Vazirnezhad et al., 2005a) by using only contextual rules is also inserted in Table 5.

Table 5 reveals that hybrid dynamic pronunciation models make the best improvements in ASR. This is due to the capacity of these models to consider the structural and word level information, while static versions do not take into account the rate of speech which is an important factor in pronunciation variation. An important point in our work is the fact that the hybrid models are designed, at the level of building blocks, to efficiently deal with data insufficiency. Generalized decision trees and contextual rules are trained with a medium-size training set of only 25 h. We have designed a hybrid of generalized decision trees and contextual rules which benefits from advantages of each of its building blocks. It should also be noted that contextual rules are able to model co-articulation phenomena efficiently but they are not basically sensitive to any word level information, while generalized decision trees in the hybrid approach compensate for this drawback too. Although the reported improvements are obtained using a Persian ASR system, a comparison between the obtained results and those reported in similar research performed on English and other languages, some of which are reported in the next section, confirms the efficiency of our approach in dealing with any medium-size training set.

9. Discussions

Performing quantitative comparison between various researches on pronunciation modeling is rather difficult as these studies are carried out in domains of various languages and are evaluated on different tasks and circumstances using different recognition engines. However, we believe that it would be useful to review the results of some valuable research in this field to clearly demonstrate the place of the present research and its specific contributions among other works.

Fukada et al. (1999) trained a neural network to generate multiple pronunciations of words. Spontaneous dialogues from a total of 230 speakers (100 males and 130 females) were used for pronunciation and acoustic model training. Canonical pronunciations with quintphone context and their corresponding realized pronunciations (about 120,000 samples in total) were used as inputs and outputs for the pronunciation network training. They achieved 3.4% WER improvement for a simple dictionary at the baseline of 34.5% WER. They also reported a 2.6% WER improvement for their expert dictionary at the baseline of 29.0% WER.

Cremelie and Martens (1999) introduced a method that relied on a pronunciation rule formalism that respects a hierarchy within a set of rules and imposes a number of constraints under the form of negative rules. They produced pronunciation variants of each word by applying these rules wherever applicable, with regard to phonemic context. Their experiments showed that the introduction of such variants in a segment-based recognizer significantly improves the recognition accuracy. On TIMIT, which is a medium-size read speech corpus containing 6300 sentences, an absolute word error rate reduction of 1.41% at a baseline of 8.39% was obtained.

Fosler-Lussier (1999a) used decision trees to capture phonetic pronunciation variations of words and syllables in spontaneous speech. He built models for syllables and words which could dynamically change the pronunciations used in the speech recognizer based on the extended context, including surrounding words, phones, speaking rate, etc. Implementation of the new pronunciation models automatically derived from data using the ICSI speech recognition system showed a 4–5% relative improvement on the Broadcast News recognition task. When the full 200 h training set was used, word trees gained 0.3–0.6% over the baseline (depending on the test set) compared to 0–0.5% gain for the syllable trees.

An absolute WER reduction of as high as 6.3% at the baseline of 54.2% WER is obtained in our work as the best result, which is equivalent to a relative WER reduction of 11.6%. This is a noticeable result among reported results in similar works. This result is obtained while we used a medium-size speech corpus of 25 h read speech as our training set. Our work is comparable to the work of Fosler-Lussier (1999a) from the point of view that in both of the works, pronunciation variation is modeled on the word level. Meanwhile, we have focused on the issue of designing models which could be trained with a far smaller corpus. This is done by implementing the idea of sharing the information on pronunciation of words with the same syllabic structures and making generalizations over pronunciation models. The other important advantage of our work is that in case of adding new words to the lexicon, that do not exist in the training set, we do not need an extra training corpus which must contain enough realizations of the new words, because the already trained models are sufficient to generate pronunciation variants of all the new words. We quantitatively evaluated the effects of some factors such as rate of speech, word predictability, syllable location and stress on pronunciation variation, and employed them effectively in modeling pronunciation variations. We introduced hybrid pronunciation models in two static and dynamic versions. Dynamic versions use information on rate of speech to dynamically update the lexicon entries in an adaptation to this factor. Results show that employing rate of speech in generating the lexicon entries, effectively improves the performance of ASR in comparison with the static version.

10. Conclusions

In this paper, we introduced a novel technique to combine decision trees and contextual rules in the framework of hybrid statistical models to generate word pronunciation variants. Hybrid statistical models consider the word structure, word level features and also co-articulation phenomena, simultaneously. The idea of generalized decision trees effectively handles blurred phone identities commonly found in conversational speech, instead of making a hard separation of phones. Results show that the pronunciation variants which are generated by these hybrid statistical models are closer to recognized variants in comparison with variants which are produced only by contextual rules. Hybrid models are combinations of contextual rules and generalized decision trees and benefit from the advantages of both. The hybrid models solve the drawback of contextual rules, which work regardless of word level information such as phonological structure of the word and stress. The word level constraints are marked by generalized decision trees, and then contextual rules will be applied only on marked regions. Pronunciation models which are designed at the word level often need a very large training corpus with sufficient number of realizations of each word. However, in our approach, the hybrid architecture is designed, in the level of its building blocks, to avoid the need for a large corpus. This was carried out by the idea of generalization over trees, quantization of code vectors of pronunciation patterns to decrease the large variety of patterns and classification of phonemes and exploiting the class of a phoneme instead of the phoneme itself in prediction process by trees. In this manner, hybrid models are able to be trained by a medium-sized corpus. Introducing variants, generated by the hybrid models, to the lexicon of an ASR system, SHENAVA, led to a relative WER reduction of as high as 11.6% which is a noticeable result. This shows the capacity of the generalized decision trees in employing word-level information to model pronunciation variation of words, while keeping the capability of using only a medium-size speech corpus to be trained. Moreover, the trained models respond to all input words and we do not need to be concerned about future new entries of the lexicon in the possible new tasks.

Acknowledgements

The authors would like to express their sincere thanks to Iran Telecommunication Research Center (ITRC) and Research Center of Intelligent Signal Processing (RCISP) for their continual support of the present work. We would also like to thank Dr. Homayounpour, Dr. Sameti, Dr. Sayadian, Dr. Seyyed Salehi, Dr. Bijankhan, Dr. Eslami and Mr. Honnibal for their motivating feedback and guidance in this work.

References

Adda Decker, M., Boula de Mareuil, P., Adda, G., Lamel, L., 2005. Investigating syllabic structures and their variation in spontaneous French. Speech Communication 46 (2), 119–139.

Almasganj, F., Seyedsalehi, S.A., Bijankhan, M., Sameti, H., Sheikhzadegan, J., 2001. SHENAVA-1: Persian spontaneous continuous speech recognizer. In: Proceedings of the International Conference on Electrical Engineering, pp. 101–106 (in Persian).

Bijankhan, M., Sheikhzadegan, M.J., 1994. FARSDAT – the Farsi spoken language database. In: Proceedings of the International Conference on Speech Sciences and Technology (2), pp. 826–829.

Bijankhan, M., Sheikhzadegan, M.J., Roohani, M.R., Zarrintare, R., Ghasemi, S.Z., Ghasedi, M.E., 2003. Tfarsdat – the telephone Farsi speech database. In: Proceedings of the EUROSPEECH 2003, pp. 1525–1528.

Black, P.E., 1999. Algorithms and Theory of Computation Handbook. CRC Press/LLC.

Chen, K., Hasegawa-Johnson, M., 2004. Modeling pronunciation variation using artificial neural networks for English spontaneous speech. In: Proceedings of the ICSLP-04, pp. 1461–1464.

Cremelie, N., Martens, J.-P., 1999. In search of better pronunciation models for speech recognition. Speech Communication 29 (2–4), 115–136.

Davel, M., Barnard, E., 2006. Bootstrapping pronunciation dictionaries. South African Journal of Science 102 (7–8), 322–328.

Fosler-Lussier, E., 1999a. Dynamic pronunciation models for automatic speech recognition. Ph.D. Thesis, University of California, Berkeley, CA.

Fosler-Lussier, E., 1999b. Contextual word and syllable pronunciation models. In: Proceedings of the 1999 IEEE Workshop on Automatic Speech Recognition and Understanding, pp. 16–23.

Fosler-Lussier, E., Morgan, N., 1999. Effects of speaking rate and word frequency on pronunciations in conversational speech. Speech Communication 29 (2–4), 137–158.

Fukada, T., Yoshimura, T., Sagisaka, Y., 1999. Automatic generation of multiple pronunciations based on neural networks. Speech Communication 27 (1), 63–73.

Haghshenas, A.M., 1995. Phonology. Agah Publishers, Tehran (in Persian).

Hazen, T., Hetherington, L., Shu, L., Livescu, K., 2005. Pronunciation modeling using a finite-state transducer representation. Speech Communication 46 (2), 189–203.

Humphries, J., 1997. Accent modeling and adaptation in automatic speech recognition. Ph.D. Thesis, University of Cambridge, Cambridge.

Imai, T., Ando, A., Miyasaka, E., 1995. A new method for automatic generation of speaker-dependent phonological rules. In: Proceedings of the ICASSP-95, pp. 864–867.

Jande, P.A., 2008. Spoken language annotation and data-driven modeling of phone-level pronunciation in discourse context. Speech Communication 50 (2), 126–141.

Jurafsky, D., Bell, A., Gregory, M., Raymond, W.D., 2001. The effects of language model probability on pronunciation reduction. In: Proceedings of the ICASSP, vol. 2, pp. 801–804.

Kovesi, B., Boucher, J.M., Saoudi, S., 2001. Stochastic K-means algorithm for vector quantization. Pattern Recognition Letters 22, 603–610.

Ladefoged, P., 2006. A course in phonetics. Thomson Learning, fifth ed.

Milanian, H., Hosseini, R., 2002. Linguistics: paper collection. Iranian Ministry of Culture, Tehran (in Persian).

Randolph, M., 1990. A data-driven method for discovering and predicting allophonic variation. In: Proceedings of the ICASSP-90, pp. 1177–1180.

Riley, M., 1991. A statistical model for generating pronunciation networks. In: Proceedings of the ICASSP-91, pp. 737–740.

Saraclar, M., Khudanpur, S., 2004. Pronunciation change in conversational speech and its implications for automatic speech recognition. Computer Speech & Language 18 (4), 375–395.

Schmid, P., Cole, R., Fanty, M., 1993. Automatically generated word pronunciations from phoneme classifier output. In: Proceedings of the ICASSP-93, pp. II-223–II-226.

Sloboda, T., 1995. Dictionary learning performance through consistency. In: Proceedings of the ICASSP-95, pp. 453–456.

Sloboda, T., Waibel, A., 1996. Dictionary learning for spontaneous speech recognition. In: Proceedings of the ICSLP-96, pp. 2328–2331.

Strik, H., Cucchiarini, C., 1999. Modeling pronunciation variation for ASR: a survey of the literature. Speech Communication 29, 225–246.

Vazirnezhad, B., Almasganj, F., Bijankhan, M., 2005a. Automatic extraction of contextual rules and generating pronunciation variants to use in automatic continuous speech recognition. Journal of Computer Science and Engineering 3 (3), 40–50 (in Persian).

Vazirnezhad, B., Almasganj, F., Bijankhan, M., 2005b. A hybrid statistical model to generate pronunciation variants of words. In: Proceedings of the IEEE NLP-KE 05, pp. 106–110.

Wester, M., 2003. Pronunciation modeling for ASR, knowledge-based and data-derived methods. Computer Speech & Language 17 (1), 69–85.

Wooters, C., Stolcke, A., 1994. Multiple-pronunciation lexical modeling in a speaker independent speech understanding system. In: Proceedings of the ICSLP-94, pp. 1363–1366.

Yu, H., Schultz, T., 2003. Enhanced tree clustering with single pronunciation dictionary for conversational speech recognition. In: Proceedings of the EUROSPEECH 2003, pp. 1869–1872.