

ASR Consortium Pronunciation Specification for Lexicon Development

Nandini Bondale¹, Joyanta Basu², Milton S. Bepari²
nandini@tifr.res.in, joyanta.basu@cdac.in, milton.bepari@cdac.in

ASR Consortium Members

IIT Madras (coordination)

¹TIFR Mumbai

IIIT Hyderabad

IIT Guwahati

IIT Bombay

IIT Kanpur

²C-DAC Kolkata

Project Name
**Speech-based Access for
Agricultural Commodity Prices
in Six Indian Languages**

Sponsored by

Technology Development in Indian Languages (TDIL)
Department Of Information Technology
Ministry of Communication and Information Technology
Govt. of India

Contents

1. Objective of ASR Project
2. ASR Standard Label set
3. Language wise Extra Phone List
4. Extra Non-speech Sound
5. Pronunciation Dictionary
6. Conclusion

Objective

To implement and deploy a speech based system to get prices of agricultural commodities using telephone/mobile for the already existing website <http://www.agmarknet.nic.in/> managed by the Ministry of Agriculture

Reading or writing skills are not required for using the proposed system.

Six systems are being built, one for each language.

Bengali (C-DAC Kolkata),

Hindi (IIT Kanpur), **Assamese** (IIT Guwahati),

Marathi (IIT Mumbai & TIFR),

Tamil (IIT Madras), **Telugu** (IIIT Hyderabad).

ASR Standard Label set

Our Approach:

- Labels be case insensitive
- Semantically simple rules
e.g. For all Suffixes, geminations used in ASR labels
- Rules for using the symbols are explicitly stated with examples
- Currently we use non alphanumeric characters, but use of only alphanumeric labels preferred to avoid processing problems
- In this consortium, we do not have any tonal language, so have not considered related labels.

ASR Standard Label set

Vowels

ASR Labels	Examples	Pronunciation	IPA Notation
a	कमल	k a m a l	/a/
ax	राष्ट्र	r aa s' t' r ax	/ə/
aa	काम	k aa m	/a/
i	किस	k i s	/i/
ii	बीच	b ii c	/i:/
u	सुमन	s u m a n	/u/
uu	झुम	jh uu m	/u:/
e	मेरा	m e r aa	/e/
ee	ate [Dravidian long e]	ee t	/e:/
ex	8 (Tamil digit) [Dravidian short e]	ex tt eu	/ě/
o	मोर	m o r	/o/
oo	onam (Dravidian festival)	oo n' a m	/o:/
ox	1 (Tamil digit) [Dravidian short o]	ox nn eu	/ö/
ae	बँट	b ae t'	/æ/

Vowels

ASR Labels	Examples	Pronunciation	IPA Notation
ei	कैसे	k ei s e	/ei/
ou	कौन	k ou n	/ou/
ea	बहन	b ea h ea n	/ea/
ai	hand in Tamil/Kannada [Dravidian ai]	k ai	/ai/
au	out [Dravidian au]	au t	/au/
eu	8 (Tamil digit) [Dravidian eu]	ex tt eu	/eu/
ao	कॉलेज (college) [Marathi]	k ao l e j	/ao/

CONSONANTS: Velars

ASR Labels	IPA Notation	Examples	Manner of Articulation
k	/k/	कमल	Unaspirated unvoiced stop
kh	/k ^h /	खाना	Aspirated unvoiced stop
g	/g/	गरम	Unaspirated voiced stop
gh	/g ^h /	घूमना	Aspirated voiced stop
ng	/ŋ/	कंगन	Voiced nasal

CONSONANTS: Palatals

ASR Labels	IPA Notation	Examples	Manner of Articulation
c	/tʃ/	चरखा	Unaspirated unvoiced stop
ch	/tʃ ^h /	छाता	Aspirated unvoiced stop
j	/dʒ/	जमना	Unaspirated voiced stop
jh	/dʒ ^h /	झरना	Aspirated voiced stop
nj	/ɲ/	अंजन	Voiced nasal

CONSONANTS:

Retroflexes

ASR Labels	IPA Notation	Examples	Manner of Articulation
t'	/t̪/	टोकरी	Unaspirated unvoiced stop
t'h	/t̪ʰ/	ठंड	Aspirated unvoiced stop
d'	/d̪/	डाली	Unaspirated voiced stop
d'h	/d̪ʰ/	ढङ्कन	Aspirated voiced stop
n'	/ɳ/	गणित	Voiced nasal

CONSONANTS: Dentals

ASR Labels	IPA Notation	Examples	Manner of Articulation
t	/t/	तलवार	Unaspirated unvoiced stop
th	/t ^h /	थाली	Aspirated unvoiced stop
d	/d/	दम	Unaspirated voiced stop
dh	/d ^h /	धोका	Aspirated voiced stop
n	/n/	नगर	Voiced nasal

CONSONANTS:

Bilabials

ASR Labels	IPA Notation	Examples	Manner of Articulation
p	/p/	पान	Unaspirated unvoiced stop
ph	/p ^h /	फूल	Aspirated unvoiced stop
b	/b/	बल	Unaspirated voiced stop
bh	/b ^h /	भूल	Aspirated voiced stop
m	/m/	मन	Voiced nasal

CONSONANTS: Alveolars

ASR Labels	IPA Notation	Examples	Manner of Articulation
y	/j/	योग	Voiced approximant
r	/r/	राम	Voiced trill
l	/l/	लोरी	Voiced lateral
w	/ʋ/	वाहन	Voiced approximant
sh	/ʃ/	शाम	Unvoiced fricative

CONSONANTS:

ASR Labels	IPA Notation	Examples	Place of Articulation	Manner of Articulation
s'	/ʂ/	पुरुष	Retroflex	Unvoiced fricative
s	/s/	साथ	Dental	Unvoiced fricative
h	/h/	हम	Glottal	Unvoiced fricative
l'	/ɭ/	बाल	Retroflex	Voiced lateral
d'q	/ɽ/	लड़का	Retroflex	Voiced tap

Additional Labels

ASR/ Sphinx	labels (Sphinx notation)	Example word / Comments
kk mm nn' ll ss	p a kk aa a mm a a nn' a a ll aa h g u ss aa	Geminated consonants: repeat symbol पञ्जा amma (mother in Tamil) anna (brother in Tamil) (suffix is after geminnation) अल्लाह गुस्सा
f z	f oo l z uu	English sounds fool zoo
d'q khq	l a d'q k aa khq a t	[Hindi nukta: suffix q] लड़का खत
aan on	k a h aan k y on	[Hindi nasalised vowel: suffix n] कहाँ क्यों [Hindi bindu at the end of word]
j-	j- a g	जग (live in Marathi) [Marathi dental affricate]
r'		Malayalam/ old{Telugu,Kannada} retroflex r

A Note on Suffixes in the Labels

1. **Aspiration:** Use suffix **h** to denote aspiration: **k** (क) versus **kh** (ख)
2. **Retroflex consonants:** Append **'** to denote a retroflex consonant: **t** (त) versus **t'** (ट).
3. **Flap:** Use suffix **q** to denote a flap (nukta in Hindi): **d' a g a r** (डगर) versus **l a d'q k aa** (लड़का).
4. **Nasalized vowel:** Use suffix **n** to denote nasalisation of a vowel: **k a h aa** (कहा) versus **k a h aan** (कहाँ).
5. **Reduced vowel:** Use suffix **x** to denote reduction of a vowel: **r aa s' t' r ax** (राष्ट्र). Also, use suffix **x** to denote the two “short” dravidian vowels: **ex, ox** (Note: Hindi short /e/ is phonetically similar to Dravidian long e /ee/).
6. **Dental affricates:** Use suffix **-** to denote dental affricate of Marathi: **च, ज, झ**.

Priorities of suffixes: in decreasing order: **' h q n x -**

A note on using nasals to represent ‘Bindu’ in the script

Place of articulation of consonant after bindu	Label sequence	Example word
velar (k,kh,g,gh)	a n g k k a n g g a n	अंक कंगन
palatal (c, ch, j, jh)	a n j c a l a n j j a n	अंचल अंजन
retroflex (t', t'h, d', d'h)	k a n' t h' a n' d' a a	कंठ अंडा
dental (t, th, d, dh)	a n t a n d h a a	अंत अंधा
labial (p, ph, b, bh)	k a m p a n a m b a a	कंपन अंबा

ASR Standard Tags for Non Speech Sounds

Tag	Explanation
<pau>	P ause or silence
<ct>	C learing of throat
<tc>	tongue c lick
 	b reath noise
<laugh>	l aughter
<burp>	burp
<cough>	cough
<sneeze>	sneeze
<ct>	C learing of throat
<sniff>	sniff
<ls>	Lip smack
<ns>	hiccups, yawns, grunts
<bs>	B ackground s peech (babble)
<horn>	H orn noise of vehicles
<ln>	L ine noise
<bang>	Sudden (<i>impulsive</i>) noise due to b anging of door

Extra phones used compared to ASR Standard Labels

Marathi: No extra phone used

Hindi: No extra phone used

Telugu: No extra phone used

Tamil:

Unicode	Phone	Grapheme	Pronunciation
ழ	Zh	pazham	p a zh a m
ன	n~	kanakaamparam	k a n~ a g aa m b a r a m
ற	r'	Paakar'kaay	p aa g a r' k aa y
த	t~	Raaman~aatapuram	r aa m a n aa t~ a b u r a m

Contd....

Extra phones used compared to ASR Standard Labels

Bengali :

Phone	Roman Script	Pronunciation	Bangla Word
aal	haalbrld'a	h aal b r l d'	হাইব্রীড
aane	saanithiyaa(2)	s aane th l aa	সাইথিয়া
aani	saanithiyaa	sh aani th i aa	সাইথিয়া
aa0	haaod'qaaa	h aao d'q aa	হাওড়া
aaU	laau	l aau	লাউ
aeo	sheod'qaaphuli	sh aeo d'q aa ph u l i	শেওড়াফুল
lu	blull	b lu l l	বিউল
oi	nainitaala	n oi n i t aa l	নৈনিতাল

Contd....

Extra phones used compared to ASR Standard Labels

Assamese:

Phone	Assamese word	Roman script	Pronunciation
~n	বেঙেনা	bengena	b e ~n en n' aan
oi	মিঠৈ	mithoi	m in th oi
a'	অমিতা	amita	a' m in t' aa
a'n	অন্য	ainya	a'i n' n' a'n
a'i	শস্য	haisya	x a'i s s a
x	শস্য	haisya	x a'i s s a
ya'	শস্য	haisya(2)	x a'i s ya'
ya'n	অন্য	ainya(2)	a'i n ya'n

Extra Non-Speech Sounds used

Marathi

Tag	Explanation
<babble>	Speech or sound generated by other humans
<bn>	Background Noise

Bengali & Assamese

Extra Tag	Explanation / examples
<ring>	Phone ringing
<bn>	General Background Noise
<bird>	Sound of Bird
<air>	Air flow
<cry>	Children Cry

Pronunciation Dictionary

ASR Consortium uses “common set” of labels as presented.

Common set along with extra phones as presented is being used for preparing the pronunciation dictionary.

Each group has prepared a Pronunciation Dictionary of the specific language for Transcription of the data.

For our task the dictionary consists of commodity names on <http://www.agmarknet.nic.in/> site, district names in that state, yes/no for confirmation and other associated words related to prices.

The dictionary also includes multiple pronunciation variants and dialectical variations.

Conclusions

- With the experience of data collection and transcription for six languages in six states, we feel that the labels used by the consortium are easy to use and not confusing.
- ASR consortium labels may be adopted for other Indian languages with additional phones where ever required.

धन्यवाद

Thank You

धन्यवाद

நன்றி