Classification of multi-class vowels and fricatives from patients having Amyotrophic Lateral Sclerosis with varied levels of dysarthria severity

Chowdam Venkata Thirumala Kumar[1], *Tanuka Bhattacharjee*[1], *Yamini Belur*[2], *Atchayaram Nalini*[2], *Ravi Yadav*[2], *Prasanta Kumar Ghosh*[1]

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

kumarcvt55@gmail.com, tanukab@iisc.ac.in, prasantg@iisc.ac.in

Abstract

Dysarthria due to Amyotrophic Lateral Sclerosis (ALS) progressively distorts the acoustic space affecting the discriminability of different vowels and fricatives. However, the extent to which this happens with increasing severity is not thoroughly investigated. In this work, we perform automatic 4-class vowel (/a/, /i/, /o/, /u/) and 3-class fricative (/s/, /sh/, /f/) classification at varied severity levels and compare the performances with those from manual classification (through listening tests). Experiments with speech data from 119 ALS and 40 healthy subjects suggest that the manual and automatic classification accuracies reduce with an increase in dysarthria severity reaching 59.22% and 61.67% for vowels and 41.78% and 38.00% for fricatives, respectively, at the most severe cases. While manual classification is better than automatic one for all severity levels except the highest severity case for vowels, the difference between the two gradually reduces with an increase in severity.

Index Terms: Amyotrophic Lateral Sclerosis, dysarthria, severity, vowel, fricative, automatic classification, manual classification, listening test

Table 1: *Literature review on works and observations involving vowels and fricatives of dysarthric speech specific to ALS*

Vowel	**Analysis:** 1. Vowel height dimension is frequently misidentified due to limited tongue height control [8]. 2. Vowel contrasts reduce in severe patients [9, 5, 10].
	Manual Classification: 1. /u/ has less vowel intelligibility than /a/, /i/, and /o/ in control group but not in severe dysarthric group; /i/ is observed to have declined intelligibility with an increase in dysarthria severity [11].
	Automatic Classification: -
Fricative	**Analysis:** 1. Articulatory differences are observed between fricatives produced by speakers with ALS and healthy controls [12]. 2. Place of articulation gets affected for lingual fricatives in men with ALS [3]. 3. Unwanted voicing is observed in the voiceless fricative /s/ [13].
	Manual Classification: -
	Automatic Classification: -

1. Introduction

Speech sounds like different vowels and fricatives have their own individual acoustic characteristics which help in discriminating them from one another. However, the level of discrimination can change due to a variety of factors like background noise, reverberation, cross-talk etc. Impairments in speech production can also affect this discriminability. Dysarthria is one such speech disorder that progressively collapses the acoustic space of the affected individuals, thereby hampering the discriminability of different speech sounds. The effect becomes increasingly prominent with increase in the severity level. This, in turn, degrades the overall intelligibility of speech.

In this paper, we focus on dysarthria caused by Amyotrophic Lateral Sclerosis (ALS) in particular. This disorder impairs the speed and/or range of movements of articulators like lips, jaw, tongue and velum [1, 2]. Regulation of tongue height control is observed to be disrupted in most of these patients [3]. Poor laryngeal control further adds to the disabilities [2]. These articulatory impairments lead to a range of acoustic abnormalities including longer vowel durations, limited formant transitional extents, shallower formant slopes etc. [4]. These, in turn, adversely impact the discriminability of different sounds. For example, the vowel space area reduces in these patients [5] making it difficult to identify different vowels. Low vowels are commonly misidentified as high vowels [6]. The forms and extents of different speech impairments vary with the degree of severity of dysarthria. Reduced tongue and lip movement variabilities are observed in cases of mild to moderate dysarthria, while significantly elevated variabilities are displayed by severely dysarthric subjects [7]. In this work, we particularly focus on analyzing the discriminability of different vowel and fricative sounds with increasing severity of ALS-induced dysarthria.

Table 1 presents an extensive review of the major works done in the literature on dysarthric vowels and fricatives secondary to ALS. The efforts are primarily restricted to the analysis category where changes, caused by dysarthria, in the articulatory and acoustic characteristics of certain vowels and fricatives are studied. The sole attempt towards classifying dysarthric vowels has been made by Lee et al. [11]. They have done manual vowel classification at varied dysarthria severity levels through listening tests. No attempt has yet been made to analyze the performance of automatic algorithms for this purpose. On the other hand, no classification work (manual/automatic) has been carried out in the domain of dysarthric fricatives. In fact, this type of work for both vowels and fricatives is very rare in the case of normal speech itself. Till date, only Dewa [14] has developed a Convolutional Neural Network (CNN) based Javanese vowel classifier using Mel-frequency spectral coefficients obtained from healthy utterances. This paper aims to explore these gaps present in the literature of dysarthric vowels and fricatives. Thus our purpose is not to propose an alternative diagnostic tool for ALS vs. Healthy Control (HC) classification but to report a scientific study on the discriminability of vowels and fricatives, as perceived by human and automatic classifiers, at different severity levels of ALS-induced dysarthria. This can help enrich our understanding about how articulation is affected at different dysarthria severity levels and for different phonemes.

We consider **two tasks** - classification of **(1)** 4 sustained vowels, e.g. /a/, /i/, /o/ and /u/, and **(2)** 3 sustained fricatives, e.g. /s/, /sh/ and /f/. Both tasks are performed at varied severity levels of ALS-induced dysarthria. We perform automatic classifications using deep neural networks with spectral and self-supervised speech representations as the inputs. The automatic classification performances are compared with manual classification accuracies evaluated through listening tests. Speech data from 119 ALS and 40 HC subjects are used and 44 listeners performed the listening tests. Experimental results suggest that, for vowels, the manual and automatic classification accuracies drop from 89% and 78% in the case of HC subjects to 59.22% and 61.67%, respectively, in the case of the most severe patients. For fricatives, the respective drops are from 85.78% and 81.22% to 41.78% and 38%. Though humans classify both vowels and fricatives better than automatic classifiers at all dysarthria severity levels, except the most severe case for vowels, the gap between the automatic and manual classification accuracies reduces with increasing severity, reaching 2.45% and 3.78% for vowels and fricatives, respectively, at the most severe cases.

2. Dataset

Sustained utterances of 4 vowels, e.g. /a/, /i/, /o/, /u/, and 3 fricatives, e.g. /s/, /sh/, /f/, were collected from 119 ALS (73M, 46F) and 40 HC (20M, 20F) subjects at the National Institute of Mental Health and Neurosciences, India. These phonemes were chosen based on the abilities of the patients to produce the target sounds. The ALS and HC populations had aged in the ranges of 23 - 81 and 22 - 55 years, respectively. The subjects had 5 different native languages - Bengali, Hindi, Tamil, Telugu and Kannada. Dysarthria severity of each ALS patient was rated by 3 speech-language pathologists following the speech component of the ALSFRS-R scale [15]. It is a 5-point (0 - 4) measure where 0 indicates complete loss of useful speech and 4 signifies normal speech. The mode of the 3 ratings was taken as the final severity. We consider ALSFRS-R scores 0 and 1 together as the *severe dysarthric group* (G1), scores 2 and 3 together as the *mild dysarthric group* (G2) and score 4 as *ALS group with no dysarthria* (G3). Moreover, the HC subjects are taken together to form the *normal speech* group (G4). 40 subjects were recruited from each group except G1 which had 39 subjects. Each group was balanced w.r.t. gender and language distributions. During data collection, the subjects were asked to take a deep breath and prolong a vowel/fricative at a comfortable pitch (only for vowels) and loudness levels. The process was repeated 1-3 times for each phoneme depending on a subject's level of comfort. More details about the data collection process can be found in [16]. The number of utterances of each phoneme obtained from different severity groups, along with the mean and Standard Deviation (SD) of the duration of the utterances, is given in Table 2. The data collection protocol was approved by the hospital ethics committee. Also, a consent form was signed by each subject prior to data collection.

3. Classification Method

3.1. Automatic classification

We explore multi-layer dense neural networks (DNN) for automatic vowel/fricative classification. Self-supervised speech representations obtained through pre-trained models are fed as the input features to the networks. In particular, we consider 7 different pre-trained models, e.g. Wav2vec [17], Wav2Vec 2.0 [18], Hubert [19], Hubert-large [19], Tera [20], NPC [21] and Decoar 2.0 [22], for extracting the speech representations. These representations encode linguistic and paralinguistic details of the speech utterances in a compact fashion, and hence, might be suitable for the classification tasks at hand. Apart from these self-supervised representations, time-frequency representations of speech captured through mel-frequency cepstral coefficients (MFCC) along with their derivatives are also considered as input features for the DNN. In addition to these DNN-based approaches, we consider a CNN-based classifier, as proposed in [14], which takes mel-spectrograms along with their derivatives as input features.

3.2. Manual classification

Manual vowel/fricative classification is performed through listening tests. Since conducting listening tests on the entire corpus is highly time-consuming and extremely tedious, we selected a subset of the corpus for manual listening. We chose 20 subjects from each severity level with an almost equal number of subjects of both genders from each native language. One utterance of each vowel was chosen from a subject resulting in a total number of 320 (4 severity levels × 4 vowels × 20 subjects) vowel utterances. Similarly, 240 (4 severity levels × 3 fricatives × 20 subjects) utterances of fricatives were selected. Each utterance was classified by 3 human listeners who also provided a confidence score in the range of [0, 100] corresponding to each decision. The mode of these three decisions was taken as the final manual classification result. In case, all three decisions for an utterance turned out to be different, the one with the highest confidence score was taken as the final manual label.

We recruited 44 listeners to conduct this manual classification experiment. Listeners had ages ranging from 22 to 40 years with native languages spanning over Bengali, Hindi, Kannada, Malayalam, Tamil, and Telugu. Each listener was presented with 41 - 52 vowel and 31 - 40 fricative utterances with an almost equal number of utterances (9 - 12) from each severity group. This variability in number of utterances for each rater was due to the condition that each utterance should be rated by the 3 different listeners. The allotment was done by picking one utterance at a time and assigning it to 3 randomly chosen rater splits which had a less total number of utterances than the maximum possible value, i.e., 52 for vowels and 40 for fricatives. Random 8 utterances for both vowels and fricatives were presented twice to assess the consistency of the listeners. Moreover, the decisions given by the listeners for the utterances of the healthy group, i.e., G4, were used to calculate their accuracies. These accuracy and consistency measures can give us an idea

Table 2: *Number and duration of utterances of different phonemes 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*

	G1 (0,1)	**G2 (2,3)**	**G3 (4)**	**G4 (HC)**
/a/	104/2.81 (2.17)	117/4.20 (2.44)	113/4.71 (2.60)	112/5.08 (1.78)
/i/	103/2.21 (1.99)	116/1.93 (1.45)	114/4.51 (2.68)	111/4.94 (2.10)
/o/	105/2.30 (2.04)	117/3.48 (2.02)	113/4.71 (2.75)	110/4.90 (2.08)
/u/	97/2.09 (2.04)	116/3.42 (2.08)	113/4.53 (2.50)	112/4.73 (1.91)
/s/	69/0.79 (0.97)	116/1.93 (1.45)	114/2.86 (1.47)	111/3.84 (1.74)
/sh/	78/0.40 (0.61)	115/1.47 (1.20)	114/2.42 (1.75)	112/3.11 (1.51)
/f/	76/0.36 (0.60)	114/1.35 (1.35)	114/1.89 (1.47)	111/2.47 (1.91)

about how good a listener is. Only the responses from listeners with at least 75% accuracy and 75% consistency in the case of vowels and at least 60% accuracy and 60% consistency in the case of fricatives were considered further. This screening is expected to discard randomly chosen labels. This process resulted in the selection of only 20 listeners out of the 44 participants.

The entire manual classification experiment was conducted through two web applications designed for vowels and fricatives. Test audios were presented one at a time to a listener and multiple choices for the possible vowel/fricative classes (4 for vowels and 3 for fricatives) were shown. The listeners had to choose the correct option and mark their confidence level for each decision. The decisions once made could not be changed later. All listeners were presented with healthy utterances as examples before performing the classifications. They were allowed to listen to those examples, as well as the test audio, as many times as needed while performing the test. Moreover, the listeners were instructed to use earphones/headphones to be able to closely listen to the audio without any noise.

4. Experimental setup

4.1. Feature extraction

S3PRL [23] speech toolkit is used to compute self-supervised speech representations using pre-trained wav2vec, wav2vec 2.0, Hubert, Hubert-large, Tera, NPC, and Decoar 2.0 models. On the other hand, 12D MFCC and 40D mel-spectrogram (excluding the energy coefficients) with delta and double-delta measures are computed from every 20 ms speech frame with 10 ms overlap using the KALDI speech recognition toolkit [24], and MATLAB R2022b respectively. Table 3 lists the dimensions of all the considered speech representations along with the stride used for their computation.

4.2. Classifiers

In this work, we use 4-layer DNNs as one set of vowel/fricative classifiers. The networks have 1024, 512 and 256 neurons in the hidden layers with ReLU activation function. The output layer has 4 neurons in the case of vowel classification and 3 neurons in the case of fricative classification. Softmax is used as the activation function of the output layer for both classification tasks. The classifiers are trained using Adam optimizer with a learning rate of 0.0001 considering cross-entropy as the loss function. The batch size is set to 16 and the models are trained up to 100 epochs. We implement early stopping with patience of 5 to avoid overfitting. For the CNN-based classifier considered in this work, we adopt the same settings as described in [14]. All DNN model trainings are done at the frame level and the CNN training is done at 80 ms chunk level. While testing, the mode of the predicted labels for all frames/chunks is considered as the final decision for the utterance. Training and testing are performed for each severity group separately. All implementations are done in Pytorch v1.11.0 [25]. An NVIDIA GeForce RTX 2080 GPU is used for training and testing the models.

Table 3: *Strides and output dimensions of different types of self-supervised speech representations*

Details	**MFCC**	**wav2vec**	**wav2vec 2.0**	**Hubert**
Stride (ms)	10	10	20	20
O/P Dim	36	512	768	768
	Hubert large	**Tera**	**NPC**	**Decoar 2.0**
Stride (ms)	20	10	10	10
O/P Dim	1024	768	512	768

4.3. Evaluation method

We perform the experimental evaluation in two phases. In the first phase, we evaluate all automatic classifiers through 5-fold cross-validation separately at each severity level for both vowels and fricatives and analyze their relative performances. Folds are created randomly with an equal number of subjects in each fold. The same fold structure is maintained across all the classifiers. In the second phase, we evaluate the performances of the classifiers on the same dataset as used for the manual listening tests. In each severity level, we have nearly 40 subjects among whom 20 are used for listening tests, or in turn, as the test set. We use 3 subjects from the remaining as the validation set while the rest of the subjects are used as the training set for the automatic classifiers. Lastly, we perform the Wilcoxon signed-rank test [26] at 1% significance level to determine if the classification performances achieved through automatic and manual modes are significantly different.

A 5-fold evaluation method is performed to check the consistency of the models as the human test set involving more subjects results in less training data. Two levels of comparison give a better understanding of the performance of models with changes in the training set size for automatic classification. The second phase of comparison is considered for all conclusions.

5. Results and discussion

5.1. Comparison of different automatic methods

Table 4 and 5 summarize the performances of all models mentioned above for the classification of vowels and fricatives. We observe a decline in the performance with an increase in severity for all speech representations except a few cases. Tera and Decoar 2.0 achieve better performance for G3 than G4 in the case of vowels. The same pattern is observed for Tera and NPC in the case of fricatives. For vowels, CNN model achieves the best performance for G1 and G2 whereas Hubert performs the best for G3 and G4. In the case of fricatives, Hubert-large shows the best performance for G1 while CNN shows the best performance for all other severity groups, i.e., G2, G3, and G4. Overall no single model achieves the best performance across all severity levels for both vowels and fricatives. Though CNN performs the best for G2, G3, and G4 for fricative classification, it fails with a notable difference for the severe group G1.

5.2. Comparing automatic and manual classification

Comparison of the automatic classification with manual classification at different severity levels is presented in Figure 2. Hubert for vowels and wav2vec 2.0 for fricatives outperform all other models on the manual test set when we consider average performance across all severity levels. Even though CNN shows the best performance during 5-fold cross-validation, it is unable to achieve similar performance on the manual test set. Considering both the 5-fold accuracies and manual test set accuracies, the best performing model for vowels turns out to be CNN, followed by DNNs with speech representations from wav2vec, Hubert and Hubert-large. For fricatives, the top 4 models include CNN, followed by DNNs with wav2vec, wav2vec 2.0 and Hubert representations.

It is observed that, for the severe group G1, automatic vowel classification with Hubert shows better performance (though

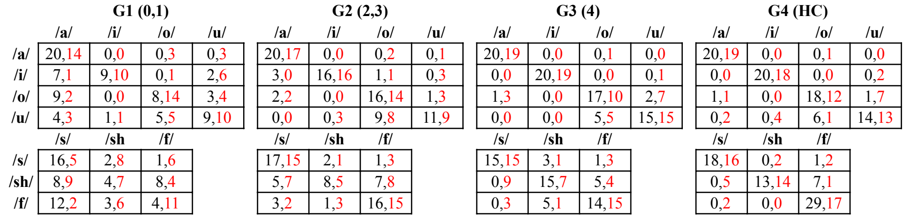
Table 4: *Mean vowel classification accuracies in % (SD in bracket) over 5-fold cross-validation obtained using different automatic classification methods*

Severity Groups	MFCC	wav2vec	wav2vec 2.0	Hubert	Hubert large	Tera	NPC	Decoar 2.0	CNN
G1 (0,1)	33.82 (5.24)	50.17 (1.79)	49.06 (7.18)	46.34 (12.36)	43.85 (6.58)	34.57 (3.05)	24.8 (4.76)	28.34 (4.58)	55.43 (14.34)
G2 (2,3)	38.67 (4.54)	64.97 (6.38)	64.59(7.41)	59.56 (15.19)	64.2 (6.25)	43.58 (4.71)	25.44 (2.67)	29.95 (4.68)	66.21 (8.01)
G3 (4)	42.58 (6.29)	66.58 (5.75)	69.16 (6.66)	74.21 (5.69)	68.54 (6.00)	54.45 (5.09)	29.52 (4.63)	44.31 (3.50)	70.63 (6.58)
G4 (HC)	41.11 (2.21)	73.44 (5.25)	72.69 (7.11)	78.39 (7.87)	73.43 (5.05)	52.19 (6.87)	29.87 (4.63)	42.58 (4.69)	77.34 (4.20)

Table 5: *Mean fricative classification accuracies in % (SD in bracket) over 5-fold cross-validation obtained using different automatic classification methods*

Severity Groups	MFCC	wav2vec	wav2vec 2.0	Hubert	Hubert large	Tera	NPC	Decoar 2.0	CNN
G1 (0,1)	31.75 (6.57)	41.88 (7.57)	34.04 (13.24)	40.53 (5.44)	44.38 (11.07)	37.64 (6.10)	30.11 (7.32)	33.05 (2.75)	32.36 (6.01)
G2 (2,3)	36.83 (7.52)	57.81 (11.19)	54.61 (8.21)	46.84 (11.38)	43.26 (8.86)	44.93 (5.04)	33.46 (33.46)	37.87 (5.90)	63.93 (6.31)
G3 (4)	40.49 (5.71)	67.37 (7.75)	54.49 (4.20)	58.27 (6.17)	48.85 (10.27)	45.19 (6.72)	35.76 (2.95)	48.09 (7.12)	70.36 (6.65)
G4 (HC)	43.86 (6.36)	73.01 (4.65)	55.2 (5.55)	65.33 (13.42)	58.11 (12.42)	41.33 (7.90)	35.4 (5.29)	48.17 (6.26)	73.6 (8.12)

Figure 1: *Confusion matrices for vowels and fricatives using manual (in black) and the best-performing automatic (in red) classification*



G1 (0,1)

	/a/	/i/	/o/	/u/
/a/	20,14	0,0	0,3	0,3
/i/	7,1	9,10	0,1	2,6
/o/	9,2	0,0	8,14	3,4
/u/	4,3	1,1	5,5	9,10

	/s/	/sh	/f/
/s/	16,5	2,8	1,6
/sh/	8,9	4,7	8,4
/f/	12,2	3,6	4,11

G2 (2,3)

	/a/	/i/	/o/	/u/
/a/	20,17	0,0	0,2	0,1
/i/	3,0	16,16	1,1	0,3
/o/	2,2	0,0	16,14	1,3
/u/	0,0	0,3	9,8	11,9

	/s/	/sh	/f/
/s/	17,15	2,1	1,3
/sh/	5,7	8,5	7,8
/f/	3,2	1,3	16,15

G3 (4)

	/a/	/i/	/o/	/u/
/a/	20,19	0,0	0,1	0,0
/i/	0,0	20,19	0,0	0,1
/o/	1,3	0,0	17,10	2,7
/u/	0,0	0,0	5,5	15,15

	/s/	/sh	/f/
/s/	15,15	3,1	1,3
/sh/	0,9	15,7	5,4
/f/	0,3	5,1	14,15

G4 (HC)

	/a/	/i/	/o/	/u/
/a/	20,19	0,0	0,1	0,0
/i/	0,0	20,18	0,0	0,2
/o/	1,1	0,0	18,12	1,7
/u/	0,2	0,4	6,1	14,13

	/s/	/sh	/f/
/s/	18,16	0,2	1,2
/sh/	0,5	13,14	7,1
/f/	0,2	0,0	29,17

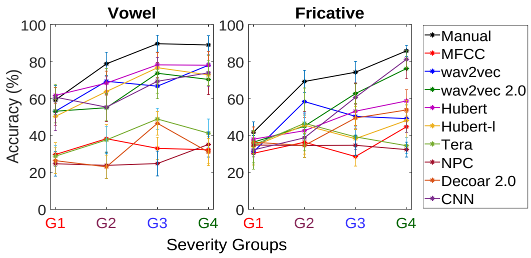


Figure 2: *Mean automatic and manual classification accuracies (SD in error bar) for vowels and fricatives at different severity levels evaluated on the manual test set*

not significant) than manual classification, whereas for other severity groups manual classification shows significantly better performance than automatic. But in the case of fricatives, manual classification shows the best performance at all severity levels. This comparison reveals that, apart from the classification of vowels for the severe group, manual classification performs better than the automatic one in all other cases. However, the gap between the performances of automatic and manual classification methods is observed to reduce with an increase in the severity level for both vowels and fricatives. From this pattern, we can conclude that the different speech representations considered here may fail to capture the discriminative information for different dysarthric vowels/fricatives even when human perception is still able to perceive the differences. Figure 1 further illustrates that, for the most severe group, both in automatic and manual classification, more confusion happens between vowels and most utterances are misclassified as vowel /a/.

5.3. Language analysis

Language-wise performances of the automatic and manual classification at different severity levels for vowels and fricatives are shown in Figure 3. For the automatic classification of vowels, we consider the best case on the manual test set which is Hubert and similarly wav2vec 2.0 for fricatives. We observe mostly similar trends in all languages for vowels and fricatives except in a few cases where the performance on a higher severity group is better than a lower one. No other notable trend is observed in the performances. This might be because we are studying isolated utterances of certain vowels and fricatives, all of which are present in all the five languages under consideration. Hence, the effect of language might be minimal.

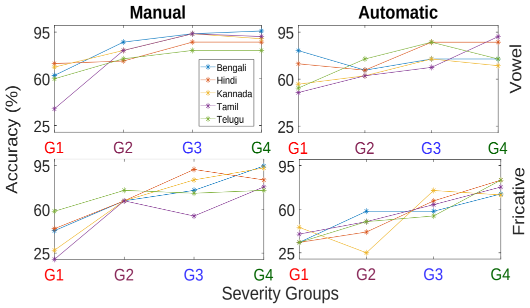


Figure 3: *Language wise accuracies of automatic and manual classification of vowels and fricatives at different severity levels*

6. Conclusion

In this work, we analyze the relative performance of manual and automatic classification of voiced vowels and voiceless fricatives for dysarthric speech secondary to ALS. Classification performances decline with an increase in severity. Manual classification is always better than automatic classification except for the highest severity case of vowels and the performance gap reduces with an increase in severity level. Vowel classification performances are always higher than fricatives concluding that voiced sounds are more differentiable than voiceless sounds in ALS. Moreover, no language-specific pattern is observed.

Acknowledgments - We sincerely thank Seena Vengalil and M S Keertipriya for their assistance during data collection. We also thank the Department of Science and Technology (DST), Govt. of India for supporting this work.

7. References

[1] 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.

[2] 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.

[3] R. D. Kent, J. F. Kent, G. Weismer, R. L. Sufit, J. C. Rosenbek, R. E. Martin, and B. R. Brooks, "Impairment of speech intelligibility in men with Amyotrophic Lateral Sclerosis," *Journal of Speech and Hearing Disorders*, vol. 55, no. 4, pp. 721–728, 1990.

[4] 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.

[5] B. Yamini, N. Shivashankar, and A. Nalini, "Vowel space area in patients with Amyotrophic Lateral Sclerosis," *Amyotrophic Lateral Sclerosis*, vol. 9, no. 1, pp. 118–119, 2008.

[6] K. Bunton and G. Weismer, "The relationship between perception and acoustics for a high-low vowel contrast produced by speakers with dysarthria," *Journal of Speech, Language, and Hearing Research*, vol. 44, no. 6, pp. 1215–1228, 2001.

[7] M. Kuruvilla-Dugdale and A. Mefferd, "Spatiotemporal movement variability in ALS: Speaking rate effects on tongue, lower lip, and jaw motor control," *Journal of Communication Disorders*, vol. 67, pp. 22–34, 2017.

[8] J. Lee, H. Kim, and Y. Jung, "Patterns of misidentified vowels in individuals with dysarthria secondary to Amyotrophic Lateral Sclerosis," *Journal of Speech, Language, and Hearing Research*, vol. 63, no. 8, pp. 2649–2666, 2020.

[9] P. Rong, "The effect of tongue–jaw coupling on phonetic distinctiveness of vowels in Amyotrophic Lateral Sclerosis," *Journal of Speech, Language, and Hearing Research*, vol. 62, no. 9, pp. 3248–3264, 2019.

[10] P. Rong, E. Usler, L. M. Rowe, K. Allison, J. Woo, G. El Fakhri, and J. R. Green, "Speech intelligibility loss due to Amyotrophic Lateral Sclerosis: The effect of tongue movement reduction on vowel and consonant acoustic features," *Clinical Linguistics & Phonetics*, vol. 35, no. 11, pp. 1091–1112, 2021.

[11] J. Lee, E. Dickey, and Z. Simmons, "Vowel-specific intelligibility and acoustic patterns in individuals with dysarthria secondary to Amyotrophic Lateral Sclerosis," *Journal of Speech, Language, and Hearing Research*, vol. 62, no. 1, pp. 34–59, 2019.

[12] K. Tjaden and G. S. Turner, "Spectral properties of fricatives in Amyotrophic Lateral Sclerosis," *Journal of Speech, Language, and Hearing Research*, vol. 40, no. 6, pp. 1358–1372, 1997.

[13] T. Antolík and F. Cecile, "Consonant distortions in dysarthria due to parkinson's disease, amyotrophic lateral sclerosis and cerebellar ataxia," *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, pp. 2152–2156, 2013.

[14] C. K. Dewa, "Javanese vowels sound classification with convolutional neural network," in *International Seminar on Intelligent Technology and Its Applications (ISITIA)*, 2016, pp. 123–128.

[15] 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.

[16] 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.

[17] S. Schneider, A. Baevski, R. Collobert, and M. Auli, "wav2vec: Unsupervised pre-training for speech recognition," *CoRR*, vol. abs/1904.05862, 2019. [Online]. Available: http://arxiv.org/abs/1904.05862

[18] 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. [Online]. Available: https://arxiv.org/abs/2006.11477

[19] 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. [Online]. Available: https://arxiv.org/abs/2106.07447

[20] 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.

[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. [Online]. Available: https://arxiv.org/abs/2011.00406

[22] S. Ling and Y. Liu, "Decoar 2.0: Deep contextualized acoustic representations with vector quantization," *arXiv preprint arXiv:2012.06659*, 2020.

[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.