Source and Filter Characteristics Based Transfer Learning for Dysarthria Severity Classification in Amyotrophic Lateral Sclerosis

Tanuka Bhattacharjee[a] , Yamini Belur[b], Atchayaram Nalini[b], Prasanta Kumar Ghosh[a]*

[a]Electrical Engineering Department, Indian Institute of Science, C V Raman Avenue, Bengaluru, Karnataka 560012, India

[b]National Institute of Mental Health and Neurosciences, Hosur Road, Bengaluru, Karnataka 560029, India

tanuka1111@gmail.com, yaminihk@gmail.com, atchayaramnalini@yahoo.co.in, prasantg@iisc.ac.in

Abstract

Dysarthria due to Amyotrophic Lateral Sclerosis (ALS) progressively affects both source and filter components of speech. We analyze the discriminative abilities of cues related to these components for automatic dysarthria severity classification specific to ALS, which is not yet well explored. We manipulate speech utterances using WORLD vocoder to retain only the desired component while removing other attributes. Temporal statistics of mel-frequency cepstral coefficients (MFCC) extracted from these modified utterances act as the input speech features. However, the primary challenge in developing dysarthria severity classification systems is the limited availability of data. Though transfer learning is often utilized to mitigate this issue, only one such effort has been reported in the context of ALS-related dysarthria. We propose to use transfer learning, specifically, fine-tuning from an auxiliary task and multi-task learning, with novel source and filter characteristics based auxiliary tasks. These tasks involve reconstructing or predicting source, filter and overall attributes from the input speech features on the same ALS speech dataset or on some auxiliary dataset comprising healthy utterances of varied speech tasks in different languages. Experimental results suggest that filter related cues capture better discriminative information providing higher severity classification performances whether transfer learning is used or not. The transfer learning approaches improve the classification accuracies, especially for the mild dysarthria class, though the optimal auxiliary task and data configuration varies with the input speech feature and transfer learning protocol used. We achieve the highest mean classification accuracy of 85.92% which is 13.29% higher than the corresponding case without transfer learning.

Keywords: Amyotrophic Lateral Sclerosis, dysarthria severity, source-filter model, transfer learning

1. Introduction

Amyotrophic Lateral Sclerosis (ALS) is a rapidly progressing neurodegenerative disease that impairs motor functions, including speech. Approximately 30% of the ALS patients experience dysarthria, a motor speech disorder, as an early sign of the disease, and nearly all patients develop it at some point as the disease progresses [1]. Though there is no cure for ALS or the associated dysarthria to date, regular treatment, and individualized disease management strategies can help slow the disease progression, and enhance the quality of life of the patients. Continuous monitoring of the disease severity is crucial for tailoring therapeutic interventions to a patient's evolving needs. Clinically, Speech-Language Pathologists (SLPs) assess dysarthria severity in ALS patients using methods such as the Frenchay Dysarthria Assessment [2] and the Revised Amyotrophic Lateral Sclerosis Functional Rating Scale (ALSFRS-R) [3]. However, these assessments can be tedious, time-consuming, and costly. Moreover, factors like the clinician's familiarity with the patient or the subject matter being spoken may influence the judgment and introduce subjective biases in the assessment [4]. Therefore, there is a pressing need for objective and consistent automatic systems for predicting dysarthria severity.

This paper aims to perform automatic 3-class dysarthria severity level classification for patients with ALS where the classes under consideration are *severe dysarthria* (SV), *mild dysarthria* (ML), and *normal speech* (NS). We propose to exploit cues related to the source and the filter components of speech for this purpose. According to the source-filter model [5], speech is produced by passing a source signal through a linear time-varying filter. The source signal models the glottal and/or supra-glottal excitation caused by the vibration, abduction, and adduction of the vocal folds, whereas, the filter models the effect of the vocal tract and the lip radiation. Dysarthria due to ALS impairs both the source and the filter components of speech utterances. Erroneous voicing, abnormal prosodic patterns like reduced pitch range, and poor voice quality are some signs suggesting source impairment [6]. Poor laryngeal control and compromised respiratory functionality are the primary causes of such source impairments [6]. On the other hand, imprecise and irregular articulations, voice nasality as well as atypical spectral characteristics of speech utterances indicate impairments in the filter component [1]. Restricted articulatory mobility and dysfunctions in the resonatory sub-system of speech give rise to such impaired filter functions [6]. Impairments creep in different speech sub-systems at different severity levels during the course of progression of dysarthria. The degree of involvement of different speech sub-systems also increases with increasing severity. As a result, the forms and extents of the impairments in the source and the filter components of speech also vary with the dysarthria severity level. Hence, the source and filter representations obtained from the speech utterances can be highly informative for automatic dysarthria severity classification.

* Corresponding author

A few efforts have been reported in the literature on speech-based automatic dysarthria severity classification for ALS. Different speech representations, like raw time-domain speech signal and log-mel spectrogram, have been explored [7, 8]. Though these representations implicitly contain both source and filter related information, no study has yet been conducted to explicitly examine the utility of source and filter specific features for ALS-related dysarthria severity classification. This paper proceeds to look into that angle.

The typical severity classification systems are trained in a supervised manner using speech data and corresponding severity labels. Researchers have used different machine learning and deep learning systems for this purpose [7, 8, 9]. A major challenge in building such systems is the limited availability of speech data from dysarthric speakers of different severity levels, particularly the more severe ones. Collecting speech samples from individuals with speech impairments is both challenging and time consuming. The process becomes even more demanding as clinical annotations for severity levels of these speech samples are required. Scarcity of data resources not only hampers the efficiency of the classifiers but also often leads to overfitting of the classifiers to the small amount of training data. This, in turn, compromises the classifiers' ability to generalize to unseen test conditions. Therefore, it is crucial to employ data-efficient techniques for training the dysarthria severity classification systems.

In the domain of deep learning, transfer learning is a widely used approach to address the issue of limited data availability. This method leverages the knowledge gained from auxiliary tasks or datasets to aid model training for the primary objective on the primary (often small) dataset. This technique has been successfully applied in various areas of dysarthric speech research, including dysarthric speech recognition [10, 11, 12, 13, 14], dysarthric speech enhancement [15], classification between dysarthric and healthy speech [16], ALS vs. healthy control (HC) classification [17], and Parkinson's Disease (PD) detection [17, 18]. Though a few studies have explored transfer learning methods to classify dysarthria severity in PD and Cerebral Palsy (CP) [19, 20], our previous work [21] is the only one which uses transfer learning for dysarthria severity classification specific to ALS. In the current work, we propose to leverage information related to the source and the filter components of speech in the transfer learning framework by introducing novel source and filter characteristics driven auxiliary tasks. Since cues related to dysarthria severity are expected to be embodied in both the source and the filter components, such auxiliary tasks can help in reinforcing discriminative information in the learned representations, which in turn, can complement the severity classification performance. A comparative analysis of the auxiliary tasks can also help us understand reinforcement of information related to which component is more beneficial for the severity classification purpose. No such effort has yet been made in the domain of dysarthria severity classification for either ALS or other diseases.

To capture the effects of either the source or the filter component in speech, we propose to manipulate the utterances using the WORLD vocoder [22] such that only the required component is retained in the modified utterances while suppressing the redundant attributes. We extract MFCC from the modified utterances. MFCC extracted from the utterances containing only the source information are denoted here as S-MFCC and those obtained from the utterances containing only the filter effects are denoted here as F-MFCC. We also consider MFCC obtained from the original utterances, referred to as O-MFCC in this paper, which incorporates information about both the source and the filter components of speech [23]. Temporal statistics of each of these three representations are utilized as the inputs for performing the 3-class dysarthria severity classification using dense neural networks (DNNs). We employ two types of transfer learning approaches, namely, fine-tuning from auxiliary tasks and multi-task learning, as well as their combinations, for aiding the classifier training. As auxiliary tasks in the transfer learning frameworks, we propose to predict the temporal statistics of one or all of S-MFCC, F-MFCC, and O-MFCC from the temporal statistics of the input representation at hand. Learning to reconstruct or predict these component-specific or overall representations as the auxiliary tasks can help in obtaining latent representations where source and filter based discriminative cues are reinforced. In one scenario, transfer learning is executed within the ALS speech dataset, without using any auxiliary dataset, by utilizing only the auxiliary tasks. In another setup, we leverage knowledge from auxiliary datasets of healthy speech through the auxiliary tasks. These approaches can improve the performance as well as the generalizability of the severity classifier. We explore different auxiliary datasets in this context to analyze the effect of using varied speech tasks and languages, same as or different from those of the ALS dataset, for performing the transfer learning.

2. Literature Survey and Contribution of Current Work

Efforts have been made in the literature to exploit acoustic properties of speech for severity classification of ALS-induced dysarthria. Suhas et al. [7] have developed a 2D Convolutional Neural Network (CNN) for this purpose. They found log-mel spectrogram to outperform O-MFCC as the input speech representation for the proposed system. Vieira et al. [8] have designed a CNN model to predict the ALSFRS-R bulbar subscore directly from raw speech signals. Wisler et al. [9] employed both acoustic and articulatory data to estimate the ALSFRS-R bulbar subscore using linear ridge regression and support vector regression.

Though very few efforts have been made to perform dysarthria severity classification specific to ALS, several other works exist which deal with severity classification of dysarthria caused by other disorders like CP. The publicly available UA-speech [24] and TORGO [25] datasets are usually considered in these works. Different spectro-temporal speech representations, like log mel spectrogram, raw magnitude and phase spectra, MFCC, constant-Q cepstral coefficients, perceptually enhanced Fourier transform spectrograms and Constant-Q transform spectrograms, have been utilized together with different machine learning and deep learning based classifiers, like Gaussian mixture model (GMM), support vector machine (SVM), DNN, CNN, gated recurrent units (GRU), long short term memory networks (LSTM), and residual neural network [26, 27, 28, 29, 30, 31]. Glottal features have also been used [32, 33]. CNN models with squeeze-and-excitation networks and residual blocks have been proposed in [34] for dysarthria severity classification using mel spectrograms. Variable continuous wavelet transform layered CNN [35] and variable short-time Fourier transform layered CNN [36] have also been explored. Javanmardi et al. [37] have used speech representations obtained using pre-trained wav2vec2-BASE [38], wav2vec2-LARGE [38], and HuBERT [39] models for automatic detection and severity classification of dysarthria. They have explored SVM and CNN as the classifiers. All these approaches suffer from the challenge of limited availability of ALS speech data for training.

Another group of researchers have explored the accuracy of off-the-shelf automatic speech recognition (ASR) systems as potential markers for grading speech intelligibility and severity of dysarthric speech [40, 41]. They have observed significant correlation between

ASR performance and speech intelligibility or severity of dysarthric speech. Since off-the-shelf ASR models are generally trained on speech data obtained from healthy individuals, the accuracy of these models degrades as the speech becomes more atypical or unintelligible with increasing dysarthria severity. A major advantage of this approach is that significantly less amount of ALS speech data are needed. However, Gutz et al. [42] claimed word error rate of ASR systems to be insufficient for grading dysarthria severity for ALS.

Several works reported in the literature have demonstrated superior severity prediction results while employing transfer learning methods. Vásquez et al. [19] have proposed a CNN-based multi-task learning framework for assessing dysarthria severity in PD patients. They have incorporated eleven auxiliary tasks including PD vs. HC classification and assessments of the degree of impairment of different articulators like lips, palate, tongue, and larynx. Soleymanpour et al. [43] have trained a 1D CNN for dysarthria severity assessment using cross-dataset transfer learning. Joshy et al. [20] have explored multi-head attention in combination with multi-task learning to classify dysarthria severity in individuals with CP. They have considered gender, age, and disorder-type identification as the auxiliary tasks. Chowdary et al. [44] have performed few-shot learning using pre-trained transformer based whisper-large-v2 model [45] for detection and severity classification of dysarthria. Different variants of Vision Transformers [46] with a classification head of 6 linear layers are fine-tuned in [47] for the severity classification purpose. In our previous work [21], we have explored the approaches of fine-tuning from auxiliary tasks and multi-task learning, as well as their combinations, to perform dysarthria severity classification for ALS. Input feature reconstruction and gender classification, on the same ALS speech dataset or other available healthy speech corpora, have been explored as the auxiliary tasks. We have used temporal statistics of O-MFCC as the input features and DNNs as models for performing the primary and auxiliary tasks. To the best of our knowledge, no effort except this work [21] has yet been made to apply transfer learning techniques for dysarthria severity classification specific to ALS. Moreover, transfer learning leveraging source and filter information of speech also remains unexplored to date in this particular domain of research.

The scope and contributions of the work in this paper, in contrast to that presented in [21], can be summarized as follows.

- This study compares the discriminative abilities of source (S-MFCC), filter (F-MFCC), and overall (O-MFCC) cues for performing dysarthria severity classification, unlike only O-MFCC explored in [21].
- We propose source and filter characteristics based auxiliary tasks here, whereas, reconstruction of input O-MFCC statistics and gender classification have been used as the only auxiliary tasks in [21].
- In this work, we explore the effects of employing single and multiple auxiliary tasks in the transfer learning approach, as opposed to only single tasks in [21].
- In [21], the auxiliary datasets under consideration had only spontaneous speech (SPON) and read speech tasks. The current work explores auxiliary datasets having additional speech tasks, e.g. sustained phoneme production (PHON), diadochokinetic rate (DDK) task, and image description (IMAG) task, along with the ones used in [21].

3. Method

3.1. Estimation of source-filter representations

The proposed method of extracting source and filter representations of speech utterances comprises four steps, namely, decomposition, modification, synthesis and feature extraction, as illustrated in Figure 1.

First, an utterance is decomposed into fundamental frequency (f_0), spectral envelope (SP) and aperiodicity (AP) components using the WORLD analyzer [22]. The estimates of the components are obtained at a frame period of 5 ms. f_0 estimates are first extracted using the Harvest algorithm [48], where 50 Hz and 450 Hz are used respectively as the floor and ceiling frequencies for the f_0 estimation range [49]. The SP and AP estimates are then computed using the CheapTrick [50] and the D4C [51] algorithms, respectively. Following modifications are then applied to the f_0, AP, and SP components to retain either the source or the filter attributes in the signal synthesized subsequently.

1. To remove the effect of filter from an utterance, the estimated SP is modified to 1s in all frequency bands. Speech is then synthesized by WORLD synthesizer using the modified SP along with the unchanged f_0 and AP. This makes the filter allpass in nature, and

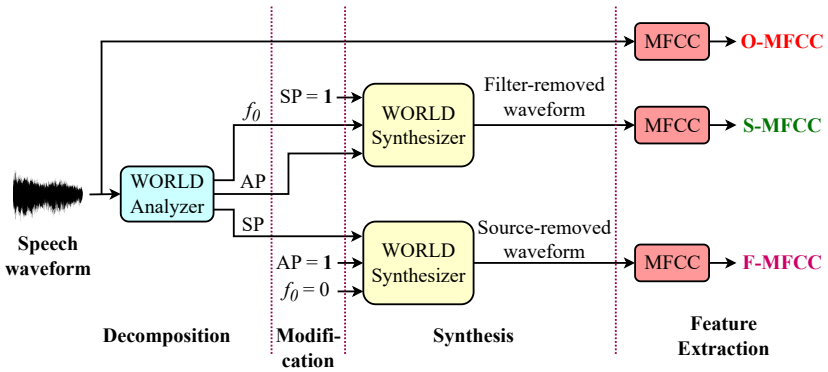


Figure 1: *Proposed method of extracting source, filter, and overall attributes from speech utterances using WORLD vocoder*

hence, the modified signal captures only source attributes without any influence of the vocal tract. We refer to this modified signal as the *filter-removed waveform*. MFCC features obtained from this modified utterance, referred to as S-MFCC, serve as the source representation.

2. To extract filter information without any influence of source, we devoice an utterance by replacing the f_0 estimates obtained from WORLD analyzer with 0s and the AP for all frequency bands with 1s [52]. Speech is then synthesized by WORLD synthesizer with the modified f_0 and AP along with the unchanged SP. This makes the source white throughout the utterance while retaining the filter characteristics. This modified signal is denoted as the *source-removed waveform*. MFCC computed from this modified signal serve as filter cues and are denoted as F-MFCC in this work.

We also extract MFCC from the original speech utterances which include information from both the source and the filter components of speech [23]. This is referred to as O-MFCC in this work.

Each of S-MFCC, F-MFCC, and O-MFCC comprise 12D MFCC (excluding energy coefficient) with delta and double-delta measures leading to 36D vectors. These are extracted from every 20 ms frame of the *source-removed waveforms*, the *filter-removed waveforms*, and the original speech, respectively, using the KALDI speech recognition toolkit [53]. A 10 ms overlap is maintained between consecutive frames. We extract temporal statistics, e.g. mean, median, Root Mean Square (RMS) value and standard deviation (SD), of these S-MFCC, F-MFCC, and O-MFCC over complete utterances. The temporal statistics vectors are concatenated to form 144-D feature vectors corresponding to each of S-MFCC, F-MFCC, and O-MFCC for every utterance. Lastly, each dimension of the feature vector undergoes z-score normalization using the mean and SD of the corresponding dimension derived from the training set.

3.2. Classification of dysarthria severity

We employ DNNs to perform the 3-class (SV vs. ML vs. NS) dysarthria severity classification. The feature vector comprising the temporal statistics of one of S-MFCC, F-MFCC, and O-MFCC extracted from a speech utterance is used as the input to the network. Figure 2 illustrates different approaches studied in this work for training the network. The paradigms are discussed briefly in our previous work [21].

3.2.1. Single task direct learning (STDL)

As a baseline framework of this study, we perform the severity classification without employing transfer learning. We refer to this approach as Single Task Direct Learning (STDL). Here, an encoder module is used to extract latent representations from the input speech features. These representations are then passed through a classifier module to predict the severity class label. The parameters of the encoder-classifier network are initialized randomly and subsequently learned jointly by minimizing the categorical cross-entropy loss of the 3-class severity classification.

3.2.2. Transfer learning protocols

This study explores 3 different transfer learning paradigms - (1) fine-tuning (FT), (2) multi-task learning (MTL), and (3) multi-task learning with pre-training and layer freezing (MTLp).

In FT, a randomly initialized network, comprising an encoder module (like STDL) followed by an auxiliary task module, is first pre-trained for an auxiliary task. The auxiliary task module is then replaced with a randomly initialized classifier module (as in STDL) while retaining the pre-trained encoder. Finally, the complete network (encoder and classifier) is fine-tuned together for severity classification by minimizing the cross-entropy loss. In this protocol, the pre-training phase may help the encoder learn some useful information by virtue of the auxiliary task which is typically related to the primary objective. Thus, this phase provides a more informed initialization to the encoder-classifier network during the fine-tuning phase and hence may benefit the severity classification performance. Moreover, pre-training has a regularization effect which helps in improving the generalizability of the severity classification network to unseen data [54].

In MTL, the latent representations obtained at the output of the encoder module (as in STDL) are processed parallely by the

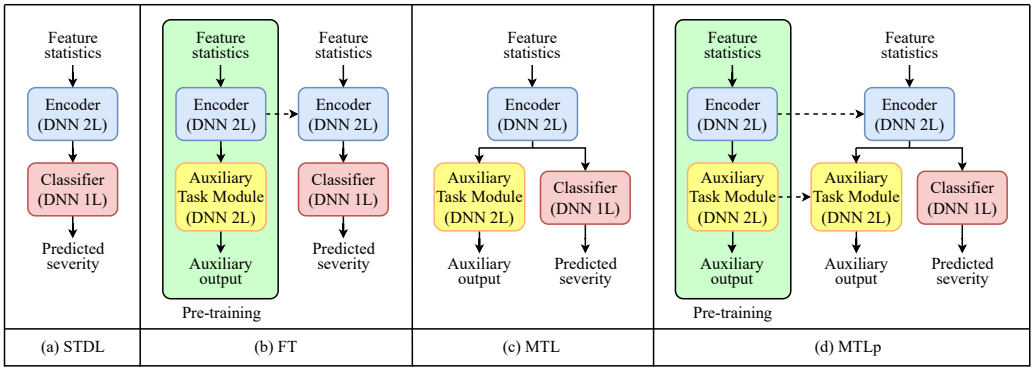


Figure 2: *Block diagram of different protocols used for network training; here, dotted arrows indicate that the weights learned for a module during pre-training are used to initialize the same module in the next step, and DNN xL stands for DNN with x layers*

classifier and the auxiliary task module for simultaneously performing severity classification and the auxiliary task. Starting from a random initialization, the complete network comprising the encoder, classifier, and the auxiliary task module is adapted by jointly minimizing the loss functions associated with severity classification and the auxiliary task. Both losses are given equal weightage. In this protocol, since the network is trained on multiple related tasks at once, the latent representation learned by the encoder can capture more robust and general features which can be shared across the tasks. Thus the auxiliary tasks act as regularizers to the primary task of severity classification, reducing the risk of overfitting to the noise or peculiarities of the feature to label mapping in the training dataset of the severity classification task [55]. As a result, the generalizability of the severity classification network to unseen data improves significantly [55].

MTLp integrates the best of both FT and MTL. First, a pre-training step is performed in the same fashion as FT. Next, a randomly initialized classifier module is added in parallel to the pre-trained auxiliary task module in the pre-trained network, forming the same architecture as in MTL. While the auxiliary task module is frozen to its pre-trained weights, the encoder and classifier modules are fine-tuned through joint minimization of the severity classification loss and the auxiliary task loss, both given equal weightage. Thus the pre-training phase provides the final network with an informed initialization, or in other words, some prior knowledge about the source and filter attributes. MTL then exploits the auxiliary tasks for the purpose of learning robust source, filter, or overall cues in the latent representation which can further benefit the primary task of severity classification. Moreover, regularization effects from both pre-training and MTL come into play, thereby improving the generalizability of the classifier.

3.2.3. Auxiliary tasks

We consider two types of auxiliary tasks in this work.

1. **Input feature reconstruction:** Here, the same speech feature that is used as the input to the network has to be reconstructed at the output. We refer to this auxiliary task as S-S, F-F, and O-O, when the input speech feature under consideration is temporal statistics of S-MFCC, F-MFCC, and O-MFCC, respectively. This task can help the encoder to concisely retain the crucial properties of the input feature in the learned latent representation, which can further aid the severity classification performance.
2. **Other feature prediction:** In this case, we propose to predict the temporal statistics of one of S-MFCC, F-MFCC, and O-MFCC which is not used as the input speech feature to the network. For example, when S-MFCC statistics are used as the input feature, we propose to predict the temporal statistics of F-MFCC or O-MFCC as the auxiliary tasks. We denote these two tasks as S-F and S-O, respectively. Similarly, the tasks of predicting the S-MFCC and O-MFCC statistics from those of F-MFCC are referred to as F-S and F-O, respectively, and the tasks of predicting the S-MFCC and F-MFCC statistics from those of O-MFCC are denoted as O-S and O-F, respectively.

In all of these auxiliary tasks, the mean squared error (MSE) between the predictions and the ground truths for the required output feature statistics is used as the auxiliary loss function. These tasks might help in reinforcing essential cues related to the source, filter or both components of speech in the learned latent representations. Since both of these components get affected progressively at different severity levels of ALS-induced dysarthria, cues related to these components when reinforced in the latent representations can aid in better classification of dysarthria severity.

Moreover, we perform transfer learning with single or multiple auxiliary tasks. In the single auxiliary task case, one of the nine auxiliary tasks S-S, S-F, S-O, F-S, F-F, F-O, O-S, O-F, and O-O is incorporated in the learning framework. In the multiple auxiliary task case, we propose to reconstruct/predict the temporal statistics of all of S-MFCC, F-MFCC, and O-MFCC from the input feature at hand. These tasks are respectively named as S-SFO, F-SFO, and O-SFO when the temporal statistics of S-MFCC, F-MFCC, and O-MFCC are used as the input features. Equal weightage is given to the primary loss and all auxiliary losses.

3.2.4. Learning within and across datasets

We explore two different setups of data usage while carrying out the transfer learning processes at hand.

1. The first setup relies solely on the ALS dataset for both severity classification as well as the auxiliary tasks. Here knowledge transfer occurs within the same dataset by leveraging the auxiliary tasks only.
2. In the second setup, auxiliary speech datasets obtained from HC subjects are used for transfer learning. For FT and MTLp, pre-training for auxiliary tasks are conducted using the auxiliary HC data. Moreover, in the cases of MTL and MTLp, a subset of the auxiliary dataset is incorporated alongside the ALS dataset during the joint learning of primary and auxiliary tasks. To integrate this additional HC data into the severity classification framework, the task is modified from a 3-class to a 4-class classification problem, with HC speech forming the additional class. Equal numbers of HC subjects as in the least-represented dysarthria severity class are selected randomly for the joint learning purpose. No auxiliary data is used during the testing phase and the ALS utterances classified as HC speech are reassigned to the NS class of ALS. Besides these, two variations of MTLp are explored in this work, as listed next.
   (a) **MTLp1:** Here, the network adaptation through multi-task learning leverages only the ALS data after pre-training is performed using the auxiliary data.
   (b) **MTLp2:** Here, pre-training is conducted using the auxiliary data while network adaptation through multi-task learning involve both the ALS dataset and the auxiliary HC data.

3.2.5. Network architecture

As shown in Figure 2, all aforementioned encoder, classifier, and auxiliary task modules are built using DNNs. The encoder is a 2-layer dense network with 128 neurons and ReLU activation in each layer. It takes the 144-D temporal statistics of S-MFCC, F-MFCC, or O-MFCC as input and generates 128-D latent representations. The classifier is a single dense layer with softmax activation which

takes the 128-D latent representations as input and predicts the severity class labels. This classifier layer has 3 neurons in all cases except MTL and MTLp with auxiliary data where this layer has 4 neurons. The auxiliary task module contains 2 dense layers with 128 neurons and ReLU activation in the first layer, and 144 neurons and linear activation in the second layer. This module takes the 128-D latent representations as input and predicts the 144-D temporal statistics of S-MFCC, F-MFCC, or O-MFCC depending on the auxiliary task at hand. For transfer learning with multiple auxiliary tasks, three units of this auxiliary task module are used in parallel to predict the temporal statistics of all three representations. Dropout layers with probability 0.3 are inserted after every layer except the output one in the networks formed using the encoder, classifier, and auxiliary task modules in order to regularize the training. Also, batch normalization is performed on the output of the first dense layer of the encoder module before applying the dropout layer.

The networks are trained using the Adam optimizer with a learning rate of 0.001 and a batch size of 32. Training continues for a maximum of 100 epochs. To prevent overfitting, early stopping is implemented with a patience of 8 determined by validation loss.

4. Dataset

4.1. ALS dataset

In this study, we perform dysarthria severity classification using spontaneous speech (SPON) utterances obtained from ALS subjects. The ALS dataset being utilized here is identical to the one used in our previous work [21]. Data collection was performed at the National Institute of Mental Health and Neurosciences (NIMHANS), Bengaluru, India. Diagnosis of ALS for the subjects was done by neurologists following the Revised El-Escorial criteria [56]. During data collection, the subjects were instructed to speak in their respective native languages for approximately one minute each on two topics, namely, *a festival they celebrate* and *a place they recently visited*. They were also given a few minutes to prepare before speaking. The subjects spoke on one or both of the topics depending on their level of comfort. A Zoom H-6 recorder with XYH-6 stereo X/Y capsule high quality unidirectional microphone [57] was used for recording the speech data. The device was placed at a distance of 2 feet from the subject. Data were recorded at 44.1 kHz and subsequently downsampled to 16 kHz for all further processing. Three SLPs rated the dysarthria severity of the subjects by listening to the recorded SPON utterances. The rating followed a 5-point scale [0 (loss of useful speech) - 4 (normal speech)], consistent with the speech function item of ALSFRS-R [3]. The final severity score was determined by taking the mode of the 3 SLPs' ratings. The subjects, who received 3 different ratings from 3 SLPs, were not included. The data collection protocol was approved by the hospital ethics committee, and each participant provided written consent before data collection.

The final dataset used in this work comprises 3.94 hours of SPON data (227 utterances) obtained from 120 ALS subjects (74 male, and 46 female). The subjects had a mean age of 54.26 years with a SD of 11.19 years. The native languages of the subjects include Bengali, Hindi, Tamil, Telugu and Kannada, with approximately equal proportion of subjects belonging to each language. The severity ratings given by all 3 SLPs agreed in the case of 53.33% of these subjects, whereas, ratings given by 2 out of the 3 SLPs matched for the remaining subjects. The inter-rater agreement, as measured by Fleiss' kappa [58], is 0.6006 indicating *moderate agreement* among the raters. We have 22 (9 male, and 13 female), 18 (12 male, and 6 female), 20 (15 male, and 5 female), 20 (11 male, and 9 female), and 40 (27 male, and 13 female) subjects with the final severity scores as 0, 1, 2, 3, and 4, respectively. We group severity scores 0 and 1 in the SV class, scores 2 and 3 in the ML class, and score 4 in the NS class. Thus, all three severity classes considered in this work have data from equal number (40) of subjects. The number of SPON utterances obtained from the three classes are 71, 78, and 78, respectively. Further details about the dataset can be found in [21].

4.2. Auxiliary datasets for transfer learning

Three datasets other than the ALS corpus are used in this work for the purposes of transfer learning.

1. **In-house HC (IHC) dataset:** This corpus contains speech data recorded in-house from 88 HC subjects (67 male, and 21 female) having mean and SD of age as 43.02 years and 9.13 years, respectively. These subjects had the same five native languages as the ALS subjects. The recording setup was also the same as that used for ALS dataset collection. Along with the SPON task, the HC subjects performed 3 other speech tasks, namely, PHON, DDK, and IMAG. In PHON, the subjects were asked to take a deep breath and produce a sustained utterance of a phoneme at comfortable f_0 and loudness levels. Four vowels, namely /a/, /i/, /o/, /u/, and three fricatives, namely /s/, /sh/, /f/, were considered for this task. In DDK, the subjects had to take a deep breath and keep repeating a monosyllabic or a tri-syllabic target as fast as they could without any interruption. Three monosyllabic targets, namely pa, ta, ka, and two tri-syllabic targets, namely pataka, badaga, were considered. Up to 3 utterances of each PHON and DDK task were recorded from a subject depending on his/her level of comfort. In IMAG, the subjects were asked to briefly describe images presented to them on a computer screen in their respective native languages. The images were selected from diverse domains like animals, food, festival, sports, daily life, natural phenomena etc. The number of images described by a subject varied depending on the subject's level of comfort. Further details about the speech tasks can be found in [59]. A total of 1596, 1189, 5762, and 175 utterances were collected from the HC subjects for the PHON, DDK, IMAG, and SPON tasks, respectively. The corresponding durations of speech data were 2.19, 1.91, 6.64, and 2.90 hours, respectively.
2. **IndicTIMIT [60]:** This dataset comprises 187360 read speech utterances in English, corresponding to 234.47 hours of speech data, collected from 80 native Indian HC speakers (39 male, and 41 female).
3. **TIMIT [61]:** This dataset contains 6300 read speech utterances in American English, corresponding to 5.38 hours of speech data, recorded by 630 native HC speakers (438 male, and 192 female).

The motivation behind exploring varieties of auxiliary datasets is to analyze how the language and the speech task of the auxiliary dataset impact the transfer learning. IHC SPON is the auxiliary dataset which is matched exactly to the ALS dataset in terms of both language and speech task. IHC PHON, IHC DDK, and IHC IMAG have speakers with same native and recording languages as those of the ALS dataset, but the speech tasks are different. IndicTIMIT differs more from the ALS dataset in the sense that, though the speakers are native Indian, the recording language is English in this dataset. Moreover, the speech task (read speech in this case) is also different

from the ALS dataset (SPON task). Lastly, we consider TIMIT as an auxiliary dataset having no similarity with the ALS dataset, i.e. both the native and recording languages of the speakers, as well as, the speech tasks performed are different. Thus exploration of this wide spectrum of auxiliary datasets can help us understand if having matched configurations of the primary and auxiliary datasets provides any added advantage for the purpose of transfer learning in our application.

5. Experimental setup

All experiments involving the ALS dataset are performed in a 5-fold cross-validation setup. The ALS dataset is divided into 5 disjoint folds. Each fold contains equal number of subjects, and hence nearly equal number of utterances, from each of the 3 severity classes. The distributions of age, gender, and language are similar across the folds. In every iteration of cross validation, data from 3 folds are used in training while data from 1 fold each are used in validation and testing. Thus, testing is done in unseen subject condition.

Different auxiliary datasets have different numbers of utterances, and hence, different number of data-points available for training and validation purposes. IHC SPON has the lowest number of utterances (i.e. 175), whereas, IndicTIMIT has the largest number (i.e. 187360). The larger datasets may have some advantages over the smaller ones in the context of transfer learning. Hence, in order to eliminate the effect of the number of data-points available, we consider the smallest auxiliary dataset IHC SPON entirely and select random subsets, containing equal number of utterances as IHC SPON, from all other auxiliary datasets. During the subset selection, the number of subjects in the subsets are maintained as close as possible to the number of subjects present in IHC SPON (i.e. 88). Since IHC PHON, IHC DDK, and IHC IMAG have same number of subjects as IHC SPON, we select at least one utterance from each subject for these auxiliary datasets. IndicTIMIT has a total of 80 subjects. In this case also, at least one utterance is chosen from each subject. On the other hand, since TIMIT comprises 630 subjects, we select the utterances from randomly chosen 88 subjects only (70 and 18, respectively, from the train and test partitions defined by the authors of the dataset). The subjects are chosen in a gender balanced manner.

All selected IHC datasets are randomly split into training and validation sets containing 70 and 18 subjects, respectively. For IndicTIMIT, the training and validation sets contain randomly chosen 64 and 16 subjects, respectively. In the case of TIMIT, the subjects chosen from the train split defined by the authors are considered to constitute the training set, while those chosen from the test split form the validation set. The entire training and validation splits of these auxiliary datasets are employed during pre-training using auxiliary data for FT and MTLp. Thus a single common pre-training step is performed for all 5 folds of cross-validation of the 3-class severity classification. During joint learning of primary and auxiliary tasks in MTL and MTLp, random subsets of these training and validation sets of the auxiliary datasets are used. The subsets contain same number of HC subjects as is present in the least represented severity class in the corresponding fold of cross-validation. No auxiliary data is used in the testing phases of cross-validation.

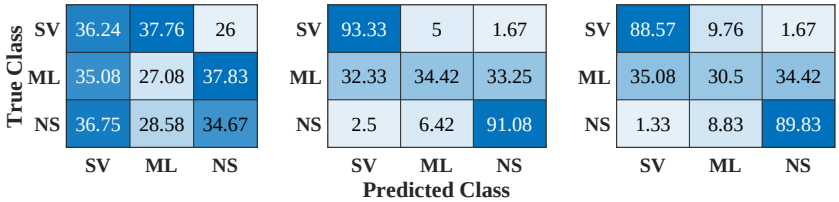
6. Results

6.1. STDL

Table 1 summarizes the classification performances obtained in the STDL framework, i.e. without leveraging any transfer learning, while using temporal statistics of S-MFCC, F-MFCC, and O-MFCC as the input features. The corresponding confusion matrices, averaged over the 5-folds of cross-validation, are illustrated in Figure 3. F-MFCC statistics achieves the highest mean severity classification accuracy of 72.63%. It is followed by O-MFCC with a 3.09% drop in the mean classification accuracy. S-MFCC statistics is observed to deliver 40.36% and 37.27% lower mean classification accuracy than F-MFCC and O-MFCC, respectively. Its mean classification accuracy is even less than the chance level accuracy for 3-class classification problems. The confusion matrices of Figure 3 further suggest that the ML class is most prone to mis-classification in all the three cases of the STDL approach. In fact, the SV and NS classes are identified with high degree of accuracy when F-MFCC and O-MFCC statistics are used as the input features, though that is not true for S-MFCC.

Table 1: *Mean classification accuracies in % (SD in bracket) obtained using different input speech representations in STDL*

S-MFCC	F-MFCC	O-MFCC
32.27 (5.78)	**72.63 (3.72)**	69.54 (4.29)



(a) S-MFCC (b) F-MFCC (c) O-MFCC

Figure 3: *Confusion matrices averaged over 5-folds of cross-validation for STDL with different input speech features; here, the entry in the cell (i, j) of each matrix indicates the % of samples of true class i which are classified as class j*

6.2. Transfer learning with single auxiliary task

Tables 2, 3, and 4 present the classification accuracies obtained using different transfer learning approaches with single auxiliary tasks where temporal statistics of S-MFCC, F-MFCC, and O-MFCC are used as the input speech features, respectively. Performances with different auxiliary data configurations are reported. For every input representation, transfer learning is observed to improve the severity classification performance as compared to the corresponding STDL case, irrespective of the transfer learning protocol, auxiliary task and auxiliary data configuration used.

6.2.1. S-MFCC statistics as input

A comparison of Tables 1 and 2 suggest that, for S-MFCC statistics as input, transfer learning results in a minimum improvement of 4.28% in the mean classification accuracy than STDL in the case of MTL with S-F as the auxiliary task and IHC SPON as the auxiliary data. The maximum improvement of 20.25% is achieved while using the MTLp2 configuration with S-F as the auxiliary task and IndicTIMIT as the auxiliary data. Figure 4(a) shows the average confusion matrix for this case. The performances on all classes are observed to improve, as compared to the STDL case with S-MFCC input (Figure 3(a)), while making the mis-classification percentages for SV and ML classes to be at par. However, even with this maximum improvement, the mean classification accuracy obtained using S-MFCC as the input (52.52%) is 20.11% and 17.02% lower than those of STDL with F-MFCC and O-MFCC as the inputs, respectively. Averaging over all transfer learning protocols, auxiliary tasks, and auxiliary data configurations, S-MFCC statistics achieves a mean severity classification accuracy of 42.81%.

Though the average performance of the four transfer learning protocols over all auxiliary task and data configurations are not largely different, the maximum mean classification accuracies for all auxiliary data configurations except IHC SPON are achieved using one of the three MTL based approaches. For all the three auxiliary tasks, as well, the highest mean classification accuracies (over all auxiliary data configurations and transfer learning protocols) are attained using the MTL based protocols only.

It can be observed further that, though the relative performance of the three auxiliary tasks varies with the transfer learning protocol and the auxiliary data configuration used, S-F achieves the highest mean classification accuracy (43.52%) when the performances are averaged over all transfer learning and auxiliary data configurations. In that case, the other two tasks, S-O and S-S, attain mean classification accuracies of 42.70% and 42.21%, respectively.

When the classification accuracies are averaged over all transfer learning protocols and all auxiliary tasks, the auxiliary dataset IndicTIMIT achieves the highest mean accuracy of 45.57%, followed by IHC PHON achieving a mean accuracy of 44.76%. All auxiliary datasets except IHC SPON are found to achieve higher mean accuracies than the case with no auxiliary data (40.62%). IHC SPON attains the mean classification accuracy of 39.73%, which is only 0.89% lower than the case with no auxiliary data.

Table 2: *Mean classification accuracies in % (SD in bracket) obtained using different transfer learning schemes with single auxiliary tasks when S-MFCC statistics act as the input speech representation; here, the entry in red indicates the case with the highest mean classification accuracy*

Auxiliary data	Auxiliary task	FT	MTL	MTLp1	MTLp2
-	S-O	38.45 (6.56)	39.24 (4.97)	43.61 (3.77)	-
	S-S	39.77 (6.69)	**40.93 (3.61)**	39.76 (5.23)	-
	S-F	**40.11 (4.91)**	36.99 (6.06)	**46.70 (4.33)**	-
IHC PHON	S-O	43.64 (5.51)	48.11 (6.05)	**47.21 (6.05)**	44.54 (5.87)
	S-S	**45.03 (6.20)**	47.69 (5.99)	40.59 (4.65)	45.05 (5.67)
	S-F	37.96 (6.00)	**49.43 (5.87)**	41.91 (4.29)	**45.91 (5.52)**
IHC DDK	S-O	**40.66 (6.08)**	**47.68 (4.90)**	**41.87 (5.55)**	41.94 (4.68)
	S-S	38.47 (5.66)	44.15 (5.45)	40.57 (5.27)	41.82 (5.76)
	S-F	38.86 (5.10)	44.12 (5.14)	41.85 (3.53)	**47.60 (1.43)**
IHC IMAG	S-O	40.08 (3.25)	45.47 (4.79)	42.72 (4.21)	44.94 (5.85)
	S-S	36.62 (2.47)	**45.49 (5.01)**	36.98 (5.76)	41.48 (5.66)
	S-F	**43.29 (5.09)**	44.61 (6.50)	**44.54 (4.68)**	**47.26 (6.29)**
IHC SPON	S-O	39.76 (5.51)	37.81 (3.43)	38.00 (5.38)	37.87 (4.99)
	S-S	38.86 (5.10)	**41.46 (5.12)**	38.35 (5.28)	38.42 (4.31)
	S-F	**44.54 (6.03)**	36.55 (5.88)	**42.34 (2.41)**	**42.78 (5.28)**
IndicTIMIT	S-O	41.89 (6.07)	44.59 (4.75)	43.20 (6.33)	48.49 (4.32)
	S-S	**45.42 (3.37)**	**47.72 (6.59)**	**49.37 (6.38)**	45.49 (5.43)
	S-F	38.71 (3.88)	45.86 (5.99)	43.63 (4.63)	**52.52 (5.77)**
TIMIT	S-O	39.71 (5.09)	45.91 (5.13)	39.26 (3.34)	**46.36 (4.57)**
	S-S	**40.58 (5.86)**	45.42 (4.78)	38.78 (2.18)	45.49 (6.43)
	S-F	39.11 (4.47)	**48.18 (6.87)**	**43.76 (6.09)**	45.89 (6.35)

6.2.2. F-MFCC statistics as input

In the case of F-MFCC, the minimum hike of 3.98% in the mean classification accuracy, as compared to STDL, is achieved in the case of MTL with F-F as the auxiliary task and IHC DDK as the auxiliary data. On the other hand, MTLp1 with F-S as the auxiliary task and IHC PHON as the auxiliary data achieves the maximum improvement of 13.29% attaining the mean classification accuracy of

85.92%, which is also the best accuracy obtained in this work. The average confusion matrix for this case is shown in Figure 4(b). The performance on the ML class is observed to improve as compared to the STDL case with the same F-MFCC statistics as input, though the performance on the NS class has declined. When averaged over all transfer learning protocols, auxiliary tasks, and auxiliary data configurations, F-MFCC statistics reaches a mean severity classification accuracy of 79.98%.

Among the four transfer learning protocols, the maximum mean classification accuracies for all auxiliary data configurations except TIMIT are achieved using one of the three MTL based approaches. Similar trend is observed among the auxiliary tasks also, except for F-O.

Table 3 further suggests that when either no auxiliary data is used or any of IHC PHON, IHC DDK, and IHC SPON is used as the auxiliary data, the auxiliary task of F-S outperforms F-O and F-F with respect to the mean classification accuracy in the cases of all four transfer learning protocols. On the contrary, F-F is found to outperform F-O and F-S in all cases where IHC IMAG, IndicTIMIT, and TIMIT are used as the auxiliary data. However, when averaged over all auxiliary data configurations and all transfer learning protocols, the three auxiliary tasks achieve similar mean severity classification accuracies (79.49% for F-O, 80.28% for F-S, and 80.16% for F-F).

When the classification accuracies are averaged over all transfer learning protocols and all auxiliary tasks, the auxiliary dataset IHC PHON and IndicTIMIT are found to outperform the case with no auxiliary data in terms of mean accuracy. IHC PHON achieves the highest mean accuracy of 82.21%, followed by IndicTIMIT with a mean accuracy of 81.04%, whereas 79.33% mean accuracy is achieved when no auxiliary data is used. The other auxiliary datasets achieve mean performances similar to the case with no auxiliary data.

Table 3: *Mean classification accuracies in % (SD in bracket) obtained using different transfer learning schemes with single auxiliary tasks when F-MFCC statistics act as the input speech representation; here, the entry in red indicates the case with the highest mean classification accuracy*

Auxiliary data	Auxiliary task	FT	MTL	MTLp1	MTLp2
-	F-O	78.39 (4.11)	77.05 (3.79)	79.74 (3.68)	-
	F-S	**81.07 (2.79)**	**81.04 (5.65)**	**81.49 (5.62)**	-
	F-F	77.13 (4.91)	80.09 (5.73)	77.94 (4.08)	-
IHC PHON	F-O	83.24 (5.78)	80.19 (5.54)	80.25 (5.33)	82.79 (3.38)
	F-S	**84.24 (5.04)**	**81.50 (4.92)**	**85.92 (4.35)**	**83.29 (5.06)**
	F-F	81.06 (4.63)	79.68 (5.95)	81.98 (5.70)	82.34 (3.00)
IHC DDK	F-O	79.33 (4.92)	79.73 (6.34)	80.63 (5.83)	77.52 (5.57)
	F-S	**80.22 (4.74)**	**80.59 (3.35)**	**81.90 (5.76)**	**81.47 (5.64)**
	F-F	78.41 (5.15)	76.61 (6.37)	76.66 (2.78)	79.30 (5.19)
IHC IMAG	F-O	80.58 (5.59)	77.95 (6.48)	79.36 (4.91)	81.07 (5.83)
	F-S	79.30 (3.55)	77.00 (4.98)	77.92 (5.20)	79.71 (6.41)
	F-F	**81.49 (6.26)**	**78.32 (5.46)**	**81.50 (3.99)**	**83.26 (5.25)**
IHC SPON	F-O	77.96 (3.74)	77.06 (5.81)	77.05 (4.26)	76.99 (5.02)
	F-S	**79.78 (4.54)**	**81.94 (4.63)**	**80.58 (5.59)**	**77.53 (3.68)**
	F-F	77.49 (4.03)	79.27 (5.80)	77.99 (5.05)	77.07 (4.21)
IndicTIMIT	F-O	82.37 (3.67)	80.12 (3.50)	82.34 (5.26)	80.59 (6.03)
	F-S	81.51 (4.33)	77.50 (4.14)	77.55 (4.04)	78.45 (4.78)
	F-F	**83.67 (5.93)**	**81.93 (3.95)**	**84.11 (4.93)**	**82.34 (3.72)**
TIMIT	F-O	81.46 (4.26)	77.08 (4.78)	77.01 (5.65)	78.36 (5.73)
	F-S	78.41 (6.19)	78.83 (5.77)	80.11 (4.69)	78.84 (6.21)
	F-F	**82.86 (4.82)**	**79.25 (4.28)**	**81.96 (4.15)**	**80.56 (3.68)**

6.2.3. O-MFCC statistics as input

Comparing Tables 1 and 4 we observe that, in the case of O-MFCC statistics as input, the minimum improvement of 4.48% in the mean classification accuracy, as compared to STDL, is achieved in the case of MTLp1 with O-S as the auxiliary task and TIMIT as the auxiliary data. The maximum improvement of 14.54% is attained in the case of MTLp1 with O-F as the auxiliary task and IHC DDK as the auxiliary data. The mean classification accuracy obtained in this case is 84.08%, which is 1.84% less than the maximum mean accuracy achieved using F-MFCC statistics with single auxiliary task (85.92% as mentioned in Table 3). The average confusion matrix for the best performing configuration of O-MFCC is shown in Figure 4(c). Similar to the case of F-MFCC, here also, the performance on the ML class improves as compared to the STDL approach with the same O-MFCC statistics as input. The performance on the SV class has also improved, though that on the NS class has declined slightly. O-MFCC statistics attains a mean severity classification accuracy of 78.32% when averaged over all configurations of transfer learning protocols, auxiliary tasks, and auxiliary data.

When averaged over all auxiliary task and data configurations, the four transfer learning protocols with O-MFCC statistics as input achieve similar mean performances. This observation matches with the cases of S-MFCC and F-MFCC statistics as inputs. Moreover, for O-MFCC, the maximum mean classification accuracies for all auxiliary data configurations except IndicTIMIT are achieved using one of the three MTL based approaches. The same trend of MTL variants achieving the best performances is also present in all of the three auxiliary tasks considered here.

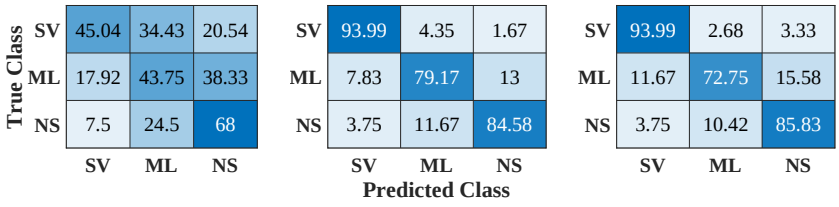
For O-MFCC, in all transfer learning paradigms for every auxiliary data configuration, O-F is found to provide higher mean classification accuracy than both O-O and O-S tasks. However, in most of the cases, the differences in the performances of the three

auxiliary tasks are not large.

When the classification accuracies are averaged over all transfer learning protocols and all auxiliary tasks, all auxiliary datasets are found to achieve slightly higher mean accuracies than the case with no auxiliary data (77.55% mean accuracy). IHC DDK achieves the highest mean accuracy of 78.87%, followed by IHC PHON with a mean accuracy of 78.70%.

Table 4: *Mean classification accuracies in % (SD in bracket) obtained using different transfer learning schemes with single auxiliary tasks when O-MFCC statistics act as the input speech representation; here, the entry in red indicates the case with the highest mean classification accuracy*

Auxiliary data	Auxiliary task	FT	MTL	MTLp1	MTLp2
-	O-O	77.08 (5.76)	76.21 (5.25)	76.20 (4.83)	-
	O-S	77.95 (5.41)	77.63 (6.04)	77.53 (1.52)	-
	O-F	**78.00 (5.95)**	**78.01 (5.86)**	**79.34 (6.23)**	-
IHC PHON	O-O	76.25 (5.00)	78.41 (2.82)	77.17 (4.75)	78.38 (5.68)
	O-S	77.56 (4.43)	75.25 (4.65)	79.74 (5.29)	80.62 (4.19)
	O-F	**79.21 (5.97)**	**79.23 (5.58)**	**81.50 (5.99)**	**81.04 (4.16)**
IHC DDK	O-O	76.63 (5.67)	78.91 (5.76)	78.86 (5.65)	78.00 (4.79)
	O-S	77.47 (6.13)	78.46 (3.67)	76.23 (5.25)	78.93 (5.90)
	O-F	**78.42 (5.01)**	**80.21 (4.13)**	**84.08 (6.08)**	**80.19 (3.48)**
IHC IMAG	O-O	75.73 (4.33)	80.16 (3.12)	77.55 (2.96)	79.74 (4.20)
	O-S	77.09 (2.98)	75.75 (4.61)	76.70 (5.54)	75.42 (5.92)
	O-F	**80.62 (4.13)**	**81.97 (4.86)**	**79.32 (5.95)**	**81.01 (4.86)**
IHC SPON	O-O	75.34 (5.14)	78.42 (4.96)	78.79 (5.87)	79.33 (4.92)
	O-S	76.11 (5.21)	77.44 (6.28)	76.24 (4.03)	77.50 (3.65)
	O-F	**79.35 (6.28)**	**79.75 (2.74)**	**83.29 (5.06)**	**80.65 (5.24)**
IndicTIMIT	O-O	79.31 (6.08)	78.46 (5.79)	78.45 (6.12)	78.88 (5.04)
	O-S	77.05 (5.93)	78.77 (6.04)	74.37 (5.51)	77.12 (5.94)
	O-F	**80.20 (4.57)**	**79.27 (4.83)**	**79.21 (5.62)**	**79.26 (4.81)**
TIMIT	O-O	78.40 (5.83)	77.94 (4.97)	77.09 (5.63)	77.06 (5.81)
	O-S	75.78 (4.50)	76.23 (5.25)	74.02 (3.34)	76.26 (5.46)
	O-F	**79.24 (5.64)**	**78.43 (5.61)**	**81.06 (5.25)**	**80.20 (5.40)**



(a) S-MFCC (MTLp2 + S-F + IndicTIMIT)

(b) F-MFCC (MTLp1 + F-S + IHC PHON)

(c) O-MFCC (MTLp1 + O-F + IHC DDK)

Figure 4: *Confusion matrices averaged over 5-folds of cross-validation for the best performing configurations of the transfer learning approaches with single auxiliary task for different input speech features; here, the entry in the cell (i, j) of each matrix indicates the % of samples of true class i which are classified as class j*

6.3. Transfer learning with multiple auxiliary tasks

Table 5 reports the classification performances obtained using different transfer learning approaches with multiple auxiliary tasks (i.e. S-SFO, F-SFO, and O-SFO) in the cases of S-MFCC, F-MFCC, and O-MFCC statistics as the input features. The obtained mean classification accuracies are found to be higher than the corresponding STDL cases with the same input feature irrespective of the transfer learning protocol and auxiliary data configuration used. For S-MFCC, in 9 out of the 27 combinations of transfer learning protocol and auxiliary data configuration explored, S-SFO is observed to achieve higher mean classification accuracies than the best performing single auxiliary task. For F-MFCC and O-MFCC, the same is true for respectively 12 and 14 cases out of the 27 combinations. When averaged over all of these 27 configurations, S-SFO, F-SFO, and O-SFO achieve mean severity classification accuracies of 44.23%, 81.13%, and 80.24%, respectively.

While using S-SFO and F-SFO as the auxiliary tasks, the respective maximum mean classification accuracies of 50.71% and 84.56% are obtained in the MTLp2 approach with IndicTIMIT as the auxiliary data. For O-SFO, as well, the maximum mean classification accuracy of 83.29% is obtained in the MTLp2 approach. However, IHC DDK is used as the auxiliary data in this case. It can be observed from Tables 2, 3, 4, and 5 that the maximum mean classification accuracies achieved using transfer learning with single

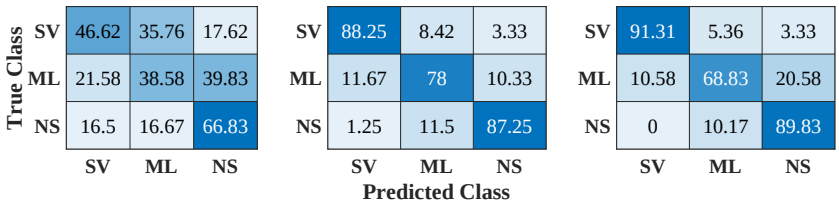
auxiliary tasks for S-MFCC, F-MFCC, and O-MFCC statistics as the input features are respectively 1.81%, 1.36%, and 0.79% higher than those of the transfer learning with multiple auxiliary tasks scenarios. The average confusion matrices for these best performing configurations of S-SFO, F-SFO, and O-SFO are shown in Figure 5(a), 5(b), and 5(c), respectively. It can be observed that the performances on the ML class improve as compared to the STDL cases, though the degree of improvements are less than those achieved in the corresponding best performing configurations of transfer learning with single auxiliary task, as shown in Figure 4. However, the performances on the SV and NS classes decline as compared to the STDL case when F-MFCC statistics is used as the input feature. For the other two input features, the performances on these two classes are better than or similar to those in the corresponding STDL cases.

Similar to the case of transfer learning with single auxiliary tasks, here also, the average performance of the four transfer learning protocols over all auxiliary task and data configurations are mostly similar. However, for all the three auxiliary tasks, the highest mean classification accuracies are obtained using the MTLp2 approach only. Moreover, for all auxiliary data configurations except IHC IMAG, the maximum mean classification accuracies are achieved using one of the MTL variants.

When the classification accuracies are averaged over all transfer learning protocols, the use of IndicTIMIT as the auxiliary dataset is observed to deliver the highest mean classification accuracies for all of S-SFO (48.18%), F-SFO (81.92%), and O-SFO (80.95%).

Table 5: *Mean classification accuracies in % (SD in bracket) obtained using different transfer learning schemes with multiple auxiliary tasks for different input speech features*

Auxiliary data	Auxiliary task	Input	FT	MTL	MTLp1	MTLp2
-	S-SFO	S-MFCC	42.79 (4.01)	39.62 (3.52)	44.06 (5.13)	-
	F-SFO	F-MFCC	81.91 (3.32)	82.83 (5.80)	81.00 (5.93)	-
	O-SFO	O-MFCC	80.66 (4.46)	80.68 (5.75)	80.61 (5.84)	-
IHC PHON	S-SFO	S-MFCC	41.02 (3.87)	48.05 (3.76)	45.03 (4.00)	49.39 (3.23)
	F-SFO	F-MFCC	80.97 (4.47)	81.94 (4.20)	79.76 (5.10)	81.53 (4.04)
	O-SFO	O-MFCC	79.75 (5.91)	80.21 (3.38)	79.38 (4.64)	80.22 (5.12)
IHC DDK	S-SFO	S-MFCC	38.35 (5.28)	44.61 (4.74)	37.44 (5.09)	43.31 (5.55)
	F-SFO	F-MFCC	79.32 (3.03)	77.48 (5.06)	78.52 (5.93)	81.08 (4.09)
	O-SFO	O-MFCC	77.57 (5.50)	82.91 (5.72)	79.36 (4.91)	**83.29 (5.06)**
IHC IMAG	S-SFO	S-MFCC	47.21 (5.31)	48.54 (5.38)	43.19 (4.16)	48.11 (5.69)
	F-SFO	F-MFCC	82.83 (4.11)	82.39 (4.52)	79.75 (5.91)	81.04 (5.65)
	O-SFO	O-MFCC	79.31 (5.16)	82.81 (5.03)	78.84 (2.72)	79.77 (3.84)
IHC SPON	S-SFO	S-MFCC	43.27 (5.22)	44.56 (4.98)	37.55 (4.07)	41.97 (5.79)
	F-SFO	F-MFCC	81.95 (5.33)	81.03 (3.19)	82.82 (3.16)	80.61 (5.07)
	O-SFO	O-MFCC	81.50 (2.17)	77.00 (5.17)	81.47 (4.59)	76.62 (5.79)
IndicTIMIT	S-SFO	S-MFCC	45.87 (5.69)	46.82 (5.74)	49.30 (5.26)	**50.71 (5.41)**
	F-SFO	F-MFCC	84.16 (4.38)	77.45 (5.47)	81.52 (3.22)	**84.56 (5.60)**
	O-SFO	O-MFCC	83.23 (5.18)	81.05 (5.13)	81.52 (4.39)	77.99 (4.84)
TIMIT	S-SFO	S-MFCC	40.98 (4.90)	48.55 (4.53)	38.83 (5.28)	45.05 (4.67)
	F-SFO	F-MFCC	79.70 (3.77)	80.58 (5.01)	80.99 (5.64)	82.74 (4.39)
	O-SFO	O-MFCC	78.84 (5.66)	81.95 (5.33)	78.46 (5.79)	81.57 (4.76)



(a) S-MFCC (MTLp2 + IndicTIMIT)

(b) F-MFCC (MTLp2 + IndicTIMIT)

(c) O-MFCC (MTLp2 + IHC DDK)

Figure 5: *Confusion matrices averaged over 5-folds of cross-validation for the best performing configurations of the transfer learning approaches with multiple auxiliary tasks for different input speech features; here, the entry in the cell (i, j) of each matrix indicates the % of samples of true class i which are classified as class j*

6.4. Performance with varying ALS data size

The primary motivation behind employing transfer learning approaches in this work is to alleviate the scarcity of speech data from dysarthric speakers of different severity levels. Figure 6 illustrates the variations in the dysarthria severity classification performance when decreasing proportion of the ALS training and validation dataset are used for building the classification system with or without

employing transfer learning. Here, we compare the best STDL configuration, i.e. STDL with F-MFCC statistics as the input, and the best transfer learning configuration, i.e. MTLp1 with F-MFCC statistics as the input, F-S as the auxiliary task and IHC PHON as the auxiliary data. It can be observed that, though the accuracies reduce as lesser percentage of data is used for training and validation in both the cases, the accuracy obtained using the transfer learning based approach with only 20% of the ALS training and validation dataset is similar (0.6% higher) to that achieved by the STDL approach with 100% of the concerned dataset. This suggests the suitability of the proposed transfer learning based configuration as the data-efficient way of building dysarthria severity classifiers.

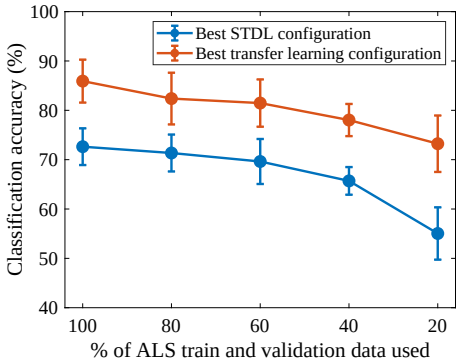


Figure 6: *Mean classification accuracies (SD in error bar) obtained with and without using transfer learning for varying size of the ALS train and validation dataset*

6.5. Analysis of generalizability

We now analyze if the transfer learning protocols under consideration helps in improving the generalizability of dysarthria severity classification. To this end, we select a random $1/3^{rd}$ segment of every utterance of the training ALS set and form a seen test set using those segments. The remaining portion of the utterances constitute the new reduced training set, where, the feature statistics are computed over the entire remaining portion of an utterance. The validation set is kept as it is, and the original test set acts as the unseen test set. This configuration is used in all iterations of the 5-fold cross validation. All models are trained and tested using this setting. Same data splits are maintained for STDL and different transfer learning approaches. Table 6 reports the classification performances obtained on the seen and unseen test sets while using different learning protocols and auxiliary tasks with F-MFCC as the input speech representations. The differences between the mean seen and unseen classification accuracies are also reported. In this analysis, we consider transfer learning approaches without using any auxiliary dataset, so as to ensure that the effect is of the learning protocol and not of the auxiliary data usage. It can be observed that, the performance gap between the seen and unseen test reduces drastically when the transfer learning protocols are employed, as compared to the case of STDL. This indicates better generalizability of the dysarthria severity classification systems trained using transfer learning. Similar trends are also observed for the other input features, namely, S-MFCC and O-MFCC statistics. All auxiliary tasks considered in this work are closely related to the primary task of dysarthria severity classification since these auxiliary tasks focus on learning source, filter, and overall attributes inherent in the utterances, which can provide useful information about the speech characteristics and the dysarthria severity levels. Transfer learning using these related auxiliary tasks has a regularization effect which in turn improves the generalizability [54, 55].

Table 6: *Mean classification accuracies (SD in bracket) on seen and unseen test subjects, along with their differences, obtained using STDL and different transfer learning methods where F-MFCC statistics act as the input speech representation; here, transfer learning is performed without using any auxiliary data*

Learning protocol	**Auxiliary task**	**Seen test accuracy (%)**	**Unseen test accuracy (%)**	**Difference in mean accuracy (%) (Seen - Unseen)**
STDL	-	90.74 (3.90)	73.04 (5.53)	17.70
FT	F-O	80.62 (2.18)	77.97 (5.58)	2.65
	F-S	87.96 (2.49)	81.49 (3.80)	6.47
	F-F	82.97 (3.38)	78.37 (4.41)	4.60
	F-SFO	83.12 (2.57)	80.53 (5.61)	2.59
MTL	F-O	79.45 (2.50)	76.18 (4.37)	3.27
	F-S	86.48 (2.39)	80.11 (5.32)	6.37
	F-F	80.18 (1.50)	79.71 (5.55)	0.47
	F-SFO	81.50 (2.44)	81.43 (3.92)	0.07
MTLp1	F-O	78.27 (5.42)	78.79 (4.66)	-0.52
	F-S	80.77 (2.62)	80.13 (5.65)	0.64
	F-F	81.94 (3.17)	78.42 (5.95)	3.52
	F-SFO	82.68 (1.57)	82.82 (4.03)	-0.14

7. Discussion

Results suggest that, though both the source and the filter components of speech are reported to get affected in ALS-related dysarthria, higher dysarthria severity classification accuracies are obtained when filter-related features are used as the inputs to the models, as compared to source-related features. This is true for both STDL and transfer learning based cases. Thus, the filter-related features seem to carry better discriminative cues, than the source-related features, for classifying different dysarthria severity levels. The slightly inferior performance of O-MFCC than F-MFCC, in most cases, might be because O-MFCC contains both source and filter related information, and the inclusion of source attributes may introduce variability that does not necessarily benefit the classification. Even with the auxiliary task of O-F, though filter attributes are expected to be reinforced in the latent representations obtained from O-MFCC statistics, it may not be possible to completely eliminate the source related information.

For transfer learning with single auxiliary task, when O-MFCC statistics are used as the inputs (Table 4), O-F achieves the highest mean classification accuracies among the three auxiliary tasks O-O, O-S, and O-F, in the cases of all transfer learning protocols for all auxiliary data configurations. This is expected as filter cues are found to carry major discriminative information for severity classification, thereby making the O-F task closely related to the primary task of severity classification. The auxiliary task of O-S is also observed to help in severity classification to a great extent, in spite of S-MFCC delivering the lowest performance individually in STDL. Though much inferior to O-F, the O-S task also might help the encoder in capturing some cues which are not easy to learn using the primary classification task only. Similarly, when F-MFCC statistics are used as the inputs in the case of transfer learning with single auxiliary task (Table 3), either of F-S or F-F performs the best. For S-MFCC, no one auxiliary task delivers better performance than others in majority of the cases. Since the accuracies obtained in the cases of S-MFCC are relatively low, we do not delve deep into this scenario. On the other hand, incorporating multiple auxiliary tasks does not provide an edge over using single auxiliary tasks in terms of the maximum mean accuracies achieved in severity classification. Trying to predict all components at once might be difficult, and hence, may introduce variabilities in the latent representations.

We observe that transfer learning aids severity classification whether any auxiliary data is used or not. Nonetheless, we observe the best average performances (over all transfer learning protocols and auxiliary tasks) in both transfer learning with single and multiple auxiliary tasks for all input speech features in the configurations where auxiliary data is incorporated. It can be observed further that, exact matching of both the languages and the speech tasks between the primary ALS dataset and the auxiliary data is not essential for the auxiliary data to deliver the best average severity classification performance. Our ALS dataset contains SPON utterances in 5 different Indian languages spoken by native speakers. When S-MFCC statistics act as the input features, the best mean classification performances in the cases of transfer learning using both single auxiliary task as well as multiple auxiliary tasks are obtained while using IndicTIMIT as the auxiliary dataset. Same is the case for transfer learning with multiple auxiliary tasks when F-MFCC statistics are used as the input features. IndicTIMIT differs from the ALS dataset in terms of both the recording language (English) and speech task (read speech). Only similarity between the two datasets is that both have native Indian speakers. In the case of transfer learning with single auxiliary task and F-MFCC statistics as inputs, the best mean performance is obtained using IHC PHON as the auxiliary dataset. For O-MFCC as input, IHC DDK delivers the best performances in both single and multiple auxiliary task configurations of transfer learning. Though both IHC PHON and IHC DDK are matched to the ALS dataset in terms of language, the speech tasks are different in both cases. Only IHC SPON has the exactly same configurations of speech task and language as the ALS dataset. However, in many cases, the highest mean classification accuracies (over all auxiliary tasks and transfer learning protocols) obtained using the other auxiliary datasets, including TIMIT (which has all of speech task, native language, and recording language different from the ALS dataset), are similar to or higher than the best mean accuracy achievable by IHC SPON. Thus, languages and speech tasks do not appear to be a barrier while performing transfer learning.

In most cases, the four transfer learning protocols are observed to perform similarly. This agrees with the observation in our previous work [21]. All transfer learning frameworks explored in this work improve the severity classification performances as compared to STDL. Similar to [21], the primary improvement is attained in the classification performance on the ML class. The highest performances achieved here using F-MFCC and O-MFCC statistics as inputs are higher than the highest performance attained in [21]. Though the performances on the SV and NS classes also improve while using transfer learning with S-MFCC statistics as input, these are observed to degrade with transfer learning in a few cases where F-MFCC and O-MFCC statistics act as the inputs. However, the decline is not remarkable.

Lastly, it is to be noted that, in the experimental setup, we employ equal number of utterances from all auxiliary datasets for transfer learning purposes so that the results are not affected by the auxiliary dataset size. Employing more utterances (if available in the auxiliary dataset) may seem to enhance the classification performances. However, we do not observe much changes in performances on repeating the experiments with all utterances present in each auxiliary dataset.

8. Conclusions

This paper analyzes source and filter related cues for 3-class dysarthria severity classification specific to ALS and finds that the filter related cues carry better discriminative information. Transfer learning approaches with novel source and filter characteristics based auxiliary tasks are explored to mitigate the low data availability issue often encountered in training dysarthria severity classification systems. We empirically confirm the utility of these approaches in improving the performance and generalizability of the classifiers. The major benefit is obtained on the classification performance for the mild dysarthria class. However, the optimal auxiliary task is found to vary depending on the input speech feature used as well as the auxiliary data configuration considered. Moreover, no significant difference is observed among the four transfer learning protocols explored. The speech task and language of the auxiliary dataset also do not seem to pose any difficulty in transfer learning. In future, we would like to explore auxiliary tasks which can bring in more explainability in the decision of the severity classifier. We would also like to predict the ALSFRS-R scores directly instead of performing the 3-class severity classification.

CRediT authorship contribution statement
Tanuka Bhattacharjee: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data Curation, Writing - Original Draft, Writing - Review & Editing, Visualization. **Yamini Belur:** Resources, Data Curation. **Atchayaram Nalini:** Resources. **Prasanta Kumar Ghosh:** Conceptualization, Validation, Writing - Review & Editing, Supervision

Declaration of competing interest
The authors have no conflict of interest to be disclosed.

Data availability
The data that support the findings of this study are available on request from the corresponding author.

Acknowledgment
This work was supported by the Department of Science and Technology (DST), Govt. of India.

9. References

[1] B. Tomik and R. J. Guiloff, "Dysarthria in amyotrophic lateral sclerosis: A review," *Amyotrophic Lateral Sclerosis*, vol. 11, no. 1-2, pp. 4–15, 2010.

[2] P. Enderby, "Frenchay dysarthria assessment," *British Journal of Disorders of Communication*, vol. 15, no. 3, pp. 165–173, 1980.

[3] 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.

[4] J. M. King, M. Watson, and G. L. Lof, "Practice patterns of speech-language pathologists assessing intelligibility of dysarthric speech," *Journal of Medical Speech-Language Pathology*, vol. 20, no. 1, pp. 1–17, 2012.

[5] G. Fant, *Acoustic theory of speech production*. Walter de Gruyter, 1970.

[6] P. Rong, Y. Yunusova, J. Wang, L. Zinman, G. L. Pattee, J. D. Berry, B. Perry, and J. R. Green, "Predicting speech intelligibility decline in amyotrophic lateral sclerosis based on the deterioration of individual speech subsystems," *PloS one*, vol. 11, no. 5, p. e0154971, 2016.

[7] 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.

[8] F. G. Vieira, S. Venugopalan, A. S. Premasiri, M. McNally, A. Jansen, K. McCloskey, M. P. Brenner, and S. Perrin, "A machine-learning based objective measure for als disease severity," *NPJ digital medicine*, vol. 5, no. 1, p. 45, 2022.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

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

[15] 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.

[16] F. Javanmardi, S. R. Kadiri, and P. Alku, "Exploring the impact of fine-tuning the wav2vec2 model in database-independent detection of dysarthric speech," *IEEE Journal of Biomedical and Health Informatics*, vol. 28, no. 8, pp. 4951–4962, 2024.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] T. Bhattacharjee, A. Jayakumar, Y. Belur, A. Nalini, 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.

[22] M. Morise, F. Yokomori, and K. Ozawa, "WORLD: a vocoder-based high-quality speech synthesis system for real-time applications," *IEICE Transactions on Information and Systems*, vol. 99, no. 7, pp. 1877–1884, 2016.

[23] P. N. Le, J. Epps, E. H. Choi, and E. Ambikairajah, "A study of voice source and vocal tract filter based features in cognitive load classification," in *20th International Conference on Pattern Recognition*. IEEE, 2010, pp. 4516–4519.

[24] H. Kim, M. Hasegawa-Johnson, A. Perlman, J. R. Gunderson, T. S. Huang, K. L. Watkin, S. Frame *et al.*, "Dysarthric speech database for universal access research." in *Proc. Interspeech*, 2008, pp. 1741–1744.

[25] F. Rudzicz, A. K. Namasivayam, and T. Wolff, "The TORGO database of acoustic and articulatory speech from speakers with dysarthria," *Language resources and evaluation*, vol. 46, pp. 523–541, 2012.

[26] M. Fernández-Díaz and A. Gallardo-Antolín, "An attention long short-term memory based system for automatic classification of speech intelligibility," *Engineering Applications of Artificial Intelligence*, vol. 96, p. 103976, 2020.

[27] H. Chandrashekar, V. Karjigi, and N. Sreedevi, "Investigation of different time-frequency representations for intelligibility assessment of dysarthric speech," *IEEE transactions on neural systems and rehabilitation engineering*, vol. 28, no. 12, pp. 2880–2889, 2020.

[28] S. Gupta, A. T. Patil, M. Purohit, M. Parmar, M. Patel, H. A. Patil, and R. C. Guido, "Residual neural network precisely quantifies dysarthria severity-level based on short-duration speech segments," *Neural Networks*, vol. 139, pp. 105–117, 2021.

[29] A. A. Joshy and R. Rajan, "Automated dysarthria severity classification: A study on acoustic features and deep learning techniques," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 30, pp. 1147–1157, 2022.

[30] Z. Yue, E. Loweimi, and Z. Cvetkovic, "Dysarthric speech recognition, detection and classification using raw phase and magnitude spectra," in *Proc. Interspeech*, 2023, pp. 1533–1537.

[31] M. Vidya and S. Ganesh Vaidyanathan, "Dysarthric severity categorization based on speech intelligibility: A hybrid approach," *Circuits Systems and Signal Processing*, vol. 43, no. 11, pp. 7044–7063, 2024.

[32] N. Narendra and P. Alku, "Automatic intelligibility assessment of dysarthric speech using glottal parameters," *Speech Communication*, vol. 123, pp. 1–9, 2020.

[33] ——, "Automatic assessment of intelligibility in speakers with dysarthria from coded telephone speech using glottal features," *Computer Speech Language*, vol. 65, p. 101117, 2021.

[34] A. A. Joshy and R. Rajan, "Dysarthria severity assessment using squeeze-and-excitation networks," *Biomedical Signal Processing and Control*, vol. 82, p. 104606, 2023.

[35] S. Sajiha, K. Radha, D. Venkata Rao, N. Sneha, S. Gunnam, and D. P. Bavirisetti, "Automatic dysarthria detection and severity level assessment using CWT-layered CNN model," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2024, no. 1, p. 33, 2024.

[36] K. Radha, M. Bansal, and V. R. Dhulipalla, "Variable STFT layered CNN model for automated dysarthria detection and severity assessment using raw speech," *Circuits, Systems, and Signal Processing*, vol. 43, no. 5, pp. 3261–3278, 2024.

[37] 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.

[38] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, vol. 33, pp. 12 449–12 460, 2020.

[39] W.-N. Hsu, B. Bolte, Y.-H. 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. 29, pp. 3451–3460, 2021.

[40] M. Tu, A. Wisler, V. Berisha, and J. M. Liss, "The relationship between perceptual disturbances in dysarthric speech and automatic speech recognition performance," *The Journal of the Acoustical Society of America*, vol. 140, no. 5, pp. EL416–EL422, 2016.

[41] A. Jacks, K. L. Haley, G. Bishop, and T. G. Harmon, "Automated speech recognition in adult stroke survivors: Comparing human and computer transcriptions," *Folia Phoniatrica et Logopaedica*, vol. 71, no. 5-6, pp. 286–296, 2019.

[42] S. E. Gutz, K. L. Stipancic, Y. Yunusova, J. D. Berry, and J. R. Green, "Validity of off-the-shelf automatic speech recognition for assessing speech intelligibility and speech severity in speakers with amyotrophic lateral sclerosis," *Journal of Speech, Language, and Hearing Research*, vol. 65, no. 6, pp. 2128–2143, 2022.

[43] 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.

[44] P. N. Chowdary, M. S. Akshay, V. S. Aravind, M. S. Aashish, G. V. N. S. L. V. Vardhan, and G. Jyothish Lal., "A few-shot approach to dysarthric speech intelligibility level classification using transformers," in *14th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, 2023, pp. 1–6.

[45] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. Mcleavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *Proceedings of the 40^{th} International Conference on Machine Learning*, 2023, pp. 28 492–28 518.

[46] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[47] A. V. S. Manoj, V. Lakshman, A. Kamuju, V. Pulagam, and G. J. Lal, "Transformer-based transfer learning for enhanced speech dysarthria severity assessment," in *15th International Conference on Computing Communication and Networking Technologies (ICCCNT)*. IEEE, 2024, pp. 1–6.

[48] M. Morise *et al.*, "Harvest: A high-performance fundamental frequency estimator from speech signals." in *INTERSPEECH*, 2017, pp. 2321–2325.

[49] 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.

[50] M. Morise, "CheapTrick, a spectral envelope estimator for high-quality speech synthesis," *Speech Communication*, vol. 67, pp. 1–7, 2015.

[51] ——, "D4C, a band-aperiodicity estimator for high-quality speech synthesis," *Speech Communication*, vol. 84, pp. 57–65, 2016.

[52] T. Bhattacharjee, J. Mallela, Y. Belur, A. Nalini, R. Yadav, P. Reddy, D. Gope, and P. K. Ghosh, "Source and vocal tract cues for speech-based classification of patients with Parkinson's Disease and healthy subjects." in *Interspeech*, 2021, pp. 2961–2965.

[53] 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.

[54] D. Erhan, A. Courville, Y. Bengio, and P. Vincent, "Why does unsupervised pre-training help deep learning?" in *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, vol. 9, 2010, pp. 201–208.

[55] S. Ruder, "An overview of multi-task learning in deep neural networks," *arXiv preprint arXiv:1706.05098*, 2017.

[56] B. R. Brooks, R. G. Miller, M. Swash, and T. L. Munsat, "El escorial revisited: revised criteria for the diagnosis of amyotrophic lateral sclerosis," *Amyotrophic lateral sclerosis and other motor neuron disorders*, vol. 1, no. 5, pp. 293–299, 2000.

[57] "Zoom XYH-6 adjustable stereo microphone capsule," https://www.zoom.co.jp/products/product-accessories/xyh-6-xy-stereo-microphone-capsule/, [Online; accessed 05-Mar-2023].

[58] J. L. Fleiss, "Measuring nominal scale agreement among many raters," *Psychological bulletin*, vol. 76, no. 5, p. 378, 1971.

[59] 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.

[60] 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.

[61] V. Zue, S. Seneff, and J. Glass, "Speech database development at MIT: TIMIT and beyond," *Speech communication*, vol. 9, no. 4, pp. 351–356, 1990.