A Transcription Scheme for Languages Employing the Arabic Script Motivated by Speech Processing Application

Shadi GANJAVI
*Department of Linguistics
University of Southern California
ganajvi@usc.edu

Panayiotis G. GEORGIOU,
Shrikanth NARAYANAN*
Department of Electrical Engineering
Speech Analysis & Interpretation
Laboratory (sail.usc.edu)
[georgiou, shri]@sipi.usc.edu

Abstract

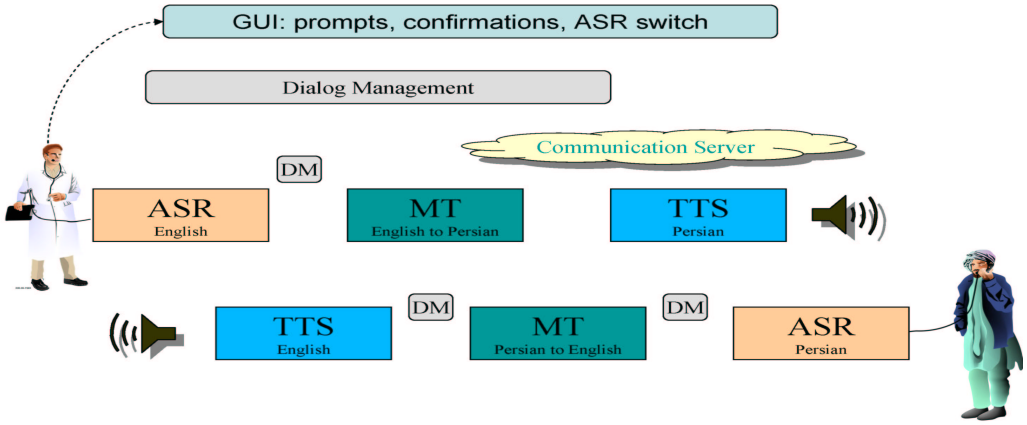
This paper offers a transcription system for Persian, the target language in the Transonics project, a speech-to-speech translation system developed as a part of the DARPA Babylon program (The DARPA Babylon Program; Narayanan, 2003). In this paper, we discuss transcription systems needed for automated spoken language processing applications in Persian that uses the Arabic script for writing. This system can easily be modified for Arabic, Dari, Urdu and any other language that uses the Arabic script. The proposed system has two components. One is a phonemic based transcription of sounds for acoustic modelling in Automatic Speech Recognizers and for Text to Speech synthesizer, using ASCII based symbols, rather than International Phonetic Alphabet symbols. The other is a hybrid system that provides a minimally-ambiguous lexical representation that explicitly includes vocalic information; such a representation is needed for language modelling, text to speech synthesis and machine translation.

1 Introduction

Speech-to-speech (S2S) translation systems present many challenges, not only due to the complex nature of the individual technologies involved, but also due to the intricate interaction that these technologies have to achieve. A great challenge for the specific S2S translation system involving Persian and English would arise from not only the linguistics differences between the two languages but also from the limited amount of data available for Persian. The other major hurdle in achieving a S2S system involving these languages is the Persian writing system, which is based on the Arabic script, and hence lacks the explicit inclusion of vowel sounds, resulting in a very large amount of one-to-many mappings from transcription to acoustic and semantic representations.

In order to achieve our goal, the system that was designed comprised of the following components:

Fig 1. Block diagram of the system. Note that the communication server allows interaction between all subsystems and the broadcast of messages. Our vision is that only the doctor will have access to the GUI and



the patient will only be given a phone handset.

(1) a visual and control Graphical User Interface (GUI); (2) an Automatic Speech Recognition (ASR) subsystem, which works both using Fixed State Grammars (FSG) and Language Models (LM), producing n-best lists/lattices along with the decoding confidence scores; (3) a Dialog Manager (DM), which receives the output of the speech recognition and machine translation units and subsequently "re-scores" the data according to the history of the conversation; (4) a Machine Translation (MT) unit, which works in two modes: Classifier based MT and a fully Stochastic MT; and finally (5) a unit selection based Text To Speech synthesizer (TTS), which provides the spoken output. A functional block diagram is shown in Figure 1.

1.1 The Language Under Investigation: Persian

Persian is an Indo-European language with a writing system based on the Arabic script. Languages that use this script have posed a problem for automated language processing such as speech recognition and translation systems. For instance, the CSLU Labeling Guide (Lander, http://cslu.cse.ogi.edu/corpora/corpPublications.html) offers orthographic and phonetic transcription systems for a wide variety of languages, from German to Spanish with a Latin-based writing system to languages like Mandarin and Cantonese, which use Chinese characters for writing. However, there seems to be no standard transcription system for languages like Arabic, Persian, Dari, Urdu and many others, which use the Arabic script (ibid; Kaye, 1876; Kachru, 1987, among others).

Because Persian and Arabic are different, Persian has modified the writing system and augmented it in order to accommodate the differences. For instance, four letters were added to the original system in order to capture the sounds available in Persian that Arabic does not have. Also, there are a number of *homophonic* letters in the Persian writing system, i.e., the same sound corresponding to different orthographic representations. This problem is unique to Persian, since in Arabic different orthographic representations represent different sounds. The other problem that is common in all languages using the Arabic script is the existance of a large number of *homographic* words, i.e., orthographic representations that have a similar form but different pronunciation. This problem arises due to limited vowel presentation in this writing system.

Examples of the homophones and homographs are represented in Table 1. The words "six" and "lung" are examples of homographs, where the identical (transliterated Arabic) orthographic representations (Column 3) correspond to different pronunciations [SeS] and [SoS] respectively (Column 4). The words "hundred" and "dam" are examples of homophones, where the two words have similar pronunciation [sad] (Column 4), despite their different spellings (Column 3).

	Persian	**USCPers**	**USCPron**	**USCPers+**
'six'	شش	SS	SeS	SeS
'lung'	شش	SS	SoS	SoS
'100'	صد	$d	sad	$ad
'dam'	سد	sd	sad	sad

Table 1 Examples of the transcription methods and their limitation. Purely orthographic transcription schemes (such as USCPers) fail to distinctly represent homographs while purely phonetic ones (such as USCPron) fail to distinctly represent the homophones.

The former is the sample of the cases in which there is a many-to-one mapping between orthography and pronunciation, a direct result of the basic characteristic of the Arabic script, viz., little to no representation of the vowels.

As is evident by the data presented in this table, there are two major sources of problems for any speech-to-speech machine translation. In other words, to employ a system with a direct 1-1 mapping between Arabic orthography and a Latin based transcription system (what we refer to as USCPers in our paper) would be highly ambiguous and insufficient to capture distinct words as required by our speech-to-speech translation system, thus resulting in ambiguity at the text-to-speech output level, and internal confusion in the language modelling and machine translation units. The latter, on the other hand, is a representative of the cases in which the same sequence of sounds would correspond to more than one orthographic representation. Therefore, using a pure phonetic transcription, e.g., USCPron, would be acceptable for the *Automatic Speech Recognizer* (ASR), but not for the *Dialog Manager* (DM) or the *Machine Translator* (MT). The goal of this paper is twofold (i) to provide an ASCII based phonemic transcription system similar to the one used in the International Phonetic Alphabet (IPA), in line of Worldbet (Hieronymus, http://cslu.cse.ogi.edu/corpora/corpPublications.html) and (ii) to argue for an ASCII based hybrid

transcription scheme, which provides an easy way to transcribe data in languages that use the Arabic script.

We will proceed in Section 2 to provide the USCPron ASCII based phonemic transcription system that is similar to the one used by the International Phonetic Alphabet (IPA), in line of Worldbet (ibid). In Section 3, we will present the USCPers orthographic scheme, which has a one-to-one mapping to the Arabic script. In Section 4 we will present and analyze USCPers+, a hybrid system that keeps the orthographic information, while providing the vowels. Section 5 discusses some further issues regarding the lack of data.

2 Phonetic Labels (USCPron)

One of the requirements of an ASR system is a phonetic transcription scheme to represent the pronunciation patterns for the acoustic models. Persian has a total of 29 sounds in its inventory, six vowels (Section 2.1) and 23 consonants (Section 2.2). The system that we created to capture these sounds is a modified version of the International Phonetic Alphabet (IPA), called USCPron(unciation). In USCPron, just like the IPA, there is a one-to-one correspondence between the sounds and the symbols representing them. However, this system, unlike IPA does not require special fonts and makes use of ASCII characters. The advantage that our system has over other systems that use two characters to represent a single sound is that following IPA, our system avoids all ambiguities.

2.1 Vowels

Persian has a six-vowel system, high to low and front and back. These vowels are: [i, e, a, u, o, A], as are exemplified by the italicized vowels in the following English examples: 'b*ea*t', 'b*e*t', 'b*a*t', 'p*u*ll', 'p*o*ll' and 'p*o*t'. The high and mid vowels are represented by the IPA symbols. The low front vowel is represented as [a], while the low back vowel is represented as [A]. There are no diphthongs in Persian, nor is there a tense/lax distinction among the vowels (Windfuhr, Gernot L.1987).

	Front	Back
High	i	u
Mid	e	o
Low	a	A

Table 2: Vowels

2.2 Consonants

In addition to the six vowels, there are 23 distinct consonantal sounds in Persian. Voicing is phonemic in Persian, giving rise to a quite symmetric system. These consonants are represented in Table 3 based on the place (bilabial (BL), lab-dental (LD), dental (DE), alveopalatal (AP), velar (VL), uvular (UV) and glottal (GT)) and manner of articulation (stops (ST), fricatives (FR), affricates (AF), liquids (LQ), nasals (NS) and glides (GL)) and their voicing ([-v(oice)] and [+v(oice)].

	BL	LD	DE	AP	VL	UV	GT
ST [-v]	p		t		k		?
[+v]	b		d		g	q	
FR [-v]		f	s	S	x		h
[+v]		v	z	Z			
AF [-v]				C			
[+v]				J			
LQ			l, r				
NS	m		n				
GL				y			

Table 3: Consonants

Many of these sounds are similar to English sounds. For instance, the stops, [p, b, t, d, k, g] are similar to the italicized letters in the following English words: '*p*otato', '*b*all', '*t*ree', '*d*oll', '*k*ey' and 'do*g*' respectively. The glottal stop [?] can be found in some pronunciations of 'bu*tt*on', and the sound in between the two syllables of 'uh oh'. The uvular stop [q] does not have a correspondent in English. Nor does the velar fricative [x]. But the rest of the fricatives [f, v, s, z, S, Z, h] have a corresponding sound in English, as demonstrated by the following examples '*f*ine', '*v*alue', '*s*and', '*z*ero', '*sh*ore', 'plea*su*re' and '*h*ello'. The affricates [C] and [J] are like their English counterparts in the following examples: '*ch*ur*ch*' and '*j*udge'. The same is true of the nasals [m, n] as in '*m*ake' and '*n*o'; liquids [r, l], as in '*r*ain' and '*l*ong' and the glide [y], as in '*y*esterday'. (The only distinction between Persian and English is that in Persian [t, d, s, z, l, r, n] are dental sounds, while in English they are alveolar.) As is evident, whenever possible, the symbols used are those of the International Phonetic Alphabet (IPA).

However, as mentioned before because IPA requires special fonts, which are not readily available for a few of the sounds, we have used an ASCII symbol that resembled the relevant IPA

symbol. The only difference between our symbols and the ones used by IPA are in voiceless and voiced alveopalatal fricatives [S] and [Z], the voiceless and voiced affricates [C] and [J], and the palatal glide [y]. In the case of the latter, we did not want to use the lower case 'j', in order to decrease confusion.

3 Orthographic Labels (USCPers)

We proceed in this section to present an alternative orthographic system for Persian, as a first step in the creation of the USCPers+ system that will be presented later. The Persian writing system is a consonantal system with 32 letters in its alphabet (Windfuhr, 1987). All but four of these letters are direct borrowing from the Arabic writing system. It is important to note that this borrowing was not a total borrowing, i.e., many letters were borrowed without their corresponding sound. This has resulted in having many letters with the same sound (homophones). However, before discussing these cases, let us consider the cases in which there is no homophony, i.e., the cases in which a single letter of the alphabet is represented by a single sound.

In order to assign a symbol to each letter of the alphabet, the corresponding letter representing the sound of that letter was chosen. So, for instance for the letter 'پ', which is represented as [p] in USCPron, the letter 'p' was used in USCPers(ian).

These letters are:

ST	FR	AF	LQ	NS
پ p	ف f	چ C	ر r	م m
ب b	ش S	ج J	ل l	ن n
د d	ژ Z			
ک k	خ x			
گ g				
ع ?				

Table 4: USCPers(ian) Symbols: Non-Homophonic Consonants

As mentioned above, this partial borrowing of the Arabic writing system has given rise to many homophonic letters. In fact, thirteen letters of the alphabet are represented by only five sounds. These sounds and the corresponding letters are presented below:

- [t] for 'ت' and 'ط';
- [q] for 'ق' and 'غ';
- [h] for 'ه' and 'ح';
- [s] for 'س', 'ص' and 'ث' and
- [z] for 'ز', 'ذ', 'ض', and 'ظ'.

In these cases, several strategies were used. If there were two letters with the same sound, the lower case and the upper case letters were used, as in table 5. In all these cases, the lower case letter is assigned to the most widely used letter and the upper case, for the other.

[t]	ت	t	ط	T
[q]	ق	q	غ	Q
[h]	ه	h	ح	H

Table 5 USCPers(ian) Symbols: Homophonic Consonants 1

In the case of the letters represented as [s] and [z] in USCPron, because the corresponding upper case letters were already assigned, other symbols were chosen. For the letters sounding [s], 's', '$' and '&' and for the letters sounding [z], 'z', '2', '7' and '#'.

[s]	س	s	ص	$	ث	&			
[z]	ز	z	ض	2	ظ	7	ذ	#	

Table 6 USCPers(ian) Symbols: Homophonic Consonants 2

These letters are not the only ambiguous letters in Persian. The letters 'ی' and 'و' can be used as a consonant as well as a vowel, [y] and [i] in the case of the former and [v], [o] and [u] in the case of the latter. However, in USCPers, the symbols 'y' and 'v' were assigned to them, leaving the pronunciation differences for USCPron to capture. For instance, the word for 'you' is written as '**v**' in USCPers, but pronounced as [**o**], and the word 'but' is written as '**v**ly' and pronounced as [**v**ali].

As is the characteristics of languages employing the Arabic script, for the most part the vowels are not represented and Persian is no exception. The only letter in the alphabet that represents a vowel is the letter 'alef'. This letter has different appearances depending on where it appears in a word. In the word initial position, it appears as 'آ', elsewhere it is represented as 'ا'. Because the dominant sound that this letter represents is the sound [A], the letter 'A' was assigned to represent 'ا', which has a wider distribution; 'V' was assigned for the more restricted version 'آ'. In Persian, like in Arabic, diacritics mark the vowels, although they are not used in writing, unless to avoid ambiguities. Therefore, in our system, we ignored the diacritics.

Borrowed Letters	USCPers Symbol	USC-Pron
أ	@	an
أ	*	a
ئ	Y	e
ء	^	no sound
ؤ	W	o

Table 7 Non-Persian Letters

Finally in creating the one-to-one mapping between the Persian alphabet and USCPers, we need to deal with the issue of "pure Arabic" letters that appear in a handful of words. We see the same situation in the borrowed words in English, for instance the italicized letters in ca*ñ*on or na*ï*ve, are not among the letters of the English alphabet, but they appear in some words used in English. In order to ensure a one-to-one representation between the orthography and USCPers, these letters were each assigned a symbol, as presented on Table7.

USCPers, therefore, provides us with a way to capture each letter of the alphabet with one and only one ASCII symbol, creating a comparable system to USCPron for the orthography.

4 USCPers/USCPron: Two Way Ambiguity

As was noted in the previous section, vowels are not usually represented in orthography and there are many homophonic letters. These two properties can give rise to two sources of ambiguity in Persian which can pose a problem for speech-to-speech machine translation: (i) in which two distinct words have the same pronunciation (homophones), like 'pair' and 'pear' in English and the Persian words like 'sd' and '$d', which are both pronounced as [sad] and (ii) in which one orthographic representation can have more than one pronunciation (homographs) similar to the distinction between the two English words convict (n) and convict (v), which are both spelled c-o-n-v-i-c-t, but different stress assignments create different pronunciations. It is important to note that English has a handful of such homographic pairs, while in Persian homographs are very common, contributing to much ambiguity. In this section, we will discuss the transcription system we have adopted in order to eliminate these ambiguities.

4.1 Homophones

The examples in Table 8 illustrate the case in (i) (the letters with the same sounds are underlined). As is evident by the last column in Table 8, in each case, the two words have similar pronunciation, but different spellings.

Gloss	USCPers	USCPron
'hundred'	$d	[sad]
'dam'	sd	[sad]
'life'	HyAt	[hayAt]
'backyard'	HyAT	[hayAt]
'Eve'	HvA	[havA]
'air'	hvA	[havA]

Table 8: Same Pronunciation, Different Spellings

The word for 'life' ends in 't', while the word for 'backyard' ends in 'T'. In the other examples, because there is no difference in the pronunciation of 'h'/'H' and 's'/'$', we get ambiguity between 'Eve'/'air' and 'hundred'/'dam'. Therefore, this type of ambiguity appears only in speech.

4.2 Homographs

The second case of ambiguity is illustrated by the examples in the following table:

Gloss	USCPers	USCPron
'lung'	SS	[SoS]
'six'	SS	[SeS]
'thick'	klft	[koloft]
'maid'	klft	[kolfat]
'Cut!'	bbr	[bebor]
'tiger'	bbr	[babr]

Table 9: Same Spelling, Different Pronunciations

Here, we see that in the middle column two words that have the same orthographic representation correspond to different pronunciations (Column 3), marking different meanings, as is indicated by the gloss. This type of ambiguity arises only in writing and not speech.

4.3 Solution: USCPers+

Because of the ambiguity presented by the lack of vowels the data transcribed in USCPers cannot be used either by MT or for language modeling in ASRs, without significant loss of information. In order to circumvent this problem, we adopted a

modified version of USCPers. In this new version, we have added the missing vowels, which would help to disambiguate. (Because this new version is USCPers + vowels, it is called USCPers+.) In other words, USCPers+ provides both the orthographic information as well as some phonological information, giving rise to unique words. Let us reconsider the examples we saw above using this new transcription system. A modified version of Table 8 is presented in Table 10.

Gloss	USCPers	USCPers+	USCPron
'hundred'	$d	$ad	[sad]
'dam'	sd	sad	[sad]
'life'	HyAt	HayAt	[hayAt]
'backyard'	HyAT	HayAT	[hayAt]
'Eve'	HvA	HavA	[havA]
'air'	hvA	havA	[havA]

Table 10: USPers+ Disambiguates Cases with Same Pronunciation & Different Spellings

Table 11 is the modified version of Table 9:

Gloss	USCPers	USCPers+	USCPron
'lung'	SS	SoS	[SoS]
'six'	SS	SeS	[SeS]
'thick'	klft	koloft	[koloft]
'maid'	klft	kolfat	[kolfat]
'Cut!'	bbr	bebor	[bebor]
'tiger'	bbr	babr	[babr]

Table 11: USCPers+ Disambiguates Cases with Same Spelling & Different Pronunciations

Data in Column 4 and Column 2 of Tables 10 and 11, respectively, show that USCPron and USCPers can give rise to ambiguity, while no ambiguity exists in USCPers+, Column 3.

The following sentence also illustrates this point, where the words 'thick' and 'maid' from Table 11 are used. Assume that ASR receives the audio input in (1) represented in USCPron:

(1) USCPron: [in koloft ast]
Gloss: thisthick is
Translation: 'This is thick'

If ASR outputs USCPers, as in (2),

(2) USCPers: Ayn klft Ast

the MT output in the English language can choose either:

(3) a. This is thick
b. This is a maid

as a possible translation. However, using USCPers+ instead of USCPers would avoid this ambiguity:

(4) USCPers+: Ayn koloft Ast (cf. (2))

As evident, there is a significant benefit by using USCPers+.

The discussion of the conventions that have been adopted in the use of USCPers+ and USCPron, e.g., not including punctuations or spelling out numbers, is beyond the scope of this paper. However, it is important to note that by adopting a reasonable number of conventions in our transcription of USCPers+ and USCPron, we have been able to provide a complete transcription convention for acoustic models and language models for the ASRs, TTSs and MTs for our English to Persian translation system.

5 Further Issue: Dealing with the Lack of Data

Despite the significant advantages of employing the USCPers+ transcription scheme, a drawback is the lack of data in this format. To address this shortcoming, semi-automated techniques of data conversion have been developed that take into consideration the statistical structure of the language. Fig. 2 depicts a network that can be inferred from a relatively small amount of humanly transliterated data. By employing statistical decoding techniques through such a model, the most likely USCPers+ sequence can be generated using minimal human intervention.

Consider for example the sentence 'SS mn drd myknd' and the network structure shown above. It is likely that the combination 'man dard' and 'dard mykonad' have been seen in the manually generated data, and thus the decoder is likely to chose the path 'man dard mykonad' as the correct transliteration.

Manual decision can be made in the cases that the system reaches a statistical ambiguity (usually in cases such as 'Ayn klft Ast') or that insufficient training data exist for the specific region of decoding.

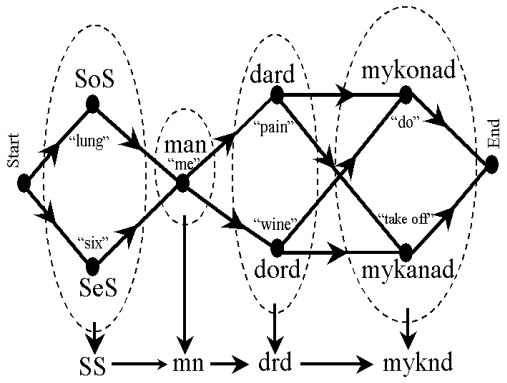


Fig 2. The possible transitions between words are probabilistically denoted in a language model, which can be employed for decoding of the most likely path, given several possibilities. Shown above are the possibilities for the decoding of the utterance "SS mn drd myknd".

The first ambiguity is rare, and usually involves short segments of text. Thus as the models improve, and we move to higher orders of decoding, the statistical ambiguity becomes less significant. Similarly, the unknown words keep decreasing as new converted data feeds back into the training corpus.

In our experiments, as the amount of training data grew from about 16k to 22k words, the precision in transliteration increased from 98.85% to 99.2%, while at the same time the amount of manual intervention was reduced from 39.6% to 22%. It should be noted that by changing the decision thresholds the intervention can fall significantly lower, to 9.4% with a training corpus of 22k words, but this has the effect of a lower precision in the order of 98.8%.

An indepth discussion of the techniques employed for the transliteration process is presented in Georgiou, et.al (2004).

6 Conclusion

This paper argues that the best way to represent data at phonological/lexical level for language modeling and MT in languages that employ the Arabic script, is by using a hybrid system, which combines information provided by orthography and includes the vowels that are not represented in orthography. The schemes proposed can significantly aid in speech-to-speech applications in a multitude of different ways: (1) the internal pronunciations of the ASR and the TTS components can employ the USCPron scheme, (2) the internal transcription of the Persian language—for purposes of language modeling and statistical machine translation among others—can employ the USCPers+ scheme and (3) in the case of a stand-alone TTS, in which case the input is pure Persian text, automated transliteration to the USCPers+ scheme, and hence to the pronunciation, can be generated with statistical language augmentation techniques, which are based on prior model training, as we describe further in Georgiou, 2004.

This would ensure a uniqueness that otherwise is not available. It has also been suggested in this paper that a modification of IPA, which would allow the use of ASCII characters, is a more convenient way to capture data for acoustic modeling and TTS. Persian data resources developed under the DARPA Babylon program have adopted the conventions described in this paper.

7 Acknowledgements

This work was supported by the DARPA Babylon program, contract N66001-02-C-6023. We would like to thank the following individuals for their comments and suggestion: Naveen Srinivasamurthy and HS, MK and SS for working with the first versions of this system and making insightful suggestions.

8 References

The DARPA Babylon program," http://darpa-babylon.mitre.org.

P. Georgiou, H. Shiranimehr and S. Narayanan (2004). Context Dependent Statistical Augmentation of Persian Transcripts for use in Speech to Speech Translation Applications. INTERSPEECH 2004-International Conference on Spoken Language Processing.

J.L. Hieronymus, ASCII Phonetic Symbols for the World's Languages: Worldbet, AT&T Bell Labs, http://cslu.cse.ogi.edu/corpora/corpPublications.html

Y. Kachru. 1987. "Hindi-Urdu," The World's Major Languages, ed. Bernard Comrie, Oxford University Press.

A.S. Kaye. 1987. "Arabic," The World's Major Languages, ed. Bernard Comrie, Oxford University Press.

T. Lander, The CSLU Labeling Guide, OGI, http://cslu.cse.ogi.edu/corpora/corpPublications.html

S. Naraynan, et. al. 2003. Transonics: A speech to speech system for English-Persian interactions.

G.L. Windfuhr. (1987). "Persian," The World's Major Languages, ed. Bernard Comrie, Oxford University Press.