Sameti et al. EURASIP Journal on Audio, Speech, and Music Processing 2011, 2011:6
http://asmp.eurasipjournals.com/content/2011/1/6

EURASIP Journal on Audio, Speech, and Music Processing
a SpringerOpen Journal

RESEARCH Open Access

A large vocabulary continuous speech recognition system for Persian language

Hossein Sameti*, Hadi Veisi, Mohammad Bahrani, Bagher Babaali and Khosro Hosseinzadeh

Abstract

The first large vocabulary speech recognition system for the Persian language is introduced in this paper. This continuous speech recognition system uses most standard and state-of-the-art speech and language modeling techniques. The development of the system, called Nevisa, has been started in 2003 with a dominant academic theme. This engine incorporates customized established components of traditional continuous speech recognizers and its parameters have been optimized for real applications of the Persian language. For this purpose, we had to identify the computational challenges of the Persian language, especially for text processing and extract statistical and grammatical language models for the Persian language. To achieve this, we had to either generate the necessary speech and text corpora or modify the available primitive corpora available for the Persian language. In the proposed system, acoustic modeling is based on hidden Markov models, and optimized decoding, pruning and language modeling techniques were used in the system. Both statistical and grammatical language models were incorporated in the system. MFCC representation with some modifications was used as the speech signal feature. In addition, a VAD was designed and implemented based on signal energy and zero-crossing rate. Nevisa is equipped with out-of-vocabulary capability for applications with medium or small vocabulary sizes. Powerful robustness techniques were also utilized in the system. Model-based approaches like PMC, MLLR and MAP, along with feature robustness methods such as CMS, PCA, RCC and VTLN, and speech enhancement methods like spectral subtraction and Wiener filtering, along with their modified versions, were diligently implemented and evaluated in the system. A new robustness method called PC-PMC was also proposed and incorporated in the system. To evaluate the performance and optimize the parameters of the system in noisy-environment tasks, four real noisy speech data sets were generated. The final performance of Nevisa in noisy environments is similar to the clean conditions, thanks to the various robustness methods implemented in the system. Overall recognition performance of the system in clean and noisy conditions assures us that the system is a real-world product as well as a competitive ASR engine.

1 Introduction

Since the start of developing speech recognizers at AT&T Bell labs in the 1950's, enormous efforts and investments were directed towards automatic speech recognition (ASR) research and development. In the 1960s, the ASR research was focused on phonemes and isolated word recognition. Later, in the 70 s and 80 s, connected words and continuous speech recognition were the major trends of ASR research. To accomplish these targets, researchers introduced linear predictive coding (LPC) and used pattern recognition and clustering methods. Hidden Markov models (HMM), cepstral analysis and neural networks were employed in the 80 s. In the next decade, robust continuous speech recognition and spoken language understanding were popular topics. In the last decade, researchers and investors introduced spoken dialogue systems and tried to implement conversational speech recognition systems capable of recognizing and understanding spontaneous speech. Machine learning techniques and artificial intelligence (AI) concepts entered into the ASR research literature and contributed considerably to fulfilling the human speech recognition needs. Up until recent years, speech recognition systems were considered as luxury tools or services and were not usually taken seriously by users. In the past 5-10 years, we have seen that ASR engines have played genuinely beneficial roles in several areas, especially in telecommunication

* Correspondence: sameti@sharif.edu
Department of Computer Engineering, Sharif University of Technology, Tehran, Iran

services and important enterprise applications such as customer relationship management (CRM) frameworks.

Several successful ASR systems having good performances are found in the literature [1-3]. The most successful approaches to ASR are the ones based on pattern recognition and using statistical and AI techniques [1,3,4]. The front end of a speech recognizer is a feature extraction block. The most common features used for ASR are Mel-frequency cepstral coefficients (MFCC) [4]. Once the features are extracted, modeling is performed usually based on artificial neural network (ANN) or HMM. Linguistic information is also used extensively in an ASR system. Statistical (n-gram) and grammatical (i.e., structural) language models [4,5] are used for this purpose.

One essential problem with putting the speech recognition systems into practice is the variety of languages people around the world speak. ASR systems are highly dependent on the language spoken. We can categorize the research areas of speech recognition into two major classes; first, acoustic and signal processing which is very much the same for ASR in every language; second, natural language processing (NLP) which is dependent on the language. Obviously, this language dependency hinders the implementation and utilization of ASR systems for any new language.

We have focused our research on Persian speech recognition during recent years. Persian ASR systems have been addressed and developed to different extents [6-10]. There are other works on the development of Persian continuous speech recognition system [11-14]. However, in the most of them, a medium vocabulary continuous speech recognition system with high word error rate is presented. Our developed large vocabulary continues speech recognition system for Persian, called Nevisa, was first introduced in [6,7] as Sharif speech recognition system. It employs the cepstral coefficients as the acoustic features and continuous density hidden Markov model (CDHHM) as the acoustic model [4,15]. A time-synchronous left-to-right Viterbi beam search, in combination with a tree-organized pronunciation lexicon is used for decoding [16,17]. To limit the search space, two pruning techniques are employed in the decoding process. Due to our practical approach in using this system, Nevisa is equipped with established robustness techniques for handling speaker variation and environmental noise. Various data compensation and model compensation methods are used to achieve this objective. Also class-based n-gram language models (LM) [18,19] with generalized phrase structure grammar (GPSG)-based Persian grammar [20] are utilized as word-level and sentence-level linguistic information. The frameworks for testing and comparing the effects of the implemented methods and also for optimizing the parameters were gradually built up. This enabled us to move towards a practical ASR system capable of being utilized as Persian dictation software also called Nevisa [10].

In the remainder of this paper, in Sect. 2, the characteristics of the Persian language, and speech and text corpora of the Persian language are reviewed. An overview of Nevisa Persian speech recognition system and overall features of this system is given in Sect. 3. This section provides a review on acoustic modeling, robustness techniques used in the system, and building statistical and grammatical language models for the Persian language. In Sect. 4 the details of the experiments and the recognition results are given. Finally, Sect. 5 gives a brief summary and conclusion of the paper.

2 Persian language and corpora

2.1 Persian language

The Persian language, also known as Farsi, is an Iranian language within the Indo-Iranian branch of Indo-European languages. It is natively spoken by about seventy million people in Iran, Afghanistan and Tajikistan as the official language. It is also widely spoken in Uzbekistan and, to some extent, in Iraq and Bahrain. This language has remained remarkably stable since the eighth century although local environments, such as the Arabic language, have influenced it. The Arabic language has heavily influenced Persian, but has not changed its structure. In other words, Persian has only borrowed a large number of lexical words from Arabic. Therefore, in spite of this influence, Arabic has not affected the syntactic and morphological forms of Persian; as a result, the language models of Persian and Arabic are fundamentally differences. Although there are several similar phonemes in Arabic and Persian, and they use similar scripts, the phonetic structure of these languages has principal differences; therefore, the acoustic models of Persian and Arabic are not the same. Consequently, the development of a speech recognition system in Arabic and Persian are different due to distinctions in their acoustic and language models.

The grammar of Persian language is similar to that of many contemporary European languages. Normal declarative sentences in Persian are structured as "(S) (O) V". This means sentences can comprise of optional subjects and objects, followed by a required verb. If the object is specific, then it is followed by the word/rə/. Despite the normal structure, there is a large potential in the language to be free-word-order, especially in preposition adjunction and complements. For example, adverbs could be placed at the beginning, at the end or in the middle of sentences, often without changing the meaning of the sentences. This flexibility in word ordering makes the task of Persian grammar extraction a difficult one. Written style of Persian is right to left and it uses Arabic script. In Arabic script, short vowels (/a/,/e/,/o/) are not

usually written. This results in ambiguities in pronunciation of words in Persian. Persian has 6 vowels and 23 consonants. Three vowels of the language are considered long (/i/,/u/,/ə/) and the other three are short vowels or diacritics (/e/,/o/,/a/). Although usually named as long and short vowels, the three long vowels are currently distinguished from their short counterparts by position of articulation, rather than by length. The phonemes of Persian are shown in Table 1 where Farsi letters, codes and IPA notations are shown, too.

Persian uses the same alphabet as Arabic with four additional letters. Therefore, the number of letters in the Persian alphabet is 32 as compared to 28 in Arabic. Each additional Persian letter represents a phoneme not present in the Arabic phoneme set, namely/p/,/tʃ/,/ʒ/ and/g/. In addition, Persian has four other phonemes (/v/,/k/,/ʔ/,/G/) which are pronounced differently from their Arabic counterpart. On the other hand, Arabic has its own unique phonemes (about ten) not defined in the Persian language. Persian makes extensive use of word building and combining affixes, stems, nouns and adjectives. Persian frequently uses derivational agglutination to form new words from nouns, adjectives and verbal stems. New words are extensively formed by compounding two existing words, as is common in German. Suffixes predominate Persian morphology, though there are a small number of prefixes. Verbs can express tense and aspect, and they agree with the subject in person and number. There is no gender in Persian, nor are pronouns marked for natural gender.

2.2 Corpora

2.2.1 Speech corpus

Small Farsdat In this paper, two speech databases, small Farsdat [21] and large Farsdat [22], are used. Small Farsdat is a hand-segmented database in the phoneme level which contains 6080 Persian sentences read by 304 speakers. Each speaker has uttered 18 randomly chosen sentences (from a set of 405 sentences) plus two sentences which are common for all speakers. The sentences are formed by using over 1,000 Persian words and are designed artificially to cover the acoustic variations of the Persian language. The speakers are chosen from ten different dialect regions in Iran and the corpus contains the ten most common dialects of the Persian language. Male to female population ratio is 2:1. The database is recorded in a low-noise environment featuring an average of 31 dB signal to noise ratio with a sampling rate of 22,050 Hz. A clean test set, called the small Farsdat test set (*sFarsdat test*), is selected from this database that contains 140 sentences from seven speakers. All the other sentences are used as train set (*sFarsdat train*). Small Farsdat, as its name indicates, is a small size speech corpus and can be used only for training and evaluating limited speech recognition systems in laboratories. This speech corpus is comparable with TIMIT corpus in English. Large Farsdat is another Persian speech database that removes some of the deficiencies of the small Farsdat.

Large Farsdat Large Farsdat [22] includes about 140 h of speech signals, all segmented and labeled in word level. This corpus is uttered by 100 speakers from the most common dialects of the Persian language. Each speaker utters 20-25 pages of text from various subjects. In contrast with small Farsdat, which is recorded in a quiet and reverberation-free room, large Farsdat is recorded in office environment. Four microphones, a unidirectional desktop microphone, two lapel microphones and a headset microphone are used to record the speech signals. All the speech signals in this corpus are recorded using two microphones simultaneously, the desktop microphone is used in all of the recording sessions and each of the other three microphones is used in about one-third of the sessions. Totally, the desktop microphone is used for about 70 h of recorded speech and the other three microphones are used for the 70 remaining hours. The average SNR of the desktop microphone is about 28 dB. The sampling rate is 16 kHz for the whole corpus.

The test set contains 750 sentences from seven speakers (four male and three female) and is recorded using the desktop microphone of the large Farsdat database. We call this set *gFarsdat test*. The average sentence length of this test set is 7.5 s. This set includes numbers, names and some grammar free sentences and contains about 5000 different words. All other speech signals in the large Farsdat recorded with the desktop microphone are used here as the train set, i.e. *gFarsdat train*. In this research only those speech les of large Farsdat that are recorded using the desktop microphone, are used in the evaluations.

Farsi noisy speech corpus To evaluate the performance of Nevisa in real applications and in noisy environments, Farsi Noisy speech (*FANOS*) database is recorded and transcribed [23,24]. This database consists of four pair sets providing four tasks. As adaptation techniques are used in our robustness methods, each task in this database includes two subsets identified as adaptation subset and test subset. Each adaptation subset is arranged as follows: 175 sentences (selected from Farsdat sentences) are uttered by seven speakers consisting of five male and two female speakers. Each speaker reads 10 identical sentences (read by all speakers) plus 15 randomly selected sentences. In addition, each test subset consists of 140 sentences uttered by five male and two female speakers, each speaker reading 20 sentences. The average length of the sentences is 3.5 s. The transcriptions are at word level for test data and at phoneme level for adaptation data. Each task demonstrates a new environment which

Table 1 Phonemes of Persian language

IPA	Char	Code	Farsi Letter	Phonetic Description
i	i	105	ای	high front unrounded
e	e	101	اِ	mid front unrounded
a	a	97	اَ	low front unrounded
u	u	117	او	high back unrounded
o	o	111	اُ	mid back unrounded
ɔ	/	47	آ	low back rounded
p͡	\	92	پ	unvoiced bilabial plosive closure
p	p	112	پ	unvoiced bilabial plosive
b͡	'	96	ب	voiced bilabial plosive closure
b	b	98	ب	voiced bilabial plosive
t͡	-	45	ت، ط	unvoiced alveolar plosive closure
t	t	116	ت، ط	unvoiced dental plosive
d͡	=	61	د	voiced dental plosive closure
d	d	100	د	voiced dental plosive
c͡	@	64	ک	unvoiced palatal plosive closure
c	c	99	ک	unvoiced bilabial plosive
k͡	*	42	ک	unvoiced velar plosive closure
k	k	107	ک	unvoiced bilabial plosive
ɟ͡	!	33	گ	voiced palatal plosive closure
ɟ	;	59	گ	voiced palatal plosive
g͡	&	38	گ	voiced velar plosive closure
g	g	103	گ	voiced velar plosive
G͡	^	94	ق، غ	voiced uvular plosive closure
G	q	113	ق، غ	voiced uvular plosive
ʔ͡	(	40	أ، ؤ، ع	glottal stop closure
ʔ	]	93	أ، ؤ، ع	glottal stop
tʃ͡	$	36	چ	unvoiced alveopalatal affricate closure
tʃ	'	39	چ	unvoiced alveopalatal affricate
dʒ͡	#	35	ج	voiced alveopalatal affricate closure
dʒ	'	44	ج	voiced alveopalatal affricate
f	f	102	ف	unvoiced labiodental fricative
v	v	118	و	voiced labiodental fricative
s	s	115	س، ث، ص	unvoiced alveolar fricative
z	z	122	ز، ذ، ظ، ض	voiced alveolar fricative
ʃ	.	46	ش	unvoiced alveopalatal fricative
ʒ	[	91	ژ	voiced alveopalatal fricative
x	x	120	خ	unvoiced uvular fricative
h	h	104	ه، ح	unvoiced glottal fricative
l	l	108	ل	lateral alveolar
r	r	114	ر	trill alveolar
m	m	109	م	nasal bilabial
n	n	110	ن	nasal alveolar
j	y	121	ی	approximant palatal

differs from the training environment. Tasks A and B are recorded in office environment with condenser and dynamic microphones, respectively with average SNR levels of 18 and 26 dB. Both tasks C and D are recorded with condenser microphone in office environment and in the presence of exhibition and car noises respectively. Corresponding SNR levels of these sets are 9 and 7 dB. Table 2 summarizes the properties of the tasks in the FANOS database.

2.2.2 Text corpus

In this research, we have used the two editions of Persian text corpus called "Peykare" [25,26]. The first edition of this corpus consists of about ten million words and it was increased to about 100 million words in the second

Table 2 The specifications of tasks in FANOS database

Task	Task A	Task B	Task C	Task D
Environment	Office	Office	Exhibition	Car Noise
Microphone	Condenser	Dynamic	Condenser	Condenser
SNR(dB)	18	26	9	7
Number of files (adapt + test)	315 (175 + 140)	315 (175 + 140)	315 (175 + 140)	315 (175 + 140)
Number of speakers (male + female)	7 (5 + 2)	7 (5 + 2)	7 (5 + 2)	7 (5 + 2)

edition [26]. All words in the first edition are annotated with part-of-speech (POS) tags. The texts of this corpus are gathered from various data sources like newspapers, magazines, journals, books, letters, hand-written texts, movie scripts, news etc. This corpus is a complete set of Persian contemporary texts. The texts are about different subjects including politics, arts, culture, economics, sports, stories, etc. The tag set of Persian Text Corpus has 882 POS tags [18,19] that are reduced to 166 POS tags in this work.

3 Nevisa speech recognition system

3.1 Overview

Nevisa is a Persian continuous speech recognition (CSR) system that integrates state-of-the-art techniques of the field. The architecture of this system including feature extraction, training and decoding (i.e. recognition) blocks is shown in Figure 1. As this figure shows, each block represents a module that can be easily modified or replaced. The modularity of the system makes it very flexible in developing CSR systems for various applications and for trying out new ideas in different modules for research works. The modules shown with dotted blocks are robustness modules and can be used optionally. The MFCC module is used as the core of feature extraction unit and is supplied with vocal tract length normalization (VTLN) [27-29], cepstral mean subtraction (CMS) [3,23] and principal component analysis (PCA) [30] robustness methods. In addition, voice activity detector (VAD) is used to separate speech segments from non-speech ones. Nevisa uses energy and zero-crossing based VAD in the pre-processing of speech signal. VAD is a useful block in the ASR systems, especially in real applications. It specifies the beginning and the end of utterance and reduces the processing cost of feature extraction and decoding blocks. The modified VAD is

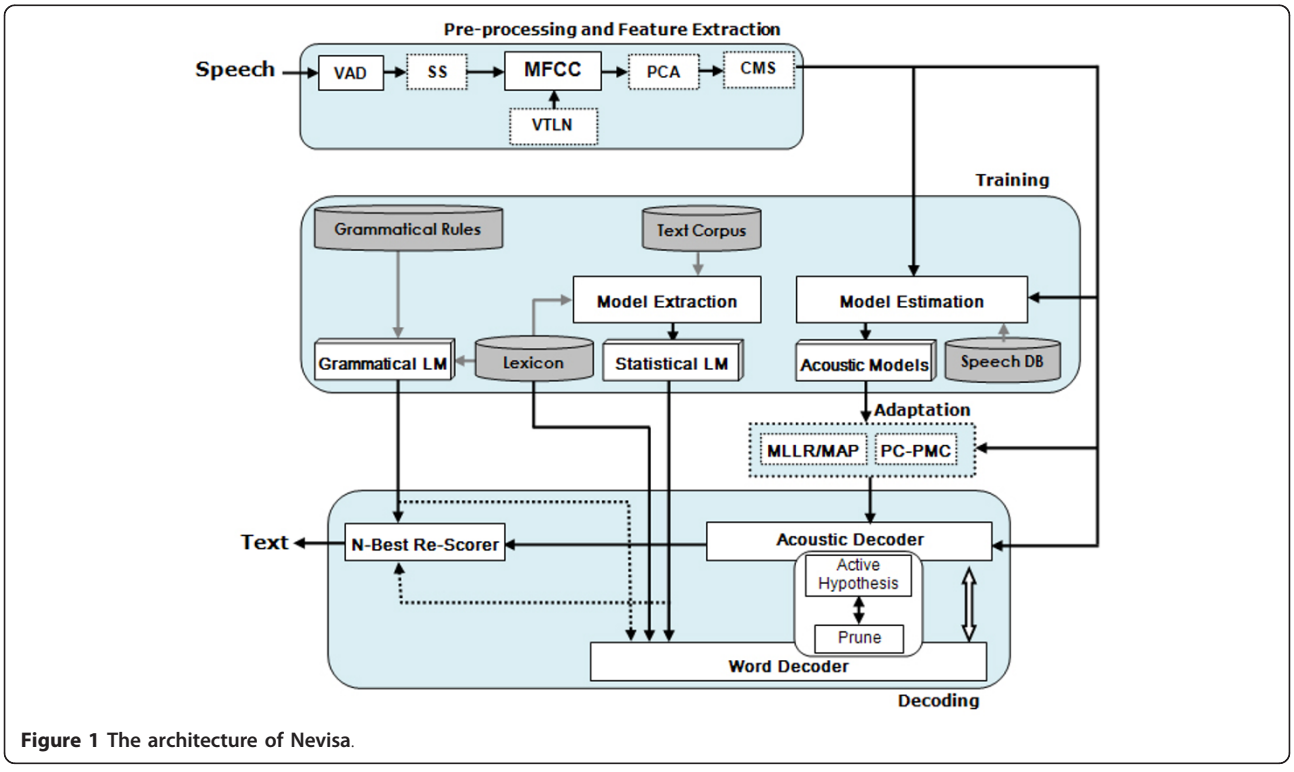

Figure 1 The architecture of Nevisa.

also used in spectral subtraction (SS) [3] and in PC-PMC [23,31,32] robustness methods to detect noise segments in the speech signal. In addition to speech enhancement and feature robustness techniques, MLLR [33], MAP [34] and PC-PMC model adaptation methods can be applied optionally on acoustic models to adapt the acoustic model parameters to speaker variations and environmental noises.

The system uses context-dependent (CD) and context-independent (CI) acoustic models that are represented by continuous density hidden Markov models. These models are mixtures of Gaussian distribution in cepstral domain. In this system, forward, skip and loop transitions between the states are allowed and the covariance matrices are assumed diagonal [6,9,10]. The parameters of the emission probabilities are trained using the maximum likelihood criterion and the training procedure is initialized by a linear segmentation. Each iteration of the training procedure consists of time alignment by dynamic programming (Viterbi algorithm) followed by parameter estimation, resulting in segmental k-means training procedure [3,4]. In decoding phase, a Viterbi-based search with beam and histogram pruning techniques are used. In this module, the recognized acoustic units are used to make active hypotheses via word decoder. The word decoder searches the lexicon tree simultaneously in interaction with the acoustic decoder and the pruning modules. The final active hypotheses are rescored using language models. Both statistical and grammatical language models can be used either in word decoder or in rescoring modules. In Nevisa, by default, statistical LM is used in the word decoder, i.e., during the search, and the grammatical model is used in n-best re-scoring module optionally. Dotted arrows in Figure 1 mean that statistical LM can be used in the rescorer module, and grammatical LM can be utilized during the search optionally.

3.2 Acoustic modeling

For acoustic modeling we employ two approaches: context-independent (CI) and context-dependent (CD) modeling. The standard phoneme set of Persian language contains 29 phonemes. This phoneme set and extra HMM models for silence, noise and aspiration are considered in the CI modeling. In sect. 4 where recognition results are given, the details of modeling process, including number of states and Gaussian mixtures, are presented.

For context-dependent modeling, we use triphones as the phone units. The major problem in triphone modeling is the trade-off between the number of triphones and the size of available training data. There are a large number of triphones in a language, but many of them are unseen or rarely used in speech corpora. So the amount of training data is insufficient for many triphones. For solving this problem, the state tying methods are used [35,36]. Two prevalent methods for state tying are data-driven clustering [35] and decision tree-based state tying [36,37]. In these methods, at the first stage, all triphones that occur in a speech corpus are trained using the available data. Then the states of similar triphones are clustered into a small number of classes (the similar triphones are the triphones that have similar middle phoneme). In the last stage, the states that lie in each cluster are tied together. The tied states are called senones [38].

Different numbers of senones and different numbers of Gaussian distributions were evaluated in the Nevisa system. The experimental results showed that clustering triphone states to 500 senones for small Farsdat and 4,000 senones for large Farsdat leads to the best WER. The evaluation results are given in Sect. 4.

3.2.1 Robustness methods

Like all speech recognizers, the performance of the Nevisa degrades in real applications and in the presence of noise [23,31,39,40]. In order to make this system robust to speaker and environment variations, many of the recent advanced methods in robustness are incorporated. Differences between speakers, in background noise characteristics and channel noises (i.e. microphones), are considered and tried to be dealt with. Nevisa uses data compensation and model compensation approaches as well as their combinations. In the data compensation approach, clean data are estimated from their noisy samples so as to make them similar to the training data. Nevisa uses spectral subtraction (SS) and Wiener filtering [23], cepstral mean subtraction (CMS) [3,23], principal component analysis (PCA) [30] and vocal tract length normalization (VTLN) [27,28,41,29] for this purpose. In the model-based approach, the models of various sounds used by the classifier are modified to become similar to the test data models. Maximum likelihood linear regression (MLLR) [33,42], maximum a posteriori (MAP) [34,24], parallel model combination (PMC) [23,31,33] and a novel enhanced version of PMC, PCA and CMS based PMC (PC-PMC) [30] are well incorporated in the system. PC-PMC algorithm takes the advantages of additive noise compensation ability of PMC and convolutional noise removal capability of both PCA and CMS methods. The first problem that is to be solved for combining these methods is that PMC algorithm requires invertible modules in the front-end of the system while CMS normalization is not an invertible process. In addition, a framework is to be designed for the adaptation of the PCA transform matrix in the presence of noise. The PC-PMC method provides solutions to these problems [30].

The integration of these robustness modules in Nevisa are shown in the Figure 1. The modularity of the system makes it very flexible to remove any one of the system

blocks, add new blocks, change or replace the existing ones.

3.3 Language modeling

Linguistic knowledge is as important as acoustic knowledge in recognizing natural speech. Language models depict the constraints on word sequences imposed by syntax, semantics or pragmatics of the language [5]. In recognizing continuous speech, the acoustic signal is too weak to narrow down the number of word candidates. Hence, speech recognizers employ a language model that prunes out acoustic alternatives by taking the previous recognized words into account. In the most applications of speech recognition, it is crucial to exploit vast information about the order of the words. For this purpose, statistical and grammatical language modeling methods are common approaches utilized in spoken human-computer interaction. These methods are used by Nevisa to improve its accuracy.

3.3.1 Statistical language modeling

In statistical approaches, we take a probabilistic viewpoint of language modeling and estimate the probability P(W) for a given word sequence $W = w_1w_2, ..., w_n$. The simplest and most successful statistical language models are the Markov chain (n-gram) source models, first explored by Shannon [43]. To build statistical language models, we have used the both first edition [25] and second edition [26] of the *Peykare* corpus. As mentioned in Sect. 2.2.2, the first edition of this corpus contains about ten million words that are annotated with POS tags. Using this corpus, we constructed different types of n-gram language models. Since the size of this edition of the corpus was not enough for making a reliable word-based n-gram language model, we built POS-based and class-based n-gram language models, in addition to the word-based n-gram model. These language models are used in the intermediate version of Nevisa. The final language model of the Nevisa has been constructed from the second edition of the *Peykare* corpus.

In building the language models using *Peykare* corpus, we faced with two problems. The first problem was orthographic inconsistency in the texts of the corpus. This problem arises from the fact that Persian writing system allows certain morphemes to appear either as bound to the host or as free affixes. Free affixes could be separated by a final form character or with an intervening space. As examples, three possible cases for the plural suffix "h/" and the imperfective prefix "mi" are illustrated in Table 3. In these examples, the tilde (~) is used to indicate the final form marker, which is represented as the control character\u200C in Unicode, also known as the zero-width non-joiner. All the different surface forms of Table 3 are found in the Persian text corpus. Another issue arises from the use of Arabic script in Persian writing, making some words have different orthographic realizations. For example three possible forms for words "mas]uliyat" (responsibility) and "majmu]eye"(the set of) are shown below in Table 4.

Table 3 Examples of different writing styles for plural suffix "h/" and imperfective prefix "mi"

Word	Attached	Intervening space	Final form
Books	کتابها	کتاب ها	کتاب‌ها
They are going	میروند	می روند	می‌روند

Another issue is the inconsistency of text encoding in Persian electronic texts. This problem arises from the use of different code pages by online publishers and people. As a result, some letters such as 'ye' and 'ke' have various encoding. For example, the letter 'ye' has three different encodings in Unicode, i.e., U+0649 and U+064A (Arabic letters 'ye') and U+06CC (Persian letter 'ye').

For solving these probleme, we must replace different orthographic forms of a word by a unique form. The main corrections that are applied on corpus texts are as below:

- All affixes that attached to the host word or separated by an intervening space are replaced with affixes separated with final form character (zero-width non-joiner character). For example, the words "ket/b h/" (the books) and "miravand" (they are going) in the examples above are replaced by "ket/b~h/" and "mi~ravand".
- Different orthographic realizations of a single word are replaced with their standard form ac-cording to the standards of APLL (Academy of the Persian Language and Literature) [44]. For example, all different forms of words "mas]uliyat" and "majmu]eye" in the above example are replaced with their standard forms (form 1 in Table 4)
- Different encodings of a specific character are changed to a unique form. For example, all letters 'ye' that are encoded by U+0649 and U+064A are changed to the letter 'ye' encoded by U+06CC.
- All diacritics (Bound graphemes) appearing in texts are removed. For example, the consonant gemination marker in the word "fann/vari" (technology) is removed resulting in the word "fan/vari"[19].

Table 4 Examples of different orthographic realizations for words "mas]uliyat" and "majmu]eye"

Word	form 1	form 2	form 3
Responsibility	مسئولیت	مسؤولیت	مسوولیت
The set of	مجموعهٔ	مجموعه‌ی	مجموعه

The multiplicity of the POS tags in the corpus was the next problem to be solved. As mentioned earlier, the tag set includes 882 POS tags. While many of them contain detailed information about the words, they are rarely used in the corpus. This results in many different tags for verbs, adjectives, nouns etc. As a solution, we decreased the number of POS tags by clustering them manually according to their syntactical similarity. In addition, for rare and syntactically insignificant POS tags, we used the IGNORE tag. A NULL tag was defined to mark the beginning of a sentence. These modifications reduced the size of the tag set to 166. Finally, the following statistics were extracted from the corpus to build the LMs [18,19]: *unigram statistics of words (The 20,000 most frequent words in the corpus were chosen as the vocabulary set); bigram statistics of words; trigram statistics of words; unigram statistics of POS tags (for 166 tags); bigram statistics of POS tags; trigram statistics of POS tags; number of assigning one POS tag to each word in the corpus (lexical generation statistics)*. After extracting the word-based n-gram statistics, the back-o trigram language model was built using Katz smoothing method [45].

In addition to the word-based and POS-based bigram and trigram models, class-based language models can be optionally used [46]. Class-based language modeling can tackle the sparseness of data in the corpus. In this approach, words are grouped into classes and each word is assigned to one or more classes. To determine the word classes, one can use the automatic word clustering methods like Brown's and Martin's algorithms [46,47]. In these clustering methods, certain information theory criteria, such as average mutual information, are used to make different classes. In Nevisa, the basic idea of Martin's algorithm [47] is used for word clustering. In this algorithm, the words are clustered initially and they are moved between classes iteratively in the direction of perplexity improvement. Although POS-based and class-based n-grams reduce the sparseness of the extracted bigram and trigram models, in many cases the probabilities remain zero or close to zero. To overcome this problem, various smoothing methods [48] such as add-one, Katz [45] and Witten-Bell smoothing [49] were evaluated on POS-based and class-based n-gram probabilities.

The various LMs mentioned above are incorporated in Nevisa in the word decoding phase (Figure 1). In this method, language model scores and acoustic model scores are combined *during the search* in a semi-coupled manner [50]. In this case, when the search process recognizes a new word while expanding different hypotheses, the new hypothesis score is computed via multiplication of following three terms: the n-gram score of new word, the acoustic model score of new word and current hypothesis score. If S_n is the current hypothesis score after recognizing the word w_n and w_{n+1} is the next recognized word after expanding the hypothesis, then the new hypothesis score in logarithm domain is as Eq. 1, where $S_{\mathrm{AM}}(w_{n+1})$ is the acoustic model score for word $w_{n+1}$and $S_{\mathrm{LM}}(w_{n+1})$ is its language model score. Since the scales of $S_{\mathrm{AM}}(w_{n+1})$ and $S_{\mathrm{LM}}(w_{n+1})$ are different, a weight parameter (α_{LM}) is usually applied as language model weight.

$$\log S_{n+1} = \log S_n + \log S_{AM}(w_{n+1}) + \alpha_{\mathrm{LM}} \cdot \log S_{\mathrm{LM}}(w_{n+1}) \quad (1)$$

The score of POS-based bigram and trigram language models are respectively computed as Eqn. 2 and Eq. 3, in which T_n and T_{n-1} are the most probable POS tags for the words w_n and w_{n-1}.

$$S_{bi}^{pos}(w_{n+1}) = \max_i [P(T_i|T_n) \cdot P(w_{n+1}|T_i)] \quad (2)$$

$$S_{tri}^{pos}(w_{n+1}) = \max_i \left[P(T_i|T_{n-1}T_n) \cdot P(w_{n+1}|T_i)\right] \quad (3)$$

In addition, the language model score for class-based bigram and trigram language models can be computed [19]. As shown in Figure 1 by dotted line, the statistical LM can be applied to the system at the end of the search by n-best re-scorer.

3.3.2 Grammatical language models

Grammar is a formal specification of permissible structures for the language that is used as another important linguistic knowledge source besides the statistical language models in speech recognition systems. In Nevisa, as in the most of the developed speech recognition systems, the output is a set of n-best hypotheses that are ordered based on their acoustic and language model scores. The output sentences do not have the true syntactic structure necessarily. For making high scored syntactic outputs a grammatical model of the language and a syntactic parser are necessary. The grammatical model includes a set of rules and syntactic features for each word in the vocabulary. The rule set describes syntactic structures of permissible sentences in the language. The syntactic parser analyzes the output hypotheses of the recognition system and rejects the non-grammatical hypotheses.

Various methods have been presented for specifying the syntactic structure of a language in the last two decades [51-53]. Generalized phrase structure grammar (GPSG) [52] is a syntactic formalism that considers language sentences as sets of phrases by assuming each phrase as a combination of smaller phrases. Using linguistic expertise and consultation, about 170 grammatical rules for Persian language using GPSG idea [20] were extracted. The employed GPSG was modified to be consistent with the Persian language. The little modified X-bar theory [54] was used for defining syntactic categories. Noun (N), verb

(V), adjective (ADJ), adverb (ADV) and preposition (P) were selected as the basic syntactic categories. These basic categories could be used as the head for larger syntactic categories like noun phrase, verb phrase, adjective phrase etc. For each syntactic category and phrase, we specify features; the features describe the lexical, syntactic, and semantic characteristics of the words. To each feature, a name and its possible values are assigned. For example, Plurality (PLU) is a binary[a] feature and its possible values are + (plural) or - (singular) and Person (PER) is an atomic[b] feature and its possible values are 1, 2, 3. After specifying categories and phrases, syntactic structures of various phrases are illustrated based on smaller syntactic categories. As an example, the following rule is one of the grammatical rules that describe noun phrases (N1) in Persian. This rule shows the noun phrase structure when the noun combines with another noun phrase as a genitive.

$$N1 \rightarrow *N1^{-}\,[GEN+, PRO-]\,N2(P2)\,(S\,[COMP+, GAP]) \quad (4)$$

In this rule, $N1^{-}$ (a noun with possibly an adjective) must have Ezafe[C] enclitic (GEN +) and non-pronoun (PRO -) head. N2 points to a complete Noun phrase (a noun with pre-modifiers and post-modifiers). It means that a complete Noun phrase can play the role of genitive for Noun. In addition, this rule shows that the other post-modifiers of noun (P2 and S) can be combined optionally. P2 points to the prepositional phrase and S[COMP +] points to the complement sentence (relative clause). The feature COMP with + value indicates that the sentence must have Persian complementizer "ke" (that, which). Similar to this rule, we write other rules for describing various syntactic structures of Persian. Furthermore, a 1,000-word vocabulary with syntactic features was annotated.

Analyzing a sentence and checking the compatibility of its structure with the grammar needs a parsing technique. Parsing algorithm offers a procedure that searches through various ways of combining grammatical rules to find a combination that generates a tree to illustrate the structure of the input sentence. This is similar to the search problem in speech recognition. A top-down chart parser [5] is incorporated in Nevisa. The grammatical language model integration in Nevisa is done in a loosely-coupled manner, as shown in Figure 1, at the end of the search process. The Parser takes the n-best list from the word decoder, analyzes each sentence according to grammatical rules and accepts the grammatically correct sentences as the output of the system.

4 Experiments and results

4.1 System parameters

In the acoustic front-end, speech signal is blocked into 20 ms frames with 12 ms overlap if sampled with 22050 Hz sampling rate, and with 25 ms of speech signal and 15 ms of overlap in the case of 16 kHz sampling rate. A pre-emphasis filter with a factor of 0.97 is applied to each frame of speech. A Hamming window is also applied to the signal in order to reduce the effect of frame edge discontinuities. After performing fast Fourier transform (FFT), the magnitude spectrum is warped according to the signal's warping factor if the VTLN option is used. The obtained spectral magnitude spectrum values are weighted and summed up using the coefficients of 40 triangular filters arranged on the Mel-frequency scale. The filter output is the logarithm of sum of the weighted spectral magnitudes. Discrete cosine transform (DCT) is then applied resulting in 13 cepstral coefficients. The first and the second derivatives of cepstral coefficients are calculated using linear regression method [23] over a window covering seven neighboring cepstrum vectors. This makes up vectors of 39 coefficients per speech frame. Finally, PCA and/or CMS are used in the cases these options are activated.

Nevisa uses phone (context independent) and triphone (context dependent) HMM modeling. All HMMs are left-to-right; forward, skips and self-loop transitions are allowed. The elements of the feature vectors are assumed uncorrelated resulting in diagonal covariance matrices. The parameters are initialized using linear segmentation and then the segmental k-means re-estimation algorithm finalizes the parameters after ten iterations. The beam width in the decoding process is 70 and the stack size is 300.

4.2 Results of language model incorporation

In this section, the evaluation results of incorporating of language models in the Nevisa system are reported. An intermediate version of Nevisa is used in the experiments of this section. The system is trained on 29 Persian phonemes with silence as the 30th phoneme. All HMMs are left-to-right and composed of six states and 16 Gaussian mixture components per state. The vocabulary size is about 1,000 words and the first edition of the text corpus is used for building the statistical language models. In these evaluations, *sFarsdat train* and *sFarsdat test* are used as train and test sets, respectively. Two different criteria were used to evaluate the efficiency of the language model variants: the perplexity and word error rate (WER) of the system.

Table 5 shows the results of Nevisa system on *sFarsdat test set* using WER as the evaluation criteria. As mentioned in Sect. 2.1, the test set contains 140 sentences from seven speakers. The Witten-Bell smoothing technique [49] was used for POS-based and class-based language models. In class-based evaluation, we used 200 classes. As the results show, the base-line (BL) with no language model, results in high WER. The word-based statistical

Table 5 Performance of Nevisa in clean condition (word level)

LM Method	WER%
BL, No LM	38.14
POS-based trigram	24.68
Class-based trigram	23.40
Word-based trigram	21.76
POS-based trigram+Grammar	18.2

LM provides higher improvement compared to other statistical LMs. Therefore, in all of the experiments in the following sections, we use the word-based LM. In the results of Table 5, the WER reduction obtained by using the grammar in the system is noticeable.

Table 6 shows the perplexity computed on the 750 sentences (about 10,000 words) of *gFarsdat test set* based on word-based n-gram model. In order to reduce the required memory size for language model, infrequent n-grams were removed from the model. The counts below which the n-grams are discarded are referred to as cutoffs [55]. Table 6 shows how the bigram and trigram cutoffs affect the size (in Mega bytes) and perplexity of a trigram language model. This table shows that the cutoffs noticeably reduce the size of language model, but do not increase the perplexity significantly. Considering Table 6, we have chosen the cutoffs 0 and 1 for bigram and trigram counts, respectively.

4.3 Results for robustness techniques

The recognition system described in section 4.2 is used to provide results for this section. Here, *sFarsdat train* is used to train phone models with six states for each model and 16 Gaussian mixture in each state. The vocabulary contains about 1,000 words and the word-based trigram language model is used. Evaluation test sets of *FANOS* database are used in these experiments.

Like all other recognition systems, the performance of Nevisa is degraded in adverse noisy conditions. Equipping this system with various compensation methods has made it robust to different noise types. Table 7 shows the recognition results of the system on four noisy tasks on *FANOS* corpus. The baseline WERs of the system on this speech corpus are very high. The recognition rates on task C and task D are negative due to the high insertion error rate. The performance of the system is considerably improved by using speaker and environment compensation methods. Table 7 shows the improvements in WER as a result of applying robustness methods. VTLN provides better compensation for less-noisy environments like tasks A and B, while PMC and PC-PMC result in higher compensation in more noisy environments. In the PC-PMC method, the number of features is reduced by 25% from 36 to 25. MLLR and MAP adapt the acoustic models to environmental conditions, microphone and speaker's signal properties. MAP results in high adaptation ability whenever the adaptation data is enough, and MLLR provides better adaptation in less-noisy conditions compared to noise-dominant conditions. The combination of PC-PMC and MLLR results in high system robustness in the presence of all noise types.

Table 6 The effect of cutoffs on the size and perplexity of a back-off trigram language model

Cutoffs (bigram)	Cutoffs (trigram)	Perplexity	Size (MB)
0	0	134.54	36
0	1	134.76	20
0	2	135.82	17
1	1	143.18	10
1	2	143.26	7.8

Table 7 Evaluation of Nevisa and the robustness methods on FANOS noisy tasks (WER% on word level)

Robustness	Task A	Task B	Task C	Task D
None	74.04	75.32	116.41	105.94
VTLN+MLLR	30.37	32.87	82.52	60.07
PMC-MAP	38.63	50.49	69.36	50.22
PC-PMC+MLLR	31.33	28.70	56.17	42.11

4.4 Final results

The final results of continuous speech recognition using Nevisa system are summarized in Table 8. According to the intermediate experiments, some of which were reported in previous sections, the final parameters of the system are optimized. The parameters of the front-end are the values described in sect. 4.1. CMS normalization is used as a permanent processing unit in the system. Context-independent (phone) and context-dependent (triphone) modeling are done using both small and large Farsdat corpus. In all experiments, the HMMs are made up using five states and eight Gaussian mixtures per state. 29 phone models and a silence model are used for the context-independent task using small Farsdat. The same acoustic models with two additional models, noise

Table 8 WER% of Nevisa on small and large Farsdat using context-independent (phone) and context-dependent (triphone) modeling

Train		Test	
Databse	Context	*gFarsdat*	*sFarsdat*
sFarsdat	Independent	29.60	25.77
sFarsdat	Dependent	20.51	16.79
gFarsdat	Independent	6.10	37.39
gFarsdat	Dependent	5.21	26.85

and breath, are used in context-independent modeling with large Farsdat. In the context-dependent modeling with small Farsdat (*sFarsdat train*) the states are tied into 500 senones while they are tied into four thousand senones in modeling with large Farsdat (*gFarsdat trai*n). In the experiments given in Table 8, word-based back-o trigram language model extracted form the second edition of the text corpus and the vocabulary size of 20,000 words are used.

As shown in Table 8, generally the performance of the system with *sFarsdat test* is lower than with *gFarsdat test*. This is due to the mismatch of the language model between the sentences of *sFarsdat test* and the text corpus. As indicated in sect. 4.1, the sentences of small Farsdat are designed artificially to cover the Persian acoustic variations and do not have a compatible language model with regular Persian texts such as the *Peykare*. Training the triphone models with small Farsdat provides higher WER in comparison with large Farsdat because the training data in small Farsdat is not enough for context-dependent modeling. Due to the small size of the *sFarsdat train*, the numbers of final tied states are reduced to 500. Furthermore, the acoustic mismatch between train and test conditions (train with *sFarsdat train* and test using *gFarsdat test* or vice versa) intensifies the increase of WER. The best performance of the system was obtained in the case of context-dependent modeling using large Farsdat database.

5 Summary and conclusion

Nevisa system was introduced as the first large vocabulary speaker-independent continuous speech recognition system for Persian language. The conventional and customized techniques for different modules of the system were incorporated. For each module, necessary modifications and parameter optimizations were performed. The parameter set for each part of the system was found by separately evaluating the performance of that part with different parameter values. The system was developed in the process of academic and industrial teamwork and was intended to be an exploitable product. Therefore, the problems of noisy environments and speaker variations had to be handled. Various robustness techniques were tried and optimized for this purpose. We also customized and utilized statistical and grammatical language models for Persian language. The general n-gram statistics of Persian were extracted and incorporated for the first time. Our evaluation results and real environ-mental tests show that the system is performing satisfactorily enough to be used by typical users.

We are now continuing our research for improved versions of Nevisa. We are using context-dependent acoustic phone units (e.g. triphones), increasing the vocabulary size and improving our language models for this purpose. We are also working on specific language models for medical, legal, banking and office automation applications.

Notes

[a]The binary features are the features that take only two possible values.

[b] The atomic features are the features that take more than two possible values.

[c] Ezafe is short vowel that makes genitives in Persian

Competing interests
The authors declare that they have no competing interests.

Received: 18 January 2011 Accepted: 5 October 2011
Published: 5 October 2011

References

1. LR Rabiner, Challenges in speech recognition and natural language processing, in *SPECOM* (June 25 2006)
2. S Furui, 50 years of progress in speech and speaker recognition research. *Trans Comput Information Technology ECTI-CIT*. **1**(2), 6474 (2005)
3. X Huang, A Acero, HW Hon, *Spoken Language Processing* (Prentice Hall, Upper Saddle River, NJ, USA, 2001)
4. L Rabiner, BH Juang, *Fundamentals of Speech Recognition* (Prentice Hall, Upper Saddle River, NJ, USA, 1993)
5. J Allen, *Natural Language Understanding* (Benjamin-Cummings Publishing Co. Inc., Redwood City, CA, USA, 1995)
6. B Babaali, H Sameti, The sharif speaker-independent large vocabulary speech recognition system, in *The 2nd Workshop on Information Technology & Its Disciplines (WITID 2004)*, (Kish Island, 2004), pp. 24–26
7. H Sameti, H Movasagh, B Babaali, M Bahrani, K Hosseinzadeh, A Fazel Dehkordi, HR Abu-talebi, H Veisi, Y Mokri, N Motazeri, M Nezami Ranjbar, Large vocabulary persian speech recognition system, in *1st Workshop on Persian Language and Computer*, 69–76 (May 24–26 2004)
8. H Movasagh, Design and implementation of an optimized search method for hmm-based persian continuous speech recognition. Ms thesis, Sharif University of Technology (2004)
9. B Babaali, Incorporating pruning techniques for improving the performance of an hmm-based continuous speech recognizer. Ms thesis, Sharif University of Technology (2004)
10. H Sameti, H Veisi, M Bahrani, B Babaali, K Hosseinzadeh, Nevisa, a persian continuous speech recognition system, in *Communications in Computer and Information Science* (Springer Berlin Heidelberg, 2008), pp. 485–492
11. M Sheikhan, M Tebyani, M Lotfizad, Continuous speech recognition and syntactic processing in iranian farsi language. *Inter J Speech Technol*. **1**(2), 135 (1997). doi:10.1007/BF02277194
12. SM Ahadi, Recognition of continuous persian speech using a medium-sized vocabulary speech corpus, in *European Conference on Speech communication and technology (Eurospeech'99)*, (Geneva, Switzerland, 1999), pp. 863–866
13. N Srinivasamurthy, SS Narayanan, Language-adaptive persian speech recognition, in *European Conference on Speech Communication and Technology (Eurospeech'03)*, Geneva (2003)
14. F Almasganj, SA Seyyed Salehi, M Bijankhan, H Razizade, M Asghari, Shenava 2: a persian continuous speech recognition software, in *The first workshop on Persian language and Computer* (Tehran, 2004), pp. 77–82
15. LR Rabiner, A tutorial on hidden markov models and selected applications in speech recognition. Proc IEEE. **77**(2), 257–286 (1989). doi:10.1109/5.18626
16. S Ortmanns, A Eiden, H Ney, Improved lexical tree search for large vocabulary speech recognition, in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'98)* (1998)
17. H Ney, R Haeb-Umbach, BH Tran, M Oerder, Improvements in beam search for 10000-word continuous speech recognition. IEEE Trans Acoust Speech, Signal Process. **2**, 353–356 (1992)

18. M Bahrani, H Sameti, M Hafezi Manshadi, A computational grammar for persian based on gpsg, in *2nd Workshop on Persian Language and Computer*, Tehran, (2006)
19. M Bahrani, H Sameti, Building statistical language models for persian continuous speech recognition systems using the peykare corpus. Intern J Comp Process Lang. **23**(1), 1–20 (2011). doi:10.1142/S1793840611002188
20. M Bahrani, H Sameti, M Hafezi Manshadi, A computational grammar for persian based on gpsg. Lang Resour Eval, 1–22 (2011)
21. M Bijankhan, J Sheikhzadegan, MR Roohani, Y Samareh, C Lucas, M Tebyani, Farsdat-the speech database of farsi spoken language, in *Proceeding of 5th Australian International Conference on Speech Science and Technology*, 826–831 (1994)
22. J Sheikhzadegan, M Bijankhan, Persian speech databases, in *2nd Workshop on Persian Language and Computer*, 247–261 (2006)
23. H Veisi, Model-based methods for noise robust speech recognition systems. Ms thesis, Sharif University of Technology (2005)
24. K Hosseinzadeh, Improving the accuracy of continuous speech recognition in noisy environments, Ms thesis (Sharif University of Technology, 2004)
25. M BijanKhan, Persian text corpus, in *1st Workshop on Persian Language and Computer*, Tehran, (2004)
26. M Bijankhan, J Sheykhzadegan, M Bahrani, M Ghayoomi, Lessons from building a persian written corpus: Peykare. Lang Resour Eval. **45**(2), 143–164 (2011). doi:10.1007/s10579-010-9132-x
27. P Zhan, M Westphal, M Finke, A Waibel, Speaker normalization and speaker adaptation- a combination for conversational speech recognition, in *European Conference on Speech Communication and Technology (EUROSPEECH'97)*, Greece, ISCA 2087–2090 (1997)
28. D Pye, PC Woodland, Experiments in speaker normalisation and adaptation for large vocabulary speech recognition, in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'97)*, Munich 1047–1050 (1997)
29. H Veisi, H Sameti, B Babaali, K Hosseinzadeh, MT Manzuri, Improving the robustness of persian large vocabulary continuous speech recognition system for real applications, in *IEEE International Conference on Information and Communication Technologies (ICTTA'06)*, 1293–1297 (April 24–26 2006)
30. H Veisi, H Sameti, The integration of principal component analysis and cepstral mean subtraction in parallel model combination for robust speech recognition. Digit Signal Process. **21**(1), 36–53 (2011). doi:10.1016/j.dsp.2010.07.004
31. MJF Gales, *Model-based Techniques for Noise Robust Speech Recognition*, (Phd thesis, University of Cambridge, 1995)
32. H Veisi, H Sameti, The combination of cms with pmc for improving robustness of speech recognition systems, in *Communications in Computer and Information Science*, (Springer Berlin Heidelberg, 2008), pp. 825–829
33. CJ Leggetter, PC Woodland, Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models. Comput Speech Lang. **9**(2), 171 (1995). doi:10.1006/csla.1995.0010
34. PC Woodland, Speaker adaptation: Techniques and challenges. IEEE Workshop on Automatic Speech Recognition and Understanding, 85–90 (1999)
35. SJ Young, PC Woodland, The use of state tying in continuous speech recognition, in *European Conference on Speech Communication and Technology (EUROSPEECH'93), ISCA*, Berlin, 2203–2206 (22–25 September 1993)
36. SJ Young, JJ Odell, PC Woodland, Tree-based state tying for high accuracy acoustic modeling, in *Proceedings of the Workshop on Human Language Technology*, Association for Computational Linguistics Morristown, NJ, 307–312 (1994)
37. JJ Odell, *The Use of Context in Large Vocabulary Speech Recognition*, (Phd thesis, Cambridge University, 1995)
38. MY Hwang, F Alleva, X Huang, Senones, multi-pass search, and unified stochastic modeling in sphinx-ii, in *European Conference on Speech Communication and Technology (EU-ROSPEECH'93)*, Berlin. ISCA (22–25 September 1993)
39. PJ Moreno, *Speech Recognition in Noisy Environments*. (Phd thesis, Carnegie Mellon University, 1996)
40. A Acero, *Acoustical and environmental robustness in automatic speech recognition*. (Phd thesis, Carnegie Mellon University, 1990)
41. L Welling, H Ney, S Kanthak, Speaker adaptive modeling by vocal tract normalization. IEEE Trans Speech Audio Process. **10**(6), 415–426 (2002). doi:10.1109/TSA.2002.803435
42. MJF Gales, PC Woodland, Mean and variance adaptation within the mllr framework. Comput Speech Lang. **10**(4), 249–264 (1996). doi:10.1006/csla.1996.0013
43. C Shannon, A mathematical theory of communication. Bell Sys Tech J. **27**, 398–403 (1948)
44. A Ashraf Sadeghi, Z Zandi Moghadam, *The dictionary of Persian orthography*, (The Acad Persian Lang Lit, 2005)
45. S Katz, Estimation of probabilities from sparse data for the language model component of a speech recognizer. IEEE Trans Acoust Speech Signal Process. **35**(3), 400–401 (1987). doi:10.1109/TASSP.1987.1165125
46. PF Brown, RL Mercer, VJ Della Pietra, JC Lai, Class-based n-gram models of natural language. Comput Linguist. **18**(4), 467–479 (1992)
47. S Martin, J Liermann, H Ney, Algorithms for bigram and trigram word clustering. Speech Commun. **24**(1), 19–37 (1998). doi:10.1016/S0167-6393(97)00062-9
48. SF Chen, J Goodman, An empirical study of smoothing techniques for language modeling, in *Proceedings of the 34th annual meeting on Association for Computational Linguistics*, Santa Cruz, California, Association for Computational Linguistics Morristown, NJ, USA, 310-318 (1996)
49. IH Witten, TC Bell, The zero-frequency problem: Estimating the probabilities of novel events in adaptive text compression. IEEE Transactions on Information Theory. **37**(4), 1085–1094 (1991). doi:10.1109/18.87000
50. MP Harper, LH Jamieson, CD Mitchell, G Ying, S Potisuk, PN Srinivasan, R Chen, CB Zoltowski, LL McPheters, B Pellom, Integrating language models with speech recognition, in *Proceedings of the AAAI-94 Workshop on the Integration of Natural Language and Speech Processing*, 139–146 (1994)
51. RM Kaplan, The formal architecture of lexical functional grammar, in *Formal Issues in Lexical-Functional Grammar*, Center for the Study of Language (CSLI), 7–28 (1995)
52. G Gazdar, E Klein, G Pullum, IA Sag, *Generalized Phrase Structure Grammar* (Harvard University Press, 1985)
53. AK Joshi, L Levy, M Takahashi, Tree adjunct grammars. Journal of Computer and System Sciences. **10**(1), 136–163 (1975). doi:10.1016/S0022-0000(75)80019-5
54. A Radford, *Transformational grammar: a first course* (Cambridge University Press, Cambridge, 1988)
55. P Clarkson, R Rosenfeld, Statistical language modeling using the cmu-cambridge toolkit, in *European Conference on Speech Communication and Technology (EUROSPEECH'97)*, ISCA, Rhodes 2707–2710 (September 22–25 1997)

doi:10.1186/1687-4722-2011-426795
Cite this article as: Sameti *et al.*: **A large vocabulary continuous speech recognition system for Persian language.** *EURASIP Journal on Audio, Speech, and Music Processing* 2011 **2011**:6.