Classification Between Patients with Amyotrophic Lateral Sclerosis and Healthy Individuals Using Hypernasality in Speech: A Low Complexity Approach

Anjali Jayakumar[1], Tanuka Bhattacharjee[1], Seena Vengalil[2], Yamini Belur[2], Atchayaram Nalini[2], Keerthipriya M[2], Darshan Chikktimmegowda[2], Prasanta Kumar Ghosh[1]

[1]SPIRE LAB, EE Dept., IISc, Bangalore, India
[2]NIMHANS, Bangalore, India

NCC 2025

Overview

Amyotrophic Lateral Sclerosis (ALS)

Overview

- A progressive neurodegenerative disorder.
- Affects motor neurons, leading to muscle weakness and eventual paralysis.

Main symptoms

- Loss of mobility and physical strength.
- Difficulty with swallowing (dysphagia) and breathing.
- **Slurred speech, weak voice, and difficulty in articulation.**

Clinical Monitoring

- Electromyography (EMG), Magnetic Resonance Imaging (MRI) scanning, blood tests.
- Regular assessments of motor functions
- There is no cure, but symptom management can improve quality of life.

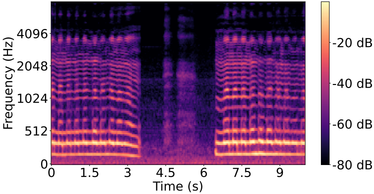
Dysarthria in ALS

Dysarthria: A motor speech disorder caused by damage to the nervous system, resulting in poor coordination of the muscles used for speech.

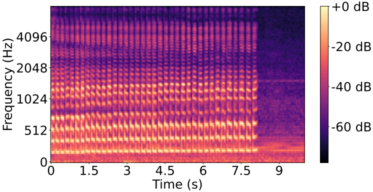
Symptoms of Dysarthria

- Slow or labored speech
- Reduced speech rate
- Difficulty with articulation of vowels
- **Increased nasalization in speech**

Increased Nasalization in ALS speech



ALS speech with increased nasalization



Healthy speech

Figure: Mel spectrogram of ALS with increased nasalization and healthy speech: rapid repetition of monosyllabic sequence ‘pa-pa-pa’

Hypernasality

SPIRE LAB

What is Hypernasality?

- Hypernasality occurs when too much air passes through the nose during speech.
- Results in a "nasal" quality of speech.

Causes of Hypernasality

- Weakness in the velopharyngeal muscles.
- Poor closure of the velopharyngeal port leads to excess nasal airflow.
- Reduced articulatory precision makes it harder to control airflow.

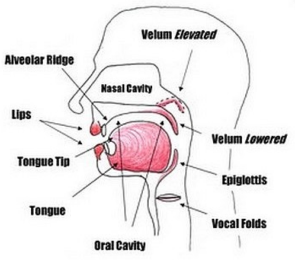


Figure: Illustration showing the velum elevating to close the nasal cavity and lowering to open it.

Image Source: Speech language resources, last accessed: January 10, 2025. [Online]. Available: https://www.speechlanguage-resources.com/speech-sound-structures.html

Motivation

- Investigate the use of speech cues to distinguish between ALS and healthy individuals leveraging **nasalization** as a key indicator.
- Explore the potential for classifying ALS speech from healthy speech by **training models solely on healthy speech data**, and using ALS data only for fine-tuning.
- Implement **low-complexity DNN models** to ensure computational efficiency while maintaining classification performance.

Literature

SPIRE LAB

Speech Based Classification and Severity Prediction of Dysarthric Speech

- **J. Mallela et al. (2020)**: CNN-BiLSTM, DNN and SVM models with MFCC features for ALS vs COT classification[1].
- **T. Bhattacharjee et al. (2023)**: DNN model with temporal statistics of MFCC for ALS dysarthria severity classification[2].
- **F. Javanmardi et al. (2024)**: HuBERT model for dysarthria severity classification[3].

[1] J. Mallela et al., "Voice based classification of patients with Amyotrophic Lateral Sclerosis, Parkinson's disease and healthy controls with CNN-LSTM using transfer learning," in ICASSP, IEEE, 2020, pp. 6784–6788.

[2] T. Bhattacharjee et al., "Transfer Learning to Aid Dysarthria Severity Classification for Patients with Amyotrophic Lateral Sclerosis," in Proc. INTERSPEECH, 2023, pp. 1543–1547.

[3] F. Javanmardi et al., "Pre-trained models for detection and severity level classification of dysarthria from speech," Speech Communication, vol. 158, p. 103047, 2024.

Literature

SPIRE LAB

Hypernasality

- **M. Saxon et al. (2019):** Proposed objective measures for hypernasality assessment using speech features from healthy and dysarthric speakers[1].
- **V. C. Mathad et al. (2020):** Investigated the correlation between nasality measures in healthy speech and speech from dysarthric speakers, emphasizing their diagnostic potential[2].
- **S. Bhattacharjee et al. (2024)**: Used HuBERT based models to identify hypernasal speech in individuals with cleft lip and palate [3].

The use of hypernasality for distinguishing ALS from COT remains unexplored.

[1] M. Saxon et al., "Objective measures of plosive nasalization in hypernasal speech," in ICASSP, IEEE, 2019, pp. 6520–6524.

[2] V. C. Mathad et al., "Deep learning based prediction of hypernasality for clinical applications," in ICASSP. IEEE, 2020, pp. 6554–6558.

[3] S. Bhattacharjee et al. "Classification of cleft lip and palate speech using fine-tuned transformer pretrained models," in Intelligent Human Computer Interaction, B. J. Choi, D. Singh, U. S. Tiwary, and W.-Y. Chung, Eds. Cham: Springer Nature Switzerland, 2024, pp. 55–61

Literature

SPIRE LAB

Low Complexity Classification

- **T. Bhattacharjee et al. (2021)**: Single-dimensional pitch is as effective as multi-dimensional MFCCs and more noise-robust for ALS and Parkinsons disease (PD) detection[1].
- **B. Akila and J. J. Vedha Nayahi (2024)**: A low-complexity approach improved PD detection by reducing feature dimensionality[2].
- **A. Jayakumar et al. (2024)**: Reducing the model complexity for ALS vs. healthy speech classification using MFCCs[3].

[1] T. Bhattacharjee et al., "Effect of noise and model complexity on detection of Amyotrophic Lateral Sclerosis and Parkinson's disease using pitch and MFCC," in ICASSP. IEEE, 2021, pp. 7313–7317.

[2] B. Akila and J. J. Vedha Nayahi, "Parkinson classification neural network with MASS algorithm for processing speech signals," Neural Computing and Applications, pp. 1–17, 2024.

[3] A. Jayakumar et al., "Low complexity model with single dimensional feature for speech based classification of amyotrophic lateral sclerosis patients and healthy individuals," in SPCOM. IEEE, 2024, pp. 1–5.

Workflow

Train for Healthy Nasal vs Non-Nasal Phoneme Classification with Models of Varying Complexity (12 layers of HuBERT for feature representation)

Evaluate on ALS vs Control (COT) Dataset 1 with High Complexity DNN Model(ALS = Nasal, COT = Non-Nasal)

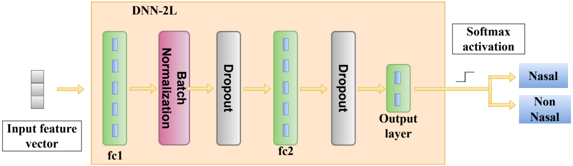
Identify Best Layer (Maximum Accuracy)

Test on ALS vs COT Dataset 2 with DNN Models of Reduced Complexity (ALS = Nasal, COT = Non-Nasal)

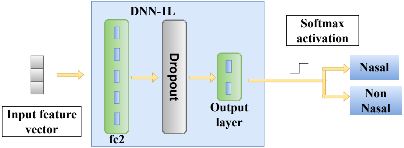
Overview

Models: Nasal vs Non-Nasal Phoneme Classification

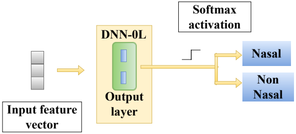
Model 1 : DNN2L



Model 2 : DNN1L



Model 3 : DNN0L



Models: Nasal vs Non-Nasal Phoneme Classification

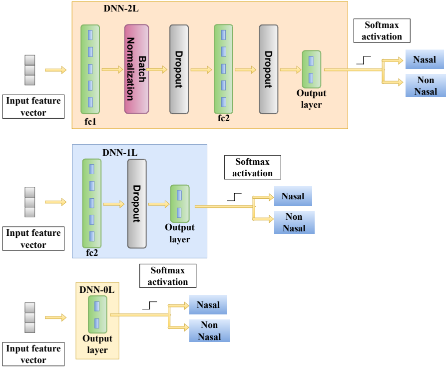


Table: Model complexity

Model	#params	FLOPs
DNN-2L	115,714	115,200
DNN-1L	98,690	98,560
DNN-0L	1,538	1,540

Overview

Training: Nasal vs Non-Nasal Phoneme Classification

Train for Healthy Nasal vs Non-Nasal Phoneme Classification with Models of Varying Complexity (12 layers of HuBERT for feature representation)

↓

Evaluate on ALS vs Control (COT) Dataset 1 with High Complexity DNN Model(ALS = Nasal, COT = Non-Nasal)

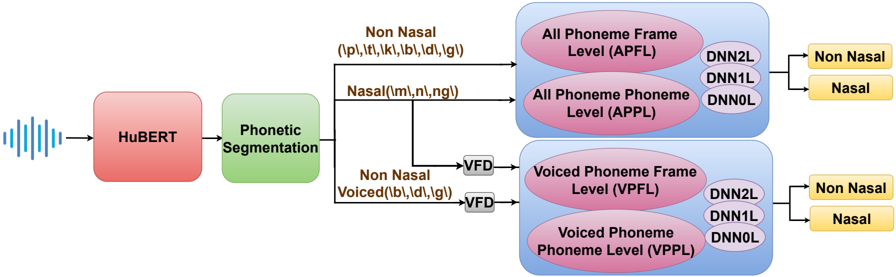
↓

Identify Best Layer (Maximum Accuracy)

↓

Test on ALS vs COT Dataset 2 with DNN Models of Reduced Complexity (ALS = Nasal, COT = Non-Nasal)

Training Pipeline



VFD: Voiced Frame Detection

Training Dataset : TIMIT and INDIC TIMIT (ITIMIT)

Table: Statistics for subsets of the TIMIT and ITIMIT dataset used in this work

Class		#Phonemes	Average Duration (SD) (s)	Total Duration (s)
		TIMIT[1]		
Nasal	TRAIN	1383	0.06 (0.02)	82.03
	TEST	624	0.06 (0.02)	37.08
Non-nasal	TRAIN	1500	0.05 (0.02)	75.51
	TEST	717	0.04 (0.02)	31.97
Non-nasal Voiced	TRAIN	1294	0.05 (0.01)	65.57
	TEST	594	0.05 (0.01)	28.14
		ITIMIT[2]		
Nasal	TRAIN	1432	0.06 (0.02)	92.86
	TEST	684	0.05 (0.02)	27.74
Non-nasal	TRAIN	1527	0.07 (0.02)	107.86
	TEST	801	0.05 (0.01)	36.17
Non-nasal voiced	TRAIN	1463	0.07 (0.02)	102.76
	TEST	421	0.06 (0.02)	24.90

[1] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, "DARPA TIMIT acoustic-phonetic continuous speech corpus CD-ROM. NIST speech disc 1-1.1," NASA STI/Recon technical report n, vol. 93, p. 27403, 1993.

[2] C. Yarra, R. Aggarwal, A. Rajpal, and P. K. Ghosh, "Indic TIMIT and Indic English lexicon: A speech database of Indian speakers using TIMIT stimuli and a lexicon from their mispronunciations," in 22nd Conference of the Oriental International Committee for the Co- ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA). IEEE, 2019, pp. 1–6

Data Processing and Feature Extraction

- **Phonetic Segmentation** Using Phonetic Boundaries
  - TIMIT : Avaliable from the dataset
  - ITIMIT : Forced alignment using KALDI speech recognition toolkit[1]
- **Voiced Frame detection**
  - Using pitch-based segmentation with Praat[2](Frame Rate 20ms)
  - Pitch Range : 50 - 450 Hz
    - Voiced frame: Pitch detected.
    - Unvoiced frame: No pitch detected.
- **Feature Extraction: HuBERT[3]**
  - Using the S3PRL toolkit[4].
  - 12 layers - Each giving 768-dimensional vector representation.
  - Frame Rate : 20ms

[1] D. Povey et al., "The KALDI speech recognition toolkit," 2011. [Online]. Available: https://api.semanticscholar.org/CorpusID: 1774023

[2] P. Boersma and D. Weenink, "Praat: doing phonetics by computer (version 5.1.13)," 2009. [Online]. Available: http://www.praat.org

[3] Y. Wang et al., "A fine-tuned Wav2Vec 2.0/HuBERT benchmark for speech emotion recognition, speaker verification and spoken language understanding," 2022. [Online]. Available: https://arxiv.org/abs/2111.02735

[4] A. T. Liu et al., "Tera: Self-supervised learning of transformer encoder representation for speech," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 29, p. 2351–2366, 2021. [Online]. Available: http://dx.doi.org/10.1109/ TASLP.2021.3095662

Overview

ALS vs COT Classification

Train for Healthy Nasal vs Non-Nasal Phoneme Classification with Models of Varying Complexity (12 layers of HuBERT for feature representation)

↓

Evaluate on ALS vs Control (COT) Dataset 1 with High Complexity DNN Model(ALS = Nasal, COT = Non-Nasal)

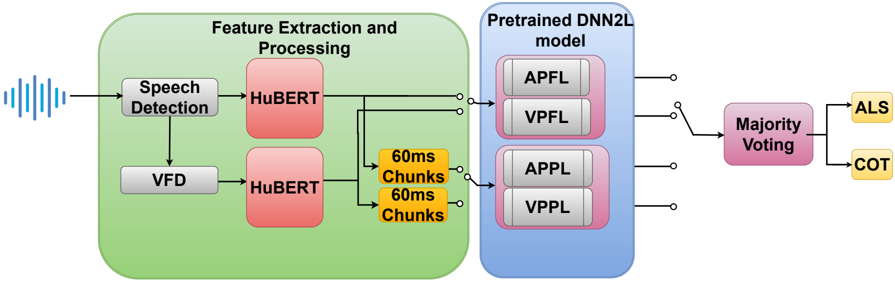
↓

Identify Best Layer (Maximum Accuracy)

↓

Test on ALS vs COT Dataset 2 with DNN Models of Reduced Complexity (ALS = Nasal, COT = Non-Nasal)

Pipeline



VFD: Voiced Frame Detection
APFL: All Phoneme Frame Level
VPFL: Voiced Phoneme Frame Level
APPL: All Phoneme Phoneme Level
VPPL: Voiced Phoneme Phoneme Level

Dataset : Recording tasks

Recording conducted at **National Institute of Mental Health and Neurosciences (NIMHANS)**, Bangalore*.

Spontaneous Speech (SPON)

- Describe a Festival
- Describe a Place
- 1 min each

Diadochokinetic Rate (DIDK)

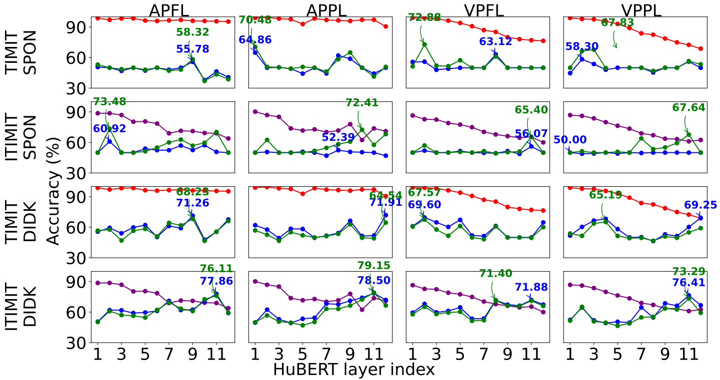
- Mono-syllabic Sequences:
  - pa-pa-pa, ta-ta-ta, ka-ka-ka
- Tri-syllabic Sequences:
  - pataka, badaga
- Upto 3 repetitions

* J. Mallela et al., "Raw speech waveform based classification of patients with ALS, Parkinson's disease and healthy controls using CNN-BLSTM," in Proc. 21st Annual Conference of the International Speech Communica- tion Association, Shanghai, China, 2020, pp. 4586–4590.

Dataset 1

Class	#Speakers	Average Duration (SD) (s)	Total Duration (min)
		SPON	
ALS	30 (18M+12F)	59.75 (19.93)	53.77
COT	30 (22M+8F)	60.04 (17.58)	59.14
		DIDK	
ALS	30 (18M+12F)	15.34 (8.09)	36.56
COT	30 (22M+8F)	18.58 (7.78)	46.14

ALS vs COT Classification Accuracies

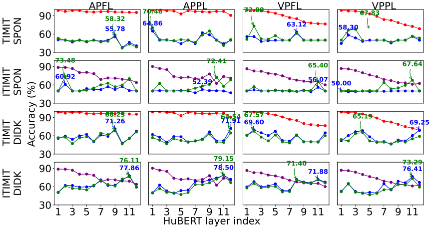


: TIMIT Accuracy on nasal vs non nasal phoneme classifictaion

: ITIMIT Accuracy on nasal vs non nasal phoneme classifictaion

: Speech Frames : Voiced Frames

Comparison of Different Conditions



—— : TIMIT Accuracy on nasal vs non nasal phoneme classifictaion

—— : ITIMIT Accuracy on nasal vs non nasal phoneme classifictaion

—— : Speech Frames

—— : Voiced Frames

Table: Max. classification accuracy (%) using various train /test conditions.

Condition	SPON	DIDK
TIMIT	72.88 (VPFL, voiced)	71.91 (APPL, speech)
ITIMIT	**73.48** (APFL,voiced)	**79.15** (APPL, voiced)
APFL	**73.48** (voiced, ITIMIT)	77.86 (speech,ITIMIT)
APPL	72.41 (voiced, ITIMIT)	**79.15** (voiced, ITIMIT)
VPFL	72.88 (voiced, TIMIT)	71.88 (speech, ITIMIT)
VPPL	67.83 (voiced, TIMIT)	76.41 (speech, ITIMIT)
Speech	64.86 (APPL, TIMIT)	78.50 (APPL, ITIMIT)
Voiced	**73.48** (APFL, ITIMIT)	**79.15** (APPL, ITIMIT)

Key Takeaway

- TIMIT achieves higher average nasal vs. non-nasal classification accuracy, of 92.50%, compared to only 74.99% for ITIMIT across all train cases.
- In terms of HuBERT layers, for ITIMIT, the higher layers perform better for the DIDK task.
- The ITIMIT dataset with voiced features and All Phonemes train case provided the highest classification accuracy for both SPON and DIDK.
- Voiced features outperformed speech features in maximum accuracy for both SPON and DIDK.
- The TIMIT generally showed lower accuracies compared to ITIMIT.

Overview

Low Complexity ALS vs COT Classification

Train for Healthy Nasal vs Non-Nasal Phoneme Classification with Models of Varying Complexity (12 layers of HuBERT for feature representation)

↓

Evaluate on ALS vs Control (COT) Dataset 1 with High Complexity DNN Model(ALS = Nasal, COT = Non-Nasal)

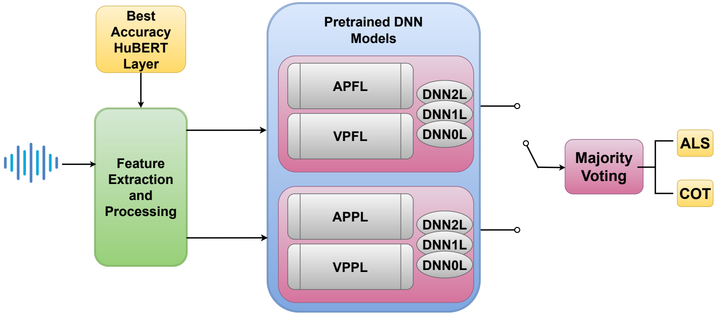
↓

Identify Best Layer (Maximum Accuracy)

↓

Test on ALS vs COT Dataset 2 with DNN Models of Reduced Complexity (ALS = Nasal, COT = Non-Nasal)

Pipeline

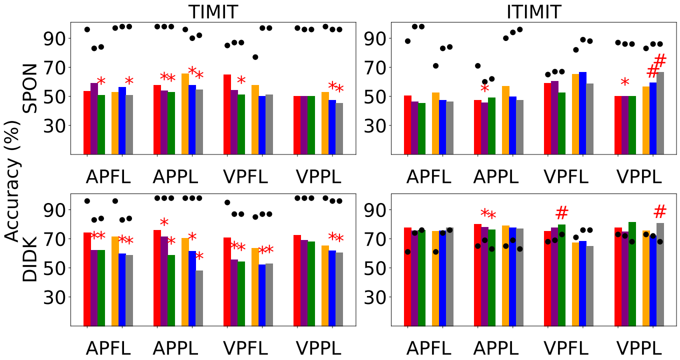


APFL: All Phoneme Frame Level
VPFL: Voiced Phoneme Frame Level
APPL: All Phoneme Phoneme Level
VPPL: Voiced Phoneme Phoneme Level

Dataset 2

Class	#Speakers	Average Duration (SD) (s)	Total Duration (min)
		SPON	
ALS	27 (17M+10F)	58.98 (15.34)	53.08
COT	25 (18M+7F)	57.91 (23.96)	48.25
		DIDK	
ALS	27 (17M+10F)	17.97 (9.87)	40.43
COT	25 (18M+7F)	19.07 (9.16)	39.72

Low Complexity ALS vs COT Classification Accuracies



DNN2L-Speech DNN2L-Voiced (#param: 115,714; FLOPs: 115,200)

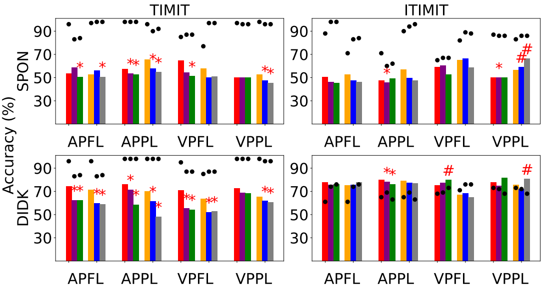
DNN1L-Speech DNN1L-Voiced (#param: 98,690; FLOPs: 98,560)

DNN0L-Speech DNN0L-Voiced (#param: 1,538; FLOPs: 1,540)

● TIMIT or ITIMIT nasal vs. non-nasal phoneme classification accuracy.

* Statistically significant performance drop # Superior performance - compared to the corresponding DNN2L model (Wilcoxon signed-rank test-1% significance level)

Comparison with DNN2L



DNN2L-Speech **DNN2L-Voiced**
(#param: 115,714; FLOPs: 115,200)
DNN1L-Speech **DNN1L-Voiced**
(#param: 98,690; FLOPs: 98,560)
DNN0L-Speech **DNN0L-Voiced**
(#param: 1,538; FLOPs: 1,540)

● **TIMIT or ITIMIT nasal vs. non-nasal phoneme classification accuracy.**
* Statistically significant performance drop # Superior performance - compared to DNN2L (Wilcoxon signed-rank test, 1% significance level)

	SPON		DIDK	
	DNN1L	**DNN0L**	**DNN1L**	**DNN0L**
Average Accuracy Drop (%)	3.06	4.43	5.44	5.90
Average Reduction in #Param (%)	14.71	98.67	14.71	98.67
Average Reduction in FLOPs(%)	14.44	98.66	14.44	98.66
#Significant Performance Drop	5/16	6/16	8/16	8/16
#Outperforming DNN2L	1/16	1/16	0/16	1/16
Best Configuration	DNN0L,ITIMIT,VPPL,Voiced frames		DNN0L,ITIMIT,VPPL,Speech frames	
Maximum Accuracy (%)	**66.48%**		**81.47%**	

Overview

Key Takeaway

- Nasality can be used as an effective indicator for ALS vs COT classification
- Training using ITIMIT provides the highest accuracies.
- Voiced frames for SPON and speech frames for DIDK provides the highest accuracies.
- Reducing model complexity has minimal impact on performance.
- DNN0L achieve the highest accuracy, with 66.48% for SPON and 81.47% for DIDK.

Future Work

- Investigating the potential of nasality as an indicator for ALS severity classification, extending beyond binary ALS vs. COT classification.
- Exploring the methodology across diverse datasets to ensure generalizability of the findings.
- Enhancing model performance through techniques such as transfer learning.

Acknowledgment

- We sincerely thank all the subjects who contributed to the speech dataset.
- We express our sincere gratitude to the Department of Science and Technology (DST), Government of India for supporting this work.

THANK YOU

Have Questions/Suggestions?
Write to us @ `spirelab.ee@iisc.ac.in`