Corpora for Speech Recognition

FDP cum Workshop on
Artificial Intelligence and Speech Technology
17th - 21th August, 2020
Indira Gandhi Delhi Technical University for Women, Delhi

SamudraVijaya K
Centre for Linguistic Science and Technology
Indian Institute of Technology Guwahati

samudravijaya@gmail.com
Outline

Resources to build ASR systems for Indian languages

1. Government of India
   a. Sponsored projects
   b. websites
2. Academic institutions, corporations
3. Organisations   ELRA   LDC   LDCIL   Open SLR
4. google translate
5. Crowdsourcing
   a. speech data acquisition
   b. annotation

samudravijaya@gmail.com
CLST, IITG
www.iitg.ac.in/samudravijaya
Spoken language resources needed for ASR, TTS

1. Audio files
2. Transcription
3. Pronunciation dictionary

metadata
Government of India agencies

- DeitY-TDIL
  - ASR Speech data: multi-speaker, words
  - Telugu
  - Tamil
  - Punjabi
  - Manipuri
  - Marathi
  - Bengali
  - Assamese

- DeitY-TDIL
  - https://www.iitm.ac.in/donlab/tts/database.ph
  - TTS speech data: 1 male + 1 female
  - 13 Indian languages

samudravijaya@gmail.com
www.iitg.ac.in/samudravijaya
CLST, IITG
http://www.newsonair.com/

samudravijaya@gmail.com
www.iitg.ac.in/samudravijaya

CLST, IITG
https://www.narendramodi.in/mann-ki-baat#0
CMU-INDIC

http://festvox.org/cmu_indic/index.html

phonetically balanced, single speaker databases with transcriptions in the language's native script

- Bengali (1),
- Gujarati (3)
- Hindi (1)
- Kannada (1)
- Marathi (2)
- Punjabi (1)
- Tamil (1)
- Telugu (3)
Interspeech 2018 Special Session: Low Resource Speech Recognition Challenge for Indian Languages

Microsoft Speech Corpus

- Telugu
- Tamil
- Gujarati

conversational and phrasal speech data + transcripts
https://msropendata.com/datasets/7230b4b1-912d-400e-be58-f84e0512985e

samudravijaya@gmail.com
www.iitg.ac.in/samudravijaya

CLST, IITG
ASR Challenge 2020

https://sites.google.com/view/asr-challenge

samudravijaya@gmail.com
www.iitg.ac.in/samudravijaya

CLST, IITG
The following data sets will be released for this challenge:

- Train set  40 hours
- Development set  5 hours
- Evaluation set  5 hours

- Lexicon
- baseline models
- results and recipes to replicate the baseline experiments will also be made available.
http://catalog.elra.info/en-us/repository/browse/ELRA-W0037/

2,627,000 words of transcribed spoken data

- Bengali
- Gujarati
- Hindi
- Punjabi
- Urdu


http://corpora.ciil.org/ldcil.htm

samudravijaya@gmail.com
www.iitg.ac.in/samudravijaya

CLST, IITG
http://catalog.elra.info/en-us/repository/browse/ELRA-S0368/

Nepali Spoken Corpus 31 hours of speech data

with phonologically transcribed and annotated texts
Gram Vaani data

http://catalog.elra.info/en-us/repository/browse/ELRA-S0405/

- 130 hours (21,000 different audio recordings)
- 4,000 unique Hindi speakers
- Bihar, Jharkhand, and Madhya Pradesh

voice-based community media platform that runs over IVR  (data connection is NOT needed)

Users can call into the system and

- listen to audio messages
- record their own message in response to messages they hear.

DOI:https://doi.org/10.1145/2909609.2909670
High quality TTS data for Bengali

High quality TTS data for Nepali.

Open-source Multi-speaker Speech Corpora for Building Gujarati, Kannada, Malayalam, Marathi, Tamil and Telugu Speech Synthesis Systems

https://www.aclweb.org/anthology/2020.lrec-1.800

Large Bengali ASR training data set

Large Nepali ASR training data set

Crowdsourced high-quality multi-speaker speech data

- Malayalam
- Tamil
- Telugu
- Kannada
- Gujarati
- Marathi

Crowd-Sourced Speech Corpora for Javanese, Sundanese, Sinhala, Nepali, and Bangladeshi Bengali

http://dx.doi.org/10.21437/SLTU.2018-11

samudravijaya@gmail.com
www.iitg.ac.in/samudravijaya
চন্দ্রলীবর কী এক কাশ্মীরী মহিলা কৈন্তলীবর কো সরপঞ্চ-জেদুন বেগম জী নে গিয়েপতার কিয়া থা।

chandraleevar ke ek kashmeeree mahila kaintrabaree ko sarapanch-zaidun begam jee ne giraphtaaar kiya tha.
https://catalog.ldc.upenn.edu/LDC2013S06

FREE samples from 20 different corpora

a. TORGO 23 hours of English speech 8 speakers with cerebral palsy or amyotrophic lateral sclerosis (dysarthric)
b. 2008 NIST Speaker Recognition Evaluation Test Set contains 942 hours of speech data.
c. CLSU kids speech: spontaneous and prompted speech from 1100 children
Speech data for a small payment

https://catalog.ldc.upenn.edu  25$  A-law transcription in UTF-8 format

1. Assamese  205hrs  https://catalog.ldc.upenn.edu/LDC2016S06
2. Bengali  215hrs  https://catalog.ldc.upenn.edu/LDC2016S08
3. Tamil  350 hrs ; 4 dialectal regions  https://catalog.ldc.upenn.edu/LDC2017S13

samudravijaya@gmail.com  CLST, IITG
www.iitg.ac.in/samudravijaya
Crowd sourcing

https://commonvoice.mozilla.org/en/languages
Annotation: crowd sourcing

https://www.iitg.ac.in/priyankoo/annotation/login.php
**ASR systems using Kaldi toolkit**

<table>
<thead>
<tr>
<th>Model name</th>
<th>Characteristics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mono</td>
<td>CI HMMs</td>
</tr>
<tr>
<td></td>
<td>13 static, 13delta-, 13delta-delta MFCCs</td>
</tr>
<tr>
<td>Tri1</td>
<td>CD HMMs</td>
</tr>
<tr>
<td></td>
<td>-do-</td>
</tr>
<tr>
<td>Tri2</td>
<td>-do-</td>
</tr>
<tr>
<td></td>
<td>13x9 MFCCs → LDA (40) → MLLT</td>
</tr>
<tr>
<td>Tri3</td>
<td>-do-</td>
</tr>
<tr>
<td></td>
<td>SAT</td>
</tr>
<tr>
<td>sGMM</td>
<td>-do-</td>
</tr>
<tr>
<td></td>
<td>subspace GMM</td>
</tr>
<tr>
<td>DNN-HMM</td>
<td>-do-</td>
</tr>
<tr>
<td></td>
<td>DNN replaces GMM</td>
</tr>
<tr>
<td>Chain</td>
<td>TDNN-HMM</td>
</tr>
<tr>
<td></td>
<td>The models are trained right from the start with a sequence-level objective function lower WER faster (due to lower frame rate)</td>
</tr>
</tbody>
</table>

samudravijaya@gmail.com  CLST, IITG  www.iitg.ac.in/samudravijaya
### Spoken Language Systems using Kaldi toolkit

<table>
<thead>
<tr>
<th>Application</th>
<th>Model/feature</th>
</tr>
</thead>
<tbody>
<tr>
<td>➢ ASR</td>
<td>DNN-HMM</td>
</tr>
<tr>
<td>➢ SAD</td>
<td>TDNN</td>
</tr>
<tr>
<td>➢ SID</td>
<td>x-vector, i-vector</td>
</tr>
<tr>
<td>➢ DIAR</td>
<td></td>
</tr>
<tr>
<td>➢ LID</td>
<td></td>
</tr>
</tbody>
</table>

Linguistic Resources in Indian languages

- in directly usable format
- some effort is needed to download from sources
  - manual or via a script
  - process speech (remove music)
  - process text (remove headers)
- some effort to acquire speech or transcription
  - crowdsourcing
  - incentive based