Comparison of Speech Features for Connected Number Speech Recognition in Indian Vernacular Languages

Mayurakshi Mukherji [1], Shreyas Kulkarni [1],

Vivek Kumar [1], Senthil Raja G [1],

Thiruvengadam Samon [1], Kingshuk Banerjee [1], Yuichi Nonaka [1]

[1] *Research and Development Centre, Hitachi India Private Limited*
Contact Email: senthil.raja@hitachi.co.in

Presenter Introduction:

Shreyas Kulkarni

shreyas.kulkarni@hitachi.co.in

Qualification:

- MTech, Systems and Control, IIT Bombay, India (2019)
- Research Engineer at Hitachi India Pvt Limited

Research Areas of Interest:

- Deep Learning for Speech Recognition
- Control for Robotic Systems
- Design and Control for Electrical Drives

Agenda:

1. Connected Number Speech Recognition
2. Speech Data Collection
3. Automatic Speech Recognition
4. Speech Feature Extraction
5. Acoustic Model Training
6. Results
7. Discussion

1. Connected Number Speech Recognition

The objective is to democratize access to financial services using voice-aided applications.

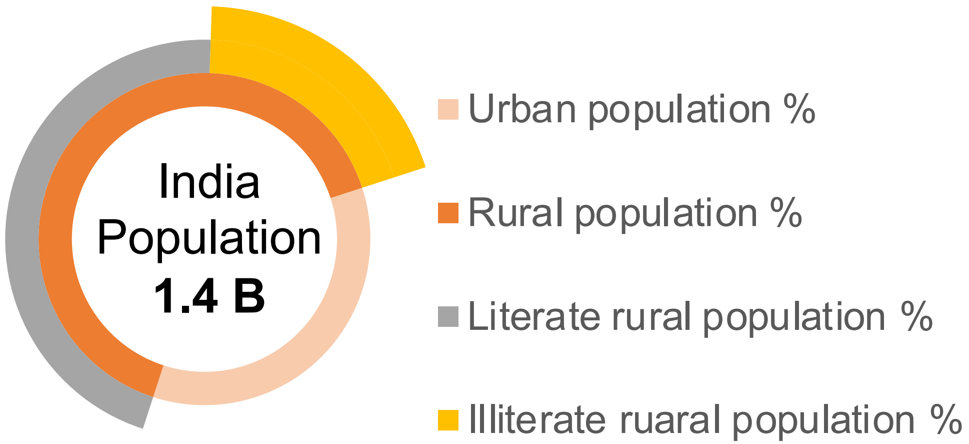


Fig. India's Populational and Literacy distribution

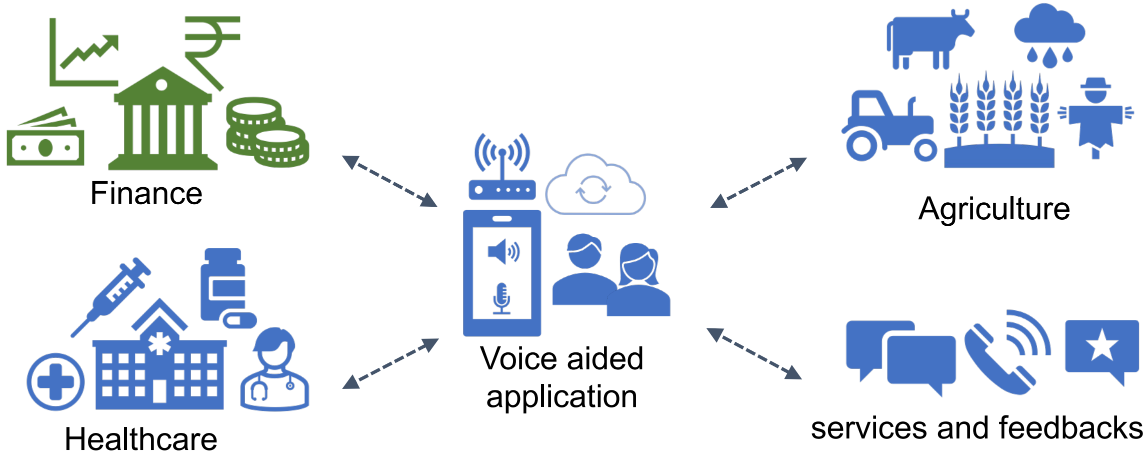


Fig. Use-cases of voice aided applications

Hitachi ASR Platform

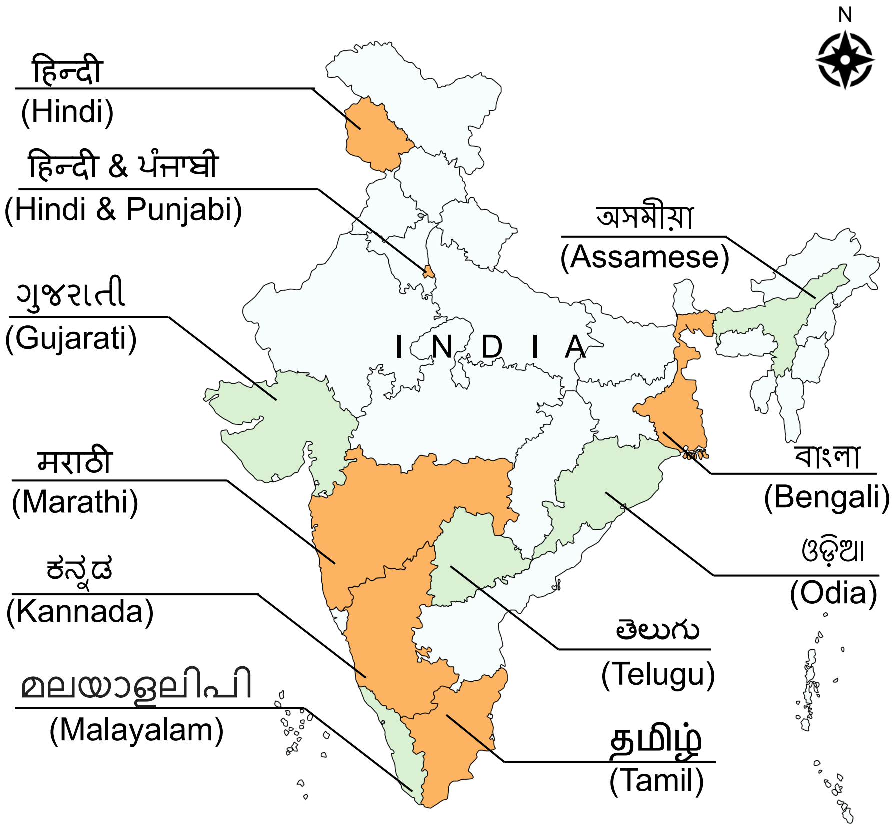
- Build domain specific and highly targeted applications with higher accuracy than general purpose Automatic Speech Recognition (ASR) systems.
- Support multiple Indian vernacular languages hence it can help bring rural empowerment.

Research activity:

- Compare and analyze the performance of ASR models trained with different speech features for Connected Number Recognition across Indian vernacular languages

2. Speech Data Collection- Connected Numbers

Speech data with diverse demography is collected for building regionally inclusive ASR systems.



🟧 Speech data is collected, and ASR models are trained

🟩 Speech data is collected

Fig. Map of India with native languages under investigation

Methodology:

Face to Face survey through purposive sampling

Target Respondent Profile:

- Gender: Both male and female
- Age group: 18 to 50 years
- Must speak native language
- Must be willing to allow to recording their voice

Sample size:

- Number of people = 1000 per language
- City (30%) and Rural (70%)
- Number of utterances per person ~ 50

Languages:

Bengali, Hindi, Tamil, Marathi, Kannada, Malayalam, Telugu, Odia, Assamese, Punjabi, Gujrati.

Speech Data: 🔊

Randomly generated Connected numbers between 0 to 100,000 in native language Ex. Hindi: “एक हज़ार चार सौ तीस” (One Thousand Four hundred thirty)

3. Automatic Speech Recognition- Workflow

ASR workflow makes use of ILSL 2.0 based transliteration scheme which is critical for model comparative study.

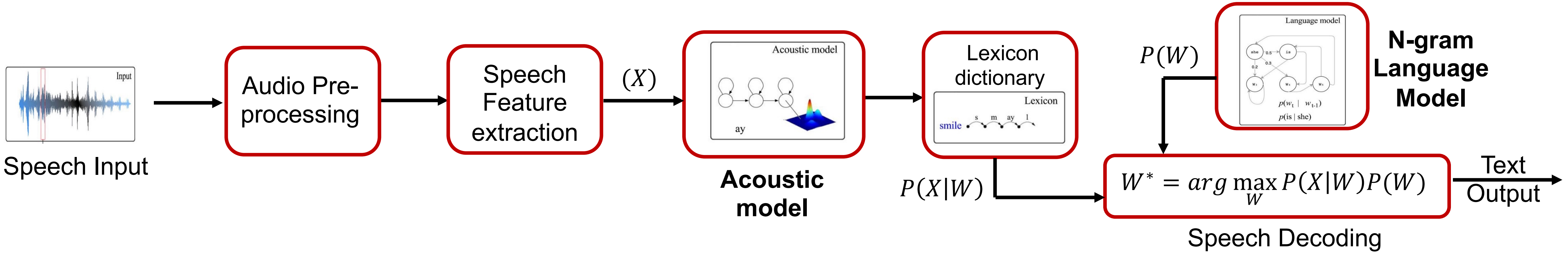


Fig. Conventional Speech Recognition Model Training Pipeline

- To compare the performances of speech recognition models in different languages, all languages' phonemes should be represented on identical platform. based on which the pronunciation dictionaries can be built.

- Generally, IPA standard is used but we used **ILSL 2.0,** a grapheme to phoneme map, specifically targeted towards creating phoneme dictionaries by introducing common representation for graphemes across multiple Indian languages [3].

Table. Sample of ILSL 2.0 standards across multiple Indian Languages [3]

ILSL	Bengali	Hindi	Marathi	Kannada	Tamil
aa	আ	आ	आ	ಆ	ஆ
kh	থ	ख	ख	க	க
dx	ড	ड	ड	ಡ	டv
y	য়	य	य	ಯ	ய
e	এ	ए	ए	ಎ	எ

4. Speech Feature Extraction

Four ASR models (per language) are trained using four combinations of speech features: MFCC, MFCC+Pitch, PLP, and PLP+Pitch

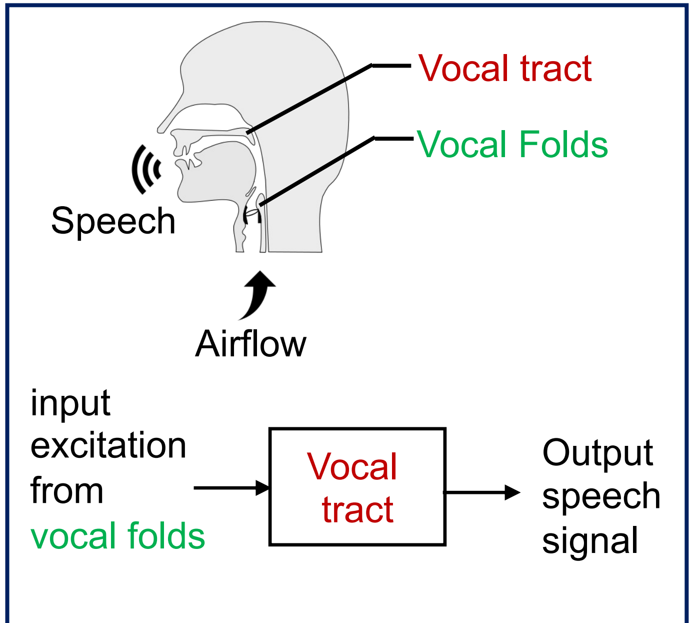
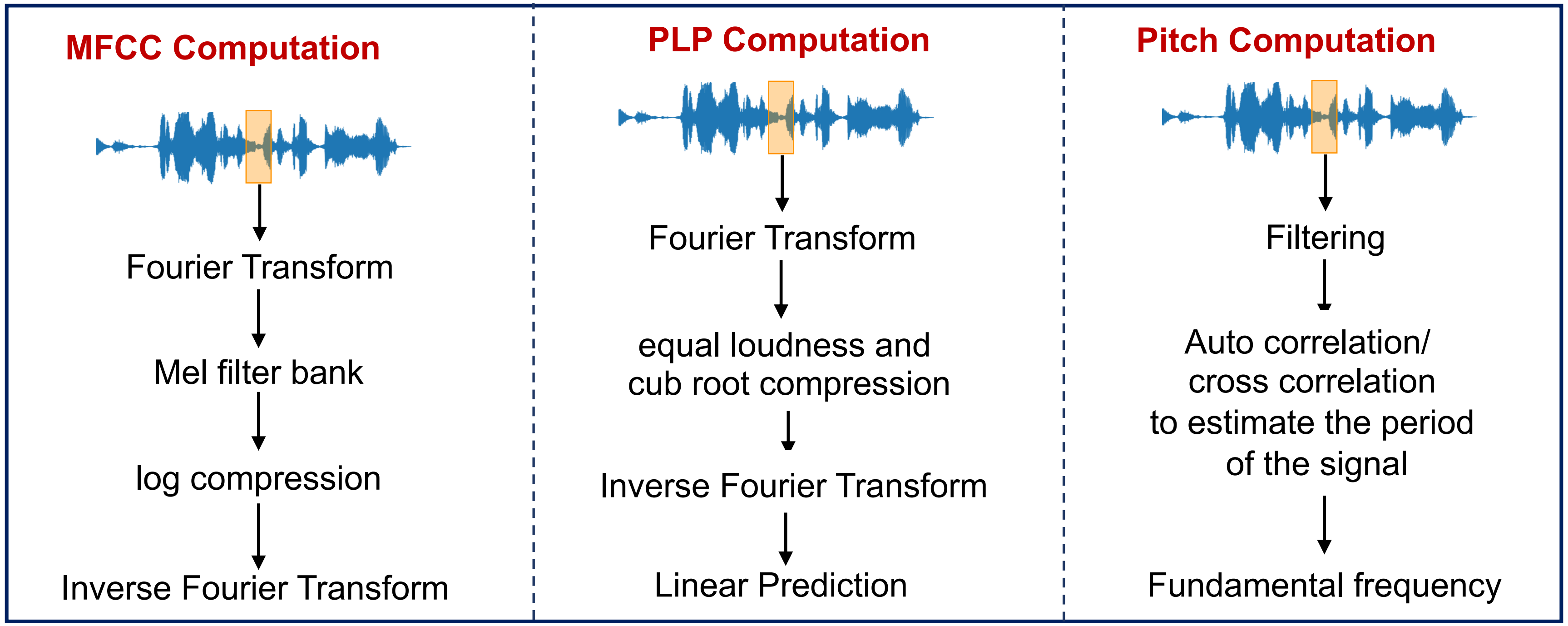


Fig. Human speech generation schematic



MFCC Computation

Fourier Transform → Mel filter bank → log compression → Inverse Fourier Transform

PLP Computation

Fourier Transform → equal loudness and cub root compression → Inverse Fourier Transform → Linear Prediction

Pitch Computation

Filtering → Auto correlation/ cross correlation to estimate the period of the signal → Fundamental frequency

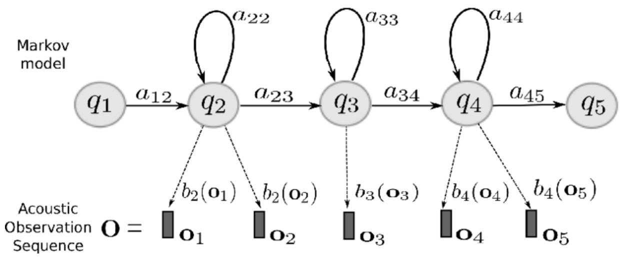
- To built automatic speech recognition systems, we should model vocal tract corresponding to speech utterances. The speech utterances can be broken up into fundamental sounds in a language.
- The MFCC and PLP features capture the significant frequencies (formant frequencies), corresponding to fundamental sounds in a language. The pitch features are important for tonal languages like Mandarin.

5. Acoustic Model Training

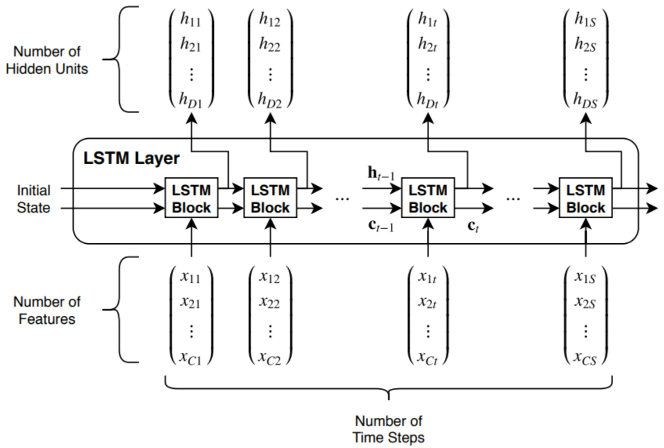
Identical training conditions are maintained for performing all the model trainings.

Audio frames
(25 ms, 10 ms overlap)

GMM-HMM training



TDNN-LSTM training

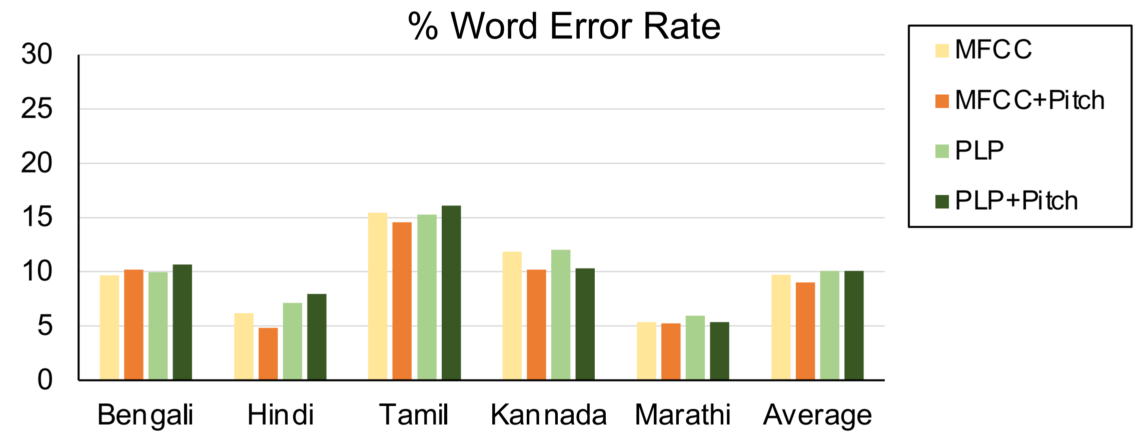

Number of Hidden Units
h_{11} h_{21} ⋮ h_{D1}
h_{12} h_{22} ⋮ h_{D2}
h_{1t} h_{2t} ⋮ h_{Dt}
h_{1S} h_{2S} ⋮ h_{DS}
LSTM Layer
Initial State
LSTM Block
LSTM Block
...
$\mathbf{h}_{t-1}$
$\mathbf{c}_{t-1}$
LSTM Block
$\mathbf{c}_t$
...
LSTM Block
Final State
Number of Features
x_{11} x_{21} ⋮ x_{C1}
x_{12} x_{22} ⋮ x_{C2}
x_{1t} x_{2t} ⋮ x_{Ct}
x_{1S} x_{2S} ⋮ x_{CS}
Number of Time Steps

Data	Distribution
Per Language: • 80% connected numbers • 20% general sentences	train set = ~60,000 samples | ~ 55 hr | 70%
	dev set = ~15,000 samples | ~ 12 hr | 20%
	eval set = ~9,000 sample | ~ 8 hr | 10%

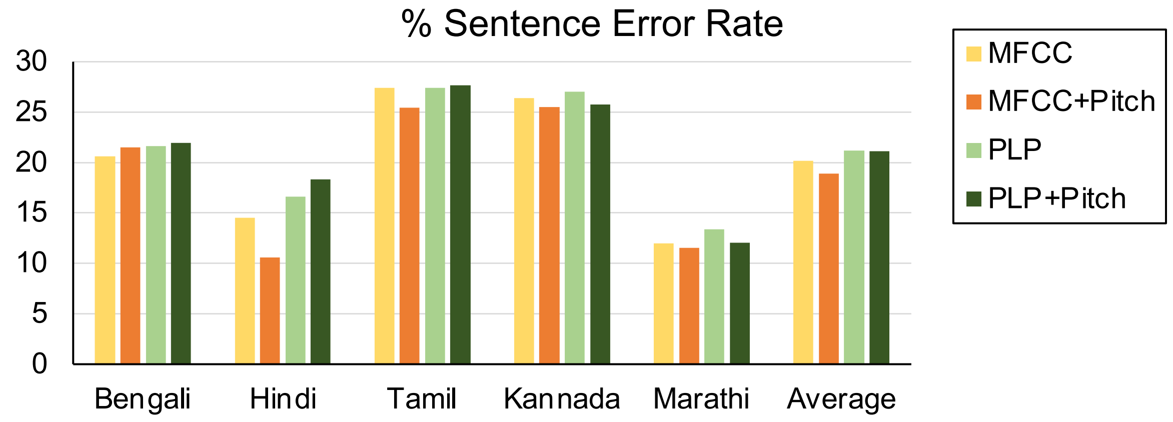
Training Stage		Model Training Parameters
GMM-HMM training	Monophone Training	num_iters = 40
	Triphone training (tri-1)	num_iters = 35
		numleaves = 2750
		totgauss = 50000
	Triphone + LDA + MLLT training (tri-2)	num_iters = 35
		numleaves = 2750
		totgauss = 50000
	Triphone + LDA + MLLT + SAT training (tri-3)	num_iters = 35
		numleaves = 2750
		totgauss = 50000
TDNN-LSTM training		epochs = 6
		hidden layers = 13
		Initial learning rate = 0.0001
		Final learning rate = 0.00001

6. Results

MFCC+Pitch feature combination has best results on average across multiple languages.



Language	% Word Error Rate			
	MFCC	MFCC+Pitch	PLP	PLP+Pitch
Bengali	9.64	10.22	9.97	10.67
Hindi	6.2	4.84	7.15	7.97
Tamil	15.46	14.56	15.29	16.08
Kannada	11.83	10.19	12.02	10.34
Marathi	5.36	5.26	5.94	5.39
Average	**9.7**	**9.01**	**10.07**	**10.09**



Language	% Sentence Error Rate			
	MFCC	MFCC+Pitch	PLP	PLP+Pitch
Bengali	20.61	21.53	21.63	21.96
Hindi	14.55	10.6	16.59	18.33
Tamil	27.42	25.42	27.4	27.65
Kannada	26.42	25.52	27.01	25.78
Marathi	12.01	11.56	13.37	12.02
Average	**20.2**	**18.93**	**21.2**	**21.12**

- Among four speech feature combinations, tested over five Indian languages, MFCC+Pitch shows the best result with a 0.68% WER improvement and 1.27% SER improvement over MFCC on average.
- MFCC+Pitch shows the best improvement in case of Hindi, where SER is reduced by 4%.

7. Discussion

This study provides Multilingual ASR tuning heuristics and language specific insights.

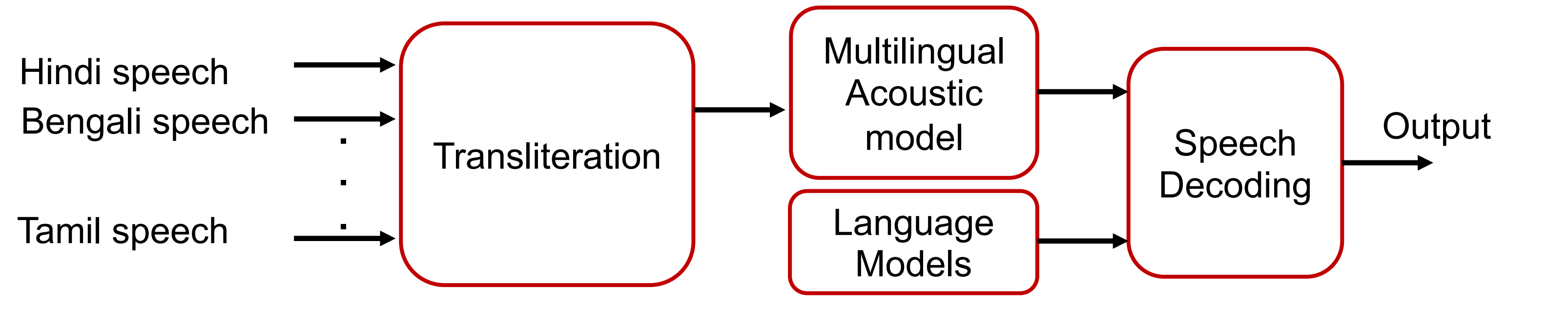


Fig. Multilingual ASR system

- The comparative results lead up to heuristics for the tuning of a multilingual ASR model to meet different recognition criteria depending on the part of the country where the model is to be deployed.
- Therefore, the model should ultimately show higher accuracy for the region-specific language, while also supporting multiple other languages.
- The Marathi and Hindi languages belong to Indo-Aryan language family, and are phonetically similar, use same script. Moreover, the ASR models show similar and relatively better performance.
- Bengali language has some elements in the grapheme-to-phoneme map which exhibit many-to-one mapping, leading to relatively poorer recognition performance.
- Tamil and Kannada languages belong to the same language family, and show similar and relatively poorer performance.

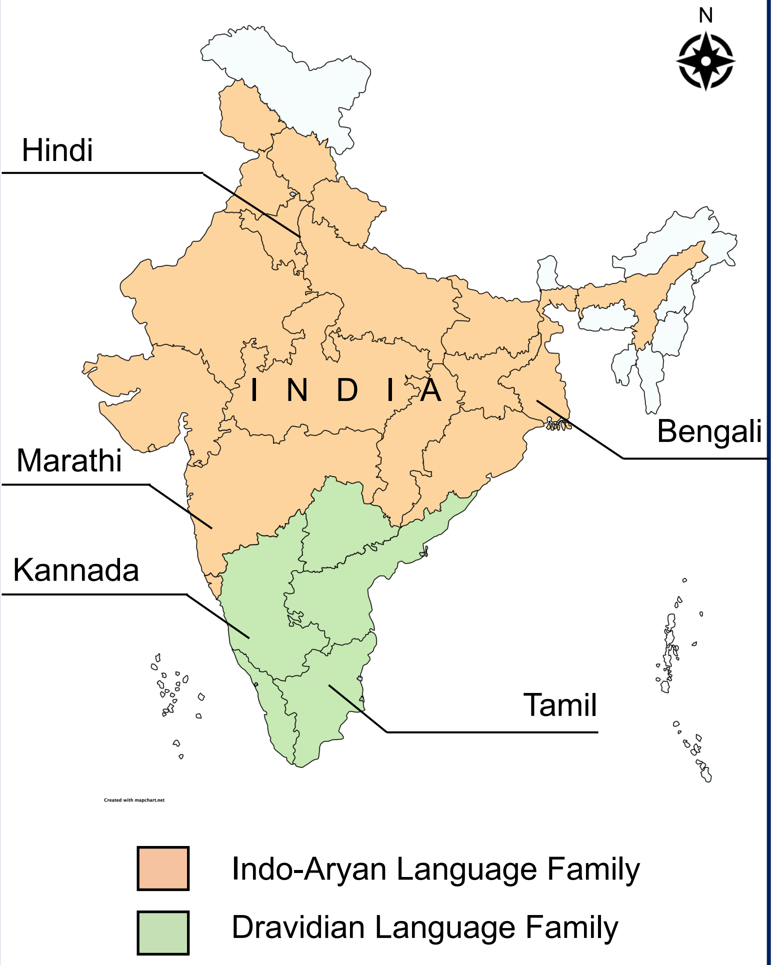

N
Hindi
I N D I A
Bengali
Marathi
Kannada
Tamil
Created with mapchart.net
🟧 Indo-Aryan Language Family
🟩 Dravidian Language Family

Fig. Map of India with dominant language families

References

[1] J. M. R. Sánchez, M. Bereau, and J. R. C. de Lara, “Feature selection for automatic speech recognition in noisy Scenarios,” In 2nd International Conference of Information Processing , vol. 40, pp. 51-71, Dec. 2019.

[2] P. Ghahremani et al., "A pitch extraction algorithm tuned for automatic speech recognition," 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 2494-2498, 2014

[3] Arun Baby, Nishanthi N.L., Anju Leela Thomas, and Hema A. Murthy, “A Unified Parser for Developing Indian Language Text to Speech Synthesizers”, Springer International Publishing Switzerland, pp. 514–521, 2016

[4] D. Namrata, “Feature extraction methods LPC, PLP and MFCC in speech recognition”, International Journal For Advance Research in Engineering And Technology (ISSN 2320-6802), vol. 1, pp. 1-6 , 2013

[5] L. Xie and Z. -q. Liu, "A comparative study of audio features for audio-to-visual conversion in Mpeg-4 compliant facial animation," 2006 International Conference on Machine Learning and Cybernetics, pp. 4359-4364, 2006, doi: 10.1109/ICMLC.2006.259085.

[6] M. Wu, D. Wang, and G. J. Brown, "A multi-pitch tracking algorithm for noisy speech," 2002 IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. I-369-I-372, 2002, doi: 10.1109/ICASSP.2002.5743731.

[7] K. Kasi and S. A. Zahorian, "Yet another algorithm for pitch tracking," 2002 IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. I-361-I-364, 2002, doi: 10.1109/ICASSP.2002.5743729.

[8] X. Lei. “Modeling lexical tones for mandarin large vocabulary continuous speech recognition.”, Ph.D. dissertation, University of Washington, 2006.

[9] F. He et al., ‘Open-source multi-speaker speech corpora for building Gujarati, Kannada, Malayalam, Marathi, Tamil and Telugu Speech synthesis systems’, Proc. The 12th Language Resources and Evaluation Conference (LREC), pp. 6494–6503, 2020.

[10] K. S. Bhogale et al., ‘Effectiveness of mining audio and text pairs from public data for improving ASR systems for low-resource languages’. arXiv, 2022.

[11] G. Chen, et al., "Data augmentation for children's speech recognition”, The "Ethiopian" System For The SLT 2021 Children Speech Recognition Challenge, arXiv preprint arXiv:2011.04547, 2020.

[12] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, ‘Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi’, Proc. Interspeech 2017, pp. 498–502, 2017.

[13] “IITM Hindi Speech Corpus: a corpus of native Hindi Speech Corpus” – Speech signal processing lab, IIT Madras, 2020.

Thank you !

धन्यवाद | நன்றி | ধন্যবাদ | ಧನ್ಯವಾದ | धन्यवाद