CENTRAL INSTITUTE OF INDIAN LANGUAGES
DEPARTMENT OF HIGHER EDUCATION
Ministry of Human Resource Development, Government of India
Manasagangotri, Mysore - 570 006

LDC-IL

Linguistic Data Consortium for Indian Languages

Standards for Speech Data Capturing and Annotation

DATA CAPTURING

1. Data in NIST format
2. Rate of sampling in the multiples of 8 kHz, depending on the purpose for which data has to be used. The purpose and the rate of sampling to be uniform for LDC-IL.

   a. For linguistic - phonetic research 48 kHz 16 bit
   and for research on speech pathology

   b. For continuous speech recognition

   Telephone landline/cellphone 8 kHz 8 bit

   All others 16 kHz 16 bit

   High quality phonetically balanced
   News reading (simulated, acquired)
   Read (via telephone)
   Read (in studio/desktop microphone)
   Application specific/domain specific
   Keyword spotting
   Spontaneous speech, meeting room
   Telephone conversation

   c. For isolated word recognition in studio environment/desktop microphone

   Application specific/domain specific
   Telephone(landline/cell phone) walkie-talkie

   d. For text-to-speech 16 kHz 16 bit

   Studio environment – rich phonetic database
   - limited domain database

ANNOTATION

Use PRAAT, Wave surfer for segmentation and annotation

1. For linguistic - phonetic research: at the layers of Segment, Allophone, Phoneme, tonal unit, intonation unit, Text (Script), Word, Phrase, Sentence. The number of levels will depend upon the target application. Examples of Segments:

   Stop closure, Plosive burst, VOT Lag, Segment - to - Segment Transition, Steady state of vowels.

2. For automatic speech recognition

   Levels such as Phone (phoneme), Syllable, Word, Sentence level to be specified in the proposal

3. For text to speech

   Phone, Diphone, Syllable, Word, Sentence level

4. Transliteration Scheme of LDC-IL

   There will be two layers of transliteration:

   a. Shallow layer: LDC-IL transliteration scheme (see LDC-IL website)
   b. Deep layer : UNICODE

Header : NIST format

Obligatory :

Name of the database id:

Speaker id:

Sampling:

Number of samples:

Big Endian:

Little Endian:

Number of bytes/samples:

Time-aligned transcription

A time-aligned transcription would be like as follows in pure text format.

Word:

```
MillisecondsPerFrame 1.00000
Language Name
END HEADER
0 200  word  a
200 560 word b
560 800 word c
```

800 900 word d etc.,
Standard LDC-IL transcription scheme/ orthography (for the specific language) MUST be used for transcriptions.

Syllable Transcription:

MillisecondsPerFrame 1.00000
Language Name
0 200 syllable 1
200 300 syllable 2
300 560 syllable 3
560 650 syllable 4
650 800 syllable 5
800 900 syllable 6 etc.,

Similar transcriptions may be given at the phonetic level.

Phonemic Transcription:

MillisecondsPerFrame 1.00000
Language Name
0 200 first phone
200 230 second phone
230 300 third phone
300 325 fourth phone
325 560 fifth phone   etc.,

Non-time aligned conventions:

This section gives the conventions for non-time aligned conventions

STANDARDS FOR RECORDING EQUIPMENTS

a. Linguistic phonetic research: Equipment should be solid state sound recorders.

b. Continuous speech recognition: PC based, telephone based, cell based.
Sound card, VHF,UHF, using various mikes eg., goose-neck, array, noise- cancellation etc.

c. Isolated word recognition

d. Text to speech: Solid state/good quality sound card. Recording in the studio environment.

DATA SUPPLY

Out-sourced institutions have to submit the data either on CDs or DVDs in the form of CD or DVD with proper labeling written in indelible ink on the top of the medium along with a written explanation of the content.

TAG SET

Language name tags

For languages listed in the 8th schedule of the Constitution and for non-scheduled languages as indicated in the Census. If need be LDC-IL can choose one if there are two or more variations.

Tags for non speech and other miscellaneous tags

1. Asterisk: Indicates cutoff speech (see example above). If beginning is cutoff, for example in me e in meeraa, then indicate as [mee]. If end is cutoff, for example ra in meeraa, then indicate as mee[raa].
2. .blip or <blip>: To indicate when the sound goes dead ï as in a line that goes silent.
3. .bn or <bn>: background noise
4. .br or <br>: breath noise
5. .laugh or <laugh>: laughter
6. .pau or <pau>: silence
7. .bs <pau>: background speech
8. .pron or <pron>: for a non standard pronunciation. If the accent can be identified as in a ñregional accent.òThen <pron-regioni> may be used. If it is not know leave it as <pron>
9. .burp or <burp>: burping
10. .cough or <cough>: coughing
11. .sneeze or <sneeze>: sneezing
12. .sniff or <sniff>: for the entire period for which the speaker sniffs
13. .sp or <sp>: if transcriber comes across an unfamiliar sound
14. .tc or <tc>: tongue click
15. .uu or <uu>: unintelligible sounds
16. .whisper or <whisper>: whispered speech
17. .ct or <ct>: clearing of throat
18. .ln or <ln>: line noise (as in telephone)
19. .glot or <glot>: if heavy glottalisation occurs
20. .bengali or <bengali>: if the language is different from the language for which the data is collected. Other languages <english>, <tamil>, <telugu>, <marathi>, <oriya>, <gujarati>, <hindi>, <bengali>, <konkani>, <tulu>, <kannada>, <Malayalam>, <kashmiri>, <urdu>, <nepali>, <punjabi>, <assamese>,é
21. If the foreign speech cannot be deciphered: <foreign text>, where text corresponds to the transcription and <foreign> indicates that the language is different.
22. .ns or <ns>: hiccups, yawns, grunts
23. .vs or <vs>: high pitched squeak
24. .female or <female>: female
25. .male or <male>: male
26. .age 40 or <age-40>: if the age can be deciphered (here age is 40)

[June 9, 2008]