Interspeech 2024
1-5 September 2024, Kos, Greece

Exploring Syllable Discriminability during Diadochokinetic Task with Increasing Dysarthria Severity for Patients with Amyotrophic Lateral Sclerosis

Neelesh Samptur[1], *Tanuka Bhattacharjee*[2], *Anirudh Chakravarty K*[1], *Seena Vengalil*[3], *Yamini Belur*[3], *Atchayaram Nalini*[3], *Prasanta Kumar Ghosh*[2]

[1]Computer Science and Engineering, PES University, India, [2]Electrical Engineering, Indian Institute of Science, India, [3]National Institute of Mental Health and Neurosciences, India

neeleshsamptur@gmail.com, tanukab@iisc.ac.in, prasantg@iisc.ac.in

Abstract

We explore the discriminability among /pa/, /ta/, and /ka/ syllables, spoken during diadochokinetic (DDK) task, at varied severity levels of amyotrophic lateral sclerosis (ALS) induced dysarthria. Though DDK rate is known to decline with increasing severity, the extent to which the discriminability among the syllables gets impacted at each severity level is not well understood. We perform manual and automatic classification of these three syllables on 100 ALS and 35 healthy subjects. Manual classification is done through listening tests. Spectral and self-supervised speech cues with deep neural classifiers are used for automatic classification. Manual classification accuracies decline from 84.07% on healthy utterances to 27.41% on utterances of the most severe patients. Automatic methods are found to outperform humans achieving 15.93% and 50.37% higher accuracies (absolute), respectively. Thus, discriminative acoustic cues seem to persist among the syllables, which automatic methods capture.

Index Terms: Amyotrophic Lateral Sclerosis, dysarthria, diadochokinetic task, syllable classification, listening test

1. Introduction

Oral diadochokinetic (DDK) tasks are widely used by clinicians during diagnosis and assessment of dysarthria prevalent in different neurological disorders like amyotrophic lateral sclerosis (ALS) [1]. These tasks are also used for speech-based automatic classification between patients suffering from such neurological diseases and healthy controls (HC) [2, 3]. DDK tasks examine how quickly and accurately one can repeat, without any interruption, a series of monosyllabic targets like ‘pa-pa-pa’ or tri-syllabic targets like ‘pataka’ [4]. Dysarthria due to ALS restricts the speed of movements of lips, jaw, tongue, and velum [5, 6] causing a reduction in the DDK rate [7]. The discriminability among the target syllables also gets compromised with increasing dysarthria severity. However, the extent to which the compromise happens at different severity levels for ALS is not well understood. This work explores that gap in the literature by analysing the degree of discriminability among /pa/, /ta/, and /ka/ syllables produced during the tri-syllabic DDK ‘pataka’ task, at varied severity levels of ALS-induced dysarthria.

Dysarthria due to ALS affects nearly all sub-systems of speech production. With increasing severity, it gradually collapses the acoustic space of the patients compromising the discriminability among different speech sounds [8]. The vowel space area is reported to reduce [8] making it difficult to discriminate between vowels. Patients often add unwanted voicing to voiceless fricatives making them sound like their respective voiced counterparts [9, 10]. Kumar et al. [11] have performed classification of different sustained vowels and different sustained fricatives at varied severities of ALS-induced dysarthria through manual listening tests as well as using automatic deep neural network approaches and reported that the classification accuracies decline drastically with increasing severity levels.

Though studies have been reported on discriminability among certain vowels and fricatives with increasing dysarthria severity for ALS patients, discriminability among syllables like /pa/, /ta/, and /ka/ remains relatively unexplored in this regard. ALS patients are often found to perform incomplete closures while uttering stop consonants [10]. This, along with impaired consonant-to-vowel formant transitions [12], may significantly distort the acoustic characteristics of the syllables under consideration. Tao et al. [13] have analysed automatic speech recognition on tri-syllabic DDK ‘pataka’ sequences for patients with traumatic brain injuries and Parkinson’s disease, but not for ALS. We explore this gap in the literature by conducting a scientific study on how humans and machines perceive the discriminability among these syllables during DDK tasks with increasing dysarthria severity for ALS.

We perform 3-class classification of /pa/, /ta/, and /ka/ syllables using manual and automatic methods at different severity levels of ALS-induced dysarthria. Manual classification is conducted through listening tests. For automatic classification, we explore dense neural networks (DNN), convolutional neural networks (CNN), and long-short term memory (LSTM) networks with spectral and self-supervised (SS) speech representations as the inputs. Though all the 3 syllables at hand constitute a voiceless stop and the vowel /a/, the automatic classification performances may not be solely driven by the phone-level classification of the stops. With a vowel in the syllable, the stop-to-vowel transition may give more information and hence better discriminability. As expected, both manual and automatic classification accuracies decline with increasing severity. However, the proposed automatic methods significantly outperform humans not only on utterances from HCs but also on dysarthric utterances of all severity levels. Automatic methods are found to achieve 15.93% and 50.37% higher accuracies (absolute) than humans on utterances from HCs and the most severe patients, respectively. This might indicate that though humans may fail to perceive the differences among these syllables with increasing dysarthria severity, distinct cues persist in the syllables which data-driven models can capture. Thus, these syllables can be explored further as potential choices of voice commands for automatic voice assistants, even for the most severe patients.

2. Dataset

Speech recordings were collected from 100 ALS (64M + 36F; age range: 28-73 years) and 35 HC (18M + 17F; age range: 31-55 years) subjects at the National Institute of Mental Health

10.21437/Interspeech.2024-1175

and Neurosciences (NIMHANS), India. Here, M and F stand for male and female, respectively. The subjects spoke one of the following native languages - Bengali, Tamil, Telugu, Hindi, and Kannada. Three speech-language pathologists (SLPs) rated the dysarthria severity of each ALS subject following the 5-point speech component [0 (complete loss of useful speech) - 4 (normal speech)] of the ALSFRS-R scale [14]. The final severity score was derived as the mode of these 3 ratings. Following Kumar et al. [11], we grouped the subjects with severity scores 0 and 1 together as the *severe dysarthric group* (SV), those with scores 2 and 3 together as the *mild dysarthric group* (ML) and the ones with score 4 as the *ALS group with no dysarthria* (ND). Lastly, a *normal speech group* (NS) was formed with all the HC subjects. All groups comprised 35 subjects, except SV which had 30 subjects. The distributions of gender and native language were similar for all these groups. The subjects performed the DDK task where they were asked to take a deep breath and keep repeating the tri-syllabic sequence ‘pataka’ as fast as possible. Up to 3 such trials were recorded from a subject depending on his/her level of comfort. Further details about the data collection procedure are given in [2]. A total of 86, 106, 105, and 99 trials were obtained from SV, ML, ND, and NS groups, respectively, with the mean and standard deviation (SD) of the durations of the trials being 3.93 (2.84), 4.55 (2.05), 6.01 (2.13), and 5.45 (2.00) sec, respectively. The hospital ethics committee approved the data collection protocol. Also, each subject signed a consent form before data collection.

3. Method

3.1. Data Preprocessing - Syllable Segmentation

We first segment individual syllables in the speech trials comprising repetitive ‘pataka’ utterances to form the syllable classification corpus. A two-phase semi-automatic method is used. First, we obtain the upper peak envelope of the waveform of a speech trial using spline interpolation over the local maxima points which are at least 5 ms apart. It is then low-pass filtered using a 2nd order digital Butterworth filter having a cutoff frequency of 15 Hz to obtain a smooth envelope of the speech waveform as shown in Figure 1. The local minima of this smooth envelope are located such that the corresponding points in the negated envelope have a minimum peak prominence of 0.2 times the maximum magnitude of the envelope. This threshold is set empirically. This constraint discards the insignificant minima. The speech segment between two consecutive minima thus identified is considered to encompass one syllable. These identified syllables are then labeled cyclically as /pa/, /ta/, and /ka/ starting from the first syllable identified in the speech trial. This automatic segmentation process is performed using MATLAB R2021a [15]. However, on manual inspection, it is found that, the automatically obtained time boundaries are not accurate in many cases. It is particularly erroneous for trials produced by severely dysarthric subjects. Moreover, the subjects, even the HCs, are observed to make mistakes while repeating the syllables, thereby generating sequences like ‘katapa’, ‘patapa’ etc. which can not be detected by this automatic method. Hence, we manually correct all erroneous segmentations by listening to each speech trial and modifying the erroneously annotated syllables using the Audacity software [16]. Figure 1 shows the automatically obtained and the manually corrected syllable segmentations for an illustrative speech trial. The total number and durations of the 3 types of syllables thus identified for the different severity groups are listed in Table 1.

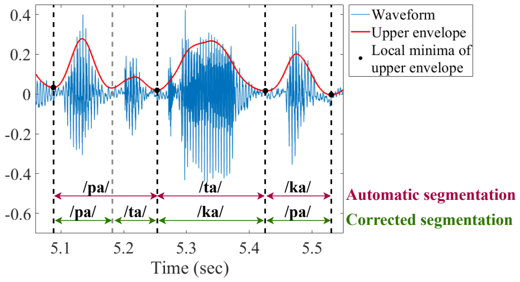


Figure 1: *Automatic and manually corrected syllable segmentations of an illustrative speech segment; the grey boundary is missed in the automatic process*

3.2. Manual Classification of Syllables

We performed listening tests following Kumar et al. [11] to carry out the manual classification process. Conducting the listening test for all data is a time-consuming and tedious job. Hence, we selected a subset of the data for this purpose. 15 subjects were chosen from each of the 4 severity groups while maintaining an almost uniform distribution of gender and native language. Two random utterances of each of /pa/, /ta/, and /ka/ were selected among the longest 50% of the respective syllables produced by a particular subject. Thus, the total number of utterances used for the listening tests were 360 (4 severity groups × 15 subjects × 3 syllables × 2 utterances).

The listening tests were carried out through a web application. First, a listener was presented with 3 example audios corresponding to the 3 syllables /pa/, /ta/, and /ka/ uttered by a healthy individual. Then, the test audios were presented sequentially. The listeners had to choose the most likely alternative among the 3 options, /pa/, /ta/, and /ka/, given for each audio. They also provided confidence scores in [0, 100] signifying their confidence in making each decision. Thus, though the listening test was designed as a forced 3-class multiple-choice test, the confidence score could capture the listener’s uncertainty in making a decision. Each syllable utterance was assigned to 3 listeners for classification. The listeners were allowed to revisit the examples and play each test audio as many times as they wanted. However, once the decision for a test audio was submitted, it could not be changed further.

27 listeners (19M + 8F; age range: 17 - 54 years) were recruited for the listening tests. The native languages of the listeners included Hindi, Kannada, Tulu, Tamil, Telugu, Malayalam, and Urdu. None of the listeners reported any hearing impairment. A total of 90 test syllable utterances with almost equal proportion from each of the 4 severity groups were assigned to every listener. 10 random utterances out of the 90 (2 or 3 utterances per severity group) were repeated making the total number of test audios to be classified by each listener equal to 100. For each participating listener, we computed 2 metrics - accuracy on the utterances belonging to the NS group and consistency on the repeated utterances. Consistency was calculated

Table 1: *Number and duration of utterances of different syllables obtained from subjects of different severity groups; each cell entry is in the form of x/y/z, where, x is the number of utterances, y is the mean duration (in sec) of the utterances, and z is SD of the durations (in sec) of the utterances*

	SV	ML	ND	NS
/pa/	294/0.33/0.16	730/0.17/0.06	1417/0.13/0.03	1349/0.11/0.03
/ta/	287/0.32/0.17	621/0.18/0.08	1323/0.12/0.03	1215/0.11/0.03
/ka/	267/0.38/0.15	716/0.25/0.12	1405/0.17/0.05	1346/0.15/0.03

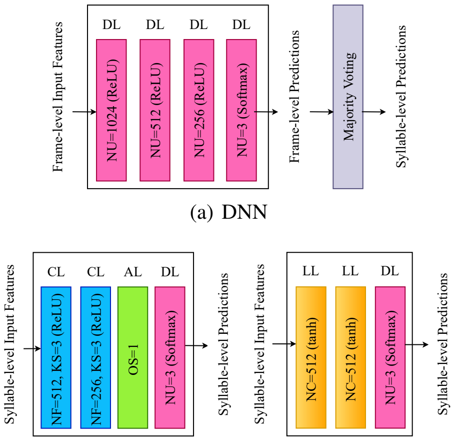
as the % of times the labels chosen for an utterance and its repeated version matched. The accuracy of the listeners ranged in 50%-100% while their consistency ranged in 40%-100%. The responses from only the listeners with at least 80% accuracy and 80% consistency were considered further. Though the listeners did not have hearing disabilities, we put this selection criteria to ensure that we considered the responses from only those listeners who were attentive and serious during the test, thus expecting to discard randomly chosen labels. This process made us select 12 listeners (9M + 3F) out of the 27 participants.

It is observed that most listeners logged low confidence scores for some of their responses. These might correspond to those syllable utterances which did not sound precisely like any of /pa/, /ta/, and /ka/, although the listener was forced to choose one option. Utterances produced by severe patients are more likely to give rise to this situation. Thus, in our subsequent calculation of manual classification accuracies, we consider a response as correct only if it matches with the ground truth and the corresponding confidence score is above a threshold. It is observed that most of the consistent responses given by the listeners on repeated utterances had confidence scores above 40, whereas, most of the inconsistent ones had the scores below 40. Hence, we set the threshold on confidence score at 40.

3.3. Automatic Classification of Syllables

Features: Since different SS speech representations are reported to be well-suited for ASR systems [17, 18, 19], we explore such representations obtained from 5 different pretrained models, namely, DeCoAR (2048D) [20], HuBERT (768D) [17], NPC (512D) [21], TERA (768D) [18], and Wav2Vec 2.0 (768D) [19] for our /pa/-/ta/-/ka/ classification task. These SS learning models differ in terms of the dimension of the latent speech representation, the speech input format, the self-supervision task performed, and the objective function considered [22]. Thus different models encode information differently in the speech representation. Hence, we explore a variety of models to exploit the best suited representation for the syllable classification task. We also use the time-frequency representation of speech captured through 12D mel-frequency cepstral coefficients (MFCC) (except the energy term) along with its delta and double-delta measures. The S3PRL speech toolkit [23] and the pretrained upstream model weights available in the toolkit are used to extract the SS speech representations, whereas, MFCC is computed for every 20 ms speech frame with 15 ms overlap using the KALDI speech recognition toolkit [24].

Classifiers: DNN, CNN, and LSTM are used as the classifiers in this work. The DNN architecture as adopted from Kumar et al. [11], along with the proposed CNN and LSTM architectures, are shown in Figure 2. DNN models are trained at the frame-level. During testing, majority voting is performed over the predictions obtained for all frames of a particular syllable to arrive at the syllable-level predictions. On the other hand, training and testing of CNN and LSTM models are done at the syllable-level. Since the syllables have varying frame counts, the feature matrices of all syllables of a speech trial are zero-padded to the length of the longest one present in that trial and fed as a single batch during training and testing of CNN and LSTM models. As reported in Table 1, SV group has much lower number of utterances of each syllable as compared to the other severity groups. Hence, to maintain uniformity, while training CNN and LSTM models at syllable-level for each of ML, ND, and NS groups, we use a randomly selected subset of the corresponding syllables such that the cardinalities of the training syllable sets for these severity groups are almost equal to that for SV. The same is not required for frame-level DNN training as the number of training frames obtained for different severity groups are found to be similar. All models are initialized randomly with a fixed seed. We train the models for a maximum of 100 epochs using Adam optimizer with a learning rate of 0.0001, along with cross-entropy loss. To avoid overfitting, early stopping is done with a patience of 5 based on the validation loss. Training and testing are performed separately for each severity group for all 3 classifiers. All implementations are done in Pytorch v1.11.0 [25]. An NVIDIA GeForce RTX 2080 GPU is used for training and testing the models.



(a) DNN

(b) CNN

(c) LSTM

Figure 2: *Architectures of different classifiers; here, DL: Dense layer, CL: 1D-Convolutional layer, AL: Adaptive average pooling layer, LL: LSTM layer, NU: Number of units in DL, NF: Number of filters, KS: Kernel size, OS: Output size of AL, NC: Number of LSTM cells*

4. Experimental Setup

Experiments are conducted in two phases. In the first phase, 5-fold cross-validation of all automatic classifiers are performed and their performances are compared for each severity group separately. The subjects of each group are equally and randomly distributed among the 5 folds and the same fold structure is maintained across all automatic classifiers. For any severity group, at every iteration, one of the folds acts as the test set and the remaining folds are used together as the train set. Random 4 subjects from the train set are chosen to form the validation set. The mean and SD of classification accuracies obtained over 5 folds of cross-validation for every severity group are used as the performance metrics. The classifier architectures, as mentioned in subsection 3.3, are tuned by optimizing the average validation performances over the 5-folds of all 4 severity groups and all input feature representations. The second phase of evaluation is carried out by evaluating the performance of the automatic classifiers on the manual listening test set. Thus, for every severity group, the utterances of the 15 subjects selected for the manual listening task are used as the test set for the automatic classifiers and the remaining subjects are used to form the train set. Random 3 subjects from the train set are used for validation. We compare the classification accuracies achieved using the automatic methods against the manual classification accuracies using the Wilcoxon signed-rank test [26] at 1% significance level. For this purpose, 30 random subsets, each containing 30 utterances, are formed out of the manual listening test set for every

Table 2: *Mean classification accuracies in % (SD in bracket) over 5-fold cross-validation obtained using different automatic classification methods for different severity groups; here, #para indicates the number of trainable parameters in the classifier*

Feature	Dim	DNN					CNN					LSTM				
		#para	NS	ND	ML	SV	#para	NS	ND	ML	SV	#para	NS	ND	ML	SV
MFCC	36	6.9L	83.69 (5.67)	74.82 (11.31)	61.12 (11.11)	50.27 (6.77)	51K	59.07 (6.81)	50.92 (10.95)	45.38 (6.26)	42.70 (6.46)	32L	76.89 (13.15)	74.74 (11.01)	63.12 (7.96)	48.68 (7.28)
NPC	512	11L	45.66 (7.95)	59.97 (20.35)	55.94 (11.69)	48.59 (3.49)	2.9L	47.79 (5.59)	43.19 (4.74)	38.41 (3.68)	35.19 (5.23)	42L	60.47 (17.97)	60.67 (8.17)	58.80 (6.21)	45.83 (6.39)
TERA	768	14L	83.91 (5.23)	90.10 (5.20)	67.80 (9.89)	61.53 (8.92)	4.2L	57.57 (5.98)	54.43 (7.89)	55.64 (5.59)	40.65 (3.07)	47L	83.32 (5.44)	90.82 (3.10)	80.64 (3.25)	47.83 (4.86)
HuBERT	768	14L	**99.03** **(0.40)**	**99.46** **(0.47)**	**97.03** **(1.45)**	79.72 (8.68)	4.2L	94.06 (6.59)	89.25 (4.25)	78.54 (4.29)	54.12 (5.11)	47L	97.98 (2.37)	98.56 (1.44)	94.52 (3.09)	**79.75** **(10.47)**
Wav2Vec 2.0	768	14L	98.13 (0.57)	97.29 (1.43)	91.21 (6.12)	71.14 (6.82)	4.2L	64.49 (14.06)	61.37 (7.25)	46.31 (5.43)	38.02 (1.37)	47L	96.23 (1.54)	94.40 (1.64)	86.40 (4.51)	68.11 (14.22)
DeCoAR	2048	27L	85.53 (7.84)	93.17 (3.47)	74.05 (5.43)	53.01 (6.04)	10L	50.53 (2.86)	49.16 (3.55)	38.84 (3.48)	36.96 (3.33)	73L	85.99 (10.03)	72.07 (19.47)	61.88 (4.21)	48.74 (5.90)

severity group and the accuracies of all automatic and manual methods on these subsets are considered.

5. Results and Discussion

Comparison of Different Automatic Classifiers: Table 2 reports the 5-fold cross validation performances obtained for the 4 severity groups using different automatic classification approaches. As expected, the classification accuracies achieved using most of the approaches drop with increasing severity levels, except a few cases. Like, TERA features with CNN perform better for ML than ND. Several approaches perform marginally better for ND than NS. This might be because ND group also does not have speech impairments like NS and may achieve better accuracies depending on the dataset considered. For NS, ND, and ML groups, HuBERT-based features with DNN achieve the best classification performance. Though for SV, HuBERT-based features with LSTM classifier turn out to be the best, DNN with the same features also attains statistically similar performance. Only HuBERT-based features are considered further due to their superior performance over others.

Automatic vs. Manual Classification: Figure 3 illustrates the automatic and manual classification performances obtained on the subjective listening test set. Here, for comparison, HuBERT representations are used as the input features for all automatic classifiers. Both manual and the highest automatic classification accuracies obtained for the NS group are at par with the literature [27, 28]. DNN and LSTM are observed to significantly outperform the manual classification approach at all severity levels, though the performance of CNN is significantly better than the manual performance only for the SV group. LSTM for NS group and DNN for ND group attain 100% mean classification accuracies with 0 SD. For ML and SV groups, LSTM and DNN, respectively, are observed to be the best performing models. Though the mean accuracies for both manual and automatic methods decline with increasing severity, the drops from NS to SV are significantly less for automatic DNN (21.11%) and LSTM (28.89%) methods than that for manual classification (56.66%). These results might indicate that though these 3 syllables become perceptually difficult to discriminate with increasing severity, the acoustic cues present in the utterances can preserve the differences to a considerable extent. The confusion matrices obtained using the manual and the best performing automatic method for each severity group (Figure 4) further suggest that humans can identify /pa/ the best at all severity levels. They confuse /ka/ the most for NS and ND but /ta/ for ML and SV. The best performing automatic method faces the highest confusion in the case of /ka/, followed by /ta/, for SV.

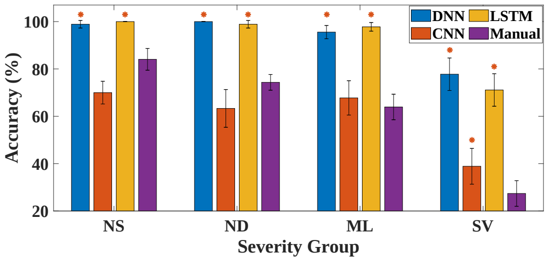


Figure 3: *Mean automatic (using HuBERT features) and manual classification accuracies (SD in error bar) obtained on the manual listening test set for different severity groups; at each severity, * indicates the automatic methods which significantly outperform humans at 1% significance level*

	/pa/	/ta/	/ka/	CS < 40
/pa/	0.91 1	0.01 0	0 0	0.08 -
/ta/	0.02 0	0.84 1	0.06 0	0.08 -
/ka/	0.12 0	0.06 0	0.77 1	0.06 -

(a) NS

	/pa/	/ta/	/ka/	CS < 40
/pa/	0.83 1	0.02 0	0.01 0	0.13 -
/ta/	0.1 0	0.73 1	0.04 0	0.12 -
/ka/	0.11 0	0.02 0	0.66 1	0.20 -

(b) ND

	/pa/	/ta/	/ka/	CS < 40
/pa/	0.8 0.97	0.03 0.03	0 0	0.17 -
/ta/	0.11 0	0.52 0.97	0.10 0.03	0.26 -
/ka/	0.08 0	0.14 0	0.6 1	0.18 -

(c) ML

	/pa/	/ta/	/ka/	CS < 40
/pa/	0.36 0.93	0.01 0.03	0.11 0.03	0.52 -
/ta/	0.08 0.13	0.15 0.73	0.16 0.13	0.62 -
/ka/	0.18 0.2	0.08 0.13	0.32 0.67	0.42 -

(d) SV

Figure 4: *Confusion matrices obtained on the manual listening test set of different severity groups using manual (in red) and the best-performing automatic (in blue) classification methods; here CS: confidence score*

6. Conclusion

We analyze the discriminability among 3 syllables /pa/, /ta/, and /ka/ at varied severities of ALS-induced dysarthria. Automatic classification methods involving SS speech representations and deep neural networks are found to be able to differentiate these syllables significantly better than humans at all severity levels, though the performances of both automatic and manual methods decline with increasing severity. In the future, we would like to incorporate voiced stops like /b/, /d/, /g/, along with the already explored voiceless /p/, /t/, /k/, in our study.

Acknowledgements - Authors thank the Department of Science and Technology (DST), Govt of India, for their support in this work.

7. 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] 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. Interspeech*, 2020, pp. 4586–4590.

[3] 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.

[4] M. Lancheros, D. Friedrichs, and M. Laganaro, “What do differences between alternating and sequential diadochokinetic tasks tell us about the development of oromotor skills? An insight from childhood to adulthood,” *Brain Sciences*, vol. 13, no. 4, p. 655, 2023.

[5] A. Illa, D. Patel, B. Yamini, M. SS, N. Shivashankar, P. K. Veeramani, S. Vengalii, K. Polavarapui, S. Nashi, N. Atchayaram, and P. K. Ghosh, “Comparison of speech tasks for automatic classification of patients with Amyotrophic Lateral Sclerosis and healthy subjects,” in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 6014–6018.

[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] B. Yamini, N. Shivashankar, and A. Nalini, “Measures of maximum performance of speech-related tasks in patients with Amyotrophic Lateral Sclerosis,” *Amyotrophic Lateral Sclerosis*, vol. 9, 2008.

[8] ——, “Vowel space area in patients with Amyotrophic Lateral Sclerosis,” *Amyotrophic Lateral Sclerosis*, vol. 9, no. 1, pp. 118–119, 2008.

[9] T. Bhattacharjee, Y. Belur, A. Nalini, R. Yadav, and P. K. Ghosh, “Exploring the role of fricatives in classifying healthy subjects and patients with Amyotrophic Lateral Sclerosis and Parkinson’s Disease,” in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[10] T. Antolík and F. Cecile, “Consonant distortions in dysarthria due to Parkinson’s disease, Amyotrophic Lateral Sclerosis and Cerebellar ataxia,” in *Proc. Interspeech*, 2013, pp. 2152–2156.

[11] C. V. T. Kumar, T. Bhattacharjee, Y. Belur, A. Nalini, R. Yadav, and P. K. Ghosh, “Classification of multi-class vowels and fricatives from patients having Amyotrophic Lateral Sclerosis with varied levels of dysarthria severity,” in *Proc. Interspeech*, 2023, pp. 146–150.

[12] R. D. Kent, R. L. Sufit, J. C. Rosenbek, J. F. Kent, G. Weismer, R. E. Martin, and B. R. Brooks, “Speech deterioration in Amyotrophic Lateral Sclerosis: A case study,” *Journal of Speech, Language, and Hearing Research*, vol. 34, no. 6, pp. 1269–1275, 1991.

[13] F. Tao, L. Daudet, C. Poellabauer, S. L. Schneider, and C. Busso, “A portable automatic PA-TA-KA syllable detection system to derive biomarkers for neurological disorders,” in *Proc. Interspeech*, 2016, pp. 362–366.

[14] 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.

[15] “MATLAB version: 9.10.0 (R2021a),” https://www.mathworks.com, The MathWorks Inc., Natick, Massachusetts, United States, 2021, [Online; accessed 10-Mar-2024].

[16] “Audacity,” https://audacityteam.org/, [Online; accessed 10-Mar-2024].

[17] W. Hsu, B. Bolte, Y. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, “HuBERT: Self-supervised speech representation learning by masked prediction of hidden units,” *CoRR*, vol. abs/2106.07447, 2021.

[18] A. T. Liu, S.-W. Li, and H.-y. Lee, “TERA: Self-supervised learning of transformer encoder representation for speech,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 2351–2366, 2021.

[19] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, “Wav2Vec 2.0: A framework for self-supervised learning of speech representations,” *CoRR*, vol. abs/2006.11477, 2020.

[20] S. Ling, Y. Liu, J. Salazar, and K. Kirchhoff, “Deep contextualized acoustic representations for semi-supervised speech recognition,” in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6429–6433.

[21] A. H. Liu, Y. Chung, and J. R. Glass, “Non-autoregressive predictive coding for learning speech representations from local dependencies,” *CoRR*, vol. abs/2011.00406, 2020.

[22] S. Liu, A. Mallol-Ragolta, E. Parada-Cabaleiro, K. Qian, X. Jing, A. Kathan, B. Hu, and B. W. Schuller, “Audio self-supervised learning: A survey,” *Patterns*, vol. 3, no. 12, 2022.

[23] W.-C. Huang, S.-W. Yang, T. Hayashi, H.-Y. Lee, S. Watanabe, and T. Toda, “S3PRL-VC: Open-source voice conversion framework with self-supervised speech representations,” in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 6552–6556.

[24] 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.

[25] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, “Pytorch: An imperative style, high-performance deep learning library,” *Advances in neural information processing systems*, vol. 32, 2019.

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

[27] S. Chandrasekar, A. Ramesh, T. Purohit, and P. K. Ghosh, “A study on the importance of formant transitions for stop-consonant classification in VCV sequence,” in *Proc. Interspeech*, 2023, pp. 4518–4522.

[28] A. Waibel, T. Hanazawa, and G. Hinton, “Phoneme recognition using time-delay neural networks,” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 31, no. 3, 1989.