Comparison of Acoustic and Textual Features for Dysarthria Severity Classification in Amyotrophic Lateral Sclerosis

Upendra Vishwanath Y. S.[1], *Tanuka Bhattacharjee*[2], *Deekshitha G*[2], *Sathvik Udupa*[3], *Kumar Chowdam Venkata Thirumala*[4], *Madassu Keerthipriya*[5], *Darshan Chikktimmegowda*[5], *Dipti Baskar*[5], *Yamini Belur*[5], *Seena Vengalil*[5], *Atchayaram Nalini*[5], *Prasanta Kumar Ghosh*[2]

[1]Hewlett Packard Enterprise, Bengaluru, India; [2]Electrical Engineering Department, Indian Institute of Science, Bengaluru, India; [3]Brno University of Technology, Brno, Czech Republic; [4]Speech Processing Lab, LTRC, International Institute of Information Technology, Hyderabad, India; [5]National Institute of Mental Health and Neurosciences, Bengaluru, India

upendravishwanath7@gmail.com, prasantg@iisc.ac.in

Abstract

We explore language-agnostic deep text embeddings for severity classification of dysarthria in Amyotrophic Lateral Sclerosis (ALS). Speech recordings are transcribed by human and ASR and embeddings of the transcripts are considered. Though speech recognition accuracy has been studied for grading dysarthria severity, no effort has yet been made to utilize text embeddings of the transcripts. We perform severity classification at different granularity (2, 3, and 5-class) using data obtained from 47 ALS subjects. Experiments with dense neural network based classifiers suggest that, though text features achieve nearly equal performances as baseline speech features, like statistics of mel frequency cepstral coefficients (MFCC), for 2-class classification, speech features outperform for higher number of classes. Concatenation of text embeddings and MFCC statistics attains the best performances with mean F1 scores of 88%, 68%, and 53%, respectively, in 2, 3, and 5-class classification.

Index Terms: Amyotrophic Lateral Sclerosis, dysarthria, severity prediction, acoustic features, textual features

1. Introduction

Amyotrophic Lateral Sclerosis (ALS) is a rapidly progressive neurodegenerative disease that affects various motor functions [1]. Speech functionalities, among others, get severely affected leading to dysarthria. Currently, there are no cures for ALS and the associated dysarthria. However, early diagnosis, methodical treatment and personalized disease management strategies can slow the disease progression and improve the quality of life of the patients. Regular monitoring of the disease severity is essential to continuously cater to the therapeutic needs of the patients. Speech-Language Pathologists (SLPs) typically examine the dysarthria severity of an ALS patient using the Frenchay Dysarthria Assessment method [2] and/or the Revised Amyotrophic Lateral Sclerosis Functional Rating Scale (ALSFRS-R) [3]. These assessments are tedious, time expensive and costly. Moreover, the clinician's familiarity with the speaker or the subject matter can influence the judgment [4]. Hence, there is an urgent need for objective and accurate automatic dysarthria severity prediction systems.

Acoustic properties of speech have been commonly explored in the literature for severity prediction of ALS-induced dysarthria. Suhas et al. [5] have performed 5-class dysarthria severity classification for ALS using a 2D Convolutional Neural Network (CNN) which takes log-mel spectrogram as input. Vieira et al. [6] have developed a CNN model for predicting the ALSFRS-R speech score from raw speech signals. Bhattacharjee et al. [7] have leveraged transfer learning approaches, like fine-tuning from auxiliary tasks and multi-task learning, to perform 3-class dysarthria severity classification for ALS. They have used temporal statistics of mel-frequency cepstral coefficients (MFCC) together with dense neural network (DnNN) based models. Along with acoustic features, articulatory cues of speech have also been explored by Wisler et al. [8] for estimating the ALSFRS-R bulbar subscore. They have used linear ridge regression and support vector regression for this purpose.

Researchers have also explored the accuracy of speech recognition as a potential marker for grading speech intelligibility and severity of dysarthric speech [9, 10, 11, 12, 13]. Word error rate (WER) of transcriptions obtained from human listeners (human speech recognition or HSR) as well as off-the-shelf automatic speech recognition (ASR) systems (such as Google Cloud ASR API[1]) have been studied. Since human listeners are more accustomed to typical speech and off-the-shelf ASR models are generally trained on only typical speech, the accuracy of both HSR and ASR degrades as the speech becomes more atypical or unintelligible with increasing dysarthria severity. Though significant correlation between ASR performance and speech intelligibility or impairment severity has been reported in a few works [10, 11], Gutz et al. [12] claimed WER of Google Cloud ASR to be insufficient for grading dysarthria severity for ALS. They found the accuracy and stability of this WER based approach to be particularly poor for the mildly impaired group. Apart from the WER-based approaches, Choi et al. [14] have used pronunciation correctness and structural prosody related features obtained from ASR transcripts of dysarthric speech for dysarthria severity prediction in stroke patients. They have fine-tuned an off-the-shelf Whisper model [15] using dysarthric speech for performing the ASR.

Though some speech recognition based approaches have been explored in the literature for dysarthria severity prediction, no effort has been made till date to utilize textual embeddings obtained from HSR or ASR transcripts for this purpose. This paper aims to study that aspect. We explore textual embeddings of HSR and ASR transcripts obtained using language-agnostic sentence embedding models, e.g. Language-Agnostic SEntence Representations (LASER) [16] and Language-agnostic BERT Sentence Embedding (LaBSE) [17], for performing severity classification at different granularity. In particular, we aim to answer the following key questions.

[1] `https://cloud.google.com/speech-to-text/v2/docs/chirp-model`

1. What is the relative performance of textual embeddings obtained from different HSR and ASR configurations?
2. How does the performance of textual features compare to those of well established acoustic features, like, MFCC, openSMILE [18] and self-supervised speech representations obtained using a Wav2Vec2 (W2V2) model [19]?
3. Do the acoustic and textual cues carry complementary information such that their fusion can yield better severity classification performance?

2. Database

Data collection was performed at National Institute of Mental Health and Neurosciences (NIMHANS), Bengaluru, India. The database contains 2280 audio files obtained from 47 ALS subjects belonging to 5 native Indian languages, e.g., Bengali, Hindi, Kannada, Tamil, and Telugu. The male:female ratio in these languages are 5:4, 2:8, 3:6, 3:6, and 4:6, respectively, with an average age of 54.51 years across subjects. The audio files contain recordings of ALS subjects describing some images presented to them. All recordings are done in the subjects' respective native languages. The recordings have an average length of 5.54 sec with a standard deviation (SD) of 2.92 sec. Dysarthria severity of each subject is rated by 3 SLPs from NIMHANS on a scale of 0-4 (0: Loss of useful speech, and 4: Normal speech). The mode of the 3 ratings is considered as the final severity. The priors of the 5 severity classes are: [0: 13.03%, 1: 15.88%, 2: 21.49%, 3: 21.49%, 4: 28.11%]. Table 1 reports the details of the database. More information about the data collection protocol can be found in [20].

Table 1: *Language and severity-wise distribution of the number of subjects/utterances*

Severity	**0**	**1**	**2**	**3**	**4**	**Total**
Bengali	2/80	2/102	1/51	2/113	2/113	9/459
Hindi	2/40	2/50	1/66	3/149	2/98	10/403
Kannada	1/42	2/103	2/102	1/51	3/161	9/459
Tamil	1/44	2/56	2/119	1/75	3/155	9/449
Telugu	2/91	1/51	3/152	2/102	2/114	10/510
Total	8/297	9/362	9/490	9/490	12/641	47/2280

3. Severity classification

As illustrated in Figure 1, the severity classification is done by extracting speech features from the raw speech signals and text features from the corresponding transcripts (obtained using ASR and HSR). We use three different feature-level configurations for performing the severity classification - (1) using speech features alone, (2) using text features alone (obtained from HSR or ASR), and (3) using a multi-modal representation created by concatenating speech and text features together. In all cases, DnNN based models are used as the classifier. In this study, three types of dysarthria severity classification experiments are conducted: (1) 2-class: [High: 0,1, Low: 2,3,4], (2) 3-class: [High: 0,1, Mild: 2,3, Normal: 4], and (3) 5-class: [Highest: 0, High: 1, Mild: 2, Low: 3, Normal: 4].

3.1. Speech Representations

We obtain speech representations using three approaches: (1) a feature set based on the mean, SD, and median of the MFCC coefficients (MFCC Stat), (2) functional features from openSMILE - emobase [18], and (3) self-supervised learning (SSL) representations from the W2V2-base-960h model [19]. Since these SSL representations are at the frame level, we compute their average across frames.

3.2. Speech-To-Text

3.2.1. Automatic Speech Recognition

We use off-the-shelf language-specific W2V2-based high-performant ASR models, available on the Huggingface platform. Each ASR model is fine-tuned separately for each individual target language. In particular, we use Indic ASR (iASR), Google ASR (gASR) and Vakyansh (vASR). Each audio file in our database is transcribed using each of the three models, thereby generating 3 ASR transcripts for each audio.

3.2.2. Human Speech Recognition

Six different human transcribers transcribe each audio, thereby generating 6 HSR transcripts for each audio. The native language of the transcribers is the same as the language of the audio. The transcribers are college or university students and do not have any hearing impairment. They are asked to write the most likely valid words for whatever they hear in their native language script. They mark filled pauses, long silences and unintelligible speech regions by the keywords <PAUSE>, <LONGSIL>, and <GARBAGE>, respectively. Click sounds originating from the recording setup are transcribed by the keyword <CLICK>. All the four keywords are written in English inside < and > symbols. Three among the six transcribers are shown the image which is being described by the speaker in the audio. This is referred to as 'with image' (WI) scenario. This is done to understand if knowing the image helps in understanding the speech content as that image itself is being described in the speech. The other three transcribers transcribe without the image information. This is referred to as 'without image' (WOI) scenario. We also consider WI and WOI transcripts together which is referred to as WOI+WI scenario. Within each scenario, the keywords are treated in 3 different ways - (1) all keywords are considered as it is for text embedding extraction (WK), (2) all keywords are removed from the transcripts before text embedding extraction (WoK), and (3) the keywords are transliterated to the corresponding language script before text embedding extraction (TrK).

3.3. Text Representation

This study examines speech data in 5 different Indian languages. To generate multilingual sentence representations, language-agnostic sentence embedding models, such as LASER and LaBSE, which support 112 and 147 languages, respectively, are utilized. These models map natural language data from different languages to a common embedding space. The embedding vectors from this space are then utilized for severity classification. We extract a sentence embedding from each transcript

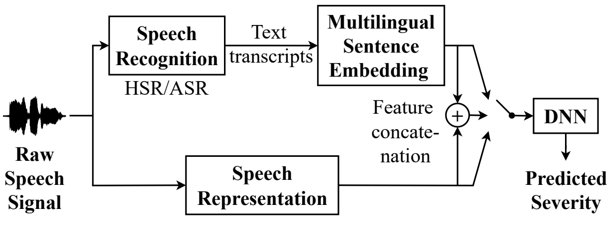

Figure 1: *Dysarthria severity classification framework*

produced by each human transcriber or ASR model. For each of WK, WoK, and TrK settings, there are multiple HSR transcripts, and hence multiple text embeddings, obtained for each audio in each of WOI, WI, and WOI+WI scenarios. Hence, we compute the average of all embeddings corresponding to a particular scenario to obtain a final representative embedding. Similarly, for ASR, we compute the average of the embeddings of the three transcripts of an audio obtained from the three ASR models to obtain a combined ASR-based embedding (CombASR).

3.4. Classifier

DnNN-based classifiers are used in this work. The DnNN models consist of 5 fully connected layers with 64, 32, 16, and 4 hidden units, along with Rectified Linear Unit (ReLU) [21] as activation function. The number of output units corresponds to the number of classes. The models are trained using cross-entropy loss function and Adam optimizer [22], with a learning rate of 0.001 and a batch size of 32. Training is conducted for a maximum of 200 epochs with early stopping (patience = 5 epochs) based on validation loss to prevent overfitting. Dropout regularization is applied after the first two layers with probabilities of 0.3 and 0.2, respectively. Batch normalization is also used after the first two layers to stabilize training[2].

4. Experimental setup

The performance of the dysarthria severity classification task is evaluated across different text and acoustic-based feature sets. Specifically, text features from LaBSE (768-dim), LASER (1024-dim), and acoustic features from statistical features of MFCC (108-dim), openSMILE (feature set = emobase, feature level = functionals, 988-dim), and W2V2-base-960h (768-dim) model are considered. MFCC features are computed with 20ms frame length and 10ms frame shift. Feature vectors are normalized using the mean and SD calculated from the training set to ensure uniformity across the train and test sets. We conduct experiments using a 5-fold cross-validation setup. Each fold contains nearly equal number of unique subjects from each severity class. Similar distributions of age, gender and language are maintained across the folds. We compute the F1 score, precision and recall on the test fold in each iteration of cross-validation. The mean and SD of these measures over the five folds are reported as the performance metrics.

5. Results and Discussion

5.1. Analysis on HSR-based Text Representations

We perform multiple experiments on HSR transcripts to study the influence of (a) keywords in transcription, and (b) prior knowledge regarding the image being described in the audio while doing the transcription, on the classification performance. The results of these experiments are presented in Table 2. From the experiments, it can be observed that LaBSE performs similar to or better than LASER in all cases except a few for 3-class classification. Hence, in the subsequent discussions, we consider the experiments based on LaBSE embeddings only.

5.1.1. Influence of Keywords

To study the influence of keywords in HSR transcripts on classification performance, text classification is performed in WK, WoK, and TrK settings. We observe from Table 2 that the WoK transcripts achieve the highest mean accuracies in most cases, but WK and TrK performances are also not significantly inferior. Thus the keywords do not seem to carry extra cues about the speech intelligibility. The classification seems to happen mostly based on the linguistic parts of the transcripts which are common to all of WK, WoK and TrK. Hence, for all further analyses, we focus on WoK setting only.

5.1.2. Influence of Image Information

To study the influence of the image information provided to human transcribers on the classification performance, experiments are done on (a) transcripts from WI group, (b) transcripts from WOI group, and (c) both groups combined. As discussed in previous subsections, focusing on WoK transcripts and LaBSE embedding, we observe that WI results in similar or lower F1 scores than WOI and WOI + WI. Moreover, WOI achieves similar or higher F1 scores than WOI + WI in these cases. Having the image information might provide the transcriber cues about the content being spoken and help in better transcribing the audios. Thus the decline in speech intelligibility with severity might be less reflected in the WI transcripts, leading to lower performance than WOI cases. Hence, we consider only WOI (WoK) with LaBSE embeddings for all further comparisons.

5.2. Analysis on ASR-based Text Representations

From the experiments on different ASRs i.e., iASR, gASR and vASR, as shown in Table 3, we can observe that the classification performance order is gASR $>$ vASR $>$ iASR. Performance on gASR is significantly higher, regardless of the feature extraction method used. Furthermore, utilizing the transcripts of all three ASRs (referred as combASR) slightly improved classification performance as shown in Table 3. Furthermore, experimental results consistently show that LaBSE outperforms LASER embeddings across all experiments.

5.3. HSR vs. ASR

The classification performance obtained on ASR transcripts is similar to that of HSR transcripts, except in the 3-class experiments, where CombASR outperforms all HSR experiments. The combined experiment using HSR (WOI, WoK) and ASR (CombASR) yields performance similar to, though slightly lower than, that of CombASR alone.

5.4. Comparison of speech and textual features

As summarized in Table 4, deep speech features extracted using the W2V2-base model outperform openSMILE-emobase features, achieving improvements of 10%, 7%, and 9% on the F1 score for classification tasks of five classes, three classes, and two classes, respectively. However, among all speech feature extraction methods, the statistical features of MFCC significantly outperform both W2V2-base and openSMILE-emobase. The speech based experiments i.e, MFCC stat and W2V2-base demonstrate superior results, outperform best text experiments (ASR:CombASR - Table 3, HSR:WOI+WoK - Table 2) in terms of average F1-score. The best speech experiments (MFCC stats) outperform the best text experiments by (7%, 7%), (9%, 6%) and (3%, 1%) for five, two and three class problems respectively. This suggests that the text information does not fully capture the characteristics of dysarthria. It is evident that for more granular classification tasks, such as the 5-class and 3-class problems, speech representations provide more insights

[2] `https://github.com/UpendraVishwanathYS/Dysarthria-Severity-Classification-in-ALS`

Table 2: *Mean ± SD of performance metrics obtained using different HSR configurations; here, bold entries indicate best performances with respect to mean F1 score in WOI, WI, and WOI+WI settings for each of 5-class, 3-class, and 2-class classification.*

Classification Setting		**Feature**	**WOI**			**WI**			**WOI + WI**		
			F1 Score (%)	**Precision** (%)	**Recall** (%)	**F1 Score** (%)	**Precision** (%)	**Recall** (%)	**F1 Score** (%)	**Precision** (%)	**Recall** (%)
5-Class	WK	LaBSE	36.35 ± 5.76	36.15 ± 4.54	42.91 ± 4.34	**35.21 ± 4.94**	38.05 ± 7.12	39.81 ± 4.72	38.87 ± 7.41	38.02 ± 8.10	45.36 ± 5.65
		LASER	32.36 ± 4.99	34.19 ± 4.36	36.66 ± 5.25	27.65 ± 5.22	25.55 ± 8.52	35.19 ± 3.67	35.12 ± 5.02	36.26 ± 4.51	38.02 ± 4.86
	WoK	LaBSE	**44.03 ± 2.16**	45.68 ± 2.52	47.48 ± 2.31	34.25 ± 3.97	36.15 ± 8.21	40.44 ± 2.20	**40.77 ± 5.37**	40.87 ± 6.09	46.47 ± 4.21
		LASER	36.59 ± 5.12	37.18 ± 6.37	39.62 ± 4.52	33.08 ± 3.44	33.85 ± 6.63	37.37 ± 1.20	37.05 ± 7.10	38.51 ± 6.91	43.69 ± 3.60
	TrK	LaBSE	30.85 ± 4.13	34.36 ± 4.51	40.02 ± 3.06	31.95 ± 4.21	33.68 ± 6.79	36.53 ± 3.49	34.00 ± 5.20	36.86 ± 4.93	40.30 ± 5.38
		LASER	22.68 ± 6.90	26.18 ± 7.77	26.02 ± 7.68	24.12 ± 6.87	22.87 ± 9.10	29.84 ± 5.34	26.81 ± 6.89	26.67 ± 8.95	31.94 ± 4.66
3-Class	WK	LaBSE	55.49 ± 4.86	65.83 ± 12.93	60.56 ± 3.61	52.93 ± 7.91	55.09 ± 9.63	56.24 ± 6.19	56.71 ± 7.46	55.19 ± 11.68	61.66 ± 5.50
		LASER	**57.24 ± 6.59**	56.41 ± 10.24	61.07 ± 4.62	**54.62 ± 6.25**	58.74 ± 9.19	56.74 ± 4.61	55.52 ± 7.44	54.18 ± 11.79	60.05 ± 5.59
	WoK	LaBSE	57.17 ± 5.36	58.09 ± 11.18	61.92 ± 3.79	50.75 ± 4.41	50.96 ± 7.73	54.85 ± 3.16	53.69 ± 2.55	50.73 ± 5.39	60.12 ± 1.49
		LASER	54.87 ± 4.12	52.68 ± 7.18	59.33 ± 2.10	51.37 ± 7.13	52.30 ± 10.05	53.82 ± 4.56	**58.38 ± 5.11**	61.43 ± 8.04	61.20 ± 3.69
	TrK	LaBSE	55.65 ± 6.56	59.51 ± 14.22	60.52 ± 4.95	48.40 ± 3.48	47.82 ± 5.19	53.31 ± 2.97	56.94 ± 8.44	57.85 ± 12.34	61.73 ± 6.65
		LASER	53.29 ± 2.27	52.20 ± 4.35	57.05 ± 1.96	49.91 ± 2.83	52.96 ± 5.28	52.60 ± 1.71	52.97 ± 2.23	53.18 ± 5.17	56.00 ± 3.03
2-Class	WK	LaBSE	85.67 ± 3.26	85.66 ± 4.41	86.56 ± 2.80	**79.43 ± 4.64**	82.43 ± 6.01	77.93 ± 4.34	84.87 ± 5.46	85.00 ± 6.33	85.46 ± 5.17
		LASER	84.51 ± 3.57	85.55 ± 5.14	84.14 ± 3.00	79.05 ± 3.65	82.58 ± 5.96	77.40 ± 3.13	83.61 ± 4.06	84.91 ± 5.76	83.08 ± 3.40
	WoK	LaBSE	**85.89 ± 4.09**	85.88 ± 4.88	86.33 ± 3.49	79.24 ± 4.72	82.93 ± 5.84	77.47 ± 4.40	**85.79 ± 4.95**	86.21 ± 6.27	86.12 ± 3.56
		LASER	84.64 ± 3.71	84.84 ± 4.68	84.95 ± 3.13	78.01 ± 4.82	80.88 ± 6.56	76.52 ± 4.57	85.25 ± 5.37	86.02 ± 6.82	85.01 ± 4.13
	TrK	LaBSE	84.46 ± 4.78	84.52 ± 5.75	85.36 ± 4.87	78.85 ± 4.65	81.45 ± 5.30	77.56 ± 4.71	84.54 ± 5.14	84.87 ± 6.09	84.81 ± 4.67
		LASER	81.91 ± 4.92	82.67 ± 5.54	82.02 ± 5.14	76.57 ± 3.89	79.14 ± 4.04	75.29 ± 4.05	81.91 ± 4.00	83.38 ± 5.78	81.49 ± 3.02

Table 3: *Mean ± SD of performance metrics for different ASR models; here, bold entries indicate best performances with respect to mean F1 score for each of 5-class, 3-class, and 2-class classification.*

Classification Setting		**Feature**	**F1 Score** (%)	**Precision** (%)	**Recall** (%)
5-Class	iASR	LaBSE	32.77 ± 11.94	35.26 ± 12.08	41.83 ± 9.32
		LASER	28.00 ± 6.56	28.20 ± 5.61	29.97 ± 9.20
	gASR	LaBSE	42.66 ± 5.24	44.31 ± 3.55	46.62 ± 5.48
		LASER	30.21 ± 2.87	28.96 ± 2.82	34.35 ± 5.09
	vASR	LaBSE	32.46 ± 7.66	37.51 ± 3.18	38.11 ± 7.92
		LASER	25.95 ± 3.75	29.57 ± 8.64	30.59 ± 5.90
	CombASR	LaBSE	**43.97 ± 4.13**	45.79 ± 3.68	48.45 ± 4.53
		LASER	35.38 ± 4.88	38.36 ± 4.86	38.33 ± 5.12
3-Class	iASR	LaBSE	53.69 ± 6.27	58.35 ± 13.29	57.78 ± 3.09
		LASER	49.50 ± 2.85	50.93 ± 8.36	54.03 ± 2.20
	gASR	LaBSE	56.86 ± 3.98	65.35 ± 6.14	61.47 ± 1.80
		LASER	55.39 ± 3.79	54.37 ± 6.73	58.75 ± 2.45
	vASR	LaBSE	55.03 ± 9.66	54.90 ± 10.72	59.19 ± 7.18
		LASER	52.86 ± 6.20	53.91 ± 10.86	56.81 ± 6.12
	CombASR	LaBSE	**61.48 ± 8.37**	67.19 ± 13.41	65.33 ± 5.58
		LASER	56.87 ± 5.29	57.81 ± 7.68	60.38 ± 3.63
2-Class	iASR	LaBSE	79.56 ± 1.72	80.69 ± 3.57	79.28 ± 1.45
		LASER	76.86 ± 4.10	77.39 ± 4.42	77.18 ± 4.71
	gASR	LaBSE	83.25 ± 4.59	82.75 ± 4.75	83.86 ± 4.33
		LASER	80.46 ± 5.81	82.32 ± 5.17	79.28 ± 6.12
	vASR	LaBSE	80.50 ± 8.33	81.20 ± 8.42	80.91 ± 8.45
		LASER	79.13 ± 7.13	79.61 ± 7.51	79.23 ± 6.76
	CombASR	LaBSE	**85.00 ± 5.08**	84.91 ± 5.32	86.31 ± 5.00
		LASER	82.97 ± 3.69	83.43 ± 4.62	83.74 ± 3.53

Table 4: *Mean ± SD of performance metrics obtained using various speech representations; here, bold entries indicate best performances with respect to mean F1 score for each of 5-class, 3-class, and 2-class classification.*

Classification Setting	**Feature**	**F1 Score** (%)	**Precision** (%)	**Recall** (%)
5-class	MFCC Stat	**51.15 ± 6.74**	52.59 ± 7.79	53.23 ± 7.83
	Opensmile	35.99 ± 7.27	39.67 ± 6.18	39.19 ± 6.98
	W2V2	45.80 ± 9.70	52.07 ± 6.92	46.77 ± 10.51
3-class	MFCC Stat	**67.58 ± 5.88**	68.79 ± 6.20	69.57 ± 4.12
	Opensmile	51.32 ± 4.52	55.23 ± 6.10	52.93 ± 4.29
	W2V2	58.37 ± 8.44	62.03 ± 8.09	58.50 ± 7.68
2-class	MFCC Stat	**87.21 ± 2.42**	86.61 ± 2.30	87.93 ± 2.61
	Opensmile	71.38 ± 6.47	76.15 ± 7.28	69.84 ± 6.13
	W2V2	79.74 ± 5.65	84.93 ± 4.08	77.90 ± 6.75

into the characteristics of dysarthria in ALS.

5.5. Multi-modal representation

We investigate the performance of multimodal representations by concatenating the best-performing speech representation and text embeddings. In this study, we choose the MFCC stats and LaBSE embeddings of CombASR and HSR (WOI, WoK) for speech and text features respectively. As shown in Table 5, we observe that combining text and speech features has enhanced classification performance slightly, outperforming models trained on individual modalities, in all cases except for 3-class classification where the MFCC Stat+ASR representation performs similar to the MFCC stat case.

Table 5: *Mean ± SD of performance metrics obtained using multi-modal feature representation; here, bold entries indicate best performances with respect to mean F1 score for each of 5-class, 3-class, and 2-class classification.*

Classification Setting	**Feature**	**F1 Score** (%)	**Precision** (%)	**Recall** (%)
5-Class	MFCC Stat + ASR	**53.25 ± 5.74**	54.86 ± 5.00	55.97 ± 4.65
	MFCC Stat + HSR	52.7 ± 4.22	53.1 ± 5.52	54.98 ± 4.85
3-Class	MFCC Stat + ASR	67.57 ± 7.05	69.96 ± 7.59	69.75 ± 5.75
	MFCC Stat + HSR	**68.3 ± 7.3**	69.74± 7.7	69.39 ± 5.12
2-Class	MFCC Stat + ASR	**88.25 ± 2.44**	87.18 ± 2.79	90.13 ± 2.41
	MFCC Stat + HSR	88.06 ± 3.81	87.63 ± 5.2	89.32 ± 2.56

6. Conclusions

The paper describes the efforts taken to categorize the severity of dysarthria in ALS patients by using the text embeddings of their speech transcripts. The study suggests that the text information carries similar discriminative information as speech cues for 2-class classification, though speech cues outperform for 3-class and 5-class dysarthria severity classification. Among the text representations, we observe that the LaBSE outperformed LASER embeddings in the majority of the experiments. The learning also revealed that in HSR, there is no significant role of the keywords and with image information. In this analysis, it is disclosed that the speech and textual cues carry complementary information and, when combined together, can outperform the individual modalities in most of the cases. Further investigations can be done for better modeling that utilizes both text and speech which can complement each other.

7. Acknowledgements

We thank Jhansi Mallela and Navaneetha G for their assistance in data preparation. We also thank the Department of Science and Technology (DST), Govt. of India for supporting this work.

8. References

[1] J. Wang, P. V. Kothalkar, M. Kim, and et al., "Automatic prediction of intelligible speaking rate for individuals with ALS from speech acoustic and articulatory samples," *Int J Speech Lang Pathol.*, vol. 20, no. 6, pp. 669–679, Nov. 2018.

[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 *et al.*, "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] B. Suhas, J. Mallela, A. Illa, B. Yamini, 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)*, 2020, pp. 1–5.

[6] 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, pp. 1–9, 2022.

[7] 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 *Proceedings of Interspeech*, 2023, pp. 1543–1547.

[8] 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.

[9] M. A. McHenry and S. M. LaConte, "Computer speech recognition as an objective measure of intelligibility," *Journal of medical speech-language pathology*, vol. 18, no. 4, pp. 99–103, 2010.

[10] 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.

[11] 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.

[12] 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.

[13] L. Moro-Velazquez, J. Cho, S. Watanabe, M. A. Hasegawa-Johnson, O. Scharenborg, H. W. Kim, and N. Dehak, "Study of the performance of automatic speech recognition systems in speakers with Parkinson's Disease," in *Proceedings of Interspeech*, 2019, pp. 3875–3879.

[14] Y. Choi, J. Lee, and M.-W. Koo, "Speech recognition-based feature extraction for enhanced automatic severity classification in dysarthric speech," in *IEEE Spoken Language Technology Workshop (SLT)*, 2024, pp. 953–960.

[15] 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 40th International Conference on Machine Learning (ICML)*, 2023, pp. 28 492–28 518.

[16] M. Artetxe and H. Schwenk, "Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond," *Transactions of the Association for Computational Linguistics*, vol. 7, pp. 597–610, 2019.

[17] F. Feng, Y. Yang, D. Cer, N. Arivazhagan, and W. Wang, "Language-agnostic BERT sentence embedding," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2022, pp. 878–891.

[18] F. Eyben, M. Wöllmer, and B. Schuller, "opensmile: The munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM International Conference on Multimedia*, 2010, pp. 1459–1462.

[19] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 12 449–12 460.

[20] 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.

[21] V. Nair and G. E. Hinton, "Rectified linear units improve restricted Boltzmann machines," in *Proceedings of the 27th International Conference on International Conference on Machine Learning (ICML)*, 2010, pp. 807–814.

[22] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.