FARSDAT – THE SPEECH DATABASE OF FARSI SPOKEN LANGUAGE

M.Bijankhan, J.Sheikhzadegan**, M.R.Roohani***, Y.Samareh*, K.Lucas****, M.Tebyani******

** The University of Tehran, Linguistics Dept., Tehran, Iran.*

*** The Tarbijat Modarres University, Tech. & Eng. Dept., Tehran, Iran.*

****The Amir Kabir University of Technology, Elect., Eng.Dept., Tehran, Iran.*

***** The University of Tehran, Elect. Eng. Dept., Tehran, Iran.*

******The Sharif University of Technology, Elect. Eng. Dept., Tehran, Iran.*

ABSTRACT- The FARSDAT has been produced for speech and speaker recognition purposes,and linguistic research as well. It consists of 386 sentences read aloud by 300 Farsi native speakers, belonging to one of the ten dialect regions. Two types of sentences are included. Type one composed of 384 phonetically balanced sentences to provide allophones in all phonetic contexts. Type two composed of two sentences to allow dialect comparison. Each speaker read twenty sentences in two sessions. All utterances were divided into two corpora: training and recognition. The speech was sampled at 44.1 Khz by 16-bit sound Blaster hardware on IBM micro computers. 2000 utterances have been manually segmented and labelled, providing a vast amount of computer files of different isolated words, syllables, allophones and some phone sequences.

INTRODUCTION

Large speech corpus is one of the key components to contribute to the progress in speech technology and promotion of speech sciences and linguistic knowledge as well. The only speech corpus in the Farsi language can be found in the Oregon Graduate Institute(OGI) Multi-Language Telephone speech corpus(Muthusamy, Cole & Oshika, 1992). FARSDAT includes a variety of the Farsi speech data uttered by 300 native speakers who differ from each other with regards to age, sex, dialect, and educational level. Each speaker uttered twenty sentences in two sessions. 100 of these speakers uttered 110 isolated words. The speech was collected in both the Linguistics Laboratory of the University of Tehran and an office room with a moderate noise environment. 6000 utterances will be segmented and labelled manually for each phoneme, syllable and some phone sequences, of which 2000 utterences have been done up to now. Several experienced phonetic transcribers performed manual segmentation and labelling using the signal editor of 16-bit sound blaster software. The ambiguities in segmentation have been solved by reference to the corresponding spectrograms extracted from DSP Sona-Graph KAY model 5500. The FARSDAT corpus was designed with acoustic-phonetic research goal and for application projects in automatic speech and speaker recognition. It is a first step toward producing Farsi speech databases to support original and advanced research in speech sciences and technology.

DATA ACQUISITION

The FARSDAT data acquisition has been performed in two phases: data selection and

data collection.

PHASE 1 –DATA SELECTION

Five activites were involved in this phase:
1- Daily newspaper texts of length about 7100000 characters were read and transmitted by 5-bit paper tapes from HP teletypes to hard-disc of an IBM PC via port RS232.
2- The most frequent words were derived by word-count processing after editing and error correcting of the read texts.
3- 384 sentences(type I sentences) were made by the first 1000 most frequent words and other phonetically rich monosyllable words(Samareh, 1977), to provide allophones in all possible left and right phonetic contexts.
4- Two sentences (type II sentences) were made consisting of all Farsi phonemes to allow dialect comparison among speakers.
5- 1272 isolated words were selected based on the most frequent words and other phonetically rich words, to provide the most frequent phonetic contexts in Farsi words.

PHASE 2 –DATA COLLECTION

Two activities with four characteristics are involved in this phase:
1- The population distribution of speakers were determined with regards to age, sex, dialects and educational level(table 1). The dialects are: Tehrani, Torki, Isfahani, Shomali, Yazdi, Jonubi, Kordi, Xorasani, Baluchi, and Lori. The population ratio of males to females is two to one.
2- Each speaker uttered twenty sentences in two sessions, consisting of eighteen type 1 sentences, selected randomly, and two type 2 sentences. There is a total of 6000 utterances in the FARADAT, which are divided into two mutually exclusive subsets: training and testing subsets. The average speaking rate is 4.2 syllables per second.
3- The speakers are seated in either the Linguistics Laboratory of the University of Tehran or in an office room both with a moderate noise environment.
4- A SONY cardioid dynamic microphone having frequency response within 80Hz to 16KHz range and effective output level of -- 61.8 dBm was used. The distance between microphone and lip position of the speaker was about 12 Cm.
5- The speech was collected using 16-bit sound Blaster hardware cards, installed in four 80486 IBM microcomputers, and sampled at 44.1KHz samples per second at 16 bit resolution. Users of the FARSDAT may decimate the data to change sampling rate as required. In this case a digital Low pass filter must be used for anti-aliasing.
6- signal to noise ratio of the FARSDAT is 30db. To provide this value of SNR, following factors have been dictated:
-- Shielded cable has been used for microphone.
-- SB16 card has been located in microcomputer such that the effects of electromagnetic noise induction on the card is minimized.
-- DC offset due to SB16 card has been reduced after sampling speech signals.
-- Other acoustic noise sources such as fluorescent lamps, fan, cooler and video TV have been turned off during speech recording.
In sum, the FARADAT is something between laboratory and real users databases (Chigier & Spitz, 1990).

This phase has been completed for 100 speakers up to now.

SEGMENTATION AND LABELLING

TABLE 2 shows IPA chart of the Farsi phonemes. A procedure for manual segmentation and labelling in the FARSDAT is based on the following three steps:

1-- Several phonetic transcribers were trained to delimitate explicitly the beginning and end sample of each syllable, some phone sequences and phoneme. Transcribers trained to detect high acoustic variability due to phonemic transitions by carefully listenning and visual examination of the speech signal using zooming technique of the SB16 signal editor.

2-- Acoustic overlap is included for each phoneme such that the phoneme is heared differently in different phonetic contexts.

3-- Since the acoustic cues are inexact in nature, ambiguities in delimitation of some phonemes, particularly sonorants and glottal stop, is arisen. To minimize the differences in the segmentation, transcribers trained to follow some steps uniformly:

-- Carefully listening to the complete signal to discern the degree of phoneme approximation.

-- Phoneme localization by listenning to the corresponding short segment and zooming on the smallest change of the speech signal.

-- Visual examination of the corresponding spectrogram by DSP Sona-Graph KAY, stored setups #04, #05, which display wideband and narrowband spectrograms plus ZCR, energy and fundamental freqency.

-- Despite of these precautions, about %20 of the phoneme boundaries differed when cross-comparing two different transcribers(Chigier & Spitz, 1990).

GLOTTAL STOP STORY

Glottal stop is one of the Farsi phonemes which has very ambiguous behaviour and made a lot of difficulties for transcribers to segment. It has three completely different acoustic realizations:

. Protoypical Glottal Stop (ʔ͡) consists of both silent and burst portions like other plosives. The frequency spectrum of burst looks like one of the following vowel. It occurs at the beginning of a stressed syllable after silence.

. Creaky Glottal Stop (ʔ̰) refers to the aperiodic and small number of glottal creaks per second. It looks like a sequence of protoypical glottal stops. It occurs at the beginning of a stressed syllable within utterances.

. Vowel-Like Glottal Stop (ʔˠ) refers to a low pitch and low amplitude kind of vowel. Some speakers do not produce it. It occurs between two unstressed syllables.

These three kinds of glottal stops are shown in the transcription of the utterance "ʔ͡aɟarʔ̰esmatraʔˠazkoni" = if you change your name (fig.1).

GENERAL STATISTICS AND DATABASE SOFTWARE

Some statistics have been drawn from the speech data of 60 speakers, which will be completed:

-- Relative frequency of each phoneme.

-- Transition probability of each phoneme to another phoneme.

– Relative frequency of each syllable.
– Transition probability of each syllable to another syllable.
– Positional probability of each phoneme in three Farsi syllable structures: CV, CVC, CVCC.
– Transitional probability of each phoneme to another phoneme in different positions of the syllable structures.
– Duration analysis of each phoneme.

Table 3 shows the mean duration of each phoneme with the corresponding standard deviation, and also the least and the most frequent phonetic context for each phoneme. As we see the mean duration of each unvoiced obstruent is greater than that of its voiced pair. High length of unvoiced plosives is as a result of aspiration.

The FARSDAT will be available on two CD ROM's. Its custom software is written by C language programming under windows. Speech files can be accessed using different fields pertaining to age, sex, dialect and educational level of speakers.

CONCLUSION

FARSDAT is the first speech database in Farsi spoken language designed thoroughly for linguistics and speech technology research purposes, as authors know. It includes a vast amount of speech files for different linguistic units of continuous speech and isolated words as well. It can be used for both training and testing phases of any Farsi speech and speaker recognition system. Future work is oriented to develop an automatic segmentation and labelling based on acoustic feature analysis derived from the FARSDAT.

REFERENCES

Chigier B. and Spitz J.(1990), Are laboratory Databases Appropriate for Training and Testing Telephone speech Recognizers? Proceedings of ICSLP'90 kobe, Japan, Vol.2, PP. 1017-1020.

Muthusamy Y.K., Cole R.A. and Oshika B.T.(1992), The OGI Multi-Language Telephone Speech Corpus proceedings of ICSLP'92, Banff, Canda, Vol.2, PP. 895-898.

Samareh Y.(1977), The Arrangement of Segmental Phonemes in Farsi, Tehran University publications, no. 1577, PP. 147-179, Tehran, Iran.

FIG.1 : TRANSCRIPTION OF THE UTTERANCE " ʔaɟarʔesmātɟaʔʸavaʒkʰōnī "

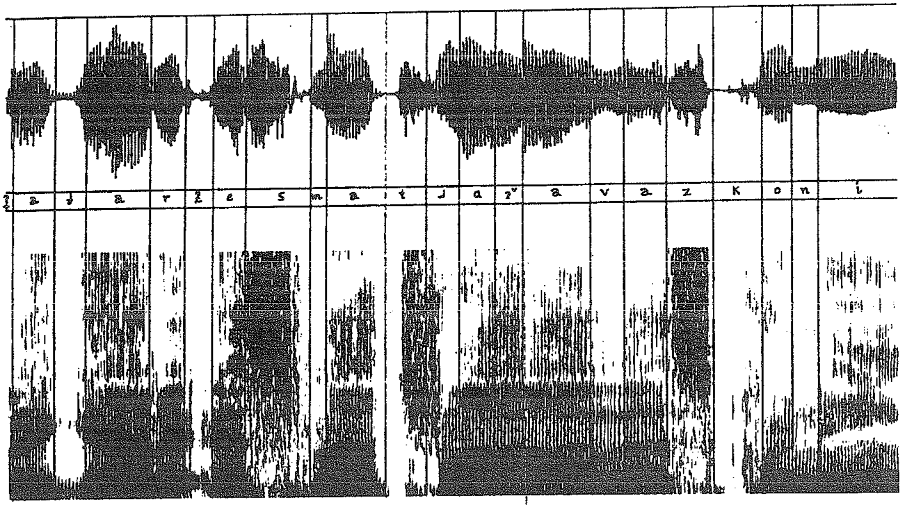

TABLE 1: POPULATION DISTRIBUTION OF SPEAKERS WITH REGARDS TO AGE, SEX, DIALECTS IN FARSDAT
(M=MALE, F=FEMALE)

AGE	SEX	DIALECTS										TOTAL
		TEHRANI	TORKI	ESFAHANI	SHOMALI	YAZDI	JONUBI	KORDI	XORASANI	BALUCHI	LORI	
10-20	M	15	4	3	2	1	1	1	1	1	1	30
	F	7	2	1	1	1	1	1	1	0	0	15
20-30	M	21	5	3	2	2	2	2	2	2	2	43
	F	10	2	2	1	1	1	1	1	1	1	21
30-40	M	22	5	4	2	2	2	2	2	2	2	45
	F	11	2	2	1	1	1	1	1	1	1	22
40-50	M	18	4	3	2	2	2	2	2	1	1	37
	F	9	2	1	1	1	1	1	1	1	1	19
50-60	M	12	4	2	1	1	1	1	1	1	1	25
	F	6	2	1	1	1	1	1	0	0	0	13
60-70	M	7	2	1	1	1	1	1	1	0	0	15
	F	3	1	1	1	1	0	0	0	0	0	7
70-80	M	3	1	1	0	0	0	0	0	0	0	5
	F	1	1	1	0	0	0	0	0	0	0	3
TOTAL		145	37	26	16	15	14	14	13	10	10	300

TABLE 2: IPA CHART OF THE FARSI PHONEMES

(V=VOICED ; UV=UNVOICED)
(symbols in paranthesis show place allophones)

MANNER OF ARTICULATION	PHONEME	PLACE OF ARTICULATION: LABIAL V	LABIAL UV	LAB.DEN V	LAB.DEN UV	DENTAL V	DENTAL UV	ALVEOL. V	ALVEOL. UV	PAL.ALV V	PAL.ALV UV	PALATAL V	PALATAL UV	VELAR V	VELAR UV	UVULAR V	UVULAR UV	LARYNG. V	LARYNG. UV
	PLOSIVE	b	p			d	t	(D)	(T)			(ɟ)	(c)	k	g	G			ʔ
	FRICATIVE			v	f			z	s	ʒ	ʃ						χ	(ɦ)	h
	AFFRICATE									ʤ	ʧ								
	NASAL	m		(ɱ)				n				(ɲ)		(ŋ)		(ɴ)			
	LATERAL							l											
	TRILL							r											
	SEMIVOWEL											j							
	VOWEL							i					u						
									e				o						
										a			ɑ						

TABLE 3: STATISTICS ON PHONEME DURATION AND PHONETIC CONTEXT FREQUENCY IN THE FARSDAT
(0=beginning of the speech)

PHONEME	DURATION		PHONETIC CONTEXT FREQUENCY	
	MEAN	STD.DEV.	LOWEST	HIGHEST
ɑ	0.125022	0.043880	f_t	d_r
s	0.118020	0.031090	a_r	a_0
a	0.116553	0.040134	χ_ʃ	k_r
ʃ	0.110566	0.026451	n_e	e_o
ʧ	0.106103	0.029776	m_i	l_e
χ	0.098897	0.026884	f_k	ɑ_ɑ
u	0.097810	0.039909	t_n	r_z
i	0.095585	0.043194	p_p	ʔ_n
o	0.094363	0.041545	d_l	ʃ_d
k	0.089507	0.024544	e_l	i_a
f	0.089421	0.029214	a_χ	o_t
ʒ	0.089025	0.032343	ʔ_ɑ	o_d
p	0.087445	0.028506	i_p	h_e
z	0.081404	0.025710	o_ʤ	a_ɑ
t	0.080346	0.027657	k_o	ɑ_a
ʤ	0.077449	0.025240	z_g	a_l
m	0.077362	0.022523	h_o	e_a
e	0.074646	0.033046	s_G	j_m
h	0.071393	0.026198	a_z	e_a
G	0.069485	0.028489	s_o	G_o
n	0.064974	0.023246	a_r	a_d
v	0.062225	0.020549	z_a	a_ɑ
b	0.061463	0.028275	0_o	o_ɑ
l	0.061087	0.018291	o_l	e_l
g	0.060079	0.022398	ʤ_e	a_a
j	0.059716	0.021934	a_u	ɑ_e
d	0.058017	0.024673	b_o	r_a
r	0.044728	0.015998	h_a	e-ɑ
ʔ	0.041953	0.021545	d_o	o_i