Severity Classification and Dysarthric Speech Detection using Self-Supervised Representations

B Sanjay[*], Priyadharshini M.K[*], Vijayalakshmi P[*], Nagarajan T[†]
[*]Department of Electronics and Communication Engineering, Sri Sivasubramaniya Nadar College of Engineering
[†]Department of Computer Science and Engineering, Shiv Nadar University
sanjay2110556@ssn.edu.in, priyadharshinimk@ssn.edu.in,
vijayalakshmip@ssn.edu.in, nagarajant@snuchennai.edu.in

Abstract

Automatic detection and classification of dysarthria severity from speech provides a non-invasive and efficient diagnostic tool, offering clinicians valuable insights to guide treatment and therapy decisions. Our study evaluated two pre-trained models—wav2vec2-BASE and distilALHuBERT, for feature extraction to build speech detection and severity-level classification systems for dysarthric speech. We conducted experiments on the TDSC dataset using two approaches: a machine learning model (support vector machine, SVM) and a deep learning model (convolutional neural network, CNN). Our findings showed that features derived from distilALHuBERT significantly outperformed those from wav2vec2-BASE in both dysarthric speech detection and severity classification tasks. Notably, the distilALHuBERT features achieved 99% accuracy in automatic detection and 95% accuracy in severity classification, surpassing the performance of wav2vec2 features.

1 Introduction

Dysarthria is a motor speech disorder caused by neurological damage that affects the mechanisms responsible for speech production. (Doyle et al., 1997) . This condition is the result of damage to the nervous system, which can occur due to either an acquired or congenital neurological illness. Common causes of dysarthria include cerebral palsy, brain tumors, traumatic brain injuries, and strokes. Additionally, it can develop as a result of neurodegenerative diseases such as Parkinson's disease, amyotrophic lateral sclerosis (ALS), or Huntington's disease, which progressively impair the body's motor functions.

A range of abnormalities in different elements of speech production typically marks speech affected by dysarthria. These can include difficulties with phonation (the ability to produce vocal sound), resonance (the quality of the voice), articulation (the clarity of speech sounds), and prosody (the rhythm and intonation patterns of speech). These combined deficits often reduce speech intelligibility, making it difficult for listeners to understand individuals with dysarthria (Duffy et al., 2012). Speech intelligibility is traditionally evaluated by speech-language pathologists in voice clinics using standardized intelligibility tests (Kent et al., 1989). Augmentative and alternative speech aids, like those in (Mariya Celin et al., 2019), show promise in enhancing dysarthric communication using confusion transducers to correct errors. However, these subjective tests are often expensive, time-consuming, and can be influenced by the pathologist's familiarity with the patient and their speech disorder, leading to potential bias (De Bodt et al., 2002). Consequently, there is a growing need for an objective method to assess dysarthric speech. The evaluation process typically involves two key steps: (1) classifying the dysarthric speech from the acoustic speech signal,(2) determining the severity of the condition. This research aims to automate the detection of dysarthria and classify its severity level based on speech signals.

In recent years, there has been significant interest in the automatic detection and severity classification of dysarthria from speech, driven by advancements in signal processing, machine learning (ML), and deep learning (DL). Dysarthria detection methods typically involve a two-stage pipeline system, which includes feature extraction followed by classification. These systems are trained in a supervised fashion using labeled speech data, where the labels (e.g., healthy vs. dysarthric) are derived from evaluations performed by speech-language pathologists. Numerous studies have explored various feature extraction techniques aimed at capturing the key characteristics of dysarthric speech production.

(Kim et al., 2015) investigated sentence-level features to examine abnormal variations in pronunciation, prosody, and voice quality. In the studies by Gurugubelli and Vuppala (2019) and Gurugubelli and Vuppala (2020), single frequency filtering-based features were examined for the detection of dysarthric speech and its classification into four intelligibility levels (very low, low, medium, and high). In studies by Rong et al. (2016) and De Bodt et al. (2002), linear weighted combinations of articulation, phonation, and prosody features were investigated for assessing the intelligibility of dysarthric speech. Glottal source features, together with openSMILE features (Eyben et al., 2010), were effectively used by Narendra and Alku (2019, 2020) to improve classification performance in both dysarthric speech detection and intelligibility classification, utilizing support vector machines (SVM) as the classifier. Xue et al. (2019) explored the usability of the extended Geneva minimalistic acoustic parameter set (eGeMAPS) (Eyben et al., 2015) in predicting phoneme intelligibility in dysarthric speech. Recently, numerous studies have investigated various spectro-temporal representations, such as spectrograms, mel-spectrograms, and MFCCs, combined with different deep learning (DL) classifiers like squeeze-and-excitation (SE) networks, convolutional neural networks (CNNs), residual neural networks (ResNets), gated recurrent units (GRUs), and long short-term memory networks (LSTMs), for estimating the intelligibility of dysarthric speech studied the effectiveness of a multi-head attention mechanism to identify severity-emphasizing periods in spectrograms, alongside a multi-task learning approach for classifying dysarthria severity levels. A comprehensive systematic review of current studies on automatic classification of dysarthria severity levels was given by Al-Ali et al. (2023).

Self-supervised representation learning has gained significant interest in the field of paralinguistics, primarily due to the smaller size of datasets in this area compared to those used for tasks like automatic speech recognition (ASR). This involves training a model in an unsupervised manner on large speech datasets, enabling it to learn speech representations directly from raw audio, which can then be applied to various tasks. Notable examples of pre-trained models include wav2vec2 (Baevski et al., 2020) and distilALHuBERT Hsu et al. (2021), which have demonstrated strong performance across several speech-related tasks, such as ASR, emotion recognition, speaker and language identification, and voice pathology detection. In this study, we investigate the effectiveness of two pre-trained models—wav2vec2 and distilALHuBERT—for the detection of dysarthria and the classification of its severity levels. Our preliminary research on dysarthria detection and severity classification using the wav2vec2 model, applied to the UA-Speech database (Javanmardi et al., 2023), produced promising results. This initial success has inspired the current extended study. The pre-trained wav2vec2 and distilALHuBERT models utilized in this research are available on HuggingFace (Wolf et al., 2020). The key contributions of this study are:

A comprehensive comparison of wav2vec2 and distilALHuBERT embeddings, using them as features for two primary tasks:

- Dysarthria detection (healthy *vs.* dysarthric speech)
- Classification of dysarthria severity into three levels (Mild *vs.* Moderate *vs.* Severe)
- Presenting novel findings on speech-based biomarkers for dysarthria, showing that distilALHuBERT features outperformed others in both dysarthria detection and severity level classification.

2 The detection and severity level classification systems

This study investigates two classification tasks: binary and multi-class classification. The existing systems employ a two-stage pipeline approach that includes a feature extraction phase followed by a classification phase, as illustrated in Fig. 1. Fig. 1(a) depicts the system designed to differentiate dysarthric speech from healthy speech (i.e., the detection task), while Fig. 1(b) presents the system used to classify dysarthria severity into three categories: Mild, Moderate, and Severe. In both tasks, the feature extraction component leverages two widely-used pre-trained models—wav2vec2-BASE (Baevski et al., 2020) and HuBERT (Wang et al., 2023)—to derive feature vectors from raw

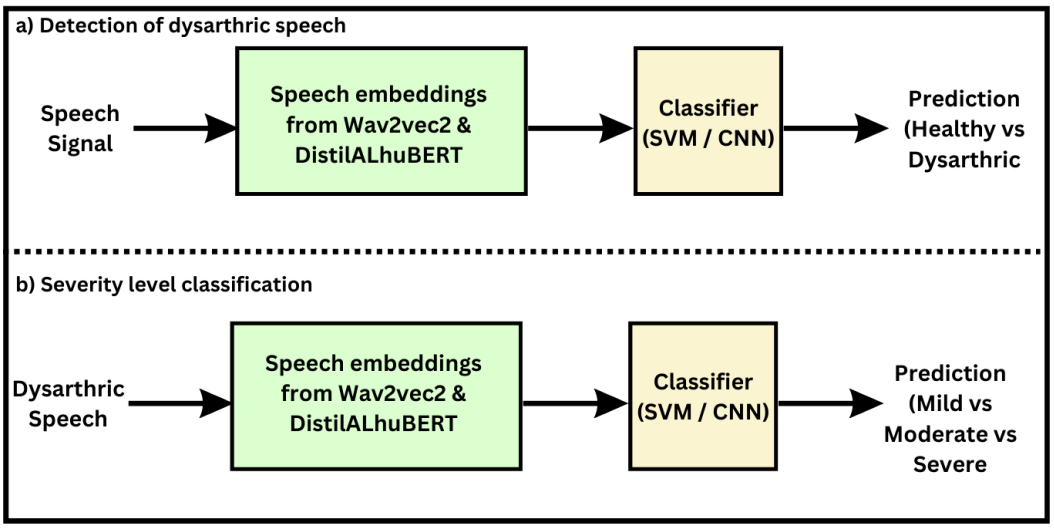

Figure 1: A schematic block diagram of the systems for (a) detection of dysarthric speech and (b) severity level classification of dysarthric speech

speech signals. The classification phase utilizes both SVM and CNN to predict the output labels.

2.1 Feature extraction using pre-trained models

Three pre-trained models are employed as feature extractors to develop the detection and classification systems: wav2vec2-BASE, and distilALHuBERT. These models were initially pre-trained on unlabeled speech data and later fine-tuned on a smaller labeled dataset for automatic speech recognition (ASR). As a result, the final layers of these models primarily capture speech representations focused on phoneme-related information (Baevski et al., 2020; Fan et al., 2020). However, the lower layers of the network retain information related to phones. These lower-level features can be effectively utilized in various downstream speech tasks, including the classification tasks examined in this study. In this work, the pre-trained models have been fine-tuned on a small labeled ASR dataset.

2.1.1 Wav2vec2

In this study, we investigated the wav2vec2-BASE model as shown in Figure.2 .The wav2vec2 architecture consists of a multi-layer CNN encoder, a context network, and a quantization module. During pre-training, the CNN encoder processes 20 ms speech segments into latent speech representations (denoted as Z). These representations are then passed through a projection layer to obtain the features to align with the inner dimension (768) of the context network. Before being fed into the context network (which consists of 12 transformer blocks in wav2vec2-BASE), a proportion p of the time steps in Z are randomly sampled. The selected time steps serve as starting points for masking, with the subsequent M time steps being masked. Relative positional embeddings are computed using grouped one-dimensional (1-D) convolution and are added to the masked representations, which are then passed through the context network and transformed into context representations (denoted as C).

The quantization module converts the latent speech representations Z into quantized representations Q, known as quantized targets. The model is optimized using a contrastive loss function L_m, which encourages the model to correctly identify the true quantized speech representation q_t from a set of candidate quantized representations $\tilde{q} \in Q_t$, including q_t and 100 distractor representations. The distractors are uniformly sampled from other masked time steps within the same utterance. Fig. 2 illustrates the wav2vec2-BASE model, featuring 12 transformer blocks.

In this study, the outputs from the transformer layers of the context network are used as features for both detection and classification tasks. For wav2vec2-BASE,the temporal average of the inputs to the first transformer layer is computed and the outputs of all 12 transformer layers for each speech signal, resulting in a total of 13 feature ma-

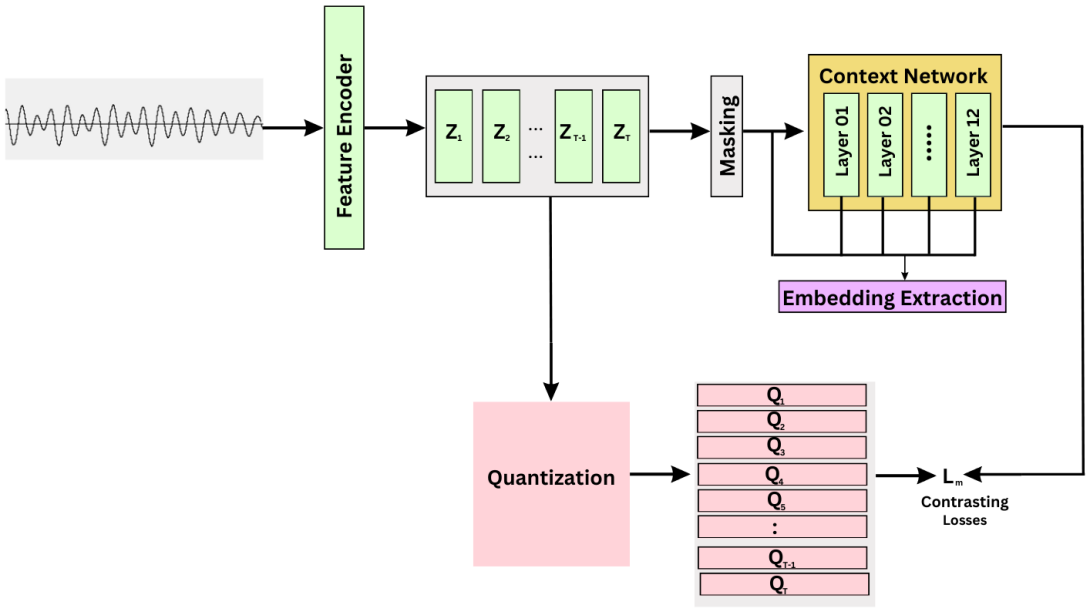

Figure 2: Block diagram of a wav2vec2 architecture with 12 transformer layers

trices. These features are averaged across frames to yield 13 768-dimensional feature vectors per speech signal. Throughout this article, these features are referred to as wav2vec2-BASE features. When specifying features from a specific layer, they are referred to as wav2vec2-BASE-N, where N indicates the transformer layer.

2.1.2 DistilALhuBERT

In this study, we focus on the HuBERT model, which is designed for audio representation learning through self-supervision shown in Figure .3 . The HuBERT model consists of a CNN-based feature extractor followed by a transformer-based encoder network. During pre-training, the speech data is subjected to K-means clustering to get the distinct classes, which then serve as the hidden units (Hu) of the speech signals. The model is trained to predict these hidden units, allowing it to learn high-level speech representations. This pre-training approach has been a key innovation in HuBERT, making it well-suited for a wide range of downstream speech tasks.

The HuBERT architecture comprises a stack of transformer layers, each consisting of a multi-head attention block and a feed-forward block. For each transformer layer i, let f_i represent the function of the i^{th} transformer layer. The output h_i is computed as:

$$h_i = f_i(h_{i-1})$$

Here, h_{i-1} represents the output from the previous layer, or the CNN feature extractor when $i = 1$. This hierarchical approach allows HuBERT to create contextual representations that capture speech signal characteristics at various levels.

In the current study, the output of the transformer layers is used as features for downstream tasks, including detection and classification. Similar to wav2vec2, the temporal average of the inputs to the first transformer layer and the outputs of each transformer layer are computed for each speech signal, providing multiple feature matrices. These features are averaged across frames, yielding a set of fixed-length feature vectors for each speech signal. This process allows us to utilize HuBERT's learned representations for various classification problems, as detailed in the following sections.

2.1.3 Classifiers

In this study, both an ML classifier (SVM) and a DL classifier (CNN) are employed for two tasks: detecting dysarthric speech and classifying its severity level (multi-class classification). The SVM utilizes a radial basis function (RBF) kernel with a regularization parameter of 1. Additionally, the gamma parameter is defined as $\gamma = 1/(D \cdot Var(X))$, where $Var(X)$ represents the variance of the training data and D is the dimensionality of the feature vectors. The RBF kernel was selected to achieve optimal accuracy.

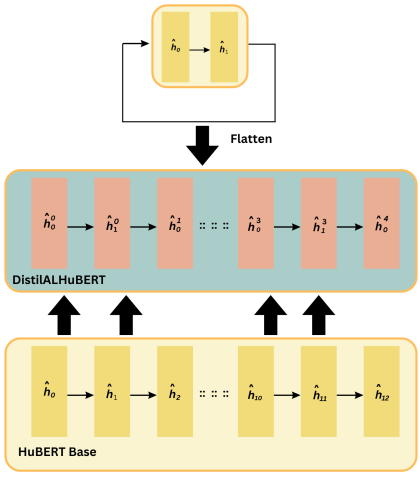

Figure 3: Block diagram of a distilALHuBERT architecture with 12 transformer layers

For the CNN classifier, the architecture consists of two sequential convolutional layers, each followed by the ReLU activation function. The output is then flattened and fed into two fully connected (dense) layers. In the final layer, binary classification (healthy vs. dysarthric speech) is used for sigmoid activation function. The CNN hyperparameters include a batch size of 64, 100 epochs 20 epochs as the early stopping, the cross-entropy loss function, and the Adam optimizer with a learning rate of 0.001.

It is worth emphasising that the same CNN architecture is used for both detection and multi-class classification. For severity level classification, the softmax activation function is used in the final dense layer to predict the severity label. The SVM and CNN classifiers were implemented using the Scikit-learn and PyTorch libraries.

3 Experimental Setup

3.1 Tamil Dysarthric Speech Corpus

The SSNCE Tamil Dysarthric Speech Corpus (TDSC) (Celin et al., 2016, 2020), developed by one of the authors is used for all the analysis. The TDSC dataset contains recordings of 20 dysarthric speakers (13 males and 7 females) diagnosed with cerebral palsy (spastic quadriplegia or bilateral paraplegia). The corpus contains recordings of 10 normal speakers (5 males and 5 females) speech data as well. Each speaker has spoken 365 words, including 103 isolated words and 262 sentences (ranging from 2 to 6 words). The dataset was designed to include sufficient examples of all Tamil phonemes. The speech dataset were collected in collaboration with the National Institute for the Empowerment of Persons with Multiple Disabilities (NIEPMD). The recording was performed using a head-mounted microphone in a laboratory environment at a sampling rate of 16kHz.

3.2 Training and Testing

The training and testing were performed by splitting the dataset into an 80-20 ratio. This split was maintained across all classes for the severity level classification experiments using the TDSC database. In both the detection and multi-class classification experiments with the CNN classifier, 10% of each speaker's training samples were randomly selected as validation data during each iteration. The training and testing process was repeated across all iterations, and the evaluation metrics were averaged over all iterations to obtain the final results.

3.3 Evaluation Metrics

The performance of the dysarthria detection systems was assessed using five common metrics: accuracy (ACC), sensitivity (SE), specificity (SP), F1-score (F1), and equal error rate (EER). For evaluating the severity level classification systems, mean accuracy and individual class accuracies were calculated. Additionally, confusion matrices were provided for both detection and multi-class classification tasks.

4 Results

This section outlines the results obtained from the features extracted using two pre-trained models (wav2vec2-BASE and distilALHuBERT) in combination with SVM and CNN classifiers. The detection experiment results are presented first in Section 4.1, followed by the severity classification experiment results in Section 4.2.

4.1 Results for detection of dysarthric speech

The performance of the classification experiments is summarized in Table.1 for both the SVM and CNN models, with results reported across five metrics: accuracy (ACC), sensitivity (SE), specificity (SP), F1-score (F1), and equal error rate (EER).

For the SVM classifier, DistilALHuBERT achieved a significant improvement, with

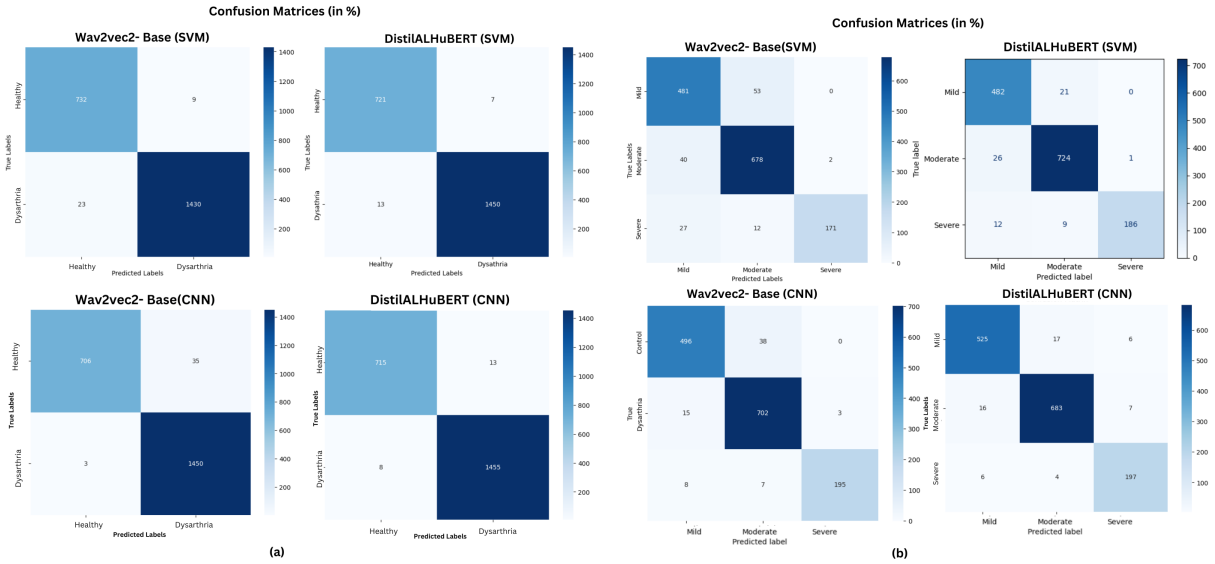

Figure 4: Confusion matrices of a) Dysarthria detection b) Severity level classification given by the SVM and CNN classifiers for wav2vec-2-Base and DistilHuBERT Features.

Classifier	Feature	ACC [%]	SE	SP	F1	EER
SVM	Wav2vec2-Base	97	97	98	99	0.03
	DistilALHuBERT	**99**	99	99	99	0.01
CNN	Wav2vec2-Base	98.3	95	95	97	0.02
	DistilALHuBERT	**99**	99	98	99	0.01

Table 1: Dysarthria detection results with wav2vec2-Base, and HuBERT features for the SVM and CNN classifiers. Here ACC refers to accuracy, SE refers to sensitivity, SP refers to specificity, and F1 refers to F1-score.

an accuracy of 99% compared to 97% from wav2vec2-BASE. Furthermore, sensitivity, specificity, and F1-score for DistilALHuBERT reached 99%, while wav2vec2-BASE performed slightly lower with 98% sensitivity and 99% specificity. The EER for DistilALHuBERT was also notably lower at 0.01 compared to 0.03 for wav2vec2-BASE.

In the CNN classifier, a similar trend was observed, with DistilALHuBERT outperforming wav2vec2-BASE across most metrics. DistilALHuBERT achieved a perfect 99% accuracy, while wav2vec2-BASE attained 98.3%. Additionally, sensitivity and specificity for DistilALHuBERT were 98% and 99%, respectively, compared to 95% sensitivity and 97% specificity for wav2vec2-BASE. The EER for DistilALHuBERT was also lower at 0.01 compared to 0.02 for wav2vec2-BASE.From the results, it can be observed that the DistilALHuBERT feature consistently outperforms the wav2vec2-BASE feature across all metrics in both classifiers.

The confusion matrices in Fig.4a show classifiers tend to misclassify dysarthric speech as healthy more often than vice versa. Wav2vec2-Base (SVM) correctly classified 1,430 of 1,490 dysarthric samples, misclassifying 23 as healthy. DistilALHuBERT (SVM) performed better, misclassifying only 13. CNN configurations showed similar patterns, with Wav2vec2-Base and DistilALHuBERT accurately classifying 1,450 and 1,455 dysarthric samples, respectively.

DistilALHuBERT's exceptional performance stems from its advanced feature extraction, identifying complex acoustic patterns in dysarthric speech across severity levels. Fine-tuning on domain-specific data enhanced its ability to distinguish between healthy and dysarthric speech, reducing misclassification rates.

Classifier	Feature	ACC [%]	ACC [%]		
		Overall	C_{mild}	$C_{moderate}$	C_{severe}
SVM	Wav2vec2-Base	68.4	55	85	47
	DistilALHuBERT	**95.28**	95.8	96.4	**89.8**
CNN	Wav2vec2-Base	94.9	97	93	93
	DistilALHuBERT	**96**	97	95	**96.34**

Table 2: Classification results for dysarthria detection using Wav2vec2-Base and DistilALHuBERT features for the SVM and CNN classifiers. Here ACC refers to accuracy for the full dataset, and C_{mild}, $C_{moderate}$, and C_{severe} refer to accuracy for each severity level of dysarthria.

4.2 Results for severity level classification of dysarthric speech

The classification results for the speech data across different classifiers and features are presented in Table. 2 The performance is measured in terms of overall accuracy (ACC) and class-wise accuracy for mild, moderate, and severe classifications.

For the SVM classifier, the Wav2vec2-BASE feature yielded an overall accuracy of 68.4%. However, the classification accuracy for mild, moderate, and severe levels was significantly lower, with scores of 55%, 85%, and 47%, respectively. In contrast, the DistilALHuBERT feature substantially improved performance, achieving an overall accuracy of 95.28%. The class-wise accuracies for DistilALHuBERT were notably high, with 95.8% for mild, 96.4% for moderate, and 89.8% for severe classifications.

The CNN classifier exhibited similar results. The Wav2vec2-BASE feature provided an overall accuracy of 94.9%, with class-wise accuracies of 97% for mild, 93% for moderate, and 93% for severe classifications. DistilALHuBERT outperformed Wav2vec2-BASE, achieving an overall accuracy of 96%. The class-wise accuracy for DistilALHuBERT was 97% for mild, 95% for moderate, and an exceptional 96.34% for severe classifications.

DistilALHuBERT outperforms Wav2vec2 in classifying dysarthric speech, thanks to its advanced feature extraction and fine-tuning on domain-specific data. Both models use self-supervised learning, but DistilALHuBERT's distilled architecture preserves HuBERT's key knowledge, enabling it to capture subtle acoustic patterns more effectively. This enhanced capability allows DistilALHuBERT to better distinguish between various severity levels of dysarthric speech, leading to fewer misclassifications and improved overall performance.

Limitations

This study focused on fine-tuning two self-supervised speech models, wav2vec 2.0 and DistilHuBERT, on a Tamil speech dataset to address the language's unique characteristics. As Tamil, a Dravidian language, differs significantly from English, fine-tuning was crucial to capture its specific nuances in phonetics, prosody, and syntax. While both models performed well particularly DistilHuBERT in severity classification and dysarthric speech prediction, the study also revealed limitations. DistilHuBERT's high accuracy can be attributed to its top layers' embeddings being rich in Tamil-specific features. However, this limits the model's generalizability across languages. The study also highlights the constraints of current self-supervised models pre-trained primarily on English or similar languages. Although fine-tuning adapts the model to Tamil, it underscores these models' inherent limitations when working with linguistically diverse datasets. They are not inherently multilingual and require substantial adaptation to perform well across various languages and dialects.

Acknowledgments

The authors would also like to thank the Ministry of Electronics and Information Technology (MeitY), Government of India, for funding the project on 'Assistive Speech technologies'under the Project titled 'National Language Translation Mission (NLTM):BHASHINI', Ref. No. 11(1)/2022-HCC(TDIL).

References

Afnan Al-Ali, Somaya Al-Maadeed, Moutaz Saleh, Rani Chinnappa Naidu, Zachariah C Alex, Prakash

Ramachandran, Rajeev Khoodeeram, et al. 2023. Classification of dysarthria based on the levels of severity. a systematic review. *arXiv preprint arXiv:2310.07264*.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. volume 33, pages 12449–12460.

Mariya Celin, T. Nagarajan, and P. Vijayalakshmi. 2016. Dysarthric speech corpus in tamil for rehabilitation research. pages 2610–2613.

T. A. Mariya Celin, T. Nagarajan, and P. Vijayalakshmi. 2020. Data augmentation using virtual microphone array synthesis and multi-resolution feature extraction for isolated word dysarthric speech recognition. *IEEE Journal of Selected Topics in Signal Processing*, 14(2):346–354.

Marc S De Bodt, Maria E Hernández-Dıaz Huici, and Paul H Van De Heyning. 2002. Intelligibility as a linear combination of dimensions in dysarthric speech. *Journal of communication disorders*, 35(3):283–292.

Philip C Doyle, Herbert A Leeper, Ava-Lee Kotler, Nancy Thomas-Stonell, Charlene O'Neill, Marie-Claire Dylke, and Katherine Rolls. 1997. Dysarthric speech: A comparison of computerized speech recognition and listener intelligibility. *Journal of rehabilitation research and development*, 34:309–316.

Joseph R Duffy et al. 2012. *Motor speech disorders: Substrates, differential diagnosis, and management*. Elsevier Health Sciences.

Florian Eyben, Klaus R Scherer, Björn W Schuller, Johan Sundberg, Elisabeth André, Carlos Busso, Laurence Y Devillers, Julien Epps, Petri Laukka, Shrikanth S Narayanan, et al. 2015. The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing. *IEEE transactions on affective computing*, 7(2):190–202.

Florian Eyben, Martin Wöllmer, and Björn Schuller. 2010. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 1459–1462.

Zhiyun Fan, Meng Li, Shiyu Zhou, and Bo Xu. 2020. Exploring wav2vec 2.0 on speaker verification and language identification. *arXiv preprint arXiv:2012.06185*.

Krishna Gurugubelli and Anil Kumar Vuppala. 2019. Perceptually enhanced single frequency filtering for dysarthric speech detection and intelligibility assessment. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6410–6414. IEEE.

Krishna Gurugubelli and Anil Kumar Vuppala. 2020. Analytic phase features for dysarthric speech detection and intelligibility assessment. *Speech Communication*, 121:1–15.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM transactions on audio, speech, and language processing*, 29:3451–3460.

Farhad Javanmardi, Saska Tirronen, Manila Kodali, Sudarsana Reddy Kadiri, and Paavo Alku. 2023. Wav2vec-based detection and severity level classification of dysarthria from speech. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Ray D Kent, Gary Weismer, Jane F Kent, and John C Rosenbek. 1989. Toward phonetic intelligibility testing in dysarthria. *Journal of Speech and Hearing Disorders*, 54(4):482–499.

Jangwon Kim, Naveen Kumar, Andreas Tsiartas, Ming Li, and Shrikanth S Narayanan. 2015. Automatic intelligibility classification of sentence-level pathological speech. *Computer speech & language*, 29(1):132–144.

T. A. Mariya Celin, G. Anushiya Rachel, T. Nagarajan, and P. Vijayalakshmi. 2019. A weighted speaker-specific confusion transducer-based augmentative and alternative speech communication aid for dysarthric speakers. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 27(2):187–197.

N P Narendra and Paavo Alku. 2020. Automatic intelligibility assessment of dysarthric speech using glottal parameters. *Speech Communication*, 123:1–9.

NP Narendra and Paavo Alku. 2019. Dysarthric speech classification from coded telephone speech using glottal features. *Speech Communication*, 110:47–55.

Panying Rong, Yana Yunusova, Jun Wang, Lorne Zinman, Gary L Pattee, James D Berry, Bridget Perry, and Jordan R Green. 2016. Predicting speech intelligibility decline in amyotrophic lateral sclerosis based on the deterioration of individual speech subsystems. *PloS one*, 11(5):e0154971.

Haoyu Wang, Siyuan Wang, Yaguang Gong, and Wei-Qiang Zhang. 2023. Distilalhubert: A distilled parameter sharing audio representation model. pages 45–50.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45.

Wei Xue, Catia Cucchiarini, RWNM van Hout, and Helmer Strik. 2019. Acoustic correlates of speech intelligibility. the usability of the egemaps feature set for atypical speech.