Language Technologies Research Center (LTRC)

IIIT, Hyderabad

LTRC - Summary

- Largest academic center of speech and language technology in India.
- Addresses the complex problem of
  - Natural language understanding
  - Natural language generation
  - Both speech and text modes.

- Research in LTRC on
  - basic and applied aspects of language technologies.

- LTRC has four labs
  - **NLP-MT Lab**
  - **Anusaaraka Lab**
  - **Information Retrieval and Extraction Lab (iREL)**
  - **Speech lab**

MT-NLP Lab

Major Research Areas

- Machine Translation
  - Statistical Machine Translation
  - Neural Machine Translation
  - Speech to speech MT
- Discourse Processing
- Dialog Agents
- Question Answering
- Natural Language Generation
- Linguistic Resources for ML

Research Areas ... Some details

1. Computational Grammatical Model
   - Treebanks [linguistically annotated corpus] for Hindi and Urdu
   - Computational Paninian Grammar framework
2. Parsing [processing sentences]
   - Constraint based parsers for Indian languages
   - Data-driven parsers for Indian languages
   - Shallow parsers, Part-of-Speech taggers, Morphological analyzers
3. Machine Translation
   - Transfer based approaches
   - Automatic learning of transfer rules
   - Statistical machine translation (English to Indian languages)
4. Semantics
   - Purpose-net
   - Unsupervised/semi-supervised word category disambiguation
5. Dialogue and Discourse analysis
   - Anaphora resolution in text [finding the referents of pronouns]
   - Building NLIDB systems

Major Projects

SWAYAM courses translation (80+ courses) funded by MoE

SSMT funded by PSA. ASR-MT-TTS pilot project

IL-IL MT funded by MEITY. Pilot on Hindi-Telugu and English-Telugu

Multilingual Document Summarization funded by DRDO.

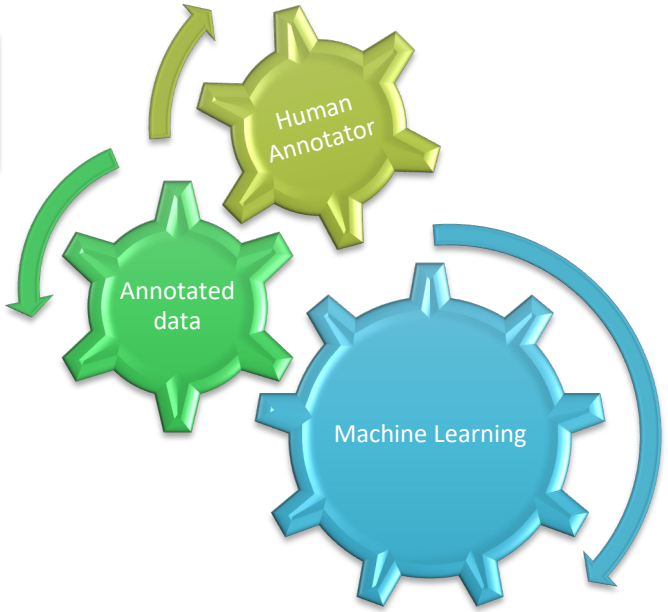
English-IL, IL-IL MT funded by MEITY - for 11 language pairs.

English-IL funded by MEITY, consortium project lead by CDAC, Pune

Discourse MT funded by MEITY, consortium project lead by AUKBC

Linguistic Resource Generation

- **Annotated Data**
  - Human Expertise required
  - Prerequisite for ML
- **Machine Learning Algorithms**
  - Cannot work without data
  - Needs noise free "gold" data
- **Human Annotator**
  - Generate annotation corpus
  - Bootstrap using ML



Resource Properties

- Multimodal
- Multilingual
- Comprehensive

Data sources

- Domain Specific
  - Judicial/Legal
  - Technical/Medical
  - Tourism
- Task Specific
  - Goal Oriented Dialogs
  - Information Extraction

Annotation Tasks

- Identify structure
  - Linguistic/ Phonetic
  - Objects
- Knowledge Representation

Corpus → Annotation → Validation → Machine Learning → Bootstrap

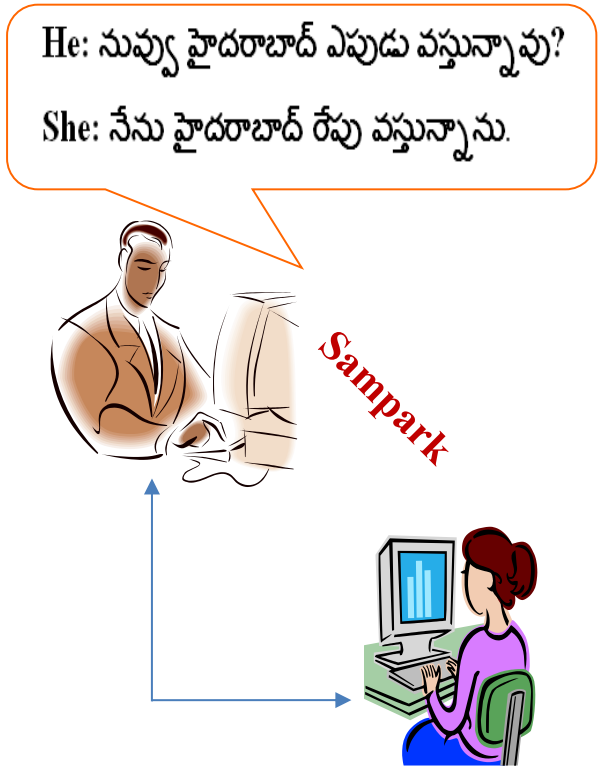
Machine Translation

Goal:

- Sampark translates automatically from one Indian language to another.

Features:

- Hybrid system which uses
  - ❖ Rule Based techniques
  - ❖ Statistical techniques
- Combines linguistics with machine learning
- Modular : A sub system for each stage.
- Pipeline architecture suitable for rapid development, exploration and teaching.
- Robust : Handles failure in levels of analysis. Always produce output.



He: तुम हैदराबाद कब आ रही हो?
She: मैं हैदराबाद कल आ रही हूँ।

http://sampark.org.in

Currently 4 systems out of 18 language pairs are online and 14 more systems to follow.

- ❖ Hindi → Punjabi
- ❖ Punjabi → Hindi
- ❖ Urdu → Hindi
- ❖ Telugu → Tamil

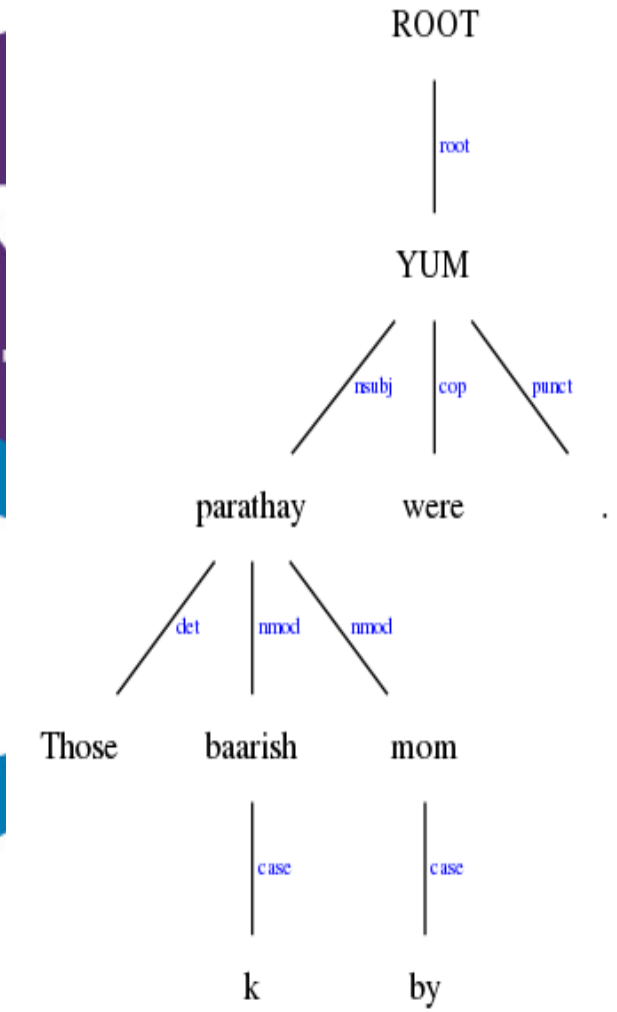
How it works?

- Analysis, Transfer, and Generation Paradigm

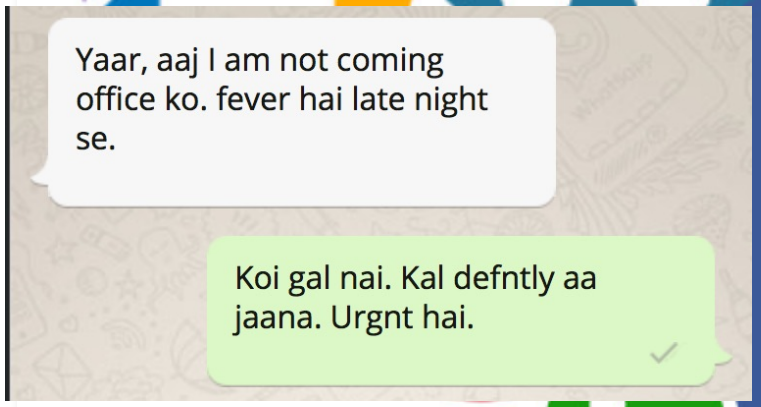
Mixed Language Analysis

- Re-engineering NLP for Mixed Language input
- Social Media mein Ubiquitous
- Next Generation of communication interfaces

"Those baarish k parathay by mom were YUM ."



Mixed Language Tools



- Language Identifier
- Parser

NLP Tasks

- Language Understanding
- Question Answering
- Automatic Translation

Conversational Agents – Dialog Systems

Paninian Applied Grammar (PAGrammar)

There are structural and lexical similarities among languages

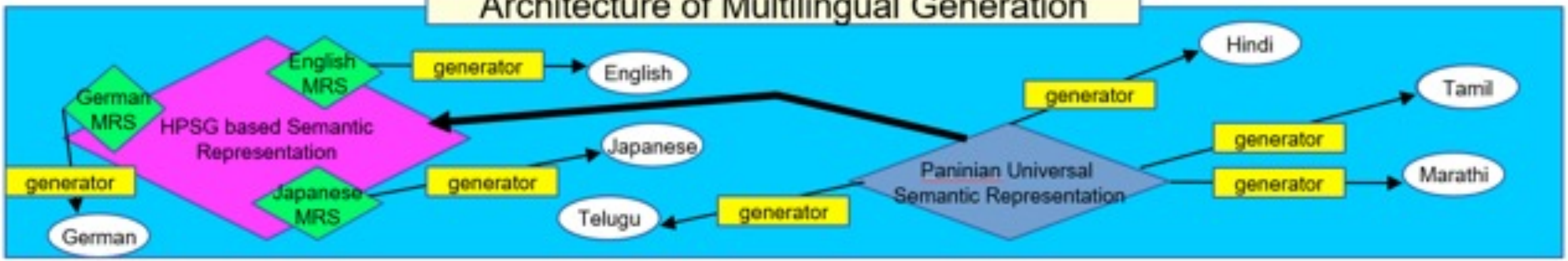
GOAL
Apply vyakarana shaastras and Generate any natural langauge With the help of technology

Shastra and technology

IMPACT
Enable Common man to comprehend any language and communicate to any language speaker using their own language With the help of technology.

Akshar Bharati

Architecture of Multilingual Generation



Computational Humor

Radhika Mamidi

Understanding humour

Event: Christmas
Illocutionary force: Greeting
Language: Bilingual
Punning technique: Compounding of two words X-mas and maska 'butter'.
Resolution: *Yule* referring to Christmas and 'You'll' as in *You'll love it*. 'It' referring to butter.

Humour Generation

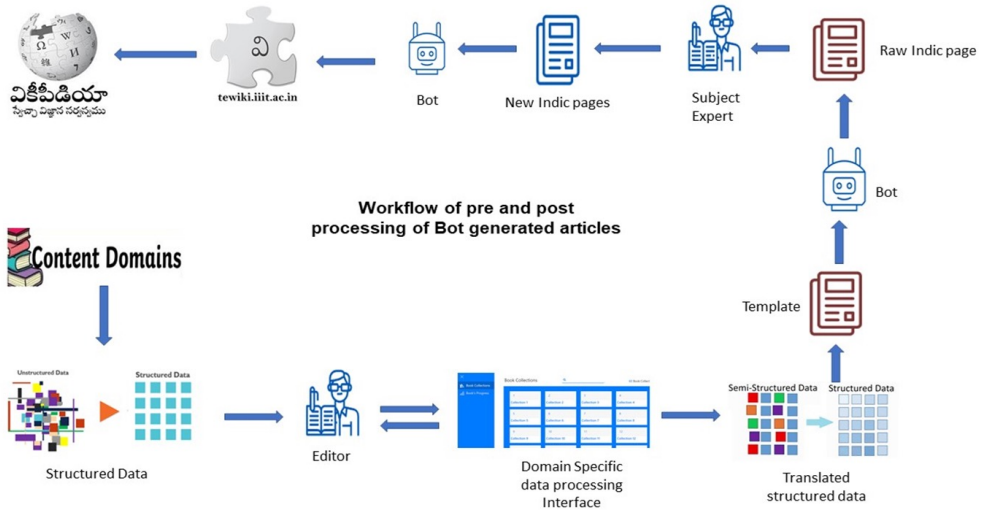
Two examples of jokes generated by our system are:

1. *Dur se dekha to Obama tha, dur se dekha to Obama tha, paas jaake dekha to pajama tha.*
2. *Dur se dekha to Mussadi tha, dur se dekha to Mussadi tha, paas jaake dekha to fissadi nikla.*

Information Retrieval and Extraction Lab (iREL)

Enriching Wikipedia in Telugu and other Indian Languages

- Innovative and ***proven*** framework for creating encyclopedic content
  - Four-pronged strategy - **Technology, Resources, Content development, Community development**.
- Partnerships - Wikimedia Foundation, TS government, and other partners
- Created more than 10,00,000 articles
- Community Creation
  - >2000 volunteers are trained
- Resource Creation
- Tools for Productivity enhancement



Auto Generated Content	Iterative Content development	Validation and Corrections	Upload to sandbox
Domain attributes Structure Identification Resource/KB/DB identification Scrapper finetuning Template identification Translations and transliterations	Categories Info boxes Pictures References Interwiki links Additional content in various sections	Corrections and other editing activity Micro Content development platform Gamification	Public/community review Integration with global Wikipedia by the community

Tools for Computational Journalism

- Discover patterns in a users' reading behaviour for news personalisation
- Leverage news content for better recommendations
- Automatically identify unconventional news (Fake News, Bizarre News)
- Summarise news in multiple languages
- Translate news from one language to another for quicker spread

Woolly memory? Sheep can recognise celebrities

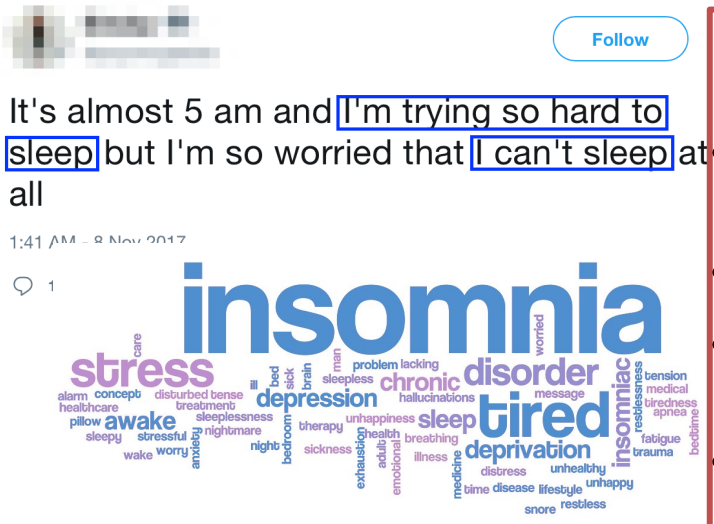
CODS-COMAD 2018, ICDM 2017

Information Retrieval and Extraction Lab

Applications

- Automatically cater to users needs using implicit information: dynamically adjusts to users changing behaviour
- Recommendation approach which is language agnostic: useful for multi-lingual news aggregators
- Prevent spread of fake news: provide users with authentic content
- Faster spread of important events across the globe

Mining Deep Domain Specific Insights from Social media: Healthcare, Finance, …



- Discover information not available in structured knowledge sources of Organisations
- Mining insights from social media conversations
- Sub-problems: Concept Normalisation, Salient Named Entity Recognition, Entity linking, …
- In collaboration with TCS, IISc, Novartis, Predera, Microsoft

What are the unknown benefits of a drug?
What are the unknown side effects of a drug?
Which community is talking more about a specific disease?
What are the most common advice on a specific condition?

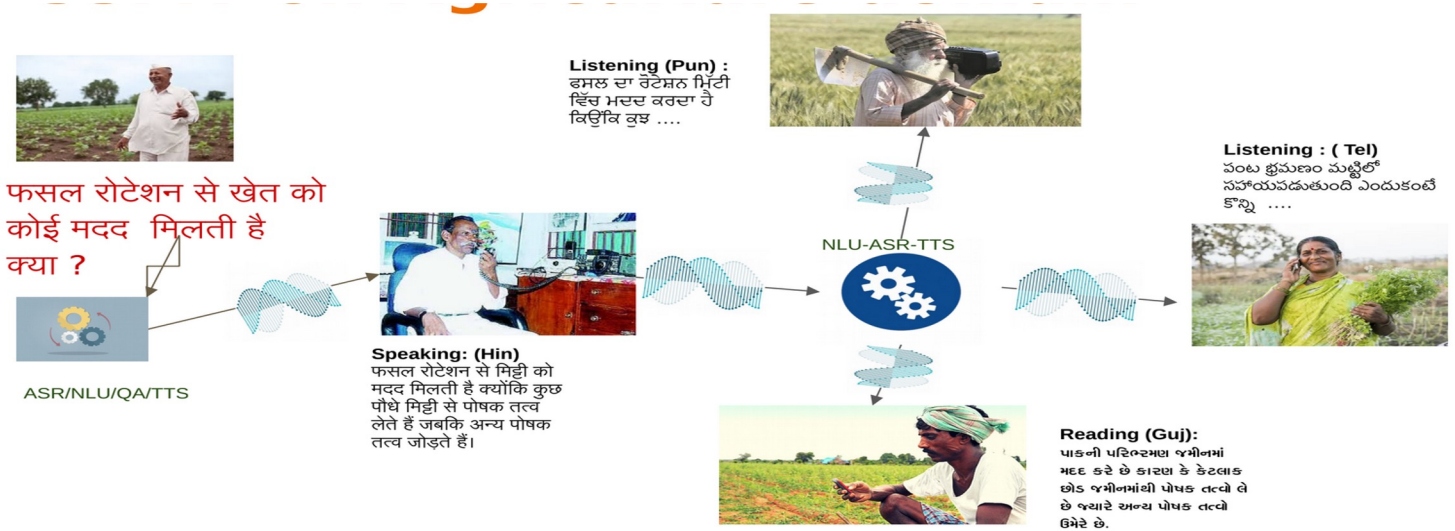
Speech Processing Lab

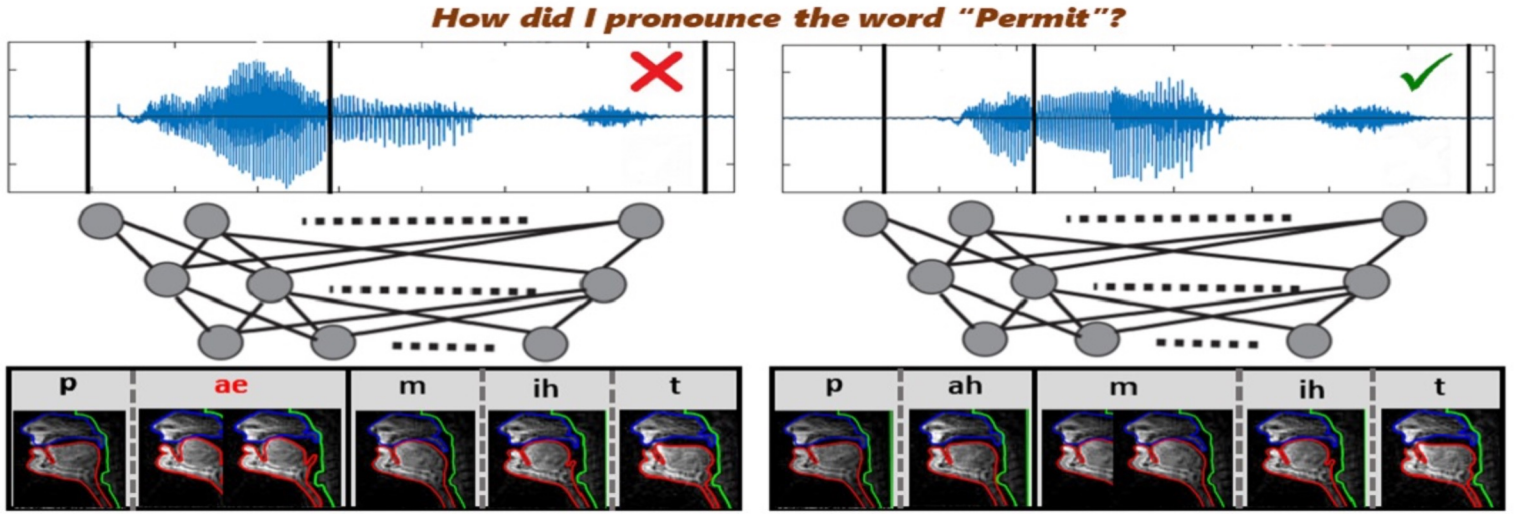
Speech Processing Lab

Research Areas:

1. Computer Assisted Language Tutoring
2. Accented Speech Recognition
3. Speaker Recognition (spoofing)
4. Language Identification
5. Pathological Speech analysis
6. Speech to speech translation in Indian context
7. Multilingual speech recognition in Indian context
8. Spoken Language Forensics and Informatics
9. Spoken Language Understanding in Indian context
10. Representation Learning for Speech

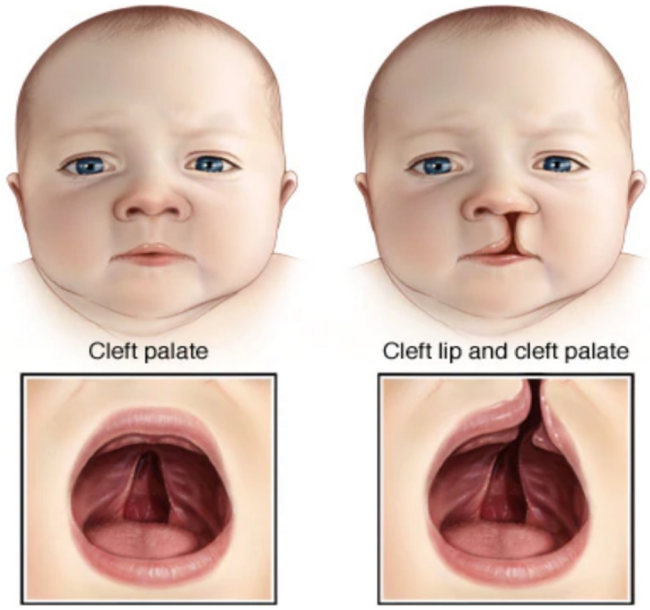
https://www.youtube.com/watch?v=CwQnaljWjvE&t=134s



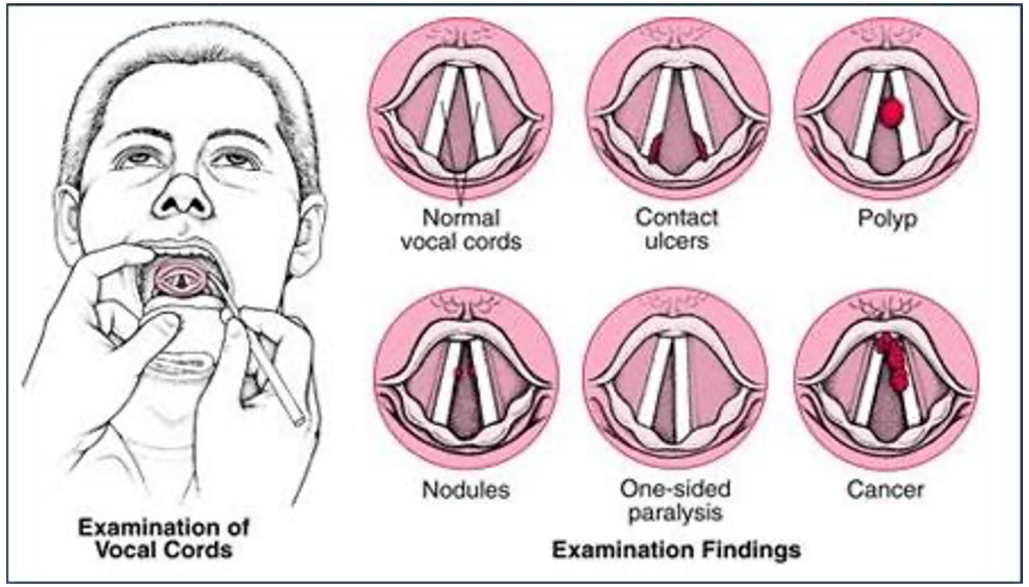


Pathological speech processing

Spastic dysarthria



Resonance disorders



Voice disorders

Stuttering

Emotions / Depression / Mood swings

- Speech disorders are detected and analysed by Speech & Language Pathologists (SLPs).
- What is the role of speech researcher here?
- Can we replace a SLP by a speech system/technology?

LTRC - Summary

- Largest academic center of speech and language technology in India.
- Addresses the complex problem of
  - Natural language understanding
  - Natural language generation
  - Both speech and text modes.

- Research in LTRC on
  - basic and applied aspects of language technologies.

- LTRC has four labs
  - **NLP-MT Lab**
  - **Anusaaraka Lab**
  - **Information Retrieval and Extraction Lab (iREL)**
  - **Speech lab**

Thank you