Transfer Learning to Aid Dysarthria Severity Classification for Patients with Amyotrophic Lateral Sclerosis

Tanuka Bhattacharjee[1], *Anjali Jayakumar*[1], *Yamini Belur*[2], *Atchayaram Nalini*[2], *Ravi Yadav*[2], *Prasanta Kumar Ghosh*[1]

[1]Electrical Engineering Department, Indian Institute of Science, Bengaluru, India
[2]National Institute of Mental Health and Neurosciences, Bengaluru, India

tanukab@iisc.ac.in, anjalij@iisc.ac.in, prasantg@iisc.ac.in

Abstract

A major challenge involved in automatic dysarthria severity classification for patients with Amyotrophic Lateral Sclerosis (ALS) is the difficulty to build a speech corpus which is large enough to train accurate and generalizable classifiers. To overcome this constraint, we employ transfer learning approaches, specifically, fine-tuning from an auxiliary task and multi-task learning. Input feature reconstruction and gender classification, on the same ALS speech dataset or other healthy speech corpora, are explored as the auxiliary tasks. We use temporal statistics of mel-frequency cepstral coefficients as the features and dense neural networks for performing the primary and auxiliary tasks. Experiments suggest that transfer learning aids severity classification with up to 11.03% absolute increase in the average classification accuracy as compared to direct single task learning. The improvement is attributed mainly to better classification of the mild class than severe/normal classes.

Index Terms: Amyotrophic Lateral Sclerosis, dysarthria severity, transfer learning, fine-tuning, multi-task learning

1. Introduction

Amyotrophic Lateral Sclerosis (ALS) is a progressive neurodegenerative disease that impairs muscle movements. Speech musculature, among others, gets critically affected leading to dysarthria. Though there is no cure for ALS or the associated dysarthria, regular therapy and personalized disease management strategies can help enhance the quality of life of these patients. Regular monitoring of the disease condition is essential for continuously modifying these strategies depending on the needs of the patients. Clinically, dysarthria severity of an ALS patient is examined by Speech-Language Pathologists (SLPs) as per the ALSFRS-R scale [1]. In spite of the merits of ALSFRS-R, this clinical assessment procedure is tedious and highly time-consuming. Moreover, subjectivity and perception being involved, the assessments may not always be consistent. Thus, accurate and consistent automatic dysarthria severity prediction systems are the need of the hour.

Though several works are present in the literature on speech-based automatic classification of ALS patients and Healthy Controls (HC) [2, 3, 4], only a few efforts have been reported in the domain of speech-based automatic dysarthria severity prediction for ALS. Suhas et al. [5] have employed a 2D Convolutional Neural Network (CNN) for this purpose and observed that log-mel spectrogram outperforms Mel-Frequency Cepstral Coefficients (MFCC) as the input speech representation for the proposed system. Wisler et al. [6] have utilized speech along with articulatory data to estimate the ALSFRS-R bulbar subscore using linear ridge regression and support vector regression. The primary challenge in developing sophisticated dysarthria severity prediction models is the scarcity of data resources. Collecting speech data from patients having speech impairments is a delicate and laborious task. Getting the collected data clinically annotated for dysarthria severity further adds to the difficulty. Hence, the dysarthria severity prediction systems to be developed should essentially be data-efficient.

Data scarcity not only reduces the efficiency of the classifiers but also affects their generalizability as the models tend to overfit to the small amount of training data. In deep learning practice, transfer learning is used frequently to deal with these issues. In this method, the learning achieved through some auxiliary tasks or from some auxiliary datasets is utilized to train the models for the primary task at hand. Transfer learning has already been adopted in various domains of dysarthric speech research including dysarthric speech recognition [7, 8, 9, 10, 11], dysarthric speech enhancement [12], ALS vs. HC classification [13] and Parkinson's Disease (PD) detection [13, 14]. Joshy et al. [15] have employed multi-head attention with multi-task learning for automatic dysarthria severity classification for patients with Cerebral Palsy. Identification of gender, age and disorder-type have been explored as the auxiliary tasks. Vásquez et al. [16] have proposed a CNN-based multi-task learning approach for dysarthria severity assessment of PD patients. Eleven different auxiliary tasks including PD vs. HC classification and assessments of the degree of impairment of different articulators like lips, palate, tongue and larynx have been considered. Soleymanpour et al. [17] have trained a 1D CNN for dysarthria severity assessment through cross-dataset transfer learning. All these works have demonstrated superior severity prediction results while employing transfer learning methods. However, to the best of our knowledge, no effort has yet been reported towards employing transfer learning approaches for dysarthria severity classification specific to ALS.

In this work, we explore the utility of two particular transfer learning approaches, namely, fine-tuning from an auxiliary task and multi-task learning, as well as their combinations for developing dysarthria severity prediction systems for ALS patients. Our primary task is to perform 3-class dysarthria severity classification [normal (N) speech, mild (M) dysarthria and severe (S) dysarthria] using spontaneous speech utterances. We consider input speech Feature Reconstruction (FR) and Gender Classification (GC) as the auxiliary tasks. In one setting, we perform transfer learning within the same ALS speech dataset which is used for severity classification by leveraging only the auxiliary tasks. This approach is expected to benefit the generalizability of the severity classifier as it regularizes the model training. In the other setup, we exploit the knowledge gained from some auxiliary datasets of healthy speech by virtue of the auxiliary tasks. In all cases, temporal statistics of MFCC are used as the input speech features and dense neural networks are used for

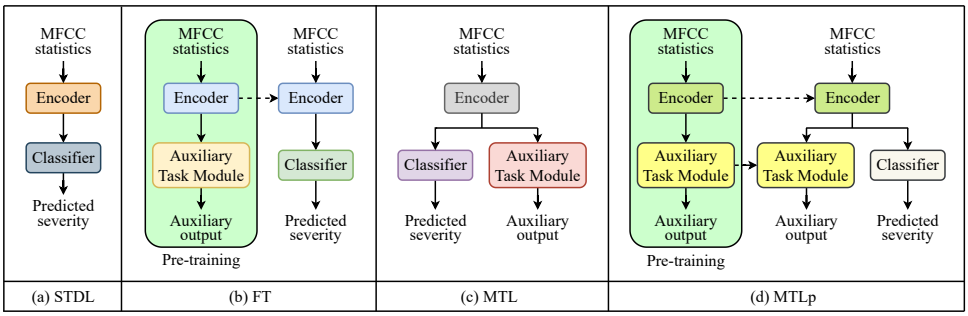


Figure 1: *Block diagram of different network training protocols used in this work; here, dotted arrows indicate that the weights learned during pre-training are used as initializations in the next step*

carrying out the primary and auxiliary tasks. Experimental validation using 120 ALS patients confirms that transfer learning indeed aids the dysarthria severity classification systems with a significant hike of upto 11.03% in the average classification accuracy as compared to the case with no transfer learning.

2. Method

We pose dysarthria severity prediction as a 3-class classification problem. According to the speech component of ALSFRS-R scale, dysarthria severity can be quantized into 5 discrete levels ranging from 0 to 4 where 4 indicates normal speech and 0 denotes complete loss of useful speech. We consider severity levels 0 and 1 together as the *severe class* (S), 2 and 3 together as the *mild class* (M) and 4 alone as the *normal class* (N). We use dense neural networks as the classifiers with temporal statistics of MFCC over an entire utterance as the input features. We analyze the effect of different transfer learning schemes in improving the classification performance of the network. Figure 1 illustrates different approaches explored in this work.

2.1. Single Task Direct Learning (STDL)

In this approach, a network with an encoder followed by a classifier module is employed. The encoder module extracts latent representations from the input features which are then utilized by the classifier module to predict the severity class. The network weights are initialized randomly and then trained to minimize categorical cross-entropy loss for our primary task of 3-class severity classification.

2.2. Transfer learning protocols

We primarily consider two paradigms of transfer learning, e.g. fine-tuning and multi-task learning, along with combinations of the two, to complement the training of the severity classifier.

1. **Fine-Tuning (FT):** A randomly initialized network with an encoder module followed by an auxiliary task module is first pre-trained for an auxiliary task by minimizing the concerned loss function. The auxiliary task module is then replaced by the classifier module (as in STDL) without altering the encoder block. Pre-trained weights from the auxiliary task are used to initialize the encoder while the classifier weights are initialized randomly. Finally, this network is fine-tuned for severity classification by minimizing the cross-entropy loss, as in STDL.
2. **Multi-Task Learning (MTL):** Here, the latent representations learned by the encoder module are fed parallelly to the classifier and the auxiliary task module. Thus, the network can perform severity classification and the auxiliary task at the same time. All weights of the network are initialized randomly and subsequently adapted through joint minimization of loss functions associated with severity classification and the auxiliary task. Equal weightage is given to both the losses.
3. **Multi-Task Learning with Pre-training and Layer Freezing (MTLp):** This approach is a combination of the previous two approaches. The first step, in this case, is same as the pre-training step of FT. In the second step, a randomly initialized classifier block is added to the pre-trained network in parallel to the auxiliary task module, thus making the network architecture same as in the case of MTL. The weights of the auxiliary task module are freezed at the pre-trained values while the pre-trained encoder weights and the randomly initialized classifier weights are fine-tuned by jointly minimizing the loss functions associated with severity classification and the auxiliary task. In this case also, equal weightage is given to both the losses.

Auxiliary tasks and losses

In all the above-mentioned schemes, input Feature (i.e. MFCC statistics) Reconstruction (FR) and 2-class Gender Classification (GC) are explored as the auxiliary tasks. The FR task might help the latent representations learned by the encoder to retain most of the important information present in the input features. On the other hand, dysarthria might cause gender-specific acoustic changes to the patients' speech. Thus performing GC as the auxiliary task might help in efficiently capturing such information in the learned latent representations. For FR, mean squared error between the input and reconstructed features is used as the auxiliary loss function, whereas, for GC, binary cross-entropy serves as the auxiliary loss.

Learning within and across datasets

All the transfer learning approaches considered in this work can be implemented with two different setups of data usage. In the first condition, only the ALS dataset is used for all auxiliary tasks as well as the primary task of severity classification. Thus, in this setup, transfer of knowledge occurs within the same dataset from one task scenario to the other. This method attempts to regularize the network by training it for multiple goals. As a result, the network becomes more generalizable. In the second condition, some auxiliary datasets of speech recordings obtained from HC subjects are employed for transfer learning purposes. Pre-trainings for auxiliary tasks, as required in the cases of FT and MTLp, are performed on the auxiliary datasets. Moreover, during joint learning of primary and auxiliary tasks, as in the cases of MTL and MTLp, a portion of the auxiliary dataset is considered along with the ALS dataset. In order to fit this auxiliary HC data in the framework of severity classification, we modify the severity classification task from being 3-class to 4-class classification problem with the 4^{th} class being speech from HCs. Equal number of HC subjects as in the ALS severity class with least subjects are randomly chosen from the

auxiliary dataset for this purpose. However, no auxiliary data is used during the testing phase. The ALS utterances which are predicted as healthy (i.e. the 4^{th} class) during testing are relabeled as normal (N) speech class of ALS. In case of MTLp, two further sub-conditions are considered while using the auxiliary datasets. In MTLp1, the pre-training is done using auxiliary datasets but the network adaptation through multi-task learning is performed using only the ALS data, whereas, in MTLp2, auxiliary data are used in both pre-training and network adaptation.

3. Dataset

ALS dataset

Spontaneous speech recordings were collected from 120 ALS subjects at the National Institute of Mental Health and Neurosciences, India. Dysarthria severity of the subjects were annotated by three SLPs according to the speech component of ALSFRS-R scale. The mode of the three ratings was considered as the final severity score. Demographic details of the subjects are given in Table 1. The subjects had five different native languages, namely, Bengali, Hindi, Tamil, Telugu and Kannada, with approximately equal proportion of subjects belonging to each language. During data collection, the subjects were asked to talk about *a festival they celebrate* and *a place that they had recently visited* in their respective native languages for around one minute each. They were given a few minutes of preparation time. Further details about the data collection procedure and the recording setup can be found in [3]. The total durations of speech data recorded from subjects of each severity level are mentioned in Table 1. The hospital ethics committee reviewed and approved the data collection protocol. A consent form was also signed by each subject prior to data collection.

Auxiliary datasets for transfer learning

Apart from the ALS dataset, three other datasets are used in this work for the purposes of transfer learning. A spontaneous speech dataset collected in-house from 88 HC subjects has been considered. The speech tasks and recording protocols used for this dataset are identical to those of the ALS dataset. Similar to the ALS subjects, the HC subjects also had the same five native languages. Besides this, TIMIT [18] and Indic TIMIT [19] datasets containing read speech data in American and Indian English, resepectively, have also been considered. Further details about these three datasets are given in Table 2.

4. Experimental setup

Feature extraction

12D MFCC (excluding energy coefficient) with delta and double-delta measures constituting a 36D vector is computed from every 20 ms speech frame with 10 ms overlap. Frame-wise energy-based Voice Activity Detection (VAD) is then performed to identify and remove the silence frames. Both these steps are performed using the KALDI speech recognition toolkit [20]. During VAD, frames having log mel energy higher than the threshold of [$\tau_1 + \tau_2$*(mean log mel energy of the utterance)] are marked as speech and the rest as silence. We set $\tau_1 = 5$ and $\tau_2 = 0.5$ which are the default values used in KALDI. Temporal statistics vectors, namely, mean, median, Root Mean Square (RMS) value and Standard Deviation (SD), of MFCC are then computed over all speech frames of an utterance. These statistics vectors are concatenated leading to a 144-D feature vector for each utterance. Lastly, z-score normalization is applied individually on each dimension of the feature vector using the mean and SD obtained from the train set.

Table 1: *Severity-wise subject demography and recorded speech data duration for ALS dataset*

Severity class	**Severe (S)**		**Mild (M)**		**Normal (N)**
ALSFRS-R	**0**	**1**	**2**	**3**	**4**
#M:#F	9:13	12:6	15:5	11:9	27:13
Mean (SD) of age (years)	58.55 (1.14)	56.63 (1.20)	51.10 (1.08)	54.45 (1.04)	52.28 (0.76)
Speech duration (min)	32.04	36.58	41.96	39.60	86.04

Model description

All neural network blocks used in this work are dense networks. The architectures of these blocks are as follows.

1. **Encoder:** It is a 2-layer dense network with each layer having 128 neurons and ReLU activation function. It takes the 144-D feature vectors of MFCC statistics as input and generates 128-D latent representations. Batch Normalization is performed on the output of the first dense layer. Moreover, for the purpose of regularization, dropout with frequency 0.3 is also added after batch normalization and the second dense layer.
2. **Classifier:** This block comprises of a single dense layer with 3 neurons and softmax activation, except in the cases of MTL and MTLp with auxiliary data where this layer constitutes 4 neurons. It takes the 128-D latent representations as input and predicts the severity class labels.
3. **Auxiliary task module:** Two different architectures of this module are used for the two auxiliary tasks of FR and GC. In case of FR, this module is designed as a 2-layer dense network where the first layer has 128 neurons with ReLU activation and the second layer has 144 neurons with linear activation. A dropout layer with probability 0.3 is inserted after the first layer. This module takes the 128-D latent representations obtained by the encoder as input and generates the reconstructed feature vectors of dimension 144. For the task of GC, a single dense layer with 2 neurons and softmax activation is employed as the auxiliary task module. This module also takes the 128-D latent representations as input and predicts the gender labels.

All networks are trained using Adam optimizer with a learning rate of 0.001. The batch size is kept at 32. For each model, training is continued till a maximum of 100 epochs, while early stopping with a patience of 8 based on validation loss is imposed to avoid overfitting. All model implementations are done using Keras v1.0 [21]. An NVIDIA GeForce RTX 2080 GPU is used for training and testing the models.

Evaluation protocol

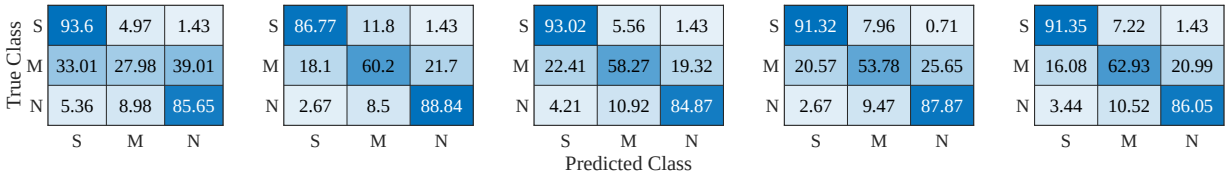
For all experiments, the ALS dataset is randomly split into training, validation and test sets containing 60%, 20% and 20% of the subjects, respectively. This random splitting is done 10 times to facilitate 10-fold validation. The mean and SD of the balanced classification accuracies for the 3-class severity classification obtained over this 10-fold validation are reported as the performance metrics. For transfer learning using Indic TIMIT

Table 2: *Subject demography and recorded speech data duration for auxiliary datasets*

Dataset	**HC data**	**Indic TIMIT**	**TIMIT**
#M:#F	67:21	39:41	438:192
Mean (SD) of age (years)	43.02 (9.13)	25.42 (6.05)	29.78 (8.09)
Speech duration (hours)	2.90	234.47	5.38

Table 3: *Mean balanced classification accuracies in % (SD in bracket) obtained over 10-folds of random validation using different network training schemes; here, * indicates the approaches which outperform STDL at 1% significance level and # indicates that FR outperforms GC as the auxiliary task at 1% significance level*

Auxiliary data	**Auxiliary task**	**STDL**			
-	-	69.08 (3.66)			
		FT	**MTL**	**MTLp1**	**MTLp2**
-	FR	77.14 (6.53)*	75.50 (3.91)*	**77.66 (3.47)***	-
	GC	74.30 (5.82)	75.44 (5.46)	73.17 (6.32)	-
HC data	FR	76.82 (4.98)*	74.88 (5.61)	76.28 (4.47)*	76.56 (6.37)
	GC	74.58 (4.39)*	73.70 (3.69)*	74.23 (7.16)	74.41 (4.78)*
Indic TIMIT	FR	**78.60 (6.52)***	75.88 (4.68)*	77.38 (3.75)*	75.41 (3.96)#*
	GC	71.22 (6.59)	75.45 (4.58)	71.56 (4.38)	71.02 (5.79)
TIMIT	FR	75.75 (5.34)*	**78.72 (6.89)***	75.75 (6.79)*	**80.11 (3.80)***
	GC	77.19 (4.26)*	77.52 (5.51)*	75.34 (3.66)*	76.60 (5.48)*



(a) STDL (b) FT (FR + Indic TIMIT) (c) MTL (FR + TIMIT) (d) MTLp1 (FR + ALS data) (e) MTLp2 (FR + TIMIT)

Figure 2: *Confusion matrices averaged over 10-folds of random validation for STDL and the best performing configurations of the transfer learning approaches; here, S: Severe class, M: Mild class, N: Normal class; the entry in the cell (i, j) of each matrix indicates the % of samples of true class i which are being classified as class j*

and HC datasets, random 80% of the subjects are chosen to form the training set while the rest 20% form the validation set. In case of TIMIT, the train-test splits given by the dataset authors are used as the training and validation sets. Thus, the train set contains 462 subjects whereas the validation set has 168 subjects. No testing is done using these auxiliary datasets. Lastly, Wilcoxon signed-rank test [22] at 1% significance level is carried out to determine if the balanced classification accuracies obtained using different approaches are significantly different.

5. Results and Discussion

Table 3 summarizes the dysarthria severity classification performances achieved using STDL and different transfer learning approaches considered in this work. All transfer learning schemes are observed to achieve higher mean accuracies than STDL. The improvements are significant at 1% significance level for 13 out of 15 configurations with FR as the auxiliary task and 7 out of 15 configurations with GC as the auxiliary task. The best average classification accuracy of 80.11% (11.03% higher than the accuracy obtained using STDL) is achieved using MTLp2 with FR as the auxiliary task and TIMIT as the auxiliary dataset. These observations indicate that transfer learning indeed aids dysarthria severity classification.

Table 3 also shows that the performances achieved using the two auxiliary tasks are statistically similar in all cases except MTLp2 with Indic TIMIT as the auxiliary dataset. In this case, FR task outperforms GC at 1% significance level. However, the average accuracies achieved using FR tasks are higher than those obtained using GC tasks in all cases except FT approach with TIMIT as the auxiliary dataset. Another observation evident from Table 3 is that, while using transfer learning, the performances obtained with or without employing the auxiliary datasets are statistically similar. Though the HC data is matched to the ALS data in terms of the speech task and languages, TIMIT and Indic TIMIT are essentially different. These datasets have read speech in American and Indian English as opposed to spontaneous speech in multiple Indian languages as present in the ALS dataset. Thus, languages and speech tasks do not appear to be a barrier while performing transfer learning. Moreover, for a particular configuration of auxiliary task and dataset, the performances of all the four transfer learning approaches are found to be statistically similar.

Figure 2 illustrates the average confusion matrices obtained using STDL and the best performing configurations of FT, MTL, MTLp1 and MTLp2 approaches. Correctly classifying the utterances belonging to the mild class appears to be the primary challenge in the case of STDL. 72.02% utterances of this class are mis-classified as the other two classes. The transfer learning configurations significantly improve the performance on the mild class, thereby driving the confusion matrices towards being more diagonally dominant. While doing so, the performances on the severe and normal classes degrade in a few cases. However, the decline is not significant.

6. Conclusion

This paper empirically confirms the suitability of transfer learning approaches to circumvent the low data availability constraint encountered in developing dysarthria severity classification systems for patients with ALS. The primary benefit is found to be obtained in terms of better classification of utterances belonging to the mild dysarthric class. Moreover, on average, FR tends to perform better than GC as the auxiliary task. No significant effect of having language or speech task mismatches between the ALS data and the auxiliary datasets is observed. As our future work, we plan to explore wider varieties of auxiliary tasks and more complex architectures of the networks. We also plan to perform 5-class dysarthria severity classification, thereby directly predicting the ALSFRS-R speech subscore.

Acknowledgements - We sincerely thank the subjects for their valuable speech contributions. We are grateful to Seena Vengalil and M S Keertipriya for their assistance during data collection. We also thank the Department of Science and Technology (DST), Govt. of India for supporting this work.

7. References

[1] 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.

[2] B. Suhas, D. Patel, N. R. Koluguri, Y. Belur, P. Reddy, N. Atchayaram, R. Yadav, D. Gope, and P. K. Ghosh, "Comparison of speech tasks and recording devices for voice based automatic classification of healthy subjects and patients with Amyotrophic Lateral Sclerosis," in *Proc. Interspeech*, 2019, pp. 4564–4568.

[3] 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.

[4] M. Vashkevich and Y. Rushkevich, "Classification of ALS patients based on acoustic analysis of sustained vowel phonations," *Biomedical Signal Processing and Control*, vol. 65, p. 102350, 2021.

[5] S. BN, J. Mallela, A. Illa, Y. BK, N. Atchayaram, R. Yadav, D. Gope, and P. K. Ghosh, "Speech task based automatic classification of ALS and Parkinson's disease and their severity using log mel spectrograms," in *International Conference on Signal Processing and Communications (SPCOM)*. IEEE, 2020, pp. 1–5.

[6] A. Wisler, K. Teplansky, J. R. Green, Y. Yunusova, T. Campbell, D. Heitzman, and J. Wang, "Speech-based estimation of bulbar regression in Amyotrophic Lateral Sclerosis," in *Proceedings of the Eighth Workshop on Speech and Language Processing for Assistive Technologies*, 2019, pp. 24–31.

[7] F. Xiong, J. Barker, Z. Yue, and H. Christensen, "Source domain data selection for improved transfer learning targeting dysarthric speech recognition," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7424–7428.

[8] T. Mariya Celin, P. Vijayalakshmi, and T. Nagarajan, "Data augmentation techniques for transfer learning-based continuous dysarthric speech recognition," *Circuits, Systems, and Signal Processing*, vol. 42, no. 1, pp. 601–622, 2023.

[9] C. Ding, S. Sun, and J. Zhao, "Multi-task transformer with input feature reconstruction for dysarthric speech recognition," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 7318–7322.

[10] Q. Yu, Y. Ma, and Y. Li, "Enhancing speech recognition for Parkinson's disease patient using transfer learning technique," *Journal of Shanghai Jiaotong University (Science)*, pp. 1–9, 2022.

[11] S. Tejaswi and S. Umesh, "DNN acoustic models for dysarthric speech," in *23^{rd} National Conference on Communications (NCC)*, 2017, pp. 1–4.

[12] D. Korzekwa, R. Barra-Chicote, B. Kostek, T. Drugman, and M. Lajszczak, "Interpretable deep learning model for the detection and reconstruction of dysarthric speech," in *Proc. Interspeech*, 2019, pp. 3890–3894.

[13] 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.

[14] Y. Li, X. Zhang, P. Wang, X. Zhang, and Y. Liu, "Insight into an unsupervised two-step sparse transfer learning algorithm for speech diagnosis of Parkinson's disease," *Neural Computing and Applications*, vol. 33, pp. 9733–9750, 2021.

[15] A. A. Joshy and R. Rajan, "Dysarthria severity classification using multi-head attention and multi-task learning," *Speech Communication*, vol. 147, pp. 1–11, 2023.

[16] J. C. Vásquez-Correa, T. Arias-Vergara, J. R. Orozco-Arroyave, and E. Nöth, "A multitask learning approach to assess the dysarthria severity in patients with Parkinson's disease." in *Proc. Interspeech*, 2018, pp. 456–460.

[17] M. Soleymanpour, M. T. Johnson, and J. Berry, "Increasing the precision of dysarthric speech intelligibility and severity level estimate," in *Speech and Computer: 23^{rd} International Conference, SPECOM, Proceedings 23*. Springer, 2021, pp. 670–679.

[18] 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.

[19] 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.

[20] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovský, G. Stemmer, and K. Veselý, "The Kaldi speech recognition toolkit," in *Workshop on automatic speech recognition and understanding*. IEEE Signal Processing Society, 2011, iEEE Catalog No.: CFP11SRW-USB.

[21] F. Chollet, "Keras," https://github.com/fchollet/keras, 2015, [Online; accessed 05-Mar-2023].

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