Tfarsdat - the Telephone Farsi Speech Database

Mahmood Bijankhan[1,2], Javad Sheykhzadegan [2], Mahmood R. Roohani [2], Rahman Zarrintare [2], Seyyed Z. Ghasemi [1,2] Mohammad E. Ghasedi[2,3]

[1] Department of Linguistics, University of Tehran, Iran
[2] Research Center of Intelligent Signal Processing, Tehran, Iran
[3] Ministry of Education, Tehran, Iran
mbjkhan@chamran.ut.ac.ir zarintareh@yahoo.com

Abstract

This paper describes an ongoing research to create an acoustic phonetic based telephone Farsi speech database, called "Tfarsdat". It is compared with two LDC Farsi corpora, OGI and Call friend in terms of corpus dialectology. Up to now, we have recorded about 8 hours of monologue calls containing spontaneous and read speech for 64 speakers belonging to one of ten dialect regions. A hierarchical annotation system is used to transcribe phoneme, word and sentence levels of speech data. User software is written to access speech and label files efficiently using a menu driven query system. We conducted two experiments to validate Tfarsdat statistically. Results showed the necessity of increasing speaker size and also quality enhancement of annotation system.

1. Introduction

Farsi (Persian language) is one member of the Iranian branch of the Indo-Iranian languages, a subfamily of the Indo-European languages. It is the language of Iran, Afghanistan, Tajikistan and the Pamirs mountain region. Farsi, as a language spoken by Iranians, has so many varieties due to cultural interactions of different nationalities within Persian land, now Iran. Telephone Farsi Database (Tfarsdat) is the first project involving creation of a fixed network telephone speech database for Persian Farsi (from now on: Farsi) taking into account dialect varieties. The primary goal of Tfarsdat creation is to provide a phone-based speech resource for telephone ASR systems in small scale. The ultimate goal is to satisfy the telephone speech resources demand of academia and industry in the domain of the large vocabulary speaker independent telephone speech recognition systems. For two reasons satisfaction of the primary goal is emphasized as a basic research activity: first the phonological understanding of the speech signal complexity in spontaneous mode of communication, and second, lack of studies of acoustic phonetic analysis of Farsi telephone speech, which is crucial for development of a large scale database to support application systems.

Two Farsi telephone speech corpora are known to be collected and annotated: The OGI multi-language telephone speech corpus [1] and Call friend Farsi [2], catalogued by LDC, both for the sake of automatic language identification. One of LDC current projects is to modify Call friend Farsi to support speech recognition of conversational Farsi.

This paper describes activities performed to reach the primary goal, as mentioned above. In section two, we explain the Tfarsdat general specifications. In section three, we describe the strategy of phone and word segmentation and labeling. Section four describes transcription properties of lexical entries. Section five explains how to get at speech data in terms of different linguistic levels via user software. Statistical validation of the database is given in section six. The paper comes to an end with a conclusion and some suggestions for future works in section seven.

2. Tfarsdat General Specifications

Research Center of Intelligent Signal Processing (RCISP) created the Tfarsdat for acoustic phonetic studies of academic centers and a reference for assessment of telephone Farsi ASR systems. The project completed in 2002 after a two-year work of six researchers. Bellow follows the general specifications of the Tfarsdat.

2.1. Corpus Dialectology

In OGI and Call friend Farsi databases, speakers were selected from Farsi native speakers inside continental United States. Therefore, both corpora suffer from appropriate speaker coverage in terms of dialect, age and socio-economic class varieties. As compared with these, Tfarsdat speaker selection strategy was managed not only based on sociolects, but also geographical distribution of Farsi dialects in Iran. Ten dialect regions were selected: Tehrani, Turkic, Isfahani, Shomali, Yazdi, dzonubi, Xorasani, Kurdish, Lori and Baluchi, as in Farsdat [3]. Sixty-four speakers were asked to carry on a very natural monologue using vocabulary and syntax of Farsi standard dialect, which is Tehrani, but with their own region accent. This is a socio-linguistic situation most often occurs in Iranian linguistic community. We notice a vast amount of differences in phonetic, morphological and syntactic patterns of Farsi dialects.

2.2. Corpus data

Speech data were of two types: spontaneous and read. A prompt sheet consisted of data items of both types were posted to speakers' living area. They were asked to call to the project headquarter in Tehran from their hometown in order to cover data variability due to variable characteristics of telephone channel frequency responses in Iran. For spontaneous speech, speakers' monologue contained greeting, personal information regarding to age, educational level, economic, cultural and political characteristics of birthplace and overall conditions of the place he/she is calling from. Each speaker at the end of spontaneous speech narrated a memoir. Tfarsdat read data consisted of speakers' utterances of all natural numbers from which all other numbers are generated, days, months and alphabets names, fifty most frequent Farsi words, six Farsdat sentences including two sentences that contain all Farsi phonemes [3], and finally all 138 Farsi CV syllables. Up to now, Tfarsdat size is about eight hours: two hours for spontaneous speech and six hours for read speech.

2.3. Speakers

Speakers were selected with regard to age, gender, educational level and belonging to one of aforementioned ten dialect regions. Population ratio of male to female was roughly eight to three. 81% of the speakers were 20-50 years old, and 19% of them 10-20 and 50-70 years old. 53% of speakers were of Tehrani and Isfahani dialects, and the remaining 47% were distributed on other dialects according to the frequency of each dialect population in Tehran.

Table 1*: IPA symbols for phonetic and phonemic transcription.*

Symbol	Farsi Alphabets	Linguistic Description
i	ـیـ، ی	high front vowel
e	ـهـ، ا، ه	mid front vowel
a	ـَـ، ا	low front vowel
u	و	high back vowel
o	ـُـ، و	mid back vowel
ɒ	ا، آ، ی	low back vowel
b˺	بـ، ب	b closure
p˺	پـ، پ	p closure
d˺	د	d closure
t˺	تـ، ت	t closure
ɟ˺	گـ، گ	ɟ closure
g˺	گـ، گ	g closure
c˺	کـ، ک	c closure
k˺	کـ، ک	k closure
G˺	قـ، ق، غ، غـ، غ	G closure
ʔ˺	ا، عـ، ع، ئـ، ء، ـعـ	ʔ closure
ʤ˺	جـ، ج	ʤ closure
ʧ˺	چـ، چ	ʧ closure
b	بـ، ب	voiced bilabial plosive
p	پـ، پ	unvoiced bilabial plosive
d	د	voiced dental plosive
t	تـ، ت، ط	unvoiced dental plosive
ɟ	گـ، گ / گیر	voiced palatal plosive
g	گـ، گ / گاری	voiced velar plosive
c	کـ، ک / کیف	unvoiced palatal plosive
k	کـ، ک / کار	unvoiced velar plosive
G	قـ، ق، غ، غـ، غ	voiced uvular plosive
ʔ	ا، عـ، ع، ئـ، ء، ـعـ	glottal stop
ʤ	جـ، ج	voiced alveopalatal affricate
ʧ	چـ، چ	unvoiced alveopalatal affricate
v	و	voiced labiodental fricative
f	فـ، ف	unvoiced labiodental fricative
Z	ذ، ز، ضـ، ض، ظ	voiced alveolar fricative
s	ثـ، ث، سـ، س، صـ، ص	unvoiced alveolar fricative
ʒ	ژ	voiced alveopalatal fricative
ʃ	شـ، ش	unvoiced alveopalatal fricative
x	خـ، خ	unvoiced uvular fricative
h	حـ، ح، هـ، ه، ـه	unvoiced glottal fricative
m	مـ، م	bilabial nasal
n	نـ، ن	alveolar nasal
r	ر	alveolar trill
l	لـ، ل	alveolar lateral
j	یـ، ی	palatal glide

2.4. Recording Environment and Equipment

Speakers called from either a home or quiet office rooms with different distances from project headquarter. Speech was collected via the fixed network using an interface fed by telephone line analog and through SB16 sound card microphone, converted to digitized sound with 11.025 kHz sampling rate and 16 bit resolution. Spontaneous and read speeches were recorded in separate sessions.

We computed following parameters for 320 files of the Tfasdat to evaluate quality of recording acoustically:

- Mean power: sum of the squares of the sample size divided by the number of samples, resulted in 1054.77 on average for all files.
- Signal to noise ratio: difference of silence mean power from noisy signal mean power divided by silence mean power times 10, resulted in 23.3 db on average for all files.
- Clipping rate: number of saturated samples divided by number of all samples, resulted in 6.8×10^{-6} on average for all files.

3. Annotation

The goal of annotation was to provide linguistic unit access for users in order to study acoustic properties of the signal and to train and test a task-oriented telephone ASR system. Tfarsdat software was written to support linguistic annotation at the levels of phoneme, word and sentence. To get rid of inter - annotator inconsistencies, one acoustic phonetic graduate student instructed to annotate the corpus manually based on following commands:

- Segment the signal into acoustic phonetic chunks of phonemes (and sub phonemes such as silent portion of plosives) and words using synchronous display of speech waveform and spectrogram (figure 1).
- Label phoneme chunks using phonetic transcription and word chunks using phonological transcription. Therefore, a phonologically transcribed word like "**vaGt**", meaning "time", may be transcribed "**vax**" phonetically with phonetic chunks of [**v**], [**a**] and [**x**] due to [**t**] deletion and [**G**] spirantization. Phonetic or phonological symbols are called "linguistic labels" (table 1). In this way, we are able to derive multiple pronunciations of a word.
- Label other sounds or noises such as mispronunciation, lip smack, cough, hesitation, telephone noise, background noise and others as described in table 2. These are called "nonlinguistic labels".
- Label word chunks phonologically, as explained above, and orthographically using Arabic characters of Tahoma fonts under windows.
- Segment Farsi suffix formative /e/, called "kasreje ezafe" in Persian, separately from the root word and label it phonetically as uttered [**e**] or [**je**] depending upon the last phoneme of the root word. This is due to its importance from language modeling point of view.
- Disambiguate phonetic segmentation and labeling based on zooming the smallest changes of formant frequencies and playing the signal to detect acoustic boundary.

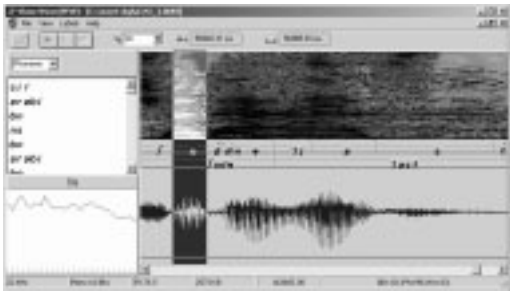

Figure 1: *Software tool for transcription of the Tfarsdat.*

4. Lexicon

Tfarsdat contains 25232 word tokens up to now, setting aside CV syllables, each with its absolute frequency in the corpus, resulting in 5092 word types. Therefore, each word occurs about 5 times on average. Each word is transcribed phonetically, phonemically and orthographically. From phonetic and phonemic transcriptions, one can capture the different alternations a word may have due to speaker's way of articulation. If a word were pronounced with free variations due to speaker's sociolect, each variation would be taken into account as a separate entry. Since speaker utterances in spontaneous speech involved phoneme approximant articulation, a modified symbol of phoneme label of the form "<phoneme label>1" was added to phoneme symbols to promote the precision of phonetic transcription. In addition, infinitive and verb compounds were segmented and labeled as separate entries.

Table 2: *Nonlinguistic labels percentage of occurrence in the Tfarsdat.*

nonlinguistic labels	linguistic Description	Percent
ls	lip smack	0.17
br	breath	1.30
uh	inter - word pause	0.02
cog	cough	0.01
bn	background noise	0.20
hes	hesitation sound	0.33
ns	non - speech sound	0.08
ln	telephone line noise	3.59
def	mispronunciation	0.09
deff	non - sense word	0.01

5. User Software

Tfarsdat is equipped with user software to support phonetic, n-gram and word search through a menu driven query system. The search can be proceeded with different values of the speaker code, age, gender, grade and dialect. Therefore, output label and wave files can be accessed as a result of the user defined variables. In addition, a user can get at the different items of the read and spontaneous speech files separately (figure 2). Accessed search items may be played, saved or retrieved.

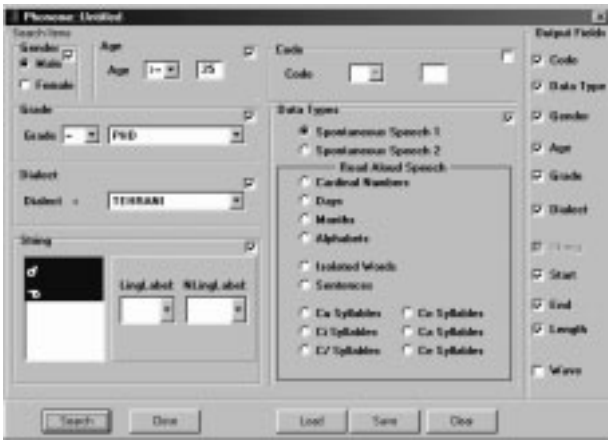

Figure 2: *A menu driven query for the string "ɑ'ɒ".*

6. Statistical Validation

To insure that Tfarsdat is phonetically and lexically rich enough to be a representative acoustic resource for Farsi speech research, two experiments were conducted to explore its statistical validity.

Experiment 1. We transcribed a text of the size 2803999 phonemes, based on a conservative type of speech, as our reference database. Then, we crosschecked the relative frequencies of Farsi phonemes and diphones for spontaneous part of the Tfarsdat and the reference database. The result showed not only appearance of all phonemes and the most frequent diphones in the Tfarsdat, but also an acceptable coincidence of the results ranks in both databases (Table 3).

Table 3: *Phoneme rank coincidence of the Tfarsdat and reference database.*

Tfarsdat symbol	Percent	Reference database symbol	Percent
a	12.33	e	11.77
e	10.27	a	11.57
ɒ	8.45	ɒ	8.66
m	5.94	r	6.30
i	5.51	d	5.30
r	5.20	n	5.21
n	4.98	i	5.17
d	4.79	ʔ	4.39
t	4.73	m	4.17
ʔ	4.00	t	4.08
s	3.85	o	3.33
h	3.71	j	3.18
o	3.47	s	3.15
l	2.99	b	3.03
b	2.69	h	2.93
ʃ	2.30	v	2.44
k	2.19	z	2.16
v	2.15	l	2.12
z	2.07	k	1.98
j	1.53	ʃ	1.93
u	1.49	u	1.39
x	1.11	G	1.09
G	1.00	f	1.05
ʤ	0.84	g	0.95
g	0.77	ʤ	0.90
f	0.75	x	0.87
p	0.48	p	0.53
ʧ	0.40	ʧ	0.30
ʒ	0.02	ʒ	0.06
	100		100

By the way, an examination of table 3 shows some inconsistencies that must be explained. For example, the most frequent phoneme in the Tfarsdat is the vowel /**a**/, but /e/ in the reference database. A statistical analysis showed that "kasreje ezafe" morpheme /e/ occurred about twice in the

reference database in comparison with spontaneous part of the Tfarsdat. That is why /e/ has highest rank in the reference database. This point must be emphasized that one general stylostatistic feature of the formal Persian writing is that Persian writers tend to generate long syntactic phrases using "Kasreje ezafe". Such phrases are generated, in English, by the preposition "of", or by left branching modification of a head noun without using any proposition, which is prohibited such structures by Persian syntax. Another evidence comes from bigram frequency. While bigram [je], another phonetic realization of the "kasreje ezafe" morpheme was in second rank with 2.42 percent frequency for the reference, it was in eighth rank with 1.63 percent frequency for the Tfarsdat.

Another source of bigram frequency inconsistency arises from the way Farsi native speakers utter the morphosyntactic definite marker "ر ه". While this marker is uttered by the same bigram [ر ه] in conservative utterance, it is produced mostly by the allomorph [o] in spontaneous speech. Thus, in the reference database, it is in twelfth rank with 1.26 percent frequency, but in rank of 32 in the Tfarsdat with 60 percent frequency.

One more bigram frequency inconsistency was due to different articulations of the noun plural suffix "ha" in conservative and spontaneous speech. Farsi native speakers tend to delete /h/ which then results in resyllabification of the phoneme sequence. Therefore, a variety of the bigrams [xa] (x denotes the last phoneme of the root noun) were generated in spontaneous speech that highly affected bigram frequency. Consequently, while "ha" is in rank 22 with 98 percent frequency in the reference database, it is in rank 30 with 78 percent frequency in the Tfarsdat.

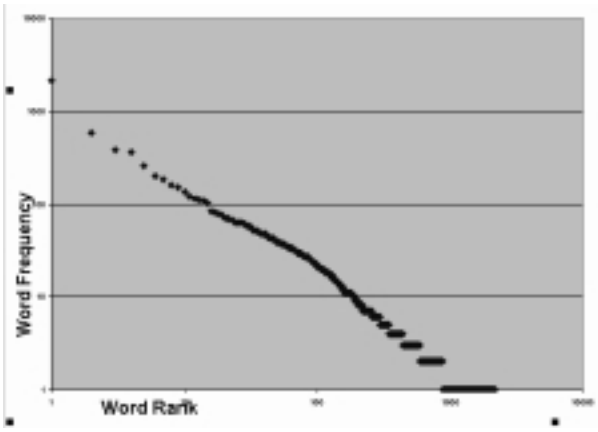

Figure 3: *Rank-frequency plot of the Tfarsdat words on doubly logarithmic axes.*

Experiment 2. We counted up how often each word of the spontaneous part of the speech occurs, and then listed the words in order of their frequency to determine the rank of each word. Figure 3 shows a plot of rank-frequency of the words on doubly logarithmic axes, that is roughly a straight line with slope –1. The result was in concordance with one of Zipf's laws, which explain the principle of least effort in speech production and comprehension [4]. The plot is a rough description of the frequency distribution of words such that there are a few very common words like function words, a middling number of medium frequency words, and many low frequency words. 91% of the words occurred less than ten times, and 70.37% of the words occurred just once.

7. Discussions and Conclusion

We described an ongoing acoustic phonetic based fixed network telephone Farsi SLR with ASR applications, Tfarsdat, which is now available with documentation in two CDs. Up to now, speech data for 64 speakers have been collected and annotated. We have provided a hierarchical linguistic annotation system for phoneme, word and sentence levels. Annotation system is exactly the same as TIMIT's [5]. User can access speech and label files directly Via Tfarsdat directory file structure or using a menu driven query system within user software. Speech data access should be based on arbitrary defined variables including speaker's code, gender, dialect and educational level. Experiments for statistical validation of the Tfarsdat showed the need for increasing of the number of speakers talking about various subjects in spontaneous part to lessen the distance of the n-gram relative frequency of the Tfarsdat from the one in the reference database.

Three activities are in mind for future works. First, increasing number of speakers and changing speech type from monologue to dialogue. Second, modification of annotation system for transcription of linguistic complexities usually occurs in dialogue type. DAMSL is proposed for this activity [6]. The third activity will be the transcription of the inflective allomorphs in speech label files and lexicon, due to the linguistic richness of Farsi noun, adjective and verb systems from inflective morphological point of view. Existing hierarchical annotation system of the Tfarsdat facilitates this activity very well.

8. References

[1] Muthusamy, Y. K. et. Al., "The OGI Multi-Language Telephone Speech Corpus, Proc. of the ICSLP ,Vol. 2, 1992: 895-898.

[2] LDC Homepage:http://WWW.ldc.upenn.edu/catalog/catalogEntry.jsp?catalogId=LDC 96S50.

[3] Bijankhan, M. et. al., "FARSDAT-The Farsi Spoken Language Database", Proc., of the 5th Int. Conf. on Speech Sciences and Technology. Vol. 2. 1994: 826-829.

[4] Manning, C. D. and Schutze H., "Foundations of Statistical Natural Language Processing, the MIT press, Cambridge, Massachusetts, 1999.

[5] Bird, S. And Liberman, M.,"Towards A Formal Framework for Linguistic Annotation", LDC Homepage: http://WWW.ldc.upenn.edu.

[6] Allen, J. and Core, M., "Draft of DAMSL: Dialog Act Markup in Several Layers", Multiparty Discourse Group, Discourse Research Initiative (DRI), 1997.