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

*

Anjali Jayakumar
Dept. of Electrical Engg.,
Indian Institute of Science,
Bengaluru, India
anjalij@iisc.ac.in

Tanuka Bhattacharjee
Dept. of Electrical Engg.,
Indian Institute of Science,
Bengaluru, India
tanukab@iisc.ac.in

Seena Vengalil
NIMHANS,
Bengaluru, India

Yamini Belur
NIMHANS,
Bengaluru, India

Nalini Atchayaram
NIMHANS,
Bengaluru, India

Keerthipriya M
NIMHANS,
Bengaluru, India

Darshan Chikktimmegowda
NIMHANS,
Bengaluru, India

Prasanta Kumar Ghosh,
Dept. of Electrical Engg.,
Indian Institute of Science,
Bengaluru, India
prasantg@iisc.ac.in

***Abstract*—Hypernasality, commonly observed in dysarthric speech due to impaired velopharyngeal function, is a key characteristic of Amyotrophic Lateral Sclerosis (ALS). The increasing demand for lightweight models capable of operating on resource-constrained platforms, such as mobile devices or general-purpose computers, underscores the importance of developing efficient methods for ALS analysis. This study explores hypernasality as a potential indicator for ALS by utilizing nasal and non-nasal phonemes from healthy speech for ALS vs. Healthy Controls (COT) classification. We use HuBERT, a pre-trained speech representation model, alongside a Dense Neural Network (DNN) for classification. We also explore reducing model complexity by minimizing the number of dense layers in a DNN model and compare its performance with that of a higher-complexity DNN model. Experiments involving 57 ALS patients and 55 COT using Spontaneous Speech (SPON) and Diadochokinetic Rate (DIDK) tasks show that hypernasality can effectively distinguish ALS from COT. The HuBERT layer that provides the highest classification accuracy is selected based on results from Test Set 1 (30 ALS and 30 COT), and the performance of the low-complexity models for this layer is evaluated on Test Set 2 (27 ALS and 25 COT). The highest classification accuracy for the SPON task is 73.47%, using features from the second layer of the HuBERT model, while the DIDK task achieves 79.15% accuracy with features from the eleventh layer. Reducing model complexity leads to minimal average accuracy loss across various train-test conditions— 3.07% and 4.43% for SPON, and 5.44% and 5.90% for DIDK—while achieving substantial reductions in model parameters (14.71% and 98.67%) and floating-point operations (FLOPs) (14.44% and 98.66%). Notably, the lowest complexity model, for DIDK, achieves 81.46% accuracy with just 1,538 parameters and 1,540 FLOPs, compared to 115,714 parameters and 115,200 FLOPs in the high-complexity model. For SPON, it achieves a maximum accuracy of 66.48%. This work demonstrates that hypernasality serves as an effective ALS indicator, and reduced model complexity provides a feasible trade-off between performance and resource efficiency.**

***Index Terms*—Amyotrophic Lateral Sclerosis, Hypernasality, HuBERT**

I. Introduction

Amyotrophic Lateral Sclerosis (ALS) is a progressive neurodegenerative disease that leads to motor neuron degeneration, causing severe muscle weakness and speech impairments, including dysarthria and hypernasality. Hypernasality, characterized by excessive nasal resonance due to velopharyngeal dysfunction, is found to be present in 75% of motor neuron disease patients [1] and 73.88% of ALS patients with bulbar onset [2], making it a potential indicator of ALS.

ALS monitoring is time-consuming and relies on specialized tests, making frequent assessments challenging [3]. Speech-based ALS vs. Healthy Control (COT) classification has been explored in several studies [4]–[6]. Recent studies have also explored transformer-based models like HuBERT [7] for classifying vowels and fricatives in ALS patients [8], detecting and classifying dysarthric speech severity [9], and other speech tasks [10].

Automatic detection of hypernasality has advanced significantly with techniques such as 1/3-octave band analysis, group delay-based signal processing and the Teager energy operator [11]–[13]. Eshghi, Marziye, et al. in [11] demonstrated that 1/3-octave band analysis can be an early and effective indicator of hypernasality. HuBERT based model for distinguishing

This research is supported by the Department of Science and Technology (DST), Government of India.

hypernasal speech in patients with cleft lip and palate has been explored in [14]. Previous studies have also explored using healthy speech to assess nasality in dysarthric speech [15], [16]. However, the use of hypernasality for distinguishing ALS from COT remains unexplored.

Low-complexity models have been widely explored for speech classification tasks such as ALS and Parkinson's disease (PD) vs. COT classification. For instance, single-dimensional pitch has been shown to offer comparable performance to multi-dimensional mel-frequency cepstral coefficients (MFCCs) while providing greater robustness to noise, making it an effective feature for low-complexity ALS and PD detection [17]. While reducing the model complexity for ALS vs. COT classification have been studied using MFCCs [18], the potential of HuBERT representations for such classification remains unexplored.

This study proposes a novel ALS vs. COT classification approach using hypernasality as an indicator of ALS speech. We extract the HuBERT representation and train a DNN model to classify phonemes as nasal or non-nasal for healthy speakers under different conditions and use this model to classify speech as ALS or COT based on the majority of frames, considering ALS as the nasal class. Additionally, we explore low-complexity models to enhance the practical utility of the classification system. This work aims to demonstrate hypernasality's potential as a reliable ALS indicator and the feasibility of low-complexity, speech-based ALS vs. COT classification.

Experiments are conducted with a total of 57 ALS patients and 55 COT subjects, divided into two sets: Test Set 1 (30 ALS and 30 COT) and Test Set 2 (27 ALS and 25 COT). The HuBERT layer providing the maximum classification accuracy is selected based on experiments conducted on Test Set 1, and the performance of low-complexity models for that layer is evaluated on Test Set 2. The Spontaneous Speech (SPON) and Diadochokinetic Rate (DIDK) tasks are used. The highest classification accuracy of 73.47% for the SPON task and 79.15% for the DIDK task, using features extracted from the second and eleventh layers of the HuBERT model, respectively, are achieved on Test Set 1. When the model complexity is reduced using the HuBERT layers that provided the highest accuracies for different train-test configurations, and tested test set 2, the average ALS vs. COT classification accuracy drops by and 3.07% and 4.43% for the SPON task, and by 5.44% and 5.90% for the DIDK task using the lower-complexity models. However, the reduction in accuracy is relatively small compared to the substantial reductions in model complexity, with a 14.71% and 98.67% decrease in the number of parameters (#param), and a 14.44% and 98.66% reduction in floating-point operations (FLOPs), respectively, compared to the high-complexity model. Notably, for the DIDK task, a maximum accuracy of 81.46% is achieved with 1,538 parameters and 1,540 FLOPs, compared to 115,714 parameters and 115,200 FLOPs for the high-complexity model. However, for the SPON task, the maximum accuracy achieved is 66.48% with the same reduction in model complexity.

II. DATASET

A. ALS and COT Dataset

Speech samples were collected from 57 ALS patients (35M + 22F) and 55 COT (40M + 15F), aged 30-76 years for ALS and 35-65 years for COT, speaking Bengali, Kannada, or Telugu as native language. Three Speech-Language Pathologists assessed dysarthria severity in ALS patients based on prerecorded SPON samples, using a 5-point scale (0 = loss of useful speech to 4 = normal speech) similar to the speech component of ALSFRS-R scale [19]. The final severity score for each patient was determined by the mode of these ratings. The severity scores were distributed as follows: 10 patients with score 0, 10 with score 1, 12 with score 2, 10 with score 3, and 15 with score 4.

For data collection, two types of speech tasks were employed: SPON and DIDK. In the SPON task, participants spoke for about one minute each on *a festival they celebrate* and *a place they had recently visited* in their native language. The DIDK task required participants to repeat monosyllabic or trisyllabic sequences such as *pa-pa-pa*, *ta-ta-ta*, *ka-ka-ka*, *pataka*, and *badaga* after taking a deep breath. Each sequence was recorded up to three times, depending on the subject's comfort. For further details on the data collection and recording setup, please refer [4]. The ALS and COT dataset was split into two balanced test sets (Test Set 1 and Test Set 2) based on age, severity score, and language, with statistics provided in Table I.

TABLE I: ALS and COT dataset statistics

Class	#Speakers	SPON		DIDK	
		Average Duration (SD) (s)	Total Duration (min)	Average Duration (SD) (s)	Total Duration (min)
		Test Set 1			
ALS	30 (18M+12F)	59.75 (19.93)	53.77	15.34 (8.09)	36.56
COT	30 (22M+8F)	60.04 (17.58)	59.14	18.58 (7.78)	46.14
		Test Set 2			
ALS	27 (17M+10F)	58.98 (15.34)	53.08	17.97 (9.87)	40.43
COT	25 (18M+7F)	57.91 (23.96)	48.25	19.07 (9.16)	39.72

B. TIMIT and INDIC TIMIT Datasets

Sentences from TIMIT [20] (with 40 male speakers and 40 female speakers) and INDIC TIMIT (ITIMIT) [21](with 40 male speakers and 40 female speakers) are segmented into individual phonemetic units using the phonetic boundaries, and are then categorized into nasal, non-nasal and non-nasal voiced phonemes for classification purposes. The nasal phonemes includes */m/*, */n/*, and */ng/*, the non-nasal phonemes includes */b/*, */d/*, */g/*, */p/*, */t/*, and */k/*, while the non-nasal voiced phonemes includes */b/*, */d/* and */g/*. The statistics of the training dataset is given in Table II. The sentences are split into train and test sets according to the standard train-test split provided by the dataset.

TABLE II: Statistics for subsets of the TIMIT and ITIMIT datasets used in this work.

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

III. Method

The classification pipeline uses the HuBERT model to extract speech features for analyzing nasal and non-nasal speech. HuBERT representations of TIMIT and ITIMIT phonemes are used as input to train a DNN model. Assuming ALS speech contains more nasal sounds than COT speech, the model classifies Test Set 1 into ALS (nasal) and COT (non-nasal). The best-performing HuBERT layer on Test Set 1 for various train-test combinations is selected to reduce classification complexity, and performance is evaluated on Test Set 2.

A. Data Processing

For training, Voice Activity Detection (VAD) is performed on nasal and non-nasal voiced phonemes of TIMIT and ITIMIT dataset using pitch-based segmentation with Praat [22], aiming to isolate segments where nasalization is most prominent. Each 20 ms frame of the speech signal is analyzed to detect the presence of pitch. If pitch is detected, the frame is labeled as 'voiced'; otherwise, it is labeled as 'unvoiced'. The voiced frames of nasal and non-nasal voiced phonemes are isolated, resulting in an average dropoff of 19.29% in the number of frames for nasal phonemes and 22.91% for non-nasal voiced phonemes for the TIMIT dataset, and 14.37% for nasal and 18.75% for non-nasal voiced phonemes for the ITIMIT dataset. We extract HuBERT representations from the phonetic units at various layers of the HuBERT model to assess their effectiveness in distinguishing between nasal and non-nasal sounds.

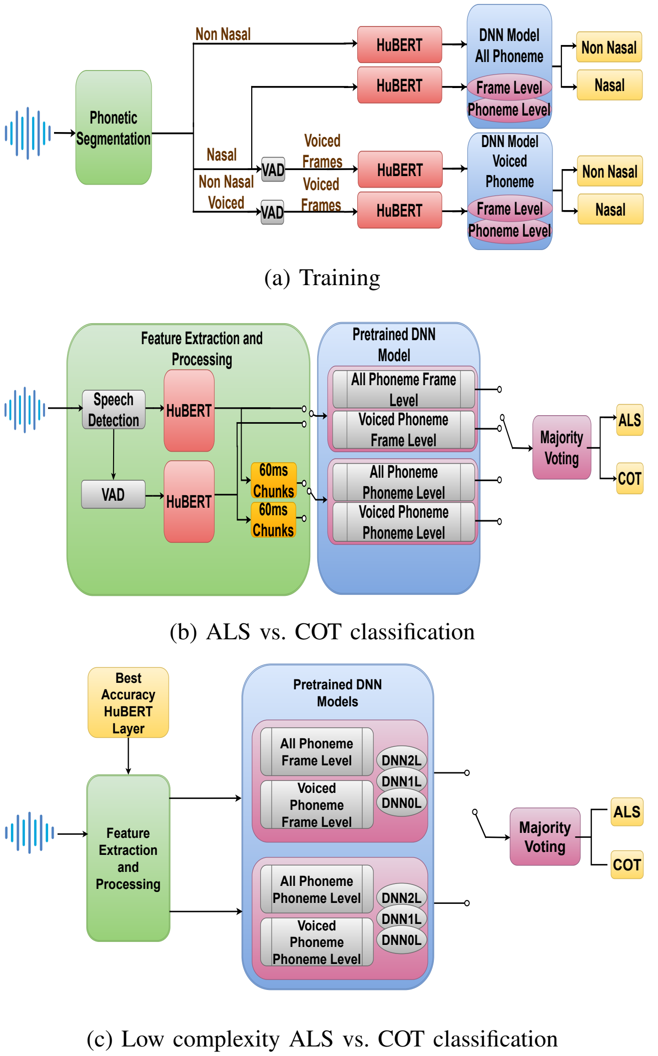
For the ALS and COT dataset, we first analyze all speech frames after removing silent segments, using the librosa library [23]. Secondly, we focus exclusively on voiced frames of the speech segments, extracted using VAD as discussed above. For both methods, the relevant frames of HuBERT representations at various layers are extracted. Extracting the voiced frames results in an average dropoff in the number of frames by 0.51% for SPON and 2.30% for DIDK for the ALS data, and 1.71% for SPON and 12.80% for DIDK for the COT data, compared to the speech frames.

B. Training

Fig1a demonstrates the training process where we use a DNN model to train on TIMIT and ITIMIT datasets separately, using four configurations of training, and the model weights are saved for further predictions on the ALS and COT dataset. Firstly, the input dataset is divided into two categories:

- **All phonemes**: The HuBERT representation of non-nasal and nasal phonemes, without isolating the voiced frames are used as the input to the classification model.
- **Voiced phonemes**: The HuBERT representation of non-nasal voiced and nasal phonemes, on isolating the voiced frames are used as the input to the classification model.

Each category is trained at two levels:



(a) Training

(b) ALS vs. COT classification

Best Accuracy HuBERT Layer
Feature Extraction and Processing
Pretrained DNN Models
All Phoneme Frame Level
Voiced Phoneme Frame Level
DNN2L
DNN1L
DNN0L
All Phoneme Phoneme Level
Voiced Phoneme Phoneme Level
DNN2L
DNN1L
DNN0L
Majority Voting
ALS
COT

(c) Low complexity ALS vs. COT classification

Fig. 1: Classification pipeline

- **Frame level**: Individual frames of each phoneme is given as input to the model, along with its corresponding label.
- **Phoneme level**: The model receives the mean representation of all frames corresponding to each phoneme, along with its corresponding label.

Combining these classifications and training levels, the four train cases are:

1) **All Phonemes - Frame Level (APFL)**
2) **All Phonemes - Phoneme Level (APPL)**
3) **Voiced Phonemes - Frame Level (VPFL)**
4) **Voiced Phonemes - Phoneme Level (VPPL)**

C. ALS vs. COT Classification

A binary classification is performed using the trained models for all HuBERT layers, with ALS considered as the nasal class and COT as the non-nasal class. As illustrated in Fig.1b, for each training case, classification is performed using the speech and voiced HuBERT frames from Test Set 1. For phoneme-level training, chunks of 60 ms are used, corresponding to the average duration of a phoneme in the TIMIT and ITIMIT datasets. The classification is first conducted at the frame or chunk level, followed by majority voting across all frames or chunks within an utterance to determine the final predicted class.

D. Low Complexity ALS vs. COT Classification

Fig.1c demonstrates the methodology for low complexity ALS vs. COT classification. The HuBERT layer that achieves the highest accuracy for each train-test combination on Test Set 1 is selected to develop a low-complexity model for ALS vs. COT classification on Test Set 2. Model complexity is reduced by reducing the number of dense layers in the DNN model. The low-complexity models are then trained on the TIMIT and ITIMIT datasets for nasal vs. non-nasal classification, using each train case and the corresponding HuBERT layer. Finally, the trained models are applied to ALS vs. COT classification on Test Set 2, and a majority voting is performed to predict the final class.

IV. Experimental setup

A. Feature extraction

The TIMIT sentences are segmented into phonetic units using the phonetic alignment timestamps provided within the dataset. For the ITIMIT sentences, forced alignment was performed using the KALDI speech recognition toolkit [24] to extract phonetic alignment.

HuBERT representations are extracted from the model across twelve layers for the TIMIT, ITIMIT, and the ALS and COT dataset, with a frame rate of 20 ms, using the S3PRL toolkit [25]. Each layer produces a 768-dimensional vector representation. Pitch analysis in Praat is used to extract voiced frames from the audio files every 20 ms, with a pitch range of 50 Hz to 450 Hz, and all other parameters are set to their default values. The librosa library is employed to separate speech frames from silence.

B. Model description

We employ a DNN model (DNN2L) that consists of an input layer which accepts the HuBERT representations from each layer as feature vectors, serving as the high-complexity model. This model includes two fully connected dense layers, each with 128 units and ReLU activation functions. Batch normalization is applied after the first dense layer, and dropout is used with a rate of 0.3 in both dense layers. The final layer is a softmax output layer with 2 units for binary classification.

For the low-complexity classifications, we use DNN1L, which consists of a single dense layer with 128 units and ReLU activation, followed by a dropout layer and an output dense layer with 2 units and softmax activation. DNN0L is a simplified model consisting solely of the softmax output dense layer with 2 units. The models' hyperparameters are tuned based on validation accuracy.

C. Training and Evaluation

The models are trained using the Adam optimizer, with a learning rate of 0.001 and binary cross-entropy as the loss function. Training utilizes a batch size of 32, and the model is trained for up to 100 epochs. The training data is split into training and validation sets with an 80:20 ratio, with the validation set used for early stopping with a patience of 8. The balanced accuracy scores on the testing datasets are reported as the performance metrics. A Wilcoxon signed-rank test [26] at a 1% significance level is used to compare classification accuracies of low-complexity models against DNN2L. Test Set 2 is divided into 12 sub-groups, and the resulting accuracy values are used for the test.

V. Result and Discussion

The classification accuracies for ALS vs. COT on each HuBERT layer are given in Fig.2. 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.

A. Comparison of Train Cases

We consider the HuBERT layer representation that yields the highest classification accuracy for each train-test combination. The corresponding highest accuracy and HuBERT layer index for each configuration is provided in Fig 2.

The maximum classification accuracy of **73.47%** for the SPON task is achieved by training on the ITIMIT dataset with **APFL** and testing on voiced frames, using the second HuBERT layer. For the DIDK task, the maximum accuracy of **79.15%** is attained by training with **APPL** on the ITIMIT dataset and testing on voiced frames, using the eleventh HuBERT layer. Considering the maximum classification accuracies, on average, the best performance is obtained with **APPL**, resulting in an average accuracy of **65.03%** for the SPON task and **73.52%** for the DIDK task, across all test conditions of both TIMIT and ITIMIT datasets. For the DIDK task, ITIMIT outperforms TIMIT with an average improvement of **7.13%** in the highest accuracy across all train-test combinations.

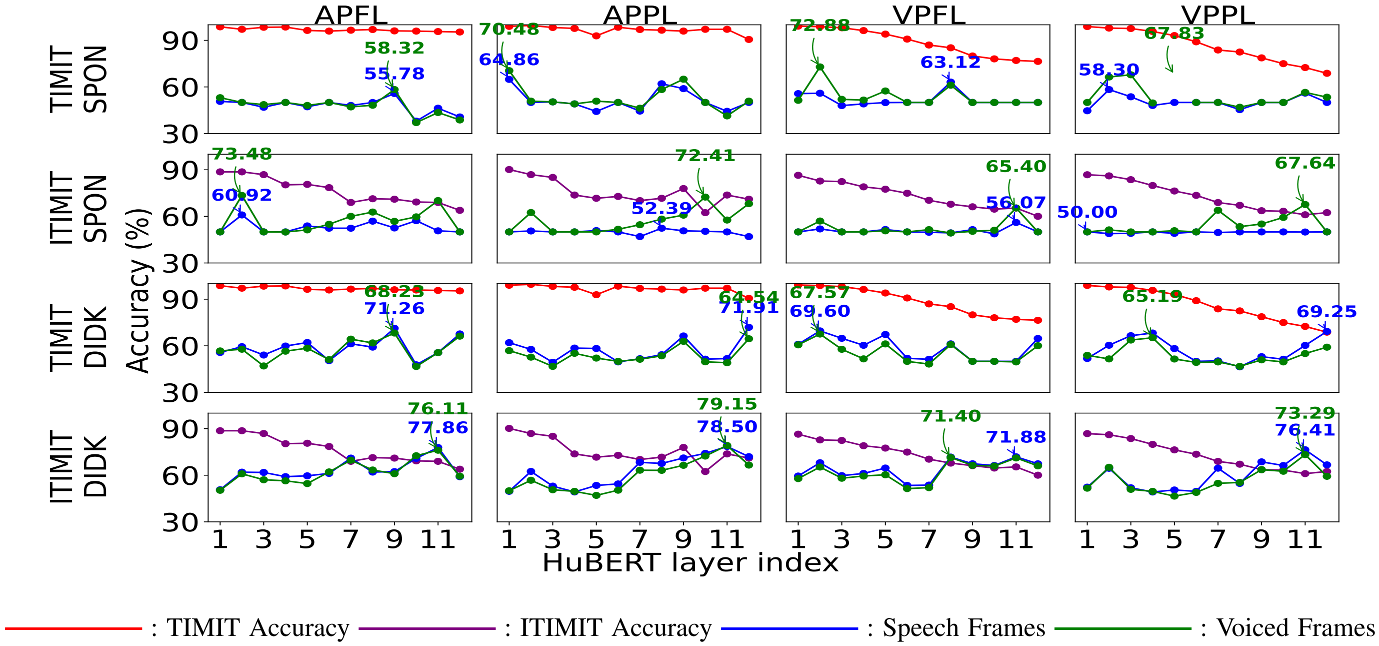


Fig. 2: Classification accuracies for ALS vs. COT under various train-test conditions on Test Set 1. The values represent the highest classification accuracy (%) achieved for each train-test combination.

B. *Comparison of Test Conditions*

The mean of the highest accuracies for each test condition across the four train cases, is shown in Table III. The DIDK task outperforms the SPON task with an average difference of **8.89%** across all train cases. Notably, providing voiced frames during testing for the SPON task improves accuracy by **10.87%**. Performance at both the frame level and chunk level are similar.

TABLE III: The mean of maximum classification accuracy (%) across the four train cases on Test Set 1

Test Condition	SPON	DIDK
Speech	57.68	**73.33**
Voiced	**68.55**	70.68
Frame	**63.24**	71.74
Chunk	62.99	**72.28**

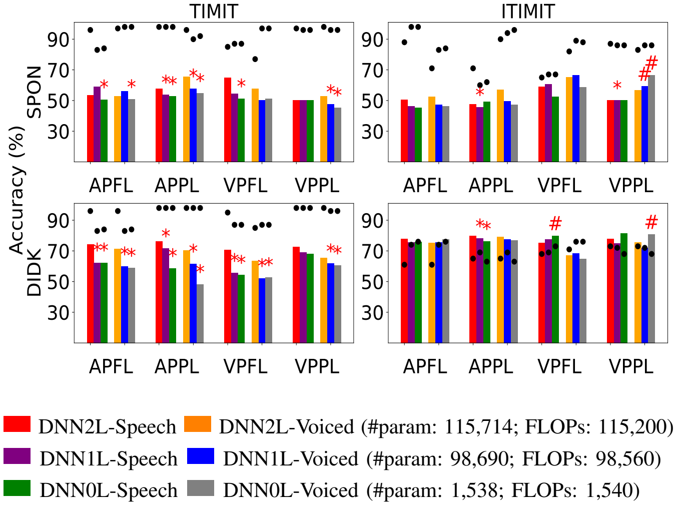


Fig. 3: ALS vs. COT classification accuracies (%) for different train-test conditions for low complexity models on Test Set 2, where ● denotes the TIMIT or ITIMIT nasal vs. non-nasal phoneme classification accuracy. Here, * indicates a statistically significant performance drop, and # indicates superior performance, compared to the corresponding DNN2L model, according to the Wilcoxon signed-rank test at the 1% significance level.

C. *Low complexity classification*

The plots for the low complexity classification accuracies on Test Set 2 are given in Fig.3. The classification accuracies for nasal vs. non-nasal phoneme in the TIMT and ITIMIT datasets are comparable across different DNN models. The average accuracies for TIMIT are **95.55%, 91.25%, and 91.50%** for DNN2L, DNN1L, and DNN0L, respectively, while for ITIMIT, the corresponding accuracies are **72.25%, 77.11%, and 76.04%**. On average, the ALS vs. COT classification accuracy for the SPON task decreases by **3.06%** and **4.43%** for DNN1L and DNN0L, respectively, and for the DIDK task, it declines by **5.44%** and **5.90%**, compared to DNN2L. Despite these accuracy reductions, DNN1L and DNN0L offer significant complexity reductions, with a **14.71%** and **98.67%** reduction in the number of parameters, and a **14.44%** and **98.66%** reduction in FLOPs compared to DNN2L. The maximum classification accuracy is obtained with DNN0L and

VPPL train case, achieving **66.48%** for the SPON task trained on ITIMIT with voiced frames as input, and **81.47%** for the DIDK task, trained on ITIMIT with speech frames as input. This indicates that the DIDK task is more suitable for ALS vs. COT classification, with hypernasality serving as an indicator of the disease. Among the 32 train-test combinations, only 11 for SPON and 16 for DIDK shows a statistically significant performance drop, while 2 for SPON and 2 for DIDK outperforms the corresponding DNN2L model at 1% significance level according to Wilcoxon signed rank test.

VI. Conclusion

This study demonstrates that hypernasality can effectively indicate ALS. The model performs reliably with minimal computational resources, especially for the DIDK task, making it suitable for deployment on resource-constrained platforms. These findings support the development of practical, non-invasive ALS detection tools for everyday use. Future work will focus on improving accuracy through diverse datasets and exploring alternative methods to enhance performance.

Acknowledgement - We sincerely thank the Department of Science and Technology (DST), Government of India, for supporting this work. We also express our gratitude to all the participants who contributed to the dataset used in this study.

References

[1] A. W. Kummer and L. Lee, "Evaluation and treatment of resonance disorders," *Language, Speech, and Hearing Services in Schools*, vol. 27, no. 3, pp. 271–281, 1996.

[2] E. Candelo, S. S. Vasudevan, D. Orellana, A. M. Williams, and A. L. Rutt, "Exploring the impact of amyotrophic lateral sclerosis on otolaryngological functions," *Journal of Voice*, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0892199724002364

[3] O. Hardiman, L. H. Van Den Berg, and M. C. Kiernan, "Clinical diagnosis and management of amyotrophic lateral sclerosis," *Nature reviews neurology*, vol. 7, no. 11, pp. 639–649, 2011.

[4] J. Mallela, A. Illa, S. BN, S. Udupa, Y. Belur, N. Atchayaram, R. Yadav, P. Reddy, D. Gope, and P. K. Ghosh, "Voice based classification of patients with Amyotrophic Lateral Sclerosis, Parkinson's disease and healthy controls with CNN-LSTM using transfer learning," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP).* IEEE, 2020, pp. 6784–6788.

[5] J. Mallela, Y. Belur, N. Atchayaram, R. Yadav, P. Reddy, D. Gope, and P. K. Ghosh, "Raw speech waveform based classification of patients with ALS, Parkinson's disease and healthy controls using CNN-BLSTM," in *Proc. 21^{st} Annual Conference of the International Speech Communication Association, Shanghai, China*, 2020, pp. 4586–4590.

[6] T. Bhattacharjee, A. Jayakumar, Y. Belur, N. Atchayaram, R. Yadav, and P. K. Ghosh, "Transfer Learning to Aid Dysarthria Severity Classification for Patients with Amyotrophic Lateral Sclerosis," in *Proc. INTERSPEECH*, 2023, pp. 1543–1547.

[7] W.-N. Hsu, B. Bolte, Y.-H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "HuBERT: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. PP, pp. 1–1, 10 2021.

[8] C. V. T. Kumar, T. Bhattacharjee, Y. Belur, A. Nalini, R. Yadav, and P. K. Ghosh, "Classification of multi-class vowels and fricatives from patients having amyotrophic lateral sclerosis with varied levels of dysarthria severity," in *Interspeech*, 2023, pp. 146–150.

[9] F. Javanmardi, S. R. Kadiri, and P. Alku, "Pre-trained models for detection and severity level classification of dysarthria from speech," *Speech Communication*, vol. 158, p. 103047, 2024.

[10] Y. Wang, A. Boumadane, and A. Heba, "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

[11] M. Eshghi, K. P. Connaghan, S. E. Gutz, J. D. Berry, Y. Yunusova, and J. R. Green, "Co-occurrence of hypernasality and voice impairment in amyotrophic lateral sclerosis: Acoustic quantification," *Journal of Speech, Language, and Hearing Research*, vol. 64, no. 12, pp. 4772–4783, 2021.

[12] P. Vijayalakshmi, M. R. Reddy, and D. O'Shaughnessy, "Acoustic analysis and detection of hypernasality using a group delay function," *IEEE Transactions on biomedical engineering*, vol. 54, no. 4, pp. 621–629, 2007.

[13] D. A. Cairns, J. H. Hansen, and J. E. Riski, "A noninvasive technique for detecting hypernasal speech using a nonlinear operator," *IEEE transactions on biomedical engineering*, vol. 43, no. 1, p. 35, 1996.

[14] S. Bhattacharjee, H. S. Shekhawat, and S. R. M. Prasanna, "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.

[15] V. C. Mathad, K. Chapman, J. Liss, N. Scherer, and V. Berisha, "Deep learning based prediction of hypernasality for clinical applications," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP).* IEEE, 2020, pp. 6554–6558.

[16] M. Saxon, J. Liss, and V. Berisha, "Objective measures of plosive nasalization in hypernasal speech," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP).* IEEE, 2019, pp. 6520–6524.

[17] T. Bhattacharjee, J. Mallela, Y. Belur, N. Atchayarcmf, R. Yadav, P. Reddy, D. Gope, and P. K. Ghosh, "Effect of noise and model complexity on detection of Amyotrophic Lateral Sclerosis and Parkinson's disease using pitch and MFCC," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP).* IEEE, 2021, pp. 7313–7317.

[18] A. Jayakumar, T. Bhattacharjee, S. Vengalil, Y. Belur, N. Atchayaram, and P. K. Ghosh, "Low complexity model with single dimensional feature for speech based classification of amyotrophic lateral sclerosis patients and healthy individuals," in *International Conference on Signal Processing and Communications (SPCOM).* IEEE, 2024, pp. 1–5.

[19] J. M. Cedarbaum, N. Stambler, E. Malta, C. Fuller, D. Hilt, B. Thurmond, A. Nakanishi, B. A. S. Group, and A. complete listing of the BDNF Study Group, "The ALSFRS-R: a revised ALS functional rating scale that incorporates assessments of respiratory function," *Journal of the neurological sciences*, vol. 169, no. 1-2, pp. 13–21, 1999.

[20] 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.

[21] 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 *22^{nd} 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.

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

[23] B. McFee, C. Raffel, D. Liang, D. Ellis, M. Mcvicar, E. Battenberg, and O. Nieto, "Librosa: Audio and music signal analysis in python," 01 2015, pp. 18–24.

[24] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. K. Goel, M. Hannemann, P. Motlícek, Y. Qian, P. Schwarz, J. Silovský, G. Stemmer, and K. Veselý, "The KALDI speech recognition toolkit," 2011. [Online]. Available: https://api.semanticscholar.org/CorpusID:1774023

[25] A. T. Liu, S.-W. Li, and H.-y. Lee, "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

[26] R. Woolson, "Wilcoxon signed-rank test," *Wiley encyclopedia of clinical trials*, pp. 1–3, 2007.