NISP: A Multilingual Multi-accent Dataset for Speaker Profiling

Authors

Shareef Babu Kalluri[1], Deepu Vijayasenan[1], Sriram Ganapathy[2], Ragesh Rajan M[1], Prashant Krishnan[2]

{shareefbabu1, deepu.senan, sriram.iisc, mrageshrajan, gillyprash29}@gmail.com

[1] *National Institute of Technology Karnataka, Surathkal, India,*
[2] *Learning and Extraction of Acoustic Patterns (LEAP) lab, Indian Institute of Science, Bangalore, India*

ICASSP-21, 6-11 June 2021 ● Toronto, Ontario, Canada

Outline

Motivation

- Many of the available datasets have partial information for speaker profiling applications.
- Datasets are limited to monolingual – Indian languages.
- Estimating the physical parameters like height and age of a speaker helps in applications like forensics and commercial scenarios.
  - Eg. In voice surveillance applications, predicting the speaker meta data from the short chunks of speech data is crucial for biometric evidence generation.

Contribution

- A new dataset (NISP) has created which has speech data from five (Hindi, Kannada, Malayalam, Tamil, Telugu) different Indian languages along with English.
- The metadata information for speaker profiling applications like
  1. Linguistic information – L1, L2
  2. Regional information – geographic location of the native place
  3. Physical characteristics of a speaker – Height, age, Shoulder size, Weight.
- This dataset is publicly made available in the following address, *https://github.com/iiscleap/NISP-Dataset*

Recording Protocol

- The speech data was collected – high quality microphone (with Scarlett solo studio, CM25 a large diaphragm condenser microphone).
- Sampling rate 44.1 kHz with a bit-rate 16 bits per sample.
- Audio recording setup – "*Speech Recorder*" [a] and with *Focusrite Scarlett solo studio* audio recording device by connecting it to a laptop.

[a] This software is available in this address, *https://www.bas.uni-muenchen.de/forschung/Bas/software/speechrecorder/*

Speech Data

- The text data used in the reading task – L1 language as well as English.
- The text provided to speakers – daily news articles
  1. Unique sentences without any contextual continuity.
     - This setting was made to avoid any prosodic continuity in the reading task.
  2. Continuous short story section to have contextual continuity.
  3. Common sentences – English (2 – TIMIT *sa1* and *sa2* sentences and 3 general news article sentences) ; L1 – 2 common sentences.
- Overall, each subject provided with
  1. 20-25 – unique sentences in L1 and English
  2. 20-25 – contextual sentences in L1 and English,
  3. 5 common sentences for English, and 2 sentences from L1.

Potential Applications

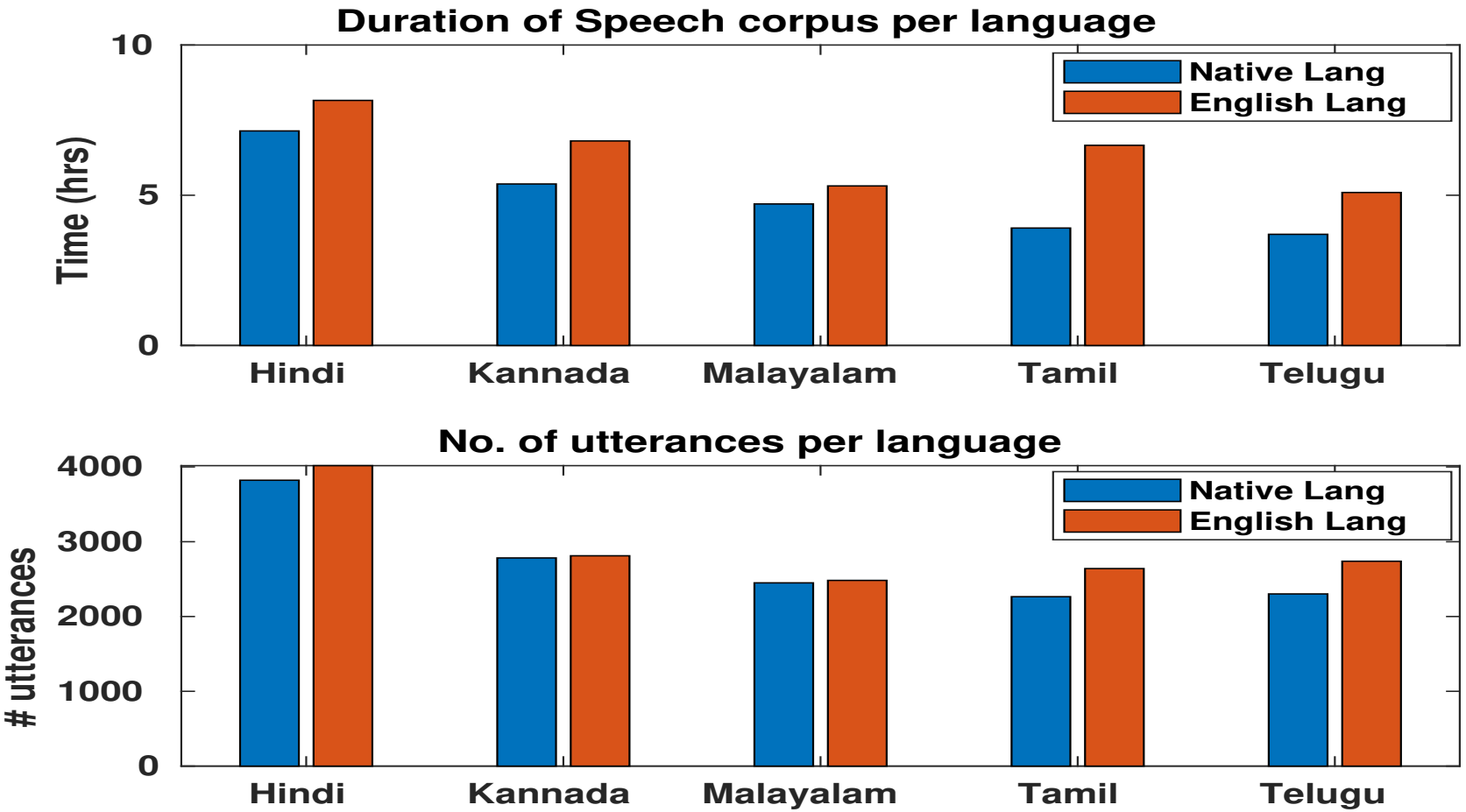
- Physical Parameter Estimation
- Accent & Language Identification
- Speaker Recognition
- Speech Recognition

Distribution of speakers per language

Table 1: Distribution of native languages', and the number of male and female speakers in the NISP dataset

Sl.No.	Native Language	Male	Female	Total
1.	Hindi	76	27	103
2.	Kannada	33	27	60
3.	Malayalam	35	25	60
4.	Telugu	35	22	57
5.	Tamil	40	25	65
Total Speakers		219	126	345

Distribution of duration of speech data & No. of Utterances per language



Distribution of speakers per region

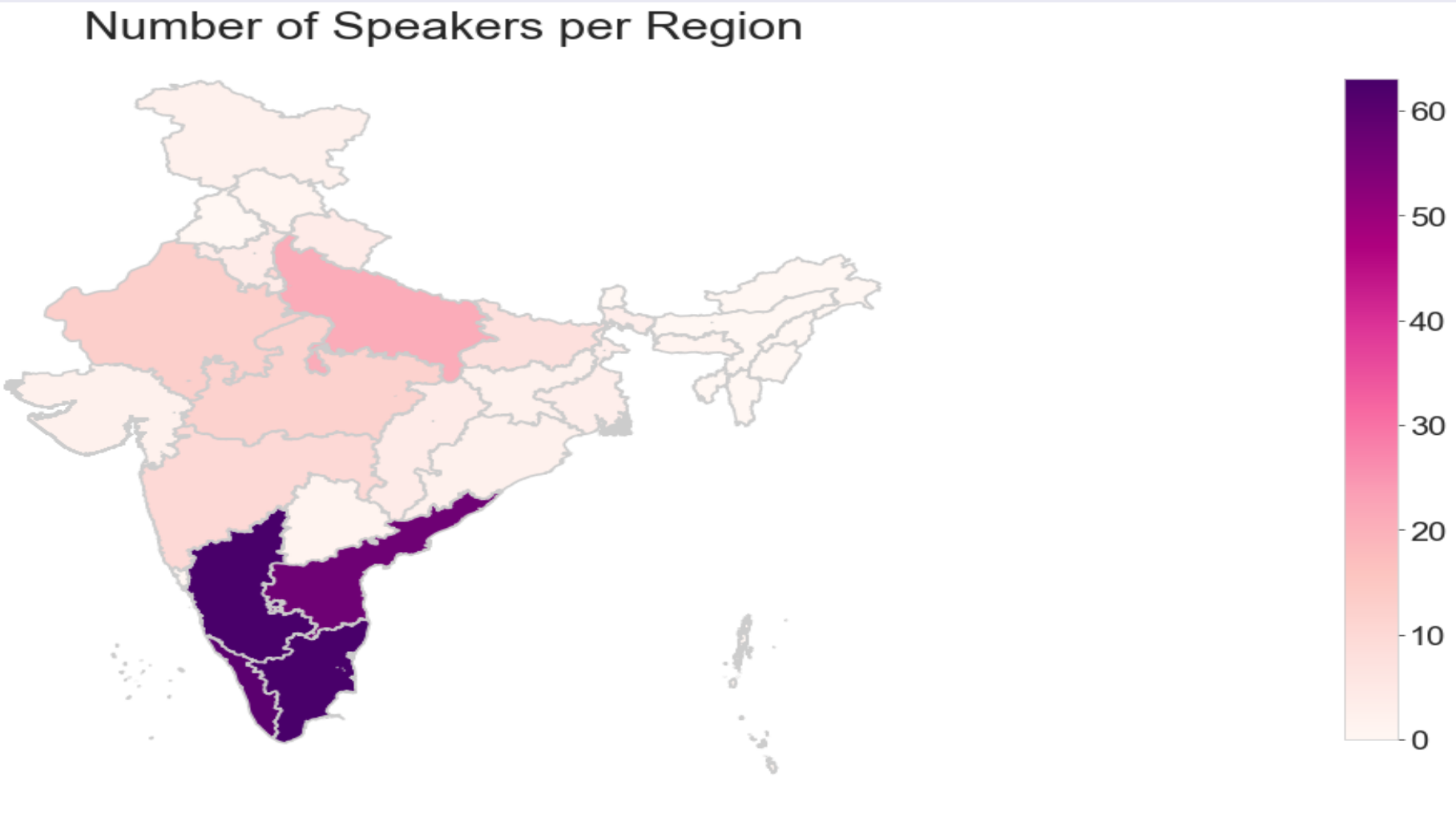


Table 2: Gender wise statistics of each physical parameter in the NISP dataset

Physical Characteristic	**Min**	**Max**	**Mean**	**Standard Deviation**
Male Speakers				
Height (*cm*)	151.0	191.0	171.6	6.7
Shoulder width (*cm*)	32.0	55.0	44.7	3.2
Weight (*kg*)	43.4	116.5	69.4	11.9
Age (*y*)	18.0	47.5	24.4	5.6
Female Speakers				
Height (*cm*)	143.0	180.0	158.9	6.8
Shoulder width (*cm*)	30.0	53.0	39.7	3.4
Weight (*kg*)	34.1	86.2	56.5	10.5
Age (*y*)	18.3	46.5	25.1	6.1
Male and Female Speakers				
Height (*cm*)	143.0	191.0	166.9	9.1
Shoulder width (*cm*)	30.0	55.0	42.9	4.0
Weight (*kg*)	34.1	116.5	64.7	13.0
Age (*y*)	18.0	47.5	24.7	5.8

Dataset split

- The training split — 210 (134 M + 76 F) speakers with 17161 (10911 M + 6933F) utterances
- The test split – 135 (85 M + 50 F) speakers with 11107 (6933 M + 4174 F) utterances.

Error Metrics

- Mean Absolute Error (MAE).
- Target Mean Predictor (TMP) – estimating the target with training data mean without considering the speech data.

Baseline Experiment

- We perform the physical parameter estimation task using three different features namely, mel filter bank features, formants and harmonics [1].
- Computed the first order statistics (Fstat) from the 40 Mel filter bank features using a 256 component diagonal covariance Gaussian Mixture Model Universal Background Model (GMM-UBM).
- The GMM was trained – 20 MFCC $+\delta + \delta = 60$ dimensional features.
- The formant and fundamental frequency features – percentiles (5,25,50,75 and 95) are computed.
- The harmonic features including both frequency locations (F-loc) and amplitude features (Amp) – same set of percentiles are computed.
- These computed statistics from each individual feature are given to linear Support Vector Regression (SVR) model to predict each physical parameter.
- x-vectors – extended TDNN model trained on voxceleb data [2].

Baseline Results

Table 3: Comparison of three feature combination – Comb -3 (Fstats + formant + harmonic features (amplitude + frequency locations)) with default predictor and x-vector model

	Height (cm) Estimation			Weight (kg) Estimation		
	Male	Female	All	Male	Female	All
	MAE	MAE	MAE	MAE	MAE	MAE
TMP	5.22	5.30	7.14	7.74	7.88	9.08
Comb–3	**5.16**	**5.30**	**5.11**	**7.06**	**6.84**	**7.06**
x-vector	5.69	6.04	5.85	8.37	7.56	8.03
	Shoulder (cm) Estimation			Age(y) Estimation		
TMP	1.98	**2.44**	2.99	4.40	4.39	4.42
Comb–3	**1.93**	2.47	**2.11**	**3.80**	**3.55**	**3.76**
x-vector	2.25	3.15	2.61	4.01	4.94	4.39

Conclusions

- A multilingual speaker profiling dataset – recorded in five different Indian native languages (Hindi, Kannada, Malayalam, Tamil, and Telugu) along with English language.
- This dataset has linguistic information, regional information and physical characteristics of a speaker – useful in commercial and forensic applications of speaker profiling.
- Overall, this dataset has 56.86 hours (24.83 –L1, 32.03 – English) of speech data.
- For speaker profiling tasks on this dataset, the baseline results performs better in MAE measure when compared to the TMP.

References

[1] Shareef Babu Kalluri, Deepu Vijayasenan, and Sriram Ganapathy. Automatic speaker profiling from short duration speech data. *Speech Communication*, 121:16–28, 2020.

[2] David Snyder, Daniel Garcia-Romero, Gregory Sell, Alan McCree, Daniel Povey, and Sanjeev Khudanpur. Speaker recognition for multi-speaker conversations using x-vectors. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5796–5800. IEEE, 2019.

Acknowledgements

- We thank the SERB under grant no: EMR/2016/007934 for funding to create the database.
- We thank student volunteers who helped in creating this dataset.


Thank you for
Listening
Questions?