SPEECH ASSESSMENT METHODS PHONETIC ALPHABET (SAMPA) : ANALYSIS OF URDU

HASAN KABIR AND ABDUL MANNAN SALEEM

ABSTRACT

An important consideration to be kept in mind, while dealing with speech processing, especially speech synthesis, is that the computer cannot process the phonemic data directly as the way we write it in IPA symbols. The IPA symbols representing different sounds are just graphical symbols with no core meaning stored in the computer. For the processing, the phonemic data is stored using machine readable phonetic alphabets comprising of some standard character sets, e.g., ASCII, UNICODE, etc. This paper provides information for representing Urdu phonemes with one such standard by the name of SAMPA (Speech Assessment Methods Phonetic Alphabets).

1. INTRODUCTION

SAMPA (Speech Assessment Methods Phonetic Alphabet) is a machine-readable phonetic alphabet. It was originally developed under the ESPRIT project 1541, SAM (Speech Assessment Methods) in 1987-89 by an international group of Phoneticians, comprising speech scientists from nine countries of the European Community. It was initiated and coordinated by John Wells of University College London, and was applied in the first instance to the European Communities languages Danish, Dutch, English, French, German, and Italian (by 1989); later to Norwegian and Swedish (by 1992); and subsequently to Greek, Portuguese, and Spanish (by 1993). The purpose of SAMPA was to form the basis of an international standard machine-readable phonetic alphabet for purposes of international collaboration in speech research. SAMPA as presently constituted covers all the symbols needed for the phonemic transcription of the principal European Union languages. One useful application is for sending phonetic symbols by e-mail (http://www.phon.ucl.ac.uk/home/sampa/home.htm).

SAMPA is an ASCII encoding of the phonemes of particular languages, based on the International Phonetic Alphabet (IPA). The general SAMPA definition maps IPA symbols to ASCII codes, while the SAMPA applications to specific languages additionally pre-suppose a specific phonemic analysis. Consequently, while agreeing on the IPA to ASCII mapping, it is possible to make different choices of phonemic analyses for different languages, and thus define different SAMPA representations. Since its first application to 7 European languages, SAMPA has been applied to a wide range of languages, with modifications and extensions suggested for reasons, which have arisen from practical use in speech technology and spoken language lexicography (http://coral.lili.uni-bielefeld.de/Documents/sampa.html).

2. BACKGROUND

SAMPA maps symbol of the IPA (International Phonetic Alphabet) onto ASCII codes in the range 37-126, the 7-bit printable ASCII characters. Associated with the coding (mapping) are guidelines for the transcription of the languages to which SAMPA has been applied. Unlike other proposals for mapping the IPA onto ASCII, SAMPA is not one single author's scheme, but represents the outcome of collaboration and consultation among speech researchers in many different countries. The SAMPA transcription symbols have been developed by or in consultation with native speakers of every language to which they have been applied, but are standardized internationally.

The IPA phonetic symbols that are also lower-case alphabet symbols naturally remain the same in SAMPA. These are the lower-case Latin letters a-z, ASCII/ANSI 97-122. SAMPA recodes all other phonetic symbols covered within the range 37-126. In the current SAMPA, ASCII 39 (‘) stands for rising tone and (`) stands for Retroflex.

2.1. SAMPROSA

In its basic form SAMPA was seen as catering essentially for segmental transcription, particularly of a traditional phonemic or near-phonemic kind. Prosodic notation was not adequately developed. This shortcoming was remedied by a proposed parallel system of prosodic notation, SAMPROSA. It is important that prosodic and segmental transcriptions be kept distinct from one another, on separate representational tiers (certain symbols have different meanings in SAMPROSA and different meanings in SAMPA: e.g. (H) denotes a labial-palatal semivowel in SAMPA, but High tone in SAMPROSA) (Wells).

2.2. X-SAMPA

John Wells has extended SAMPA to cover the entire International Phonetic Alphabet, and Dafydd Gibbon has introduced extensions for morphological boundary marking.

A recent proposal for an extended version of the segmental alphabet, X-SAMPA, would extend the presently agreed conventions so as to make provision for every symbol on the Chart of the International Phonetic Association, including all diacritics. In principle, this would make it possible to produce a machine-readable phonetic transcription for every known human language.

A SAMPA transcription is designed to be uniquely parsable. As with the ordinary IPA, a string of SAMPA symbols does not require spaces between successive symbols. SAMPA is language independent (http://www.phon.ucl.ac.uk/home/sampa/x-sampa.htm).

The present SAMPA recommendations (as devised for the basic six languages) are set out in the following table. All IPA symbols that coincide with lower-case letters of the Latin alphabet remain the same; all other symbols are recoded within the ASCII range 37-126.

TABLE 1 Vowels, Diphthongs and Triphthongs. Taken from http://victorian.fortunecity.com/vangogh/555/Spell/sampa.htm

Short	*Long*	*Diphthongs*	*With Schwa*
æ {	**a:** A	**Ai** aI	**aə** a@ | **aiə** aI@
at, ax, ask, cat	alms, want	eye, ice, bite	are / ire, fire
e E	**3`** 3	**Ei** eI	**eə** e@
edje, get, elbow	her, girl, urban	ace, ape, vein	air, care, there
ɪ I	**i** i	**Oi** oI	**iə** i@
it, in, index, ill	eel, east, very	Oil, boy, loyal	ear, fear, deer
ɔ Q	**o** o	**Ou** ou	**oə** o@
ox, cot	awe, call, cost	Oh, oat, low	for, four, floor, more
ʊ U	**u** u	**Ju** yu	**uə** u@
hook, put, book	ooze, zulu, zoo	you, few, fuse	your, sure
^ V	**ə** @	**Au** Au	**auə** Au@
up, cut	ago, sofa, unit	out, down	our, flower, power

3. LITERATURE REVIEW

The IPA symbols representing different sounds are just graphical symbols with no core meaning stored in it. The attempts at a standard ASCII form of the IPA resulted into TIMITBET, MRPA, SAMPA, etc; (Hieronymus p. 1).

There is another standard being followed by phoneticians called *OGIbet*. It was developed by TIMIT, Texas Instruments (TI) and Massachusetts Institute of Technology (MIT). The phonemes in this alphabet are spelled using lower-case English alphabetics [a-z]. As

OGI moved into labeling of numerous languages, it was felt necessary to adopt a more international labeling convention, and so *Worldbet* by Jim Heironomous of AT&T Bell Labs was chosen as the basis for future labeling. In addition, older OGI corpora were relabeled to Worldbet using a combination of direct transliteration. Worldbet uses various ASCII special symbols quite liberally. Some of the ASCII special symbols used in Worldbet are: digits, :, @, _, >, &, (, =, ~, *, ., and ? (http://cslu.cse.ogi.edu/toolkit/old/old/documentation/cslurp/wincslurp/node16.html).

World bet is an attempt to have a phonetic alphabet, which covers all of the world's languages in a systematic fashion. It is an ASCII version of IPA plus a number of symbols, which were found useful in database labeling, which are not currently in the official IPA set. It is designed for a large set of languages including Indian, Asian, African and European Languages as proposed by Hieronymus (p. 2).

4. RESULTS

As mentioned earlier, SAMPA is not single author's scheme, but represents the outcome of collaboration and consultation among speech researchers in many different countries. Till now, any authentic mapping of Urdu sounds on SAMPA is not being done. SAMPA can be used to standardize Urdu sounds. Dr. Sarmad Hussain, National University of Computer & Emerging Sciences talked to John Wells and he suggested following mentioned things for Urdu sounds mapping into SAMPA. The results of this section comprises repository of the SAMPA mapping for all the Urdu vocalic and consonantal sounds. These are listed in table 2 and table 3 in Appendix.

4.1. Manners of Urdu Sounds

As there could not be a single character mapping for all the IPA symbols, for that matter we need some special mechanism to incorporate all the existing sounds. SAMPA uses some special notations to represent those sounds whose single character mapping is not available. This special mechanism is applied on some classes of sounds and these classes are formulated according to manner of articulation. Here we have discussed some manner of articulation of Urdu sounds for which SAMPA proposes some special notations as suggested by Wells.

Aspiration:
For aspiration, in SAMPA mapping a simple _h following the symbol for the consonant is used, e.g. n_h is the SAMPA mapping of Urdu sound نھ whose IPA equivalent is [n^{h}].

Dental:
For dental sounds, in SAMPA mapping a simple _d following the symbol for the consonant is used, e.g. t_d is the SAMPA mapping of Urdu sound ت whose IPA equivalent is [t̪].

Nasalization:
For the mapping of nasalized sounds SAMPA uses both @~ and ~ following the symbol for the vowel and consonant, but we are using ~ just for the sake of simplicity e.g. the SAMPA mapping of Urdu sound ا is A~ whose IPA equivalent is [ɑ̃].

Retroflex:
Urdu has many retroflex as compared to English. To incorporate the retroflex sounds SAMPA uses (`) notation following the symbol for the consonant, e.g. the SAMPA mapping of Urdu sound ٹ is t` whose IPA equivalent is [ʈ].

5. REFERENCES

Hieronymus, James. L.; "ASCII Phonetics Symbols for the World's Languages:

Worldbet"; AT&T Bell Laboratories, Murray Hill, NJ 07974, USA.

Wells, J. C.; "Computer-coding the IPA: A Proposed Extension of SAMPA", downloaded from
http://www.phon.ucl.ac.uk/home/sampa/ipasam-x.pdf; London: University College.

http://www.phon.ucl.ac.uk/home/sampa/home.htm; "SAMPA: Computer Readable Phonetic Alphabet".

http://www.phon.ucl.ac.uk/home/sampa/x-sampa.htm; “Computer-coding the IPA: A Proposed Extension of SAMPA”.

http://coral.lili.uni-bielefeld.de/Documents/sampa.html; “SAMPA Definitions and Information”

http://cslu.cse.ogi.edu/toolkit/old/old/documentation/cslurp/wincslurp/node16.html; “IPA, Worldbet and OGIbet”.

http://victorian.fortunecity.com/vangogh/555/Spell/sampa.htm; “SAMPA: Computer Readable Phonetic Alphabet”.

6. Appendix

6.1. Urdu Vowels

TABLE 2 Urdu vowels along with their SAMPA mappings

IPA	*SAMPA*	*Letter*	*Example*
i	I	ی	bin
e	E	ے َ	beʃərəm
æ	{	ے ِ	bæl
u	}	ُو	pura
o	o	و	kʰonɑ
ɔ	O	وَ	pɔɖa
ɑ	A	ا	bɑn
ɪ	I	ِ	ɖɪn
ɛ	E		sɛhər
ʊ	U	ُ	sʊn
ə	@	َ	kəlɪ
ĩ	i~	یں ِ	pɛhnĩ
ẽ	e~	یں	kəhẽ
æ̃	{~	یں َ	hæ̃
ũ	}~	ُوں	kʰa͡ũ
õ	o~	وں	kʰɑnõ
ɑ̃	A~	اں	ləɽkɪɑ̃

6.2. Urdu Consonants

TABLE 3 Urdu consonants along with their SAMPA mappings

IPA	*SAMPA*	*Letter*	*Example*
p	p	پ	pɑn
b	b	ب	bɑp
pʰ	p_h	پھ	pʰəndɑ
bʰ	b_h	بھ	bʰap
m	m	م	məkɽɪ
mʰ	m_h	مھ	ʈʊmʰẽ
ʈ	t_d	ت, ط	bɑʈ
ɖ	d_d	د	ɖal
ʈʰ	t_d_h	تھ	ʈʰali
ɖʰ	d_d_h	دھ	ɖʰəmal
n	n	ن	nɑn
nʰ	n_h	نھ	nʰɑnɑ
ŋ			səŋ
ʈ	t`	ٹ	kɑʈɑ
ɖ	d`	ڈ	ɖɑlɑ

ʈʰ	t`_h	ٹھ	ʈʰanɪ
ɖʰ	d`_h	ڈھ	ɖʰal
k	k	ک	kəlim
g	g	گ	gana
kʰ	k_h	کھ, که	kʰaja
gʰ	g_h	گھ	gʰa͡o
q	q	ق	qələm
ʔ	?	ع	ʔəlim
f	F	ف	faʃ
v	V	و	vaɖɪ
s	S	س, ص, ث	sarɪ
z	Z	ذ,ظ, ض, ز	zahɪɖ
ʃ	S	ش	ʃaɖɪ
ʒ	Z	ژ	ʒala
ɣ	7	غ	ɣərib
x	x	خ	xana
h	h	ھ, ح	hamɪ, hala
tʃ	tS	چ	tʃa͡e
tʃʰ	tS_h	چھ	tʃʰal
dʒ	dZ	ج	dʒa͡o
dʒʰ	dZ_h	جھ	dʒʰag
r	r\	ر	mʊqərər
rʰ	r_h	رھ	
ɽ	r`	ڑ	baɽ
ɽʰ	r`_h	ڑھ	
j	J	ی	gaja
l	L	ل	lana
lʰ	L_h	لھ	lɛhlʰana