


Members Present:


1. L Ramamoorthy (Reader cum Research Officer, Head, LDCIL)
2. Hema A Murthy (IIT Madras)
3. K Samudravijaya (TIFR, Mumbai)
4. A G Ramakrishnan (IISc, Bangalore)
5. Rekha Sharma (Retd., CIIL, Mysore)
6. M Venkatesan (LDCIL)
7. Arundhati (LDCIL)
8. Mona Parakh (LDCIL)
9. Richa (LDCIL)
10. A Bharathraju (LDCIL)

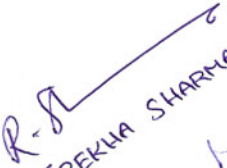
1 Summary

The committee reviewed the data collected so far. Data has been collected using a recorder EDIROL R09, in different environments (fan noise, lab, office, home, schools, ...) as per the directions given. The data was collected at a sampling rate of 48 KHz and 16 bits/sample.

1. Number of Languages: 18
2. Number of speakers: 300-450
3. Approximately 50% male and 50% female.
4. Three to Four dialectal variations for each language.
5. Variation of age: 16-20 (8 male, 8 female for each dialect), 21-50 (45 male. 45 female for each dialect) 50+ (22 male. 22 female for each


(SAMUDRAVIJAYA K)


[HEMA A MURTHY]


[REKHA SHARMA]


[L. RAMAMOORTHY]

6. Approximately about 80 hours data has been collected for each of the languages.

- (a) News: 10-35 hours
- (b) Constructed text: 5-20 hours
- (c) Isolated words: 30-50 hours
- (d) Phonetically balanced sentences: 2-15 hours

Sixteen language data has been segmented at the utterance (where an utterance is paragraph, sentence, phrase or word).

The committee discussed the data collected and its usability during the forenoon of 11th Oct 2010. The following observations were made:

1. For ASR purposes databases of the following types are required:
 - (a) Read speech (preferably news corpora)
 - (b) Telephone/cellphone speech
 - (c) Disfluent conversational speech
2. As part of the first phase, database of read speech has been collected. After an analysis of the database the following conclusions were drawn:
 - (a) The data can be used for perceptual studies, prosody analysis, speaker/language recognition and analysis.
 - (b) This data can be used to explore the building of automatic speech recognition systems.
 - (c) Can aid in the development of MULTILINGUAL phonetic engine.
3. The data in the present form, can not be used directly for recognition purposes. It requires some additional annotation:
 - (a) Inter word, Intra word pauses must be marked
 - (b) Major speech disfluencies as detailed in the guideline document must be marked.
4. As a first step, the annotation (as discussed above) will be taken up for one language, namely Tamil.

5. LDCIL will target 2 months to give IIT Madras data annotated as per the REVISED guidelines.
6. Once the annotation is completed for Tamil, IIT Madras will explore the possibility of building a speech recognition system using the database.
7. Although in the proposal, the labeling was expected to be done at syllable and phoneme levels, with progress in ASR training methodology, the committee recommends that utterance level transcriptions (sequence of words and disfluencies that make up the utterance) are made available.

Action Points:

1. Given the proliferation of cellphones across the countryside, collection of data over the cellphone must be taken up immediately. Currently the ASR consortium of TDIL is building a system for “Agricultural commodity price information by voice.” The data collection can be taken up after the pilot study is completed by the ASR consortium (funded by TDIL, DIT – coordinator Umesh Srinivasan (IIT Madras)).
2. Given the progress that has been made by the TTS consortium (vocabulary independent screen reader for six Indian languages) (funded by TDIL, DIT – coordinator Hema A Murthy (IIT Madras)), data can be collected for scaling the TTS effort to other Indian languages. Additional voice data may be collected for the existing languages.
3. Details for data collection for TTS will be sent by IIT Madras and IISc, Bangalore.
4. Data for TTS will be labeled at syllable and phoneme levels.
5. A meeting to be held to finalise data collection for TTS.
6. LDCIL and “speech groups” across the country should work in synergy to ensure that data collected LDCIL is utilised for both research and development.
7. Inclusion of meta data in the waveform (NIST format) – IIT Madras will do some prototype files and send the software for converting MS wav files to .sph files.

8. It is RECOMMENDED that LDCIL participate in O-COCOSDA to held in Khatmandu, Nepal from November 23-25, 2010.

2 Changes to Annotation

The following changes were made to the guidelines. It was decided to keep the guidelines simple:

1. Nonspeech sounds need to be marked.
2. Three different types of silences need to be marked.

After a discussion, the following guidelines were arrived at:

1. Annotation of silences: Any silence shorter than 50 ms NEED NOT be marked.

short silence (possibly intraword) (sil1):	silences of length around 50-150 ms
medium silence (possibly interword) (sil2):	silences of length between 150-300 ms
long silence (possibly interphrase) (sil3) :	silences greater than 300 ms

2. Annotation of noises: Noises can be speech-like and not speech. Two different types of noises are defined: human noise and background noise.

(a) Noises (only background):

background speech (.bs): while there is a foreground speech
(dominant), there is some background speech
vocal noise (.vn) : any speech-like noise by the speaker
background noise (.bn) : any background noise that is present


(b) Background noise along with speech: [.bn-bhaarat mahaan]


(c) Background noise along with speech and silence: [.bn-bhaarat silx mahaan] (any of sil1, sil2, sil3.

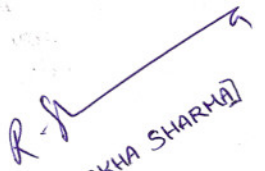
3. Annotation of speech disfluency: Only restarts/false starts need to be marked.


For example the speaker intends to speak "bengaluru" but speaks be bengaluru. Then mark this as be-bengaluru.

4. Utterances should be no longer than 15secs. So the annotator should find a long silence around 500 ms and split the sentence appropriately.


(SAMUDRAVISAYA K)


[AENA R MORTHY]


[REKHA SHARMA]


[L. RAMAMORTHY]