Journal of Zhejiang University SCIENCE A
ISSN 1673-565X (Print); ISSN 1862-1775 (Online)
www.zju.edu.cn/jzus; www.springerlink.com
E-mail: jzus@zju.edu.cn

Speaker adapted dynamic lexicons containing phonetic deviations of words

Bahram VAZIRNEZHAD[†1,3], Farshad ALMASGANJ[1], Seyed Mohammad AHADI[2], Ari CHANEN[3]

(*[1]Biomedical Engineering Department, Amirkabir University of Technology, Hafez Avenue, Tehran, Iran*)

(*[2]Electrical Engineering Department, Amirkabir University of Technology, Hafez Avenue, Tehran, Iran*)

(*[3]Language and Knowledge Management Research Lab, School of Information Technologies, University of Sydney, NSW, Australia*)

[†]E-mail: bvazirnezhad@aut.ac.ir

Received Dec. 18, 2008; Revision accepted Apr. 4, 2009; Crosschecked July 20, 2009; Online published Aug. 14, 2009

Abstract: Speaker variability is an important source of speech variations which makes continuous speech recognition a difficult task. Adapting automatic speech recognition (ASR) models to the speaker variations is a well-known strategy to cope with the challenge. Almost all such techniques focus on developing adaptation solutions within the acoustic models of the ASR systems. Although variations of the acoustic features constitute an important portion of the inter-speaker variations, they do not cover variations at the phonetic level. Phonetic variations are known to form an important part of variations which are influenced by both micro-segmental and suprasegmental factors. Inter-speaker phonetic variations are influenced by the structure and anatomy of a speaker's articulatory system and also his/her speaking style which is driven by many speaker background characteristics such as accent, gender, age, socioeconomic and educational class. The effect of inter-speaker variations in the feature space may cause explicit phone recognition errors. These errors can be compensated later by having appropriate pronunciation variants for the lexicon entries which consider likely phone misclassifications besides pronunciation. In this paper, we introduce speaker adaptive dynamic pronunciation models, which generate different lexicons for various speaker clusters and different ranges of speech rate. The models are hybrids of speaker adapted contextual rules and dynamic generalized decision trees, which take into account word phonological structures, rate of speech, unigram probabilities and stress to generate pronunciation variants of words. Employing the set of speaker adapted dynamic lexicons in a Farsi (Persian) continuous speech recognition task results in word error rate reductions of as much as 10.1% in a speaker-dependent scenario and 7.4% in a speaker-independent scenario.

Key words: Pronunciation models, Continuous speech recognition, Lexicon adaptation
doi:10.1631/jzus.A0820761 **Document code:** A **CLC number:** TP391.42

INTRODUCTION

The speech signal is highly variable. Some of these variations are speaker-dependent and others are not. Generally, any kind of variation decreases the performance of an automatic speech recognition (ASR) system, if not treated carefully. In this paper, we will introduce a way for compensating for some of these variations through creating pronunciation variants of the words involved in an ASR lexicon. This is done by employing a hybrid word pronunciation variation generator scheme which considers some of the phonetic deviations that occur because of intra-speaker variations. This phenomenon is a result of differences between speakers in respect to their voice quality, speaking style, dialect, etc. These factors cause partial deviations of acoustic features. These partial deviations cause systematic errors in the phone recognizer module and as a consequence, phonetic deviations occur in the recognized strings of phones. To the extent that the source of variability is systematic, this phenomenon can be described and modeled, which should lead to ways to handle it successfully.

Phonetic variations occurring because of speaker varieties are a well known phenomenon arising because of different mental and physical specifications of the speakers that affect their speech generation.

The degree to which this phenomenon occurs will vary depending on various factors including speaker specifications, speaking styles, and rate of speech (ROS) (Strik and Cucchiarini, 1999). Moreover, words are strung together in continuous speech and many sorts of interactions may take place between words, resulting in various phonological processes. This causes partial alteration in the pronunciations of words from their isolated form.

In ASR systems, one way to overcome this problem is to develop a lexicon containing the most likely phonetic realizations of words to describe how the entries can be recognized. In the literature, this method is referred to as explicit phonetic variation modeling. A distinction should be made between techniques that model pronunciation, in the sense that words may be pronounced with various allowed phonetic forms, and those that model phone recognizer errors resulting from partial pronunciation variation in acoustic features. Fig.1 shows the two mechanisms which cause phonetic deviations arising from speaker pronunciation.

In this paper, the introduced method has the capability to model both these mechanisms of phonetic deviations at the level of words. This may be considered in the development of lexicons used in ASR systems. Our previous work (Vazirnezhad *et al*., 2009) focused on modeling the first deviation mechanism (Fig.1). However, in this paper, we are going to introduce pronunciation models which consider both deviation mechanisms, simultaneously. The developed models use acoustic features of the speakers to predict phonetic deviations of the recognized phone strings, while still using the factors shown previously to influence pronunciation.

Recently, researchers have developed techniques to automatically generate phonetic realizations of words corresponding to the first mechanism (Fig.1). Experiments have shown that using a lexicon which contains appropriate pronunciation variants of lexical entries improves the performance of ASR systems (Cremelie and Martens, 1999; Fosler-Lussier, 1999; Fukada *et al*., 1999). In data-driven methods, the hand-labeled phonetic transcriptions of the utterances are aligned with their corresponding phonemic transcriptions obtained by concatenating the transcriptions of individual words. Alignment is achieved by means of a dynamic programming (DP) algorithm. The resulting DP alignments can then be used to train statistical models, including decision trees, artificial neural networks, etc., to predict the phonetic pronunciations of words.

In the early 1990s, some medium-size databases (e.g., TIMIT and Resource Management) were transcribed phonetically, which motivated several researchers to study phonetic pronunciation variation (Randolph, 1990; Riley, 1991; Wooters and Stolcke, 1994). Although these approaches were developed to automatically capture pronunciation rules, manual transcription was still required. Researchers later solved this problem by employing phone recognizers which enable one to transcribe a large corpus (Schmid *et al*., 1993; Imai *et al*., 1995; Sloboda, 1995; Humphries, 1997). Recently, large vocabulary continuous speech recognition (LVCSR) systems have the ability to recognize spontaneous and conversational speech, such as the Switchboard corpus. Word pronunciation variations occur more frequently in conversational speech than in read speech, as a result pronunciation modeling has become more important (Fosler-Lussier and Morgan, 1999; Saraclar and Khudanpur, 2004).

Some researchers extracted contextual rules to model word-level phone variations using data-driven approaches, to generate pronunciation variants of the words from their phonemic transcriptions (Cremelie and Martens, 1999; Vazirnezhad *et al*., 2005a). Different types of features, including phoneme context, speaking rate, speaker specifications, etc., have been used to train the decision trees (Fosler-Lussier, 1999;

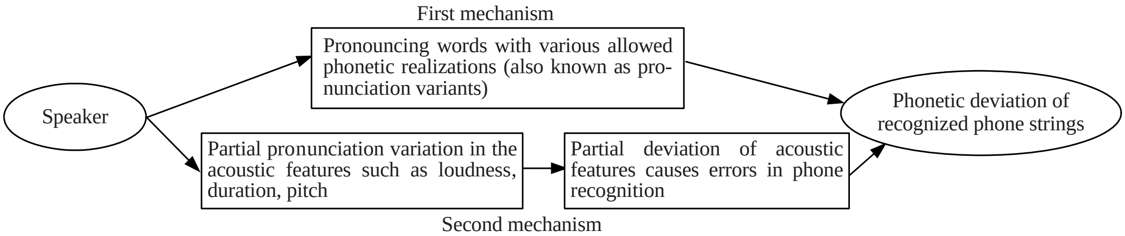
First mechanism
Pronouncing words with various allowed phonetic realizations (also known as pronunciation variants)
Speaker
Phonetic deviation of recognized phone strings
Partial pronunciation variation in the acoustic features such as loudness, duration, pitch
Partial deviation of acoustic features causes errors in phone recognition
Second mechanism

Fig.1 Two mechanisms involved in phonetic deviation of the recognized phone strings arising from speaker pronunciation

Vazirnezhad *et al*., 2005b; Jande, 2008). It has been 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 (Fosler-Lussier, 1999). These models use speaking rate and *n*-gram information to generate pronunciation variants in a dynamic framework. It has been 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 is used to predict pronunciation variants of word segments (Fukada *et al*., 1999). Moreover, it has been shown that the pitch accent can improve the prediction of pronunciation variation (Chen and Hasegawa-Johnson, 2004). Other approaches have also been proposed. For example, Hazen *et al*.(2005) used a finite-state transducer 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 finding the word variants.

In this paper, the hybrid statistical structure for automatic generation of pronunciation variants of words introduced by Vazirnezhad *et al*.(2009) is modified to be able to generate lexicons adapted to specific clusters of speakers. In other words, we involve acoustic features of each of the specific speaker clusters in the process of predicting phonetic deviations of words. The newly configured hybrid models are composed of decision trees and adapted contextual rules. The term ‘adapted contextual rules’ refers to the fact that a distinct set of contextual rules is extracted for each of the speaker clusters. Decision trees predict regions in words that are susceptible to phonetic change, and appropriate speaker adapted contextual rules are applied to the permissible regions to generate the pronunciation variants of words. Hybrid models take into account the syllabic structure of the word and ask questions about phone identities, rate of speech, unigram statistics, and the position of the stressed syllable, simultaneously. In the final step, phonemic context information is considered by contextual rules. It should be emphasized that in our proposed method, each decision tree is not assigned to only one word, as in Fosler-Lussier (1999), but to a group of words with similar phonological structures. Hence, to describe them better, they are called ‘generalized decision trees’. However, while describing our approach, we occasionally use the short term ‘decision tree’ or ‘tree’ instead of ‘generalized decision tree’, for the sake of simplicity. By using such generalized decision trees, we do not need speech data prepared distinctively for each of the investigated words to train its pronunciation model.

The remainder of this paper is organized as follows. An overview of the hybrid models is provided in the next section. The speech corpus ‘Large-FARSDAT’, which is used in training and experiments, and the ‘SHENAVA’ Farsi ASR system, which is used as our experimental workbench, are then introduced. Consequently the speaker clustering procedure is discussed. The influence of the number of clusters on the intra-cluster phonetic variations and the tradeoff problem on this issue are then addressed. Speaker adapted contextual rules and approaches in learning them are presented in the following section. The next section discusses how the set of adapted dynamic lexicons can be used in a speech recognition task. Experiments, discussion and conclusions are considered in the last three sections.

OVERVIEW OF THE HYBRID MODEL

The proposed pronunciation models are hybrids of generalized decision trees and sets of speaker adapted contextual rules. Models use the phonological structure of words, rate of speech, word unigram statistics and the location of the stressed syllable in lexical entries in addition to phonemic context information, to generate phonetic deviations. The adapted contextual rules are the sets of phonological rewrite rules which are extracted for each category of speakers separately, and which contain information on the speaking style of the speaker category at the phonetic level.

Each generalized decision tree is designed for all words with the same phonological structure as described by Vazirnezhad *et al*.(2009). For example, the pronunciation variants of all lexical entries which have the same arrangement of consonants (represented by C) and vowels (represented by V), like the string of ‘CVCVC’, are used to train the same decision tree. This approach is taken to solve the problem of sparse training data, as the training database does not need to contain all words supposed to be in the lexicon. Also, no new extra corpus is needed for new words. In fact, by sharing information of pronunciation variation in structurally similar words, the amount of necessary training data is decreased and the problem of introducing 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 to highlight the major differences between our approach and those of previous studies. Comprehensive details of generalized decision trees and training algorithms are provided in Vazirnezhad *et al*.(2009).

Hybrid statistical models generate phonetic deviations of words in two main steps (Fig.2). In the first step, the decision tree, corresponding to the syllabic structure of the input word, identifies the phonemes in the word which are likely to be substituted or deleted, and the locations where an insertion can take place. Phonological structure and arrangements of the consonants and vowels of the input word are considered to choose the corresponding tree. For instance, the word 'كتاب', which has the phonemic transcription of /*ketɒb*/, meaning 'book' in Farsi, is a word with two syllables with a 'CVCVC' phonemic pattern. Hence, the corresponding generalized decision tree would be the 'CVCVC' tree, which is trained to predict pronunciation variation of the words with the same syllabic structure. To predict the pronunciation patterns, the generalized decision tree asks for the identity of consonants and vowels defined by their membership to different phonetic categories, the speaking rate, the logarithm of unigram probability of the word and the location of the stressed syllable. Pronunciation pattern defines which phonemes are likely to be substituted or deleted and also determines where an insertion can take place. Each of the predicted pronunciation patterns is accompanied with a likelihood determined by the tree. We have set a cut-off threshold of probability to limit the number of accepted variation patterns.

In the second step, the set of adapted contextual rules, corresponding to a speaker cluster, is applied to the phonemes that are candidates to be altered, to generate the adapted lexicons. There are two types of contextual rules: substitution rules and insertion rules. Substitution rules can be represented as $L\underline{F}R \rightarrow LOR$, where phonemic string F can be recognized as string O because of explicit pronunciation or phone recognizer errors, when it is surrounded by a left context, L, and a right context, R. Insertion rules are formalized as $LR \rightarrow LOR$, where string F is empty and string O can be inserted in an LR context. The substitution rules will be applied to the regions that are chosen by the tree for substitution, and the insertion rules can be applied to the regions in which insertion is predicted. Finally, the probabilities of the created variants of a word are normalized to have a summation equal to unity. Applying various sets of adapted contextual rules results in a number of adapted lexicons for various speaker clusters.

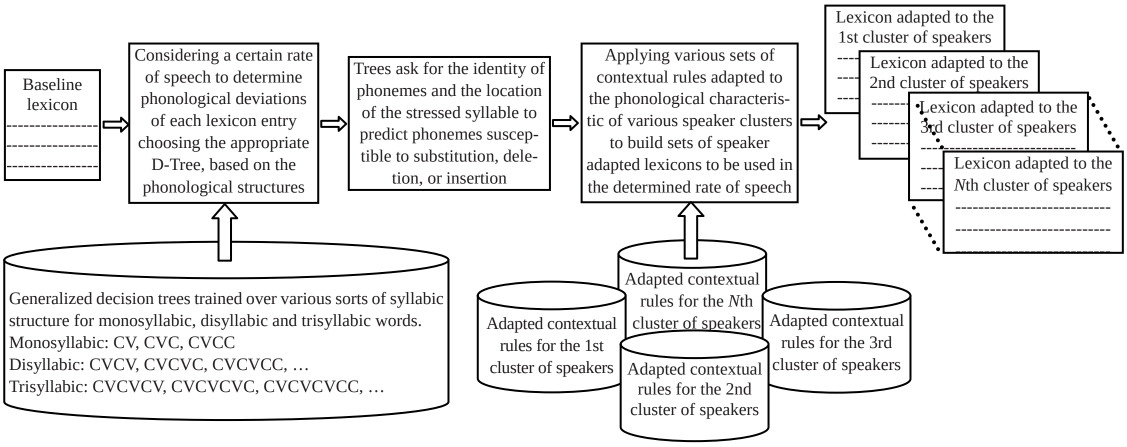
Baseline lexicon
Considering a certain rate of speech to determine phonological deviations of each lexicon entry choosing the appropriate D-Tree, based on the phonological structures
Trees ask for the identity of phonemes and the location of the stressed syllable to predict phonemes susceptible to substitution, deletion, or insertion
Applying various sets of contextual rules adapted to the phonological characteristic of various speaker clusters to build sets of speaker adapted lexicons to be used in the determined rate of speech
Lexicon adapted to the 1st cluster of speakers
Lexicon adapted to the 2nd cluster of speakers
Lexicon adapted to the 3rd cluster of speakers
Lexicon adapted to the Nth cluster of speakers
Generalized decision trees trained over various sorts of syllabic structure for monosyllabic, disyllabic and trisyllabic words.
Monosyllabic: CV, CVC, CVCC
Disyllabic: CVCV, CVCVC, CVCVCC, …
Trisyllabic: CVCVCV, CVCVCVC, CVCVCVCC, …
Adapted contextual rules for the 1st cluster of speakers
Adapted contextual rules for the Nth cluster of speakers
Adapted contextual rules for the 3rd cluster of speakers
Adapted contextual rules for the 2nd cluster of speakers

Fig.2 Block diagram of the hybrid structure used to generate phonetic deviations of lexical entries

There are a number of adapted contextual rules, equivalent to the number of speaker clusters, whose application will result in a number of adapted lexicons

We have clustered speakers in the training dataset by considering their acoustic features. When a new speaker begins to utter words, a fast algorithm extracts the acoustic vector and identifies the speaker cluster to which the new speaker is most closely matched. The adapted lexicon, in which the adapted set of contextual rules related to this cluster was used in generating lexical entries, will then be used to decode utterances of the current speaker.

LARGE-FARSDAT—FARSI SPEECH CORPUS

The corpus used in this research is Large-FARSDAT. This is a Farsi speech database (Bijankhan and Sheikhzadegan, 1994; Bijankhan *et al*., 2003). In this research, only half of the Large-FARSDAT data was used for training pronunciation models. This portion of Large-FARSDAT includes 50 speakers with a total of 25 h of speech. Large-FARSDAT contains only phonemic transcriptions, and is not phonetically labeled. To have access to the recognized phone strings of utterances, we applied speech utterances to the phone recognizer of the SHENAVA-2 ASR system. The procedure was explained in detail in Vazirnezhad *et al*.(2009). Comparison of aligned pairs of phonemic and recognized transcriptions of speech utterances shows that the training algorithm can capture phonetic deviations arising from both speakers' pronunciations and errors of the SHENAVA-2 phone recognizer simultaneously.

SHENAVA-2—EXPERIMENTAL WORKBENCH

SHENAVA-2 is a Farsi ASR system with a vocabulary containing 1200 words. The front-end section in this system consists of a hybrid of two multilayer perceptron (MLP) neural networks and a rule-based engine. SHENAVA-2 has a lexicon search based on a semi-viterbi algorithm to find the best 100 recognized phrases. Finally, an *N*-best rescoring block, which uses hidden markov models (HMMs) of Farsi phones, finds the best output phrase. SHENAVA-2 is completely introduced in Almasganj *et al*.(2001). We use a version of the SHENAVA-2 ASR system as our experimental workbench in this paper, which does not consider any language model.

SPEAKER CLUSTERING

Mel frequency cepstral coefficients (MFCCs) and its derivatives are commonly used features in speech recognition tasks and many other speech processing applications (Davis and Mermelstein, 1980; Skorik and Berthommier, 2000; Padrell *et al*., 2005). A filter bank is applied to the spectrum of each speech frame to extract these coefficients. The center frequencies of the filters are linearly distributed on the 'Mel' scale. The coefficients are the discrete cosine transform coefficients of the logarithm of the filters' output energies. In this study, we have used 13 MFCCs. The length and overlap of frames are 256 and 128 samples, respectively (11.6 ms and 5.8 ms with a sampling rate of 22050 Hz). The first derivatives of MFCCs for consecutive frames in a segment are given by

$$\begin{gathered}\Delta MFCC_i[n]=C_i[n+1]-C_i[n],\\ 1\le n\le N-1,\ 1\le i\le 13,\end{gathered}\tag{1}$$

where $C_i(n)$ is the *i*th coefficient of the *n*th frame in an utterance and *N* is the total number of frames. Consequently, MFCCs and their first derivatives are averaged over utterances corresponding to each speaker to give 26 MFCC-based components. The mean values for logarithm of energy and its derivative are the final components appending to produce a 28-dimensional feature vector for each speaker. These feature vectors are the basis for speaker clustering. A simple *K*-means algorithm is then used to cluster 50 speakers in the training corpus.

INTRA-CLUSTER PHONETIC VARIATION CRITERION

The intra-cluster phonetic variation criterion is defined to measure the average level of diversity among speakers within clusters in terms of phonetic pronunciation habits. This criterion is variable owing to variation in the number of clusters. Where there are fewer clusters, there will be more diversity among speakers within clusters and therefore the intra-cluster variation criterion will be higher. The approach in this paper is to build adapted lexicons, specified to certain clusters of speakers, to represent phonetic pronunciation habits for speakers belonging to each of the clusters. Therefore, to train distinctive adapted lexicons, it is desirable to have the least possible variation within a cluster, or the highest similarity among speakers in the cluster.

In theory, training a lexicon for every single speaker would be an ideal approach but it is hardly practical as it is inefficient to provide sufficient training data for each of the speakers, and moreover this approach may work only for speaker-dependent systems. The alternative idea examined in this paper is to build lexicons over clusters of speakers rather than for each individual speaker. In this approach, finding the optimum number of clusters is an important issue; the optimum number of clusters should result in an acceptably low intra-cluster variation criterion and should also allow sufficient training material for each cluster. Thus, finding a reasonable number of clusters is a trade-off problem between the requirements of low intra-cluster variation and sufficient training data. Determining a reasonable number of clusters and training adapted lexicons over resulting clusters lead to a lower word error rate (WER) as shown by experiments. In the following subsections, the intra-cluster variation criterion is first defined; next, the evolution of this parameter against the number of clusters is studied; finally, the trade-off problem, which plays the main role in determining a reasonable number of clusters, is discussed.

Algorithm for computation of the intra-cluster phonetic variation criterion

In this subsection, the intra-cluster phonetic variation criterion is defined. This criterion is a measure of the average speakers' diversity within clusters in the sense of phonetic deviations. The criterion is defined based on phone confusion matrices of speakers, and measures the diversity of speakers by quantifying patterns of phonetic deviations from standard pronunciation.

For each of the speakers in the training corpus, a confusion matrix is calculated by comparing the aligned standard phones and recognized phones corresponding to the utterances for the speaker. There are 29 phonemes in Farsi; considering deletion and insertion, the phone confusion matrix will be 30×30 in size. Therefore, the phone confusion matrix for every speaker consists of probabilities of pronouncing phonemes in standard pronunciations as alternative phones. These matrices are transformed to 900-component vectors by concatenating the rows of the confusion matrix. We call this vector the 'confusion vector'. The intra-cluster variation criterion can be calculated after clustering speakers by the given number of clusters. Four steps in the computation of the intra-cluster variation criterion are mentioned here.

In the first two steps, the components in the confusion vectors that have a significant amount of speaker-dependent variability are identified to be used in the computations. Components with low variance across speakers are not believed to be speaker-dependent, and are not involved in the algorithm. Discarded components include those cells which correspond to rare substitutions, for example, the component corresponding to the substitution of /a/ to /s/. The probability values for these components are almost always close to zero and hence their variances will be also zero. The first two steps are as follows:

1. 900 variances are calculated corresponding to each of the 900 components in the speakers' confusion vectors, over all of the speakers in the whole training dataset.

2. Variances are compared against a certain threshold, T_{var}. Only the components with variances above T_{var} are used in the remaining computations.

In Step 3, the average value for variances within clusters is calculated for the given clusters. In Step 4, this value is normalized to the value of the same parameter when the number of clusters is 1; this is actually equivalent to not clustering.

3. The sum of variances of the surviving components of the confusion vectors is calculated in each cluster and these sums are then averaged over selected components and over clusters. This is shown in Eq.(2), where the average value for variances for a given number of clusters, K, and a given number of selected components, N, is called $APPD(K)$, in which $APPD$ stands for 'average phonetic pronunciation diversity'.

$$APPD(K) = \frac{1}{K \times N} \sum_{k=1}^{K} \sum_{n=1}^{N} \sum_{m=1}^{M_{\{k\}}} \frac{1}{M_{\{k\}}} \left(x_{\{kn\}}[m] - \mu_{\{kn\}} \right). \tag{2}$$

In Eq.(2), k is the index for the cluster, K is the number of clusters, n is the index for component in the confusion vectors, N is the number of selected components, m is the speaker index within a certain cluster, and $M_{\{k\}}$ is the number of speakers within a certain cluster that is dependent on the chosen cluster, k. $x_{\{kn\}}[m]$ is the value of the nth component in the confusion vector for a certain speaker within cluster k for the mth speaker or confusion vector. $\mu_{\{kn\}}$ is the average value of the nth component of the confusion vector over speakers in cluster k.

4. Finally, the intra-cluster variation criterion is normalized by dividing $APPD(K)$ by the distortion ceiling, $APPD(1)$, which is the intra-cluster distortion obtained if all speakers are in the same cluster, as shown in Eq.(3):

$$Criterion = \frac{APPD(K)}{APPD(1)}. \qquad (3)$$

Based on Eq.(3), the intra-cluster variation criterion is dependent on the number of clusters, K. $APPD(1)$ is the average phonetic pronunciation diversity when K=1. The intra-cluster variation criterion should be low to support the training of cluster-specific lexicons.

Influence of the number of clusters on the intra-cluster variation criterion

To investigate the influence of the number of speaker clusters on the intra-cluster phonetic variation criterion, 50 speakers in the training data were clustered using their acoustic feature vectors. The clustering was done using the conventional K-means algorithm. The number of clusters was varied from 50 to 1, i.e., from one cluster per speaker to no clustering. Fig.3 shows how the proposed intra-cluster variation criterion evolves as the number of clusters is changed.

The threshold T_{var} used in this experiment was set to be 0.0001. T_{var} is chosen experimentally to capture components from the confusion vectors which are believed to be speaker-dependent. These components have high variance across speakers. Most of the components which correspond to phone substitutions independent of speaker variability are discarded and are not included in computations. Discarded components include mostly those elements which correspond to rare phone substitutions with very low probability values (almost 0) in confusion matrices. Using this threshold, about 200 components from the confusion vectors are chosen to be used in the computation of the intra-cluster phonetic variation criterion.

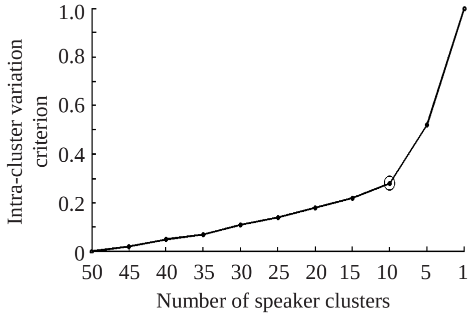

Fig.3 Evolution of the intra-cluster variation criterion vs the number of speaker clusters

These observations can be used in assessing the usability of various numbers of clusters, K. Increasing the number of clusters has the positive effect of decreasing the intra-cluster phonetic variation. However, having more clusters is not necessarily a good choice, because it means that the training data will be divided among more clusters. Sometimes, this leads to insufficient data for some clusters, which may result in an improperly trained lexicon for those clusters. Moreover, having more clusters means more lexicons, which is computationally costly. In addition to these two main disadvantages, having a high number of speaker clusters during a real-time speech recognition task creates difficulties in finding the nearest cluster for the speaker and in using the most appropriate lexicon in recognition.

Number of clusters and the trade-off between intra-cluster phonetic variations and training data requirements

The approach introduced in this paper is to train distinctive lexicons specified to certain clusters of speakers. Training a distinctive lexicon for a cluster is based only on the portion of training data from speakers belonging to that cluster. To have lexicons representing specific groups of speakers, low intra-cluster variation is required, so that lexicons may be trained more specifically. Thus, the number of clusters should not be too low, because in this case diverse speakers will be put in the same cluster and this would significantly increase the intra-cluster variation.

Training a separate lexicon for each of the speakers may be better in terms of specificity, but having insufficient training data for every single speaker makes this approach impractical. Moreover, this method works only for speaker-dependent systems with a closed set of speakers. Therefore, the number of clusters should also not be very high because it would require the training data to be divided into a large number of clusters. Hence the portion of the training data for each cluster would be smaller and insufficient to train a proper lexicon.

Therefore, there is a trade-off problem in choosing a reasonable number of clusters. The trade-off is between specificity of lexicons and sufficiency of the portion of training data assigned to every cluster. Specificity of lexicons can be obtained by having lower intra-cluster variation which is achievable having more clusters, but sufficient training data can be obtained more easily with fewer clusters.

Fig.3 depicts how intra-cluster phonetic variation evolves as a function of the number of clusters. Decreasing numbers of clusters increase the intra-cluster variation criterion. However, the rate of increase for the criterion is lower when clusters decrease from 50 towards 10 than when clusters decrease from 10 to 1. Thus, decreasing the number of clusters from 50 to 10 increases the criterion to only less than one-third of its highest value, while decreasing clusters to lower than 10 considerably increases the intra-cluster variation criterion. It is reasonable to conclude that with fewer than 10 clusters, diverse speakers will be put into the same clusters, which is a drawback in terms of specificity in training lexicons. It appears that 10 clusters would be a good compromise in the trade-off between intra-cluster phonetic variation and sufficiency of training data for each cluster. In the EXPERIMENTAL EVALUATION section, a set of ASR experiments is conducted to examine the influence of the number of clusters over the whole process of training adapted lexicons and using them in a complete speech recognition task. The experiments validate the prediction of 10 clusters as a suitable number of clusters. Of course, this number applies to the present situation in which there are 50 speakers in the database. In a different situation, i.e., another speech database with a different number of speakers, another number of clusters may be found to be a better choice.

APPROACHES TO LEARNING SPEAKER ADAPTED CONTEXTUAL RULES

Speaker adapted contextual rules are the second module of the hybrid architecture introduced in this paper and are used to produce a phonetic realization of words. Our approach is to use rules which best describe the most frequent pronunciation mechanisms in sets of speakers with similar confusion vectors. For each speaker cluster, a separate set of adapted contextual rules is derived considering the portion of training data which is uttered by the speakers belonging to that cluster. Two approaches are designed to find sets of adapted contextual rules. In the first approach, contextual rules for a cluster are obtained by modifying the contextual rules derived from the whole corpus, in a manner to conform to the specifications of the speakers of that cluster. This is done by considering the phone confusion matrix of the cluster. In the second approach, for each speaker cluster, contextual rules have been extracted independently from the data corresponding to the training material from speakers in that cluster and not from the whole corpus. Before describing the approaches employed in adapting rules to clusters, we will explain in the following subsection the more general algorithm used to extract contextual rules, how these contextual rules are formed, the motivation for their use and how many of these rules are available.

Contextual rules

Contextual rules are a set of context-dependent rewrite rules which are defined in the form of $L\underline{F}R \rightarrow LOR$, in which *L* and *R* represent left and right single phone contexts of the focus string *F*, respectively. *F* represents the phonemic string and can be recognized as *O* owing to pronunciation or the phone recognizer error. *F* and *O* can be more than a single phone, and *F* can also be an empty string, in which case the rule becomes an insertion rule in the form of $LR \rightarrow LOR$. The combination of *LFR* is called the condition of the rule because it involves the contextual condition of applying the rule, i.e., the existence of the string *LFR* 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 by Cremelie and Martens (1999). Contextual rules represent the important context-sensitive pronunciation mechanisms which are variable between various groups of speakers. The learning algorithm compares the 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 and represents how many times a specific rule could occur in the whole database, regardless of whether or not it has happened. 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 the probability of occurrence of the pronunciation event that the rule is describing, when the conditions of the application are met. Pronunciation rules with low application likelihood are not statistically significant and such rules are better to be pruned and not to be used in generating pronunciation variants (Vazirnezhad *et al.*, 2009). The following equation defines the application likelihood for the *i*th rule r_i in which $count_2(r_i)$ is the number of times the rule has occurred and $count_1(r_i)$ is the coverage of r_i:

$$p_{\mathrm{AL}}(r_i) = p(O \mid LFR,\ r_i) = \frac{count_2(r_i)}{count_1(r_i)}. \qquad (4)$$

The rule $L\underline{F}R \rightarrow LOR$ with application likelihood of p_{AL} means that in the context of *L* and *R* the phone string *F* can be recognized as *O* with a probability of p_{AL}. Thus, it is better that the lexicon takes account of such a deviation to achieve a better match with the recognized phone string.

The algorithm for learning contextual rules resulted to 9325 statistically meaningful rules from the whole corpus. The procedure is the same as in our previous works and has been described in more detail in Vazirnezhad *et al.*(2005b). So far, we have derived contextual rules which are not yet adapted to the speakers' clusters. The main purpose of this research is to build lexicons which represent speakers' pronunciation habits. Contextual rules form an important block of the hybrid pronunciation model in which the adaptation takes place. In the following two subsections, we will explain two approaches employed for adapting contextual rules to the specifications of speakers' clusters. The adapted set of rules can be used as the second module of the hybrid model after adaptation. In a different approach, cluster-specific rules can be learned separately, for use in hybrid models.

First approach—adaptation of contextual rules from the whole training data

After clustering speakers according to their confusion vectors, a confusion matrix can be calculated and assigned to each cluster of speakers, using the training material related to the speakers in the cluster. This is done by considering aligned phonemic and recognized phone strings of the corresponding utterances out of the training data related to the cluster. In this approach, the contextual rules derived from the whole training data are modified based on the phone confusion matrix of a cluster, to obtain an adapted list of contextual rules for that cluster of speakers. This is done by pruning and strengthening of the rules considering the confusion matrix of the speaker cluster.

The approach in revising the initial set of rules is as follows. Regarding the rule $L\underline{F}R \rightarrow O$, and considering the cluster phone confusion matrix, if the probability of recognizing *F* as *O* in that matrix was greater than 0.9, the rule will be converted to the general rule $\underline{F} \rightarrow O$, which says *F* can be recognized as *O* in any context or independent of context. If the probability of recognizing *F* as *O* in that matrix was less than 0.05, the rule will be deleted from the list. The upper and lower threshold values are set such as to achieve a reasonable and significant number of modifications. With these thresholds, for a given cluster an average of around 20% of rules are subject to the modifications.

The adaptation process consists of two modification processes: one is to strengthen likely-to-happen rules while the other is to delete unlikely rules in a given cluster of speakers. The philosophy behind these modifications is to use the information on the pronunciation habits of speakers in the cluster in adapting rules based on the phone confusion matrix which statistically represents the phonetic deviation patterns. Strengthening the rules which are likely to happen is done by generalizing them to any context or making them context-independent, so they will be applied more frequently in the lexicon generation process. Deleting unlikely rules in a given cluster of speakers is again based on the corresponding phone confusion matrix. The aim of deleting rules which are not common in the cluster, besides adaptation, is to decrease confusability in the lexicon. These modifications are designed based on the fact that Farsi speakers' pronunciation habits include mainly explicit phone conversions among certain types of speakers owing to regional dialects, socio-economic level and educational class (Haghshenas, 1996).

Second approach—learning cluster-specific contextual rules

In this approach, for each of the clusters, a separate set of rules is derived from data corresponding to speakers in the cluster. The algorithm for learning rules is exactly the same as the general procedure already described. The advantage of this approach is that it learns contextual rules directly from data related to the target speakers rather than from the modification of general rules derived from the whole corpus. The drawback of this approach is that having a large number of clusters, the amount of available training data for each cluster will be insufficient as there is not enough coverage of all phonetic contexts. As a result, some rules representing important pronunciation mechanisms will not be learned.

Using adapted contextual rules in generating adapted dynamic lexicons

In this subsection, using an example, we explain how the sets of adapted contextual rules are employed in creating the lexicons needed in an ASR system. The numbers of clusters and sets of adapted contextual rules are the same. To generate lexicons containing phonetic deviations, we use hybrids of generalized decision trees and speaker adapted contextual rules as described in Section 2. In the example depicted in Fig.4, we have considered five ranges of rate of speech and therefore the speaker-adapted lexicons must appear in five groups corresponding to these rates. Fig.4 shows the procedure for generating dynamic adapted lexicons. According to Fig.4, in the case of introducing five ranges of speech rates and three speaker clusters, we would have 15 lexicons in

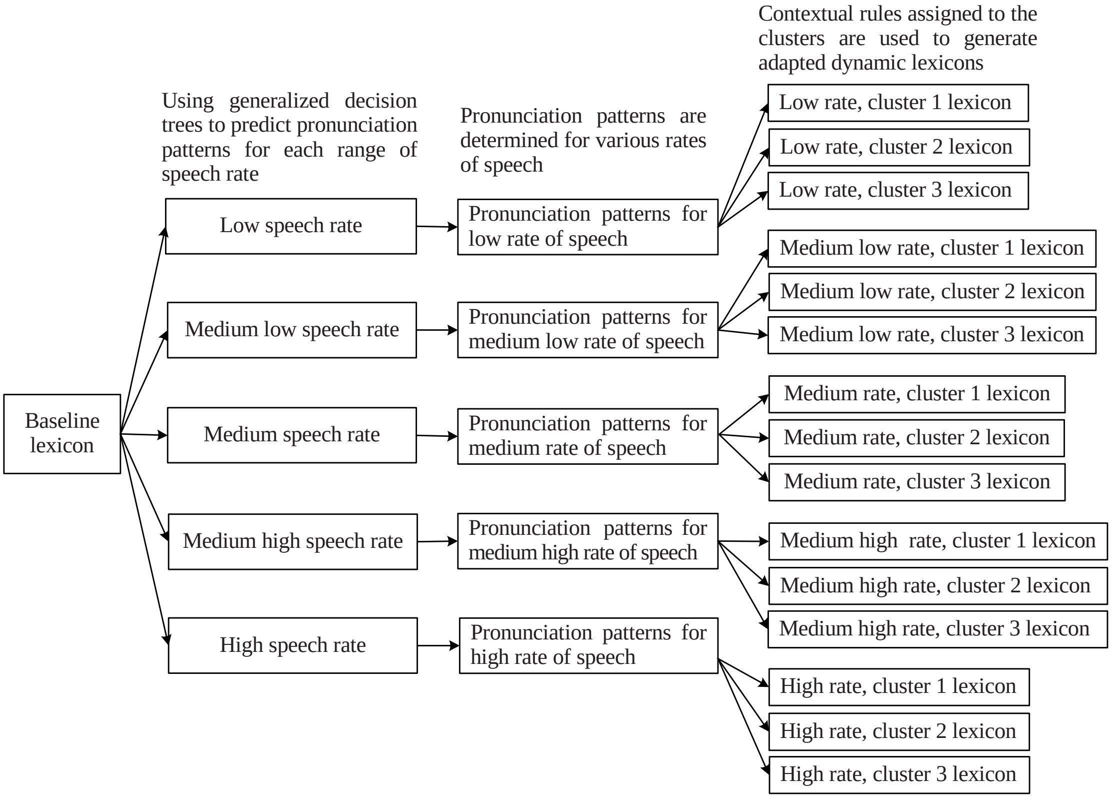
Contextual rules assigned to the clusters are used to generate adapted dynamic lexicons
Using generalized decision trees to predict pronunciation patterns for each range of speech rate
Pronunciation patterns are determined for various rates of speech
Baseline lexicon
Low speech rate
Medium low speech rate
Medium speech rate
Medium high speech rate
High speech rate
Pronunciation patterns for low rate of speech
Pronunciation patterns for medium low rate of speech
Pronunciation patterns for medium rate of speech
Pronunciation patterns for medium high rate of speech
Pronunciation patterns for high rate of speech
Low rate, cluster 1 lexicon
Low rate, cluster 2 lexicon
Low rate, cluster 3 lexicon
Medium low rate, cluster 1 lexicon
Medium low rate, cluster 2 lexicon
Medium low rate, cluster 3 lexicon
Medium rate, cluster 1 lexicon
Medium rate, cluster 2 lexicon
Medium rate, cluster 3 lexicon
Medium high rate, cluster 1 lexicon
Medium high rate, cluster 2 lexicon
Medium high rate, cluster 3 lexicon
High rate, cluster 1 lexicon
High rate, cluster 2 lexicon
High rate, cluster 3 lexicon

Fig.4 A case of generating sets of dynamic adapted lexicons in which lexicons are generated for five ranges of speech rates and three speaker clusters

the set of dynamic adapted lexicons. We have conducted experiments on ASR using adapted lexicons and the results are reported in the EXPERIMENTAL EVALUATION section. The lexicons generated for various ranges of speech rates and speaker clusters are pruned to reduce confusability among different words. The pruning procedure, which reduces confusability inside the lexicons, is the same as that introduced by Vazirnezhad *et al*.(2009).

SELECTING THE ACTIVE LEXICON

If the utterance of an unknown speaker is going to be recognized, the first step is to estimate his/her rate of speech. This rate estimator is simply a speech activity detector which determines the duration of the speech parts inside the input signal, in addition to an autocorrelation function which is applied to the speech parts. Its clear peaks show the vowels placed inside the speech signal. Because in Farsi every vowel is the core of a syllable and the numbers of vowels and syllables are the same, the rate of speech can be calculated simply in the number of syllables per second.

The second step is to find the appropriate speaker cluster. For this purpose, for each cluster, a codebook of acoustic feature vectors (MFCC+deltas) is generated. The codebook of each cluster is determined by applying a *K*-means algorithm to the set of extracted vectors from each individual frame of the speech material corresponding to the speakers of that cluster. An unknown speaker is assigned to a cluster by first calculating the distance of the new speaker vectors to each of the vectors stated in the codebook of a cluster. Finally, the cluster which has the minimum overall distance to the unknown speaker is chosen as the cluster corresponding most closely to the unknown speaker.

After finding the appropriate speaker cluster and determining the rate of input speech, the most appropriate adapted lexicon will be chosen from the set of dynamic adapted lexicons and will be used in the speech recognition process. Fig.5 shows how the set of dynamic adapted lexicons is used in the process of speech recognition. The best lexicon is chosen based on the utterance rate and the speaker cluster closest to the unknown speaker.

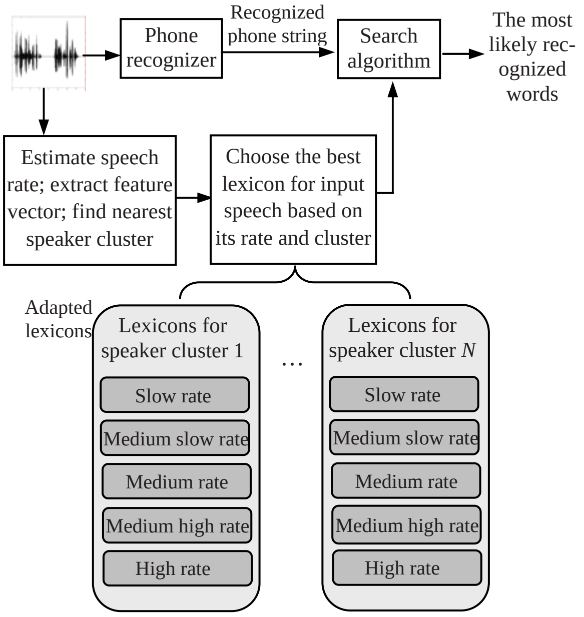
Phone recognizer
Recognized phone string
Search algorithm
The most likely recognized words
Estimate speech rate; extract feature vector; find nearest speaker cluster
Choose the best lexicon for input speech based on its rate and cluster
Adapted lexicons
Lexicons for speaker cluster 1
Slow rate
Medium slow rate
Medium rate
Medium high rate
High rate
…
Lexicons for speaker cluster *N*
Slow rate
Medium slow rate
Medium rate
Medium high rate
High rate

Fig.5 The procedure for finding the most appropriate lexicon from the set of adapted dynamic lexicons to be used in the speech recognition task

Fig.5 shows the entire process of employing the dynamic adaptive lexicon in an ASR process. In a real time application, choosing the best lexicon needs at least several seconds of input speech to measure the rate of speech and determine the speaker cluster best matched to the input speaker.

EXPERIMENTAL EVALUATION

The evaluation was conducted on a speech recognition task employing the sets of adapted lexicons for speaker clusters. The test set used in the experiments was completely separate from the training data and was chosen from the held out data. The experimental workbench was the SHENAVA-2 speech recognizer. The experiments were conducted in two main scenarios: a speaker-dependent task and a speaker-independent task. In the speaker-dependent task, the speakers in the test data were the same as speakers in the training data, while in the speaker-independent task, the speakers in the test data were not among the speakers in the training data. The following two subsections discuss experiments in speaker-dependent and speaker-independent scenarios respectively.

Experiments in a speaker-dependent scenario

The first set of experiments was conducted in a speaker-dependent scenario. This means that the speakers used in the clustering and generating adapted lexicons were the same as those used in the test data. Although the speakers in training and testing were the same, the speech material in the test set consisted of held out data, in terms of spoken utterances, which were not used in the training procedure.

Fig.6 depicts WERs obtained in experiments for the speaker-dependent scenario. Each point in Fig.6 demonstrates the WER obtained in the ASR task by using a set of adapted lexicons trained over a certain number of speaker clusters. In these experiments, the effect of the number of clusters was examined over the whole process of constructing adapted lexicons and employing them in the ASR task.

The solid line in Fig.6 shows the results obtained after applying the first approach for adapting contextual rules. The adaptation algorithm in this case considers the phone confusion matrix in the modification of contextual rules. The dashed line in Fig.6 shows the results obtained after applying the second technique for extracting adapted contextual rules. This approach uses the speech material from the same cluster for learning cluster-specific rules.

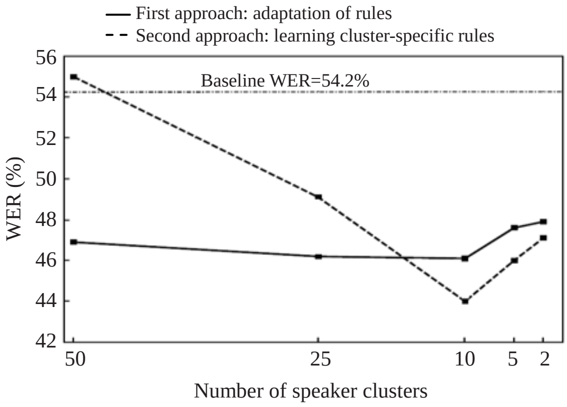

Fig.6 Word error rates (WERs) obtained in the speaker-dependent scenario

Fig.6 shows that the best result is obtained when the second approach in adaptation is used. The minimum WER of 44.1% is equivalent to an absolute WER reduction of as much as 10.1% in comparison to the baseline system. The baseline system uses a lexicon containing only the phonemic strings of words. Using the first approach results in a minimum WER of 46.1%. The best results are obtained when 10 clusters are considered in generating adapted lexicons.

The experiments described in this subsection were conducted on the same speakers of the training database. In this case, the lexicon should be adapted to speakers whose pronunciation behaviors are already learned. The set of experiments in this subsection proves the effectiveness of the approach in adapting lexicons to known speakers. However, as in many ASR applications, the test speakers are not known a priori. In the next subsection, we have evaluated the system by running experiments with different speakers for training and testing.

Experiments in a speaker-independent scenario

The second set of experiments was conducted in a speaker-independent scenario. In these experiments the speakers in the test set were different from the speakers in the training procedure. There are 50 speakers in the whole corpus. As the corpus is not huge in terms of the number of speakers, a five fold cross-validation algorithm was used to obtain statistically significant results. This means that in each test, utterances from 40 speakers were used in clustering and creating adapted lexicons, and speech material from 10 held-out speakers was used for evaluating the constructed adapted lexicons. In this way we have investigated whether or not the proposed method generalizes to speaker-independent recognition.

Fig.7 shows the results for the speaker-independent scenario. The solid line shows results obtained by the first approach and the dashed line shows those obtained using the second approach to adaptation. Each point corresponds to tests for a certain number of clusters. For each point in Fig.7, a five fold cross-validation was conducted and thus five WER values were obtained for each point.

In the speaker-independent scenario, the first approach in adapting rules showed a slightly better result (Fig.7). The minimum WER was 46.8%, which is equivalent to a 7.4% improvement in terms of absolute WER in comparison to the baseline system. The minimum WER obtained by employing the second approach was 47.1%. The best results in both approaches were obtained when the number of clusters was 10.

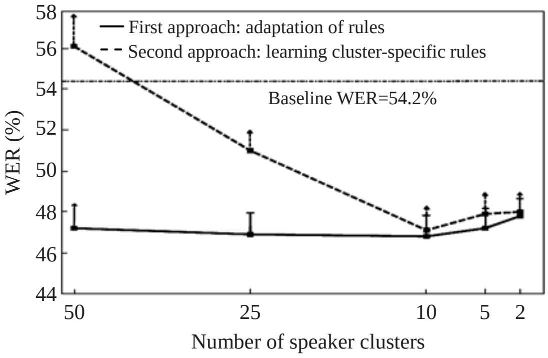

Fig.7 Word error rates (WERs) from using sets of adapted lexicons in the speaker-independent scenario
For each point, a five fold cross-validation was conducted. The mean values are depicted by squares, and the bars represent standard deviations for the five WER values obtained for each point

Comparison of results

Table 1 summarizes the best results obtained using various pronunciation models. Dynamic and static hybrid models are state-of-the-art models, in terms of architecture and components, and were explained comprehensively by Vazirnezhad *et al*.(2009). Adaptive models introduced in this paper have the same architecture as dynamic hybrid models. Moreover, adaptive models benefit from having sets of adapted contextual rules for each cluster rather than a static set of rules for any speaker.

Table 1 Word error rate reduction (WERR) and relative WERR obtained using various models in generating lexicons

Models employed to generate phonetic deviations of lexical entries	Absolute WERR (%)		Relative WERR (%)	
	SD	SI	SD	SI
Adaptive models by the 1st adaptation approach	8.1	7.4	14.9	13.7
Adaptive models by the 2nd adaptation approach	10.1	7.1	18.6	13.1
Dynamic hybrid models	6.3		11.6	
Static hybrid models	4.4		8.1	
Contextual rules	2.9		5.4	

SD: speaker dependent; SI: speaker independent

Table 1 shows that as more factors influencing phonetic deviations are introduced to the models, better results are obtained. Using adapted lexicons generated from the framework of clustering speakers significantly improved the accuracy of the recognition process in the speaker-dependent scenario. Adaptation was also effective in the speaker-independent scenario, but to a lesser extent.

DISCUSSION

The idea behind the adaptation approach introduced in this paper is that the speaker behavior in pronunciation can be predicted based on speaker acoustics—talkers with particular acoustic characteristics usually tend to produce predictable pronunciation patterns, and generate similar patterns of ASR phone recognition errors (Saraclar and Khudanpur, 2004). There are some well-known pronunciation patterns among speakers with the same accent or style of speech in Farsi who can be classified well in the acoustic feature space (Haghshenas, 1996). Rate of speech is another factor which affects both acoustic and spectral features, as well as pronunciation at the phonetic level (Zheng and Franco, 2003).

Speaker adapted lexicons have been shown to be effective in both speaker-dependent and speaker-independent scenarios. In the speaker-dependent scenario, speakers in the test set are the same as speakers involved in the training algorithm. In this case, the pronunciation habits of the test speakers are already learned by the models. Therefore, a greater improvement is expected from the adaptation, and this was confirmed by experiments. In the speaker-independent scenario, speakers in the test set are different from those used for training and the pronunciation model is supposed to predict the pronunciation patterns and phonetic deviations based on speaker acoustic features. Although the improvement in the speaker-independent scenario was less than that in the speaker-dependent scenario, the results were still significantly better than those obtained using dynamic models without adaptation.

Two approaches were employed in this study to extract cluster adapted contextual rules. The first approach is an adaptation technique and the second is an algorithm for learning cluster-specific rules from data corresponding to the cluster. For the speaker-dependent scenario, the second approach is more effective than the first. This can be explained considering the fact that in the speaker-dependent scenario the speakers in the test are the same as those used in training. In this case, using cluster-specific rules which contain the exact pronunciation habits of the test speaker is more effective. While for the speaker-independent scenario, the first adaptation approach shows slightly better results. This is because the first adaptation approach is a smoother adaptation technique and works better for new test speakers not seen previously in the training procedure. Furthermore, in the speaker-independent scenario using cluster-specific rules, there is no advantage from having exactly the same pronunciation habits for the new test speaker.

Experiments in all cases showed that the best results are obtained when the number of clusters is 10. This is in agreement with the observations using the variation criterion which show that the intra-cluster phonetic variation changes slowly when the number of clusters reduces from 50 towards 10, but after this point the rate abruptly increases. Fewer clusters have many advantages in terms of computational load. Furthermore, finding the matching cluster for the test speaker is much easier when fewer clusters exist. Moreover, as the number of clusters increases, the portion of the training data for each cluster decreases, causing difficulties in the procedure for extracting valid and significant contextual rules. As a result, the adapted lexicons lose their validity owing to insufficient training data. This is especially a drawback for the second approach which learns cluster-specific rules. In Figs.6 and 7, the results for the second approach represented by dashed lines show large deteriorations when the number of clusters is high. This is because the sets of adapted contextual rules are extracted from insufficient data. Note that the optimum number of clusters is obtained for the database employed and may be different in another situation.

According to Table 1, modelling the speaker variability in the lexicon from the framework of generating adapted lexicons leads to meaningful improvements in both speaker-dependent and speaker-independent scenarios. The results were significant enough to prove that lexicon adaptation is a worthwhile approach in the development of ASR systems.

CONCLUSION

In this paper, a state-of-the-art approach in generating phonetic deviations of lexicon entries is introduced. The models effectively consider the syllabic structure of a word, rate of speech, unigram probability of a word, stress and syllable position in the word, phonemic context and speaker specifications in the process of generating lexicons. Moreover, information on speaker specifications is injected into models through a speaker clustering scheme. These models dynamically consider the rate of speech and speaker characteristics simultaneously to generate lexicons containing phonetic deviations of words. Modelling speaker variability at the level of the lexicon is a novel approach in ASR systems. The developed dynamic adapted hybrid models are composed of generalized decision trees and sets of adapted contextual rules. Each of these modules has the ability to be trained with a moderate amount of training data. As a result, dynamic adapted models can be trained with a medium-size corpus such as Large-FARSDAT. In experiments conducted in a speaker-dependent scenario, a WER reduction of as much as 3.8% was obtained in comparison to the dynamic hybrid models. This is equivalent to an absolute WER reduction of 10.1% in comparison to the baseline lexicon containing only phonemic strings of lexical entries. The more general experiments conducted in the speaker-independent scenario showed an improvement of as much as 1.1% and 7.4% in comparison to the usage of lexicons generated by dynamic hybrid models and the baseline lexicon, respectively.

ACKNOWLEDGEMENTS

We would like to express our sincere thanks to the Iran Telecommunication Research Center (ITRC) and the Research Center of Intelligent Signal Processing (RCISP) for their continual support. We would also like to thank Dr. Bijankhan, Dr. Eslami, Dr. Homayounpour, Dr. Sayadian and Dr. Sameti for their guidance on this work.

References

Almasganj, F., Seyedsalehi, S.A., Bijankhan, M., Sameti, H., Sheikhzadegan, J., 2001. SHENAVA-1: Persian Spontaneous Continuous Speech Recognizer. Proc. Int. Conf. on Electrical Engineering, p.101-106 (in Farsi).

Bijankhan, M., Sheikhzadegan, M.J., 1994. FARSDAT—The Farsi Spoken Language Database. Proc. Int. Conf. on Speech Sciences and Technology, **2**: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. European Conf. on Speech Communication and Technology, p.1525-1528.

Chen, K., Hasegawa-Johnson, M., 2004. Modeling Pronunciation Variation Using Artificial Neural Networks for English Spontaneous Speech. Int. Conf. on Spoken Language Processing, p.1461-1464.

Cremelie, N., Martens, J.P., 1999. In search of better pronunciation models for speech recognition. *Speech Commun.*, **29**(2-4):115-136. [doi:10.1016/S0167-6393(99)00034-5]

Davis, S., Mermelstein, P., 1980. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Trans. Acoust. Speech Signal Process.*, **28**(4):357-366. [doi:10.1109/TASSP.1980.1163420]

Fosler-Lussier, E., 1999. Dynamic Pronunciation Models for Automatic Speech Recognition. PhD Thesis, University of California, Berkeley, CA, USA.

Fosler-Lussier, E., Morgan, N., 1999. Effects of speaking rate and word frequency on pronunciations in conversational speech. *Speech Commun.*, **29**(2-4):137-158. [doi:10.1016/S0167-6393(99)00035-7]

Fukada, T., Yoshimura, T., Sagisaka, Y., 1999. Automatic generation of multiple pronunciations based on neural networks. *Speech Commun.*, **27**(1):63-73. [doi:10.1016/S0167-6393(98)00066-1]

Haghshenas, A.M., 1996. A Course in Phonetics. Agah Publications, Tehran, Iran (in Farsi).

Hazen, T., Hetherington, L., Shu, L., Livescu, K., 2005. Pronunciation modeling using a finite-state transducer representation. *Speech Commun.*, **46(**2):189-203. [doi:10.1016/j.specom.2005.03.004]

Humphries, J., 1997. Accent Modeling and Adaptation in Automatic Speech Recognition. PhD Thesis, University of Cambridge, Cambridge, UK.

Imai, T., Ando, A., Miyasaka, E., 1995. A New Method for Automatic Generation of Speaker-dependent Phonological Rules. Int. Conf. on Acoustics, Speech, and Signal Processing, p.864-867. [doi:10.1109/ICASSP.1995.479831]

Jande, P.A., 2008. Spoken language annotation and data-driven modeling of phone-level pronunciation in discourse context. *Speech Commun.*, **50**(2):126-141. [doi:10.1016/j.specom.2007.07.004]

Padrell, J., Macho, D., Nadeu, C., 2005. Robust Speech Activity Detection Using LDA Applied to FF Parameters. Int. Conf. on Acoustics, Speech, and Signal Processing, p.557-560. [doi:10.1109/ICASSP.2005.1415174]

Randolph, M., 1990. A Data-driven Method for Discovering and Predicting Allophonic Variation. Int. Conf. on Acoustics, Speech, and Signal Processing, p.1177- 1180. [doi:10.1109/ICASSP.1990.116176]

Riley, M., 1991. A Statistical Model for Generating Pronunciation Networks. Int. Conf. on Acoustics, Speech, and Signal Processing, p.737-740. [doi:10.1109/ICASSP.1991.150446]

Saraclar, M., Khudanpur, S., 2004. Pronunciation change in conversational speech and its implications for automatic speech recognition. *Comput. Speech Lang.*, **18**(4):375-395. [doi:10.1016/j.csl.2003.09.005]

Schmid, P., Cole, R., Fanty, M., 1993. Automatically Generated Word Pronunciations from Phoneme Classifier Output. Int. Conf. on Acoustics, Speech, and Signal Processing, p.223-226. [doi:10.1109/ICASSP.1993.319275]

Skorik, S., Berthommier, F., 2000. On a Cepstrum-based Speech Detector Robust to White Noise. SPECOM. St. Petersburg, Russia, p.1-5.

Sloboda, T., 1995. Dictionary Learning Performance through Consistency. Int. Conf. on Acoustics, Speech, and Signal Processing, p.453-456. [doi:10.1109/ICASSP.1995.479626]

Strik, H., Cucchiarini, C., 1999. Modeling pronunciation variation for ASR: a survey of the literature. *Speech Commun.*, **29**(2-4):225-246. [doi:10.1016/S0167-6393(99)00038-2]

Vazirnezhad, B., Almasganj, F., Bijankhan, M., 2005a. Automatic extraction of contextual rules and generating pronunciation variants to use in automatic continuous speech recognition. *J. Comput. Sci. Eng.*, **3**(3):40-50 (in Farsi).

Vazirnezhad, B., Almasganj, F., Bijankhan, M., 2005b. A Hybrid Statistical Model to Generate Pronunciation Variants of Words. Proc. IEEE Natural Language Processing and Knowledge Engineering, p.106-110. [doi:10.1109/NLPKE.2005.1598716]

Vazirnezhad, B., Almasganj, F., Ahadi, M., 2009. Hybrid statistical pronunciation models to be trained with a medium-size corpus. *Comput. Speech Lang.*, **23**(1):1-24. [doi:10.1016/j.csl.2008.02.001]

Wooters, C., Stolcke, A., 1994. Multiple-pronunciation Lexical Modeling in a Speaker Independent Speech Understanding System. Int. Conf. on Spoken Language Processing, p.1363-1366.

Zheng, J., Franco, H., 2003. Modeling word level rate of speech variation in large vocabulary conversational speech recognition. *Speech Commun.*, **41**(2-3):273-285. [doi:10.1016/S0167-6393(02)00122-X]