Data collection guidelines for ASR

The objective of this effort is to collect speech data in different Indian languages to build high quality speech recognition systems that can be used in various applications. The data collected **should be conversational in nature**:

a) Extempore speech (monolingual) – about 5 mins/speaker. About 1500 speakers
   ii) Age group 18-60
   iii) Number of males and females should be balanced.
   iv) Different domains – weather, different types of news, entertainment, health, agriculture, education, jobs.
   v) BPO type of data – this can be a mockup or it could be actual data (if we can get this from any agencies) – also monolingual (at best a few words in English) – Aadhar, passport, ration card, services – banking, repairs.
   vi) Collection from different major dialects – **at least about 4 major dialects** for every language.
   vii) **About 300 hours (a mix of standard voice calls (narrowband) and voice over IP).**

b) 2-4 people chatting about a topic of interest (for example, sports, news, weather, entertainment, politics, business, everyday problems like public transport, e-governance, government schemes like aadhar, dhan yojana, ayush bharat etc)... They should prepare a little before the conversation. Many times, in the whole conversation they keep saying: “you tell”, “what’s up”, “okay” “what shall i say:.. And no real conversation. So please ask them to have some points in their mind before the start of conversation.
   i) At least **10 different domains** must be chosen.
   ii) Each conversation could last about 3-5mins between different speakers.
   iii) The number of males and females should be more or less equal in multiple sessions.
   iv) 200 hours of data must be collected for each language.
   v) About a total of **2500-3000 conversations per language** must be collected with about **1000-1500** speakers.
   vi) The conversations can be made across various smartphones using voice(narrow band) and data calls (wideband).
   vii) In addition – the data should be collected from various places – a railway station, an airport, driving in a car/vehicle – low priority
   viii) The conversations should also vary over networks 2G-4G, 5G? – apparently the coding schemes are different for all the networks. You can collect this from different

handsets – what is meant is the following: sometimes different varieties of Gs are available in the bowels of India. We need to account for this, that is all – low priority

ix) **About 500 conversations from f)** above should be multilingual – both code mixed and code switched – prefer if it is **MIX OF English, and another local language**. (e.g. Tamil+English (Tanglish) or Hindi+English (Hinglish) etc)

c) Some general rules for both a) and b)
   i) The data must be transcribed verbatim, and time stamps must be accurate to about 300-500ms.
   ii) Declassification of data – the identity of the speaker must be changed.
   iii) Permissions obtained from all speakers

The number of languages is 10:
Tamil, Bengali, Marathi, Kannada, Malayalam, Gujarati, Odia, Punjabi, Kashmiri, Santali(300), Konkani (300 hours), Maithili (300 hours), Khasi (50 hours medical domain), Assamese (50 hours medical domain), Manipuri (50 hours medical domain).

@Umesh: I guess we do not need to collect Hindi and Telugu? Since both Ekstep and IIITH have collected large amounts of data?

Data Labeling guidelines:

1. Transcribe “verbatim.” Do not correct for grammatical errors.
2. Use a dictionary for spelling when in doubt.
3. Avoid abbreviations e.g Ft George – transcribe as Fort George, % as percent (प्रतीशत, சதவிதம்). Similarly for rupees, dollars, paise etc.
4. Terminate each utterance by a . | or whatever symbol is used in a particular language.
5. All numbers written in English?
6. Separate words by spaces.
7. Use a comma instead of … or – or some other notation when a speaker changes his/her thoughts midway through the conversation.
8. Non typical sounds must be put in brackets [noise]
9. If there are multiple speakers in a conversations, then:
   <Speaker 1>: verbatim of transcript of Speaker 1 in UTF8 of the corresponding language
   <Speaker 2>: verbatim of transcript of Speaker 2 in UTF8 of the corresponding language
   You may also additional tags like <Speaker x angry> <Speaker x sad> etc to

Indicate the emotion if you can identify in the speech.

10. Keep English words in English – do not transliterate to native language.
11. Spell out letter and number sequences – e.g 1983 in the particular language as spoken by the speaker.
12. If a speaker does not complete a word, and you can guess part of the word that is spoken. For example, “I was not to- “ (where perhaps the prediction is told).
13. Indicate long silences mid sentence by a comma.
14. If there are multiple non-speech sounds, and you can identify them mark them in square brackets (as indicated above) [laughter][coughing]
15. If multiple languages are used, and you do not know a language that is being spoken, then mark this. For example, if the speaker spoke in Tamil and then in Bangla:
    நான் ஊருக்கு செல்கிறேன். অমি ৰুষ্টপরি না

    Then transcribe this, if you know the script of the language,

    If you do not know the script but are definite about the language then transcribe as:

    நான் ஊருக்கு செல்கிறேன். <start time> <Bangla> <end time>

    If you do not know the language or the script then transcribe as:

    நான் ஊருக்கு செல்கிறேன். <start time> <unknown_language> <end time>

    Please do mark the time where a different language was spoken, and ended especially if you can not transcribe it.
    If the language is English, transcribe into English. Timing information is not required if you can transcribe it in the given language.
16. Wherever possible indicate when two people spoke simultaneously
    # Bye, bye # if you are able to make out the words spoken. Or alternatively just mark it as <start time> <multi_speaker> <end time> – please mark the start and end time if you are not able to transcribe.