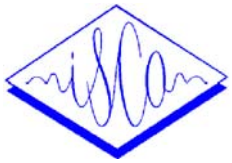

6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

RECOGNITION OF CONTINUOUS PERSIAN SPEECH USING A MEDIUM-SIZED VOCABULARY SPEECH CORPUS

S.M. Ahadi
Electrical Eng. Dept., Amirkabir University
Hafez Ave., Tehran 15914, Iran.
Email: sma@cic.aku.ac.ir

10.21437/Eurospeech.1999-210

ABSTRACT

Speech recognition in Persian (Farsi) has recently been addressed by a few native speaking researchers and some approaches to isolated word and phoneme recognition have been reported. A main bottleneck in this research field is the lack of a recognition-specific speech corpus. In this work, a phonetically balanced speech database of Persian has been modified and used in continuous speech recognition. A basic continuous speech recognizer using HMMs has been designed for this language and recognition tests have been performed. Using mixture-Gaussian monophone models, a word recognition rate of about 68% in no-grammar tests were obtained while word-pair grammar tests increased this rate to an unexpectedly high value of 99.5%. The reason is found to be the low grammar perplexity of the database which is not suitable for recognition applications. This obviates the need for a Persian speech corpus specifically designed for such tasks.

1. INTRODUCTION

Recent advances in speech recognition have emphasized the usefulness of this field of research in today's advanced world of information interchange. However, only few languages have been addressed by the researchers in this field. Persian, as the native language of a few of the Middle-eastern countries, is spoken by tens of millions of people. Speech recognition of Persian (Farsi) language has recently been addressed by a few native (Persian speaking) researchers. Most of the work done so far was concentrated on isolated word recognition [1][2]. However, recently a few approaches to Persian phoneme recognition have been proposed [3] and continuous speech recognition of Persian language has been paid attention. A main bottleneck in this field of research has been found to be the lack of a speech corpus, specifically designed for speech recognition tasks.

In order to be able to initiate a research in this field, an already available phonetically balanced continuous speech corpus of Persian language was utilized. This speech corpus was designed as a general purpose speech research tool and consisted of a medium-sized vocabulary. However, special needs of a speech recognition database were not implemented in its design.

In this work, a large amount of effort was made to overcome the problems of this database for the speech recognition task. Using the modified database, a continuous density HMM-based speech recognition system was built and tests carried out on both single and mixture Gaussian models. The following sections discuss the available database, its modification, the structure of the recognition system, its implementation and the results obtained. In the conclusion, we discuss the reasons for obtaining such results and make some suggestions for the future work.

2. THE DATABASE

The database, called *Farsdat* [4] is consisted of about 6000 Persian utterances, uttered by 300 speakers (20 utterances per speaker). In total, about 400 sentences exist in the database. Each speaker has uttered 18 randomly chosen sentences of the above mentioned set plus two sentences which were common for all speakers. The recording of the utterances for every speaker was carried out in two different sessions. The sentences were formed by using over 1000 most frequent Persian words chosen from various types of articles published in daily Persian newspapers. Some other words were then added to the above set to form sentences, which include all the possible combinations of the Persian phonemes. This results in the availability of all Persian allophones in this database.

The speakers were chosen from 10 different geographical regions of Iran so that 10 most common dialects of the Persian language were available in the database. Having a male to female population ratio of two to one, the database was recorded in a low noise environment, featuring a 34 dB signal to noise ratio.

3. MODIFICATION OF THE DATABASE

The first step before the implementation of a speech recognizer, was the modification of the database so that it would suit the requirements for speech recognition. To perform this, firstly all the utterances of the database were extracted and saved separately. In order to omit the effect of dialects in this first stage of recognizer implementation, the speakers were confined to those from Tehran origin (Tehrani accent). This limited the number of speakers to 147.

All the resulting 2940 files were then listened to and the corrupt utterances, those with recording problems and those of the speakers with speaking deficiencies were marked and deleted. This resulted in a total of 2707 utterances from 136.5 speakers, of which 1615 utterances (about 60%) belonged to male speakers and 1092 utterances (about 40%) belonged to female speakers. A survey on the remaining part of the database showed that most of the existing sentences were uttered between 3 to 10 times by different speakers.

The partitioning of the available data into training and evaluation sections was carried out by randomly selecting one of every 3 speakers as test speakers and allocating the remaining as training speakers to the training section. In total, 1814 utterances were allocated to training section, while 893 sentences were named as evaluation sentences. The ratios of the population of male to female speakers in both sections were about 60 to 40.

4. THE RECOGNITION SYSTEM

The first step in the building process of the recognition system was to introduce a set of models. The models for these experiments were based on the set of basic Persian phonemes. This set is shown in Table 1 and consists of 30 basic phonemes. As can be seen, simpler pronunciation structure of the Persian language has resulted in a smaller basic set of phonemes. A special item in this set is the glottal stop (/ʔ/) which might occur in many occasions at the start, in the middle or sometimes even at the end of a word in Persian (mostly in the words originating from Arabic). Another special case is the diphthong /ow/. Although this sound is characterized differently by Persian linguists (e.g. refer to [5]), we have defined it as a diphthong and included it in the list of our models. This was due to its frequent appearance in Persian language and unavailability of the sound /w/ in the contemporary Persian. A set of 32 monophone models for the recognition system using this basic set of phonemes plus silence and between-word space models was constructed. All the models, except

Table 1. The phonemes of the Persian language and examples of their usage.

Phonetic Group	Phoneme	Example
Vowels	ɑ	xɑb
	æ	sæbr
	ɛ	tʃɛrɑɢ
	u	ruz
	i	diruz
	o	ɡozæʃt
	ow	rowʃæn
Liquids	r	rɑh
	l	lɑlɛ
Glide	j	mɛjdɑn
Nasals	m	mærdom
	n	nɑmɛ
Plosives	b	bɑzɑr
	p	pɑrɛ
	t	tærtib
	d	dæʃt
	ɢ	ɢævi
	k	kɛtɑb
	ɡ	ɡoruh
	ʔ	mæʔlum
Fricatives	f	fɑrsi
	s	særmɑ
	v	vɑrzɛʃ
	x	xɑnɛ
	z	zɛræng
	ʒ	ʒɑlɛ
	ʃ	ʃoʔlɛ
	h	huʃ
Affricates	dʒ	dʒɑmɛd
	tʃ	tʃirɛ

that of the between-word space, were 3-state left-to-right continuous density HMMs with either single- or mixture-Gaussian pdfs. The model for the between-word space consisted of a HMM with only one emitting state. Also, null transitions were included in the composite networks to allow for the bypass of the between-word spaces if required.

The vocabulary size of the database in this stage was 1112 words which consisted of the main words chosen during the building stage of the database plus their derivatives as pronounced in the sentences. It should be noted that due to the special structure of the Persian language, when sentences are built using the original words, the pronunciation of some of the words might

change according to the sentence structure, resulting in a larger set of items in the word pronunciation dictionary.

In order to perform model training, firstly word-level transcriptions for all the sentences were prepared using the sentence definitions. A monophone word dictionary for all the words available in the vocabulary was created and used to build phone-level transcriptions for all the training sentences.

Before starting the training phase, a set of seed models were needed. For the initialization of these seed models, time-aligned transcriptions for a fairly small part of the database (119 utterances) were extracted manually. Using these utterances, their time-aligned transcriptions and a set of single-Gaussian prototype models, an initialization of the models was performed by uniform segmentation of training observations and repeated re-estimation of the seed model parameters.

A Baum-Welch re-estimation [6] was then carried out on the models by building a composite network of models for each given training utterance and using all the remaining utterances in the training set for this purpose. This resulted in the trained models, which can be used in the recognition process. The recognition process made use of a standard Viterbi search algorithm to find the appropriate network of words for the given speech files [6][7].

5. IMPLEMENTATION

In the implementation phase, all the available speech data in the training and evaluation sections defined were coded into mel cepstral coefficients. The speech data was originally sampled using a 44.1KHz sampling frequency and the samples were 16 bit wide. A downsampling was performed to convert the sampling rate of the utterance files from the original 44.1KHz to 16KHz. Consequently, all the time-aligned transcriptions had to be modified to support this sampling rate. The resultant speech signal was then applied to a preemphasis filter with a coefficient of 0.95 and blocked into frames of 25 msec. with 15 msec. of overlap. A hamming window was also applied to the signal to reduce the effects of frame edge discontinuities. Mel-cepstral coefficients were then computed using a bank of 24 triangular filters distributed on a mel-scale, resulting in 12 cepstral coefficients. A cepstral weighting (liftering) was then applied to the cepstral coefficients and delta and delta-delta cepstral and log energy coefficients were also added to the above to make up vectors of 38 coefficients per speech frame [6].

A set of 32 seed models was then built and initialized as explained in section 4. The models in this stage of the experiments were chosen to be single-Gaussians. 20 iterations of forward-backward reestimation was carried out using the smaller set of training data with time-aligned transcriptions, updating means and variances of the model pdfs. In all training iterations, a requirement of minimum 3 examples of model data among the training data was set in order to inhibit the training of the models with insufficient training data. The whole set of training data was then used to update model parameters in 4 iterations of further reestimation.

In order to build mixture-Gaussian models, a mixture incrementing procedure was followed [8]. This was done by splitting the mixture component with the largest weight in every step. Hence, the number of mixture components (say N) was increased by one and 4 iterations of training were performed on the resultant models to lead to an N+1-component system. Further steps were then taken if larger number of mixture components were desired. The splitting procedure consisted of dividing the mixture weight by two to generate two identical mixture components with half the original weights and perturbing the means of the two new components in different directions.

Table 2. System test results (recognition word accuracy) without the application of any grammar (NG) and with word-pair grammar (WPG) using different numbers of mixture components.

	No. of mixture components					
	1	2	3	4	5	6
NG	34.5	47.4	55.7	61.6	65.1	68.0
WPG	94.3	98.2	98.8	99.2	99.4	99.5

6. RESULTS AND DISCUSSION

The above system was implemented on a Pentium-based PC compatible machine. Two types of recognition tests were carried out, i.e. one without application of any grammar (specified as NG) and another one with a word-pair grammar (specified as WPG). Table 2 illustrates the recognition results obtained. In all the recognition tests performed, all available test sentences, i.e. 893, were used. As can be seen, in both cases (NG and WPG) increasing of the mixture components leads to a better result. However, the single-Gaussian no-grammar case resulted in a low 34.5% recognition rate, while increasing the mixture components, consistently improved the results leading to a 68.0% recognition rate with 6-component models.

The introduction of a word-pair grammar to the recognition process increased the recognition accuracy of all available mixture-Gaussian systems to unexpectedly high values of 94.3% to 99.5% for single-component to 6-component systems respectively. These abnormally

high recognition rates were found to be due to the extremely low perplexity of the database for this language model. This is due to the fact that the database was originally designed with the aim of providing the researchers with a phonetically balanced database, including all possible allophones of different phones of the Persian language. Hence no attention was paid to such matters as grammar perplexity of the task. One other point to note is that the number of mixture components was not increased further in these experiments, although some improvement could still be expected, especially from no-grammar case. This was decided taking into account the results of word-pair grammar case, where the improvement of the results with further increase in the number of mixture components was expected to be negligible.

7. CONCLUSION

The results of the no-grammar tests suggest that further work should be carried out to enable us achieve better recognition rates. However, the results of the word-pair grammar tests are already saturated. Hence, further boost to recognition accuracy, especially when a language model is used, cannot be achieved with the available database and the need for a database specifically designed for speech recognition purposes is obvious.

A natural extension of the above recognition system should be the implementation of the context-dependent models. However, we will only be able to assess the improvements when no-grammar tests are performed and word-pair or other grammars do not seem to be able to demonstrate any improvements obtained from context-dependent modeling. Thus, the availability of an appropriate database is of paramount importance at this stage. Obviously, in the design of such a database, other important issues such as larger vocabularies can also be taken into consideration.

8. REFERENCES

[1] Sh. Rostamzadeh, S.M. Ahadi and H. Sheikhzadeh, "Speaker-Independent, Isolated-Word Persian Speech Recognition Using CDHMMs" (in Persian), *in Proc. 6th Iraninan Conference on Elect. Eng.*, Tehran, May 1998.

[2] F. Fekri, M.R. Nakha'i and M. Tebyani, "Speech Recognition by Computer" (in Persian), *in Proc. 2nd Iraninan Conference on Elect. Eng.*, Tehran, May 1994.

[3] J. Shirazi, M. Mirsalehi, "Recognition of Persian Speech Consisting of Eighteen Phonemes Using Self-Organizing Kohonen Neural Network" (in Persian), *in Proc. 4th Iraninan Conference on Elect. Eng.*, Tehran, May 1996.

[4] M. Bijankhan *et al.*, "FARSDAT - The Speech Database of Farsi Spoken Language", in *Proc. of the fifth Australian Int. Conf. On Speech Science and Tech.*, Perth, Dec. 1994, Vol. II.

[5] Y. Samareh, *Phonology of the Persian Language*. (in Persian), 4th Ed., Tehran: University Publication Center, 1995.

[6] L.R. Rabiner, B.H. Juang, *Fundamentals of Speech Recognition*. Prentice-Hall, New Jersey, 1993.

[7] S.J. Young, N.H. Russel and J.H.S. Thornton. Token Passing: A Simple Conceptual Model for Connected Speech Recognition Systems. Technical Report F_INFENG/TR38, Cambridge University Eng. Dept., 1989.

[8] S.J. Young. *HTK: Hidden Markov Model Toolkit V1.4, Reference Manual*. Cambridge University Eng. Dept., 1992, Cambridge, England.