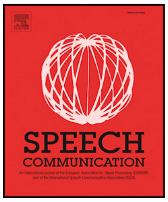
Speech Communication 158 (2024) 103047

Contents lists available at ScienceDirect

Speech Communication

journal homepage: www.elsevier.com/locate/specom



Pre-trained models for detection and severity level classification of dysarthria from speech

Farhad Javanmardi *, Sudarsana Reddy Kadiri, Paavo Alku

Department of Information and Communications Engineering, Aalto University, Finland

ARTICLE INFO

Keywords:
Dysarthria
Severity level classification
Pre-trained models
Wav2vec2
HuBERT
OpenSMILE
eGeMAPS

ABSTRACT

Automatic detection and severity level classification of dysarthria from speech enables non-invasive and effective diagnosis that helps clinical decisions about medication and therapy of patients. In this work, three pre-trained models (wav2vec2-BASE, wav2vec2-LARGE, and HuBERT) are studied to extract features to build automatic detection and severity level classification systems for dysarthric speech. The experiments were conducted using two publicly available databases (UA-Speech and TORGO). One machine learning-based model (support vector machine, SVM) and one deep learning-based model (convolutional neural network, CNN) was used as the classifier. In order to compare the performance of the wav2vec2-BASE, wav2vec2-LARGE, and HuBERT features, three popular acoustic feature sets, namely, mel-frequency cepstral coefficients (MFCCs), openSMILE and extended Geneva minimalistic acoustic parameter set (eGeMAPS) were considered. Experimental results revealed that the features derived from the pre-trained models outperformed the three baseline features. It was also found that the HuBERT features performed better than the wav2vec2-BASE and wav2vec2-LARGE features. In particular, when compared to the best-performing baseline feature (openSMILE), the HuBERT features showed in the detection problem absolute accuracy improvements that varied between 1.33% (the SVM classifier, the TORGO database) and 2.86% (the SVM classifier, the UA-Speech database). In the severity level classification problem, the HuBERT features showed absolute accuracy improvements that varied between 6.54% (the SVM classifier, the TORGO database) and 10.46% (the SVM classifier, the UA-Speech database) compared to the best-performing baseline feature (eGeMAPS).

1. Introduction

Dysarthria is a neuro-motor disorder resulting from neurological damage to the motor component of speech production (Doyle et al., 1997). Dysarthria is generally caused by an acquired or congenital neurological illness such as cerebral palsy, brain tumor, brain injury, stroke, or neurodegenerative disease such as Parkinsons's disease, amyotrophic lateral sclerosis, or Huntington's disease. Dysarthric speech is often characterized by abnormalities in the phonatory, resonatory, articulatory, and prosodic aspects of speech production that all impact speech intelligibility (Duffy, 2019).

The speech intelligibility assessment is conventionally performed by speech-language pathologists in voice clinics using standard intelligibility tests (Kent et al., 1989). However, subjective intelligibility tests are costly and laborious, and subject to pathologists' biases because of familiarity with patients and their speech disorders (De Bodt et al., 2002a). Therefore, the development of an objective method is important for the assessment of dysarthric speech. The assessment of dysarthric speech includes the following two steps: (1) identifying the presence of dysarthria from the acoustic speech signal and (2) estimating the severity level of the disease. These two steps are crucial diagnostic tasks that can assist in determining proper clinical decisions in pharmacological treatment and patient therapy. This work considers the automatic detection and severity level classification of dysarthria from speech.

In recent years, automatic detection and severity level classification of dysarthria from speech has received considerable attention due to progress in signal processing as well as in machine learning (ML) and deep learning (DL). Techniques used in dysarthria detection mainly correspond to using the so-called pipeline system that consists of two main stages (feature extraction and classification). These systems are trained in a supervised manner using collected speech data and labels (healthy vs. dysarthric). The (binary) labels are obtained from the assessments conducted by speech-language pathologists. Many studies have investigated various feature extraction methods that characterize the salient aspects of the production of dysarthric speech (Kim et al., 2008; Falk et al., 2012; Kain et al., 2007; Kim et al., 2015). In Kim et al.

* Corresponding author.
E-mail address: farhad.javanmardi@aalto.fi (F. Javanmardi).

https://doi.org/10.1016/j.specom.2024.103047
Received 22 June 2023; Received in revised form 2 January 2024; Accepted 9 February 2024
Available online 14 February 2024

(2015), the authors used sentence-level features to explore abnormal variations of pronunciation, prosody, and voice quality. In Gurugubelli and Vuppala (2019, 2020), the single frequency filtering-based features were explored in the detection and 4-class intelligibility classification of dysarthric speech (very low, low, medium, and high level of speech intelligibility). The authors of Kadi et al. (2016) proposed distinctive auditory features for the assessment of dysarthria, and showed that the combination of auditory features with mel-frequency cepstral coefficients (MFCCs) improves the estimation of speech intelligibility in dysarthria. In Rong et al. (2016), De Bodt et al. (2002b), linear weighted combinations of articulation, phonation, and prosody features of speech were investigated in the intelligibility assessment of dysarthric speech. Glottal source features together with the openSMILE features (Eyben et al., 2010) were successfully used in Narendra and Alku (2020, 2018, 2019, 2021) to enhance the classification performances in both the detection of dysarthric speech and the classification of intelligibility of dysarthric speech using support vector machine (SVM) as classifier. In Xue et al. (2019), the usability of the extended Geneva minimalistic acoustic parameter set (eGeMAPS) (Eyben et al., 2016) was explored in predicting the phoneme intelligibility of dysarthric speech. Recently, many studies have explored different spectro-temporal representations, such as the spectrogram, mel-spectrogram and MFCCs with various DL-based classifiers, such as the squeeze-and-excitation (SE) networks, convolutional neural networks (CNNs), residual neural networks (ResNets), gated recurrent units (GRUs) and long short-term memory networks (LSTMs) in the estimation of intelligibility of dysarthric speech (Chandrashekar et al., 2019, 2020; Fernández-Díaz and Gallardo-Antolín, 2020; Gupta et al., 2021; Joshy and Rajan, 2021, 2022, 2023a). The effectiveness of using a multi-head attention mechanism to identify severity-highlighting periods in spectrograms was studied together with a multi-task learning approach for dysarthria severity level classification in Joshy and Rajan (2023b). A systematic review of the existing studies on the automatic classification of the severity level of dysarthria was given in Al-Ali et al. (2023).

Self-supervised representation learning has become a topic of increasing interest in the field of paralinguistics as most datasets in this study area are relatively small compared to, for example, datasets used in automatic speech recognition (ASR). Self-supervised representation learning corresponds to training a model in an unsupervised manner on large speech datasets. The model learns speech representations from the raw audio to be used in the required task. Examples of pre-trained models are wav2vec2 (Baevski et al., 2020) and HuBERT (Hsu et al., 2021), which have shown good performance in various speech technology tasks such as ASR, emotion recognition, speaker and language identification, as well as voice pathology detection (Baevski et al., 2020; Hernandez et al., 2022; Mohamed and Aly, 2021; Vaessen and Van Leeuwen, 2022; Fan et al., 2021; Grósz et al., 2022; Tirronen et al., 2023a,b). In the current study, we explore the effectiveness of two pre-trained models (wav2vec2 and HuBERT) in speech-based detection and severity level classification of dysarthria. Our initial study in the detection and severity level classification of dysarthria showed promising results in utilizing the pre-trained wav2vec2 model for the UA-Speech database (Javanmardi et al., 2023), and this encouraging evidence motivates the current, extended investigation of the topic. The wav2vec2 and HuBERT pre-trained models used in the present study are available at HuggingFace (Wolf et al., 2019).

The main contributions of this study are:

- Conducting a layer-by-layer comparison between the wav2vec2 and HuBERT embeddings by using these embeddings as features in the following two problems:
  - detection of dysarthria (healthy *vs.* dysarthric)
  - classification of the severity level of dysarthria into four classes (very low *vs.* low *vs.* medium *vs.* high).
- Conducting a systematic comparison between three widely-used acoustic feature sets (MFCCs, openSMILE, and eGeMAPS) and the wav2vec2 and HuBERT features for two dysarthria databases (UA-Speech and TORGO) using SVM and CNN classifiers.
- Presenting new results on speech-based biomarking of dysarthria showing that the HuBERT features performed best both in the detection and severity level classification of the disease.

2. The detection and severity level classification systems

This study investigates two types of classification problems: binary and multi-class classification. The systems are built using the popular two-stage pipeline approach consisting of a feature extraction stage and a classifier stage, as shown in Fig. 1. Fig. 1(a) shows the system for distinguishing dysarthric speech from healthy speech (i.e., the detection problem), and Fig. 1(b) shows the system for classifying the severity level of dysarthria into four classes (very low, low, medium, and high level of disease severity). In both problems, the feature extraction block uses three popular pre-trained models (wav2vec2-BASE (Baevski et al., 2020), wav2vec2-LARGE (Baevski et al., 2020), and HuBERT (Hsu et al., 2021)) to extract feature vectors from raw speech waveforms. In the classifier block, one ML-based classifier (SVM) and one DL-based classifier (CNN) are used to predict the output labels. The feature extraction and classifier are described in the next sub-sections.

2.1. Feature extraction using pre-trained models

Three pre-trained models are used as feature extractors to build the detection and classification systems. These selected pre-trained models are wav2vec2-BASE, wav2vec2-LARGE, and HuBERT. These models were originally pre-trained on a large quantity of unlabeled speech data and fine-tuned on a small labeled set for ASR. Therefore the final layers of the pre-trained models have mainly learned speech representations that contain phoneme-related information (Baevski et al., 2020; Fan et al., 2021). The learned speech representations from the lower layers of the network, however, contain information related to phone. These lower layers of the network can therefore be utilized as features in various down-stream speech tasks, including the classification problems investigated in this study (Hernandez et al., 2022; Tirronen et al., 2023a; Sheikh et al., 2022; Gauder et al., 2021). The selected models used in the current study were fine-tuned on a small labeled set for ASR.

2.1.1. Wav2vec2

In this study, we have investigated two wav2vec2 models, namely wav2vec2-BASE and wav2vec2-LARGE. The wav2vec2 architecture consists of a multi-layer CNN encoder, a context network, and a quantization module. During the pre-training, a CNN encoder transfers speech segments of 20 ms into latent speech representations (denoted by Z). These representations are fed to a feature projection layer to match the inner dimension (of 768) of the context network. Before sending Z into the context network (which contains 12 transformer blocks in the wav2vec2-BASE model and 24 transformer blocks in the wav2vec2-LARGE model), a certain proportion p of all time steps of Z are randomly sampled. These sampled time steps act as starting indices for masking where the subsequent M consecutive time steps from every sampled index are masked. Then, the relative positional embeddings are computed by grouped one-dimensional (1-D) convolution and added to the masked representations. These masked representations are then passed to the context network and mapped to context representations (denoted by C). The quantization module is used to discretize latent speech representations Z into quantized representations Q (known as quantized targets). Finally, the context representations (C) together with the quantized representations Q are used to optimize the model using a contrastive loss objective function L_m. The contrastive loss makes the model identify the true quantized

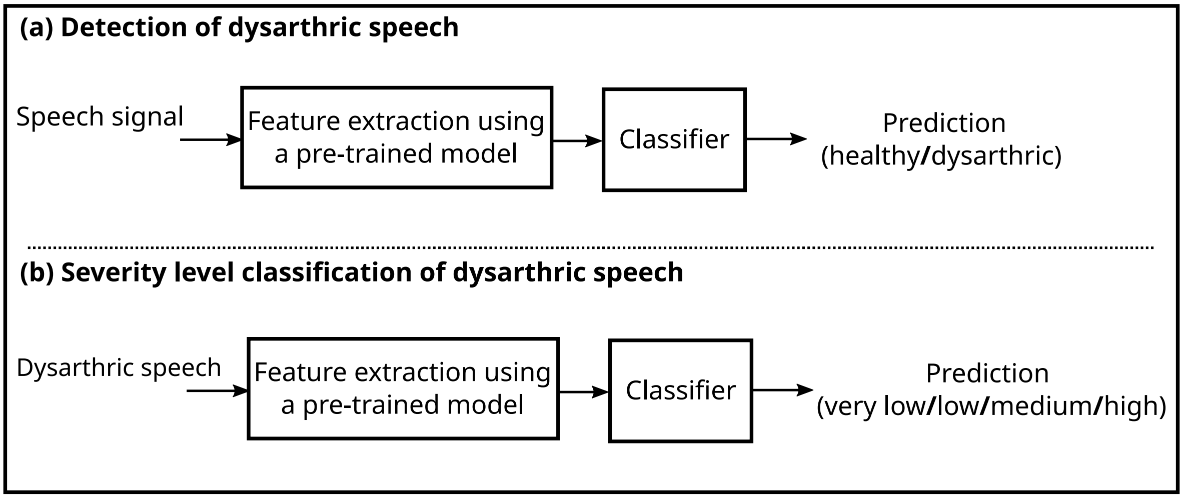

(a) Detection of dysarthric speech
Speech signal
Feature extraction using a pre-trained model
Classifier
Prediction (healthy/dysarthric)
(b) Severity level classification of dysarthric speech
Dysarthric speech
Feature extraction using a pre-trained model
Classifier
Prediction (very low/low/medium/high)

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

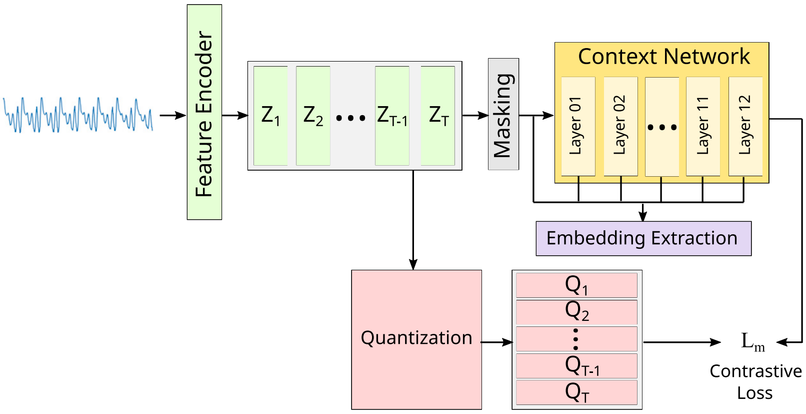


Fig. 2. Block diagram of a wav2vec2 architecture with 12 transformer layers.

latent speech representation q_t from a set of quantized candidate representations $\tilde{q} \in Q_t$ which include q_t and 100 distractors. The distractors are uniformly sampled from other masked time steps of the same utterance. Fig. 2 shows the wav2vec2-BASE model, which contains 12 transformer blocks. More details about wav2vec2 can be found in Baevski et al. (2020).

In the current study, the outputs of the transformer layers of the context network are used as features for detection and classification systems. For the wav2vec2-BASE model, the temporal average of the inputs to the first transformer layer and the output of each of the 12 transformer layers are computed for each speech signal, which results in extracting a total of thirteen feature matrices. Finally, the features are averaged across all frames to obtain a total of thirteen 768-dimensional feature vectors for each speech signal. For the wav2vec2-LARGE model, the temporal average of the inputs to the first transformer layer and the outputs of each of the 24 transformer layers are computed for each speech signal, which results in extracting a total of twenty-five feature matrices. By averaging the features across all frames, twenty-five 1024-dimensional feature vectors are obtained. In the remaining sections of this article, these features are referred to as the wav2vec2-BASE and wav2vec2-LARGE features, or alternatively as wav2vec2-BASE-N and wav2vec2-LARGE-N when specifying the features from the Nth layer.

2.1.2. HuBERT

The HuBERT model architecture contains a multi-layer CNN encoder, a context network, a projection layer, and a code embedding layer. Pre-training of the HuBERT model is carried out in two steps. In the first step, k-means clustering with 100 clusters is performed on 39-dimensional MFCCs (including the first and second derivatives) derived from the raw speech signal to generate the hidden units as targets. Then, each hidden unit target is mapped to its corresponding hidden unit embedding E. In the second step, the raw speech signal is first transformed into the latent speech representations using a CNN encoder and then projected to the correct embedding size of 1024. Next, the resulting embeddings are partially masked and fed to the context network, containing 24 transformer layers. Finally, the resulting context representations C together with the hidden unit embeddings E are used to optimize the model using the cross-entropy loss function. The pre-training process is repeated for three iterations. The k-means clustering of the second iteration is performed on the features extracted from the 6th transformer layer of the first iteration model instead of clustering the MFCC features. In the third iteration, clustering the features extracted from the 9th transformer layer of the second iteration model is carried out. Fig. 3 shows the HuBERT model architecture with 24 transformer blocks.

In this study, the temporal average of the inputs of the first transformer layer and the output of each of the 24 transformer layers are extracted for each speech signal which results in twenty-five feature matrices. These features are then averaged across all frames, yielding a total of twenty-five 1024-dimensional feature vectors. In the remaining sections of this article, these features are referred to as the HuBERT features, or alternatively as HuBERT-N when specifying the features from the Nth layer.

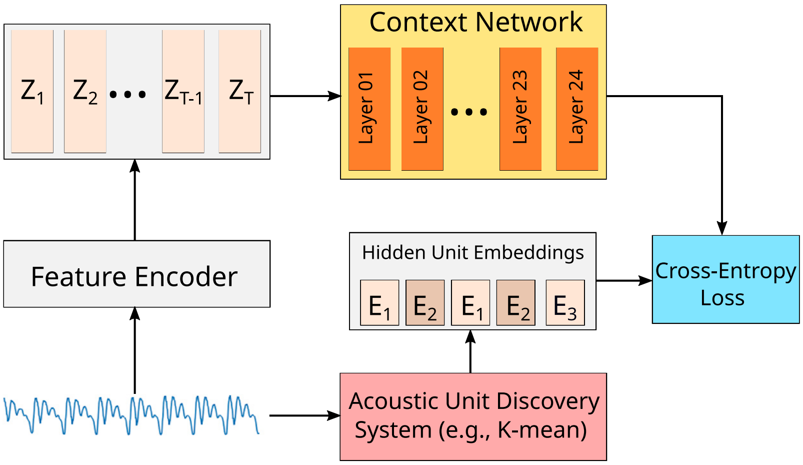


Fig. 3. Block diagram of a HuBERT architecture with 24 transformer layers.

2.2. Classifiers

In the current study, one ML classifier (SVM) and one DL classifier (CNN) are used for both the detection problem (i.e., detection of dysarthric speech) and the multi-class classification problem (i.e., severity level classification of dysarthric speech). SVM with the radial basis function kernel and a regularization parameter value of 1 was used. Moreover, the value of gamma used in SVM was defined as $\gamma = 1/(D \cdot Var(X))$, where $Var(X)$ is the variance of the training data, and D is the dimension of the feature vectors. To find the best parameters, preliminary detection experiments were first conducted by using both the linear and radial basis function (RBF) kernels with five regularization parameter (c) values $(0.01, 0.1, 1, 10, 100)$. Then, the parameter that gave the best detection accuracy was chosen.

For the CNN classifier, the architecture was built as follows: First, the input is passed through three sequential convolutional layers, each followed by batch normalization and the ReLU activation function. Then, the resulting output is flattened and delivered to the next two dense layers. Finally, the second dense layer performs the detection (healthy *vs*. dysarthria) with the sigmoid activation function. This CNN architecture was chosen based on preliminary experiments. In our prior studies in voice pathology detection (Javanmardi et al., 2022), the similar CNN architecture (consisting of two convolutional layers followed by flattening and two fully-connected layers) was found to be most effective. The hyper-parameters used for the CNN classifier are: a batch size of 64, 100 epochs with 20 epochs as the early stopping, the cross entropy as the loss function, and the Adam optimizer with a learning rate of 0.001. These hyper-parameters gave the best detection accuracy in preliminary experiments in which the CNN was first tuned with different learning rates $(0.1, 0.01, 0.001)$, different batch sizes $(8, 16, 32, 64, 128)$ and different number of convolutional layers.

It is worth emphasizing that the same CNN architecture is used in the current study for both detection and multi-class classification tasks, except that in the latter, the output dense layer with the softmax activation function predicts the output severity level label. The SVM and CNN classifiers were implemented using the Scikit-learn library (Pedregosa et al., 2011) and Pytorch library (Paszke et al., 2019).

3. Experimental setup

This section first describes the details of the dysarthria databases used in the current study. Then, the baseline features used for comparison are explained. Next, the training and testing of the classification systems for both detection and multi-class classification problems are provided. Finally, the evaluation metrics are described.

3.1. Databases of dysarthric speech

The present study was carried out using two publicly available dysarthria databases: the universal access speech (UA-Speech) database (Kim et al., 2008) and the TORGO database (Rudzicz et al., 2012).

3.1.1. UA-Speech

This database consists of 765 isolated words collected from 15 dysarthric speakers (4 females and 11 males) diagnosed with cerebral palsy and from 13 healthy controls (4 females and 9 males). The patients' age varies from 18 years to 58 years old. The dysarthric speakers were grouped into four severity categories based on their speech intelligibility ratings: very low (4 speakers), low (3 speakers), medium (3 speakers), and high (5 speakers). The 765 isolated words were recorded in three blocks (B1, B2, and B3). Each block contains 255 words, of which 155 words are common to all three blocks, and the remaining 100 words differ across the blocks. The 155 common words contain 19 computer commands (e.g., 'line', 'enter'), 26 letters of the international radio alphabet (e.g., 'Alpha', 'charlie'), 10 digits (e.g., 'one', 'two'), and the 100 most common words in the Brown corpus of written English (e.g., 'the', 'of'). The 100 uncommon words of each block were selected from Project Gutenberg novels (e.g., 'naturalization', 'faithfulness') (Kim et al., 2008). An eight-microphone array was used to record speech. Speech was sampled at 16 kHz and each microphone was spaced at intervals of 1.5 inches. In this study, the speech utterances from all three blocks of each speaker recorded by microphone number six of the array were used. More details on the UA-Speech database can be found in Kim et al. (2008).

3.1.2. TORGO

This database was recorded from 8 patients (3 females and 5 males) diagnosed with cerebral palsy or amyotrophic lateral sclerosis (ALS) and from 7 healthy control speakers (3 females and 4 males). The patients' age varies from 16 years to 50 years old. The database consists of three categories of speech signals (i.e., non-words, words, and sentences). The non-words category contains 5-10 repetitions of */iy – p – ah/*, */ah – p – iy/*, */p – ah – t – ah – k – ah/*, and vowels uttered over 5 sec at high and low pitches. The words category consists of 50 words selected from the word intelligibility section of the Frenchay Dysarthria Assessment (Enderby, 1980) and 360 words from the word intelligibility section of the Yorkston–Beukelman Assessment of Intelligibility of Dysarthric Speech (Yorkston et al., 1984). The sentences are selected from three pre-selected phoneme-rich sentence sets (i.e., 162 sentences from the sentence intelligibility section of the Yorkston–Beukelman Assessment of Intelligibility of Dysarthric Speech (Yorkston et al., 1984), 460 sentences from the MOCHA database (Wrench, 1999), the Grandfather passage from the Nemours database (Menendez-Pidal et al., 1996), and spontaneously elicited descriptive texts). More details on the selection of words and sentences can be found in Rudzicz et al. (2012). Speech was recorded simultaneously using a head-mounted microphone and an array microphone and sampled at 16 kHz. The current study was carried out using the recordings from the array microphone.

3.2. Baseline features used for comparison

Three standard acoustic feature sets (MFCCs, openSMILE, and eGeMAPS) are considered for comparison as they were shown in Gurugubelli and Vuppala (2020), Narendra and Alku (2019), Xue et al. (2019) to provide good discrimination between dysarthric and healthy speech. In this work, the first feature set (MFCCs) was computed using the Hamming windowed frames of 25 ms with a shift of 5 ms. By computing the first 13 cepstral coefficients (including the 0th coefficient) and their delta & double-delta coefficients, a 39-dimensional MFCC feature vector was obtained per utterance. MFCCs were computed using the Librosa toolkit (McFee et al., 2015). The second set (referred to in this work as openSMILE) is the INTERSPEECH 2016 Computational Paralinguistics Challenge (ComParE) feature set consisting of energy-related low-level descriptors (LLDs) such as loudness and RMS energy, spectral-related LLDs (e.g., MFCCs, psychoacoustic spectral sharpness, spectral harmonicity and spectral flux), and voicing-related LLDs, including voice quality features (jitter and shimmer), logarithmic harmonic-to-noise ratio (HNR), and Viterbi smoothing for fundamental frequency (F0). All these acoustic features together with their statistical functionals form a 6373-dimensional feature vector per utterance. The last of the three sets (eGeMAPS) consists of a limited set of features that were chosen based on their proven effectiveness in past studies, their potential to detect physiological changes in voice production, as well as their demonstrated theoretical significance. The eGeMAPS feature set includes prosodic, excitation, vocal tract, spectral, and cepstral descriptors such as logarithmic F0, jitter, formant frequencies, shimmer, loudness, HNR, alpha ratio, harmonic difference, MFCCs, and spectral flux, as well as their statistical functionals. Altogether, the eGeMAPS feature set contains 88 features. The openSMILE and eGeMAPS feature sets were computed using the freely available feature extraction openSMILE toolkit (Eyben et al., 2013).

3.3. Training and testing

The binary detection experiments were conducted using the leave-one-speaker-out (LOSO) cross-validation strategy, where one speaker was used as testing data and the remaining speakers were considered as training data for both SVM and CNN classifiers. The z-score normalization was applied to both training and testing data using the mean and standard deviation of the training data. In each iteration, the evaluation metrics were saved. This training and testing process was repeated until each speaker was used once as testing data, and finally the evaluation metrics were averaged over all iterations.

In the severity level classification experiments using the UA-Speech database, three dysarthric speakers were left out in order to have a balanced number of speakers in each class. The removed three dysarthric speakers had a “very low” level of intelligibility (one male speaker) and a “high” level of intelligibility (two male speakers). After removing these three speakers, experiments were conducted using the 12 remaining dysarthric speakers. In each iteration, four speakers (one from each class) were considered as testing data and the remaining speakers were used to train the SVM classifier. By considering one speaker from each severity class for testing, a total of 81 ($3 \times 3 \times 3 \times 3$) iterations were performed. Similarly, for the TORGO database, there were a total of 18 ($3 \times 3 \times 2$) iterations (training and testing process) as TORGO contains three levels of intelligibility from 8 speakers (i.e., three speakers for “very low”, two speakers for “low”, and three speakers for “medium”). Both in the detection and multi-class classification experiments using the CNN classifier, 10% of each speaker’s samples from the training data was randomly selected as validation data in each iteration. The training and testing process was repeated for all the iterations (81 for UA-Speech and 18 for TORGO), and finally, the evaluation metrics were averaged over all iterations.

3.4. Evaluation metrics

The performance of the dysarthria detection systems was evaluated using the following five standard metrics: accuracy (ACC), sensitivity (SE), specificity (SP), F1-score (F1), and equal error rate (EER). In order to assess the performance of the severity level classification systems, mean accuracy and class-wise accuracies were computed. We also present confusion matrices for both detection and multi-class classification problems.

4. Results

This section reports the results obtained using the features derived from the three pre-trained models (wav2vec2-BASE, wav2vec2-LARGE, and HuBERT) and the three baseline features (MFCCs, openSMILE, and eGeMAPS) using the SVM and CNN classifiers. First, the results of the detection experiments are presented in Section 4.1 and then the results for the severity classification experiments are reported in Section 4.2.

4.1. Results for detection of dysarthric speech

Performance (in accuracy) of the detection experiments for the UA-Speech database are shown in Figs. 4 and 5 for the SVM and CNN classifiers, respectively. From the baseline features, it can be observed that openSMILE performed better than MFCCs and eGeMAPS. From the results obtained using the wav2vec2-BASE, wav2vec2-LARGE, and HuBERT features, it can be seen that the HuBERT features achieved better detection accuracies compared to the wav2vec2-BASE and wav2vec2-LARGE features (except for the wav2bec2-BASE-1 and the first four wav2vec2-LARGE features, which were better than the HuBERT features). Compared to the best-performing baseline feature (openSMILE), it can be observed that almost all the HuBERT features (except for a few features derived from the final layers), and a few wav2vec2-LARGE features derived from the first layers showed better performance. For the wav2vec2-BASE features, only wav2vec2-BASE-1 outperformed the openSMILE feature.

Detection results in all five metrics are reported in Table 1 for the three baseline features and the best-performing wav2vec2-BASE, wav2vec2-LARGE, and HuBERT features for the SVM and CNN classifiers using the UA-Speech database. It can be seen in the results obtained for the baseline features that openSMILE outperformed the other two features in all the metrics for both classifiers. The results also show that HuBERT-12 performed better than wav2vec2-BASE-1 and wav2vec2-LARGE-4 in all the metrics for the SVM classifier. More specifically, HuBERT-12 showed an absolute improvement of 2.86% in accuracy compared to the best baseline feature (openSMILE). In the case of the CNN classifier, wav2vec2-LARGE-2 outperformed the other two features in all the metrics (except for specificity, where HuBERT-8 showed the same performance). Compared to openSMILE, wav2vec2-LARGE-2 gave an absolute accuracy improvement of 3.27%. As all the layers of HuBERT performed consistently better than the other features, only the HuBERT features are considered in the remainder of this article.

Confusion matrices of the detection experiments obtained for the UA-Speech database are shown in Fig. 6 for the best baseline feature (openSMILE) and the best performing HuBERT features for both the SVM and CNN classifiers. It can be observed that there are fewer confusions between healthy and dysarthric speech for the HuBERT features compared to the openSMILE, both for SVM and CNN.

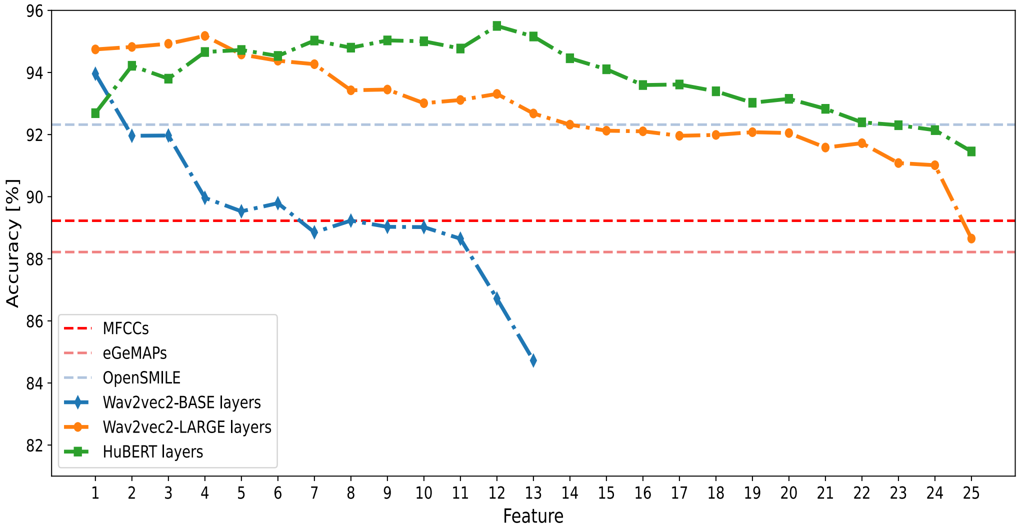


Fig. 4. Dysarthria detection accuracy given by the SVM classifier for the three baseline features (MFCCs, openSMILE, and eGeMAPS) and all 13 wav2vec2-BASE, 25 wav2vec2-LARGE, and 25 HuBERT features for the UA-speech database. The dashed lines represent the mean accuracy for the three baseline features. The bold blue, orange, and green dots represent the mean accuracy for features derived from the wav2vec2-BASE, wav2vec2-LARGE, and HuBERT models, respectively. The numbers on the *x*-axis indicate the index of the corresponding layer.

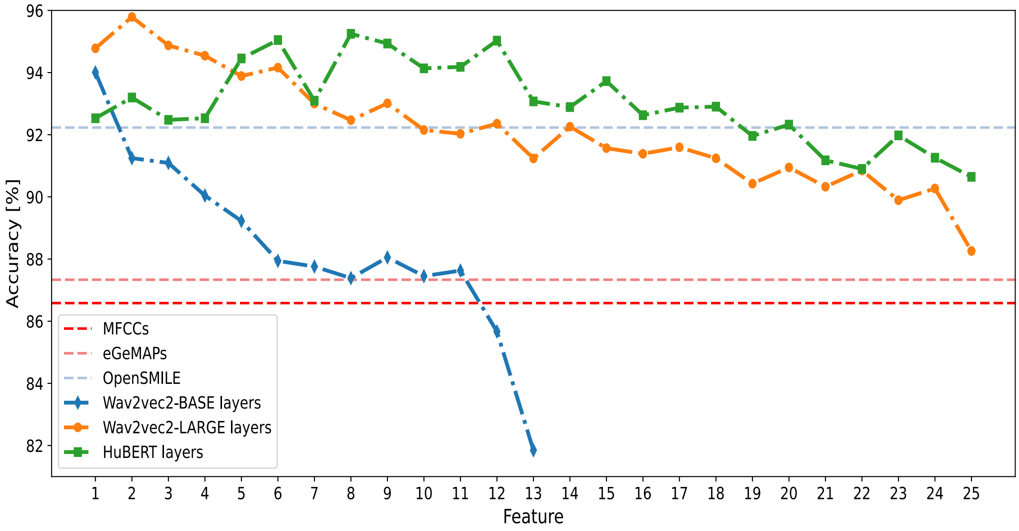


Fig. 5. Dysarthria detection accuracy given by the CNN classifier for the three baseline features (MFCCs, openSMILE, and eGeMAPS) and all 13 wav2vec2-BASE, 25 wav2vec2-LARGE, and 25 HuBERT features for the UA-speech database. The dashed lines represent the mean accuracy for the three baseline features. The bold blue, orange, and green dots represent the mean accuracy for features derived from the wav2vec2-BASE, wav2vec2-LARGE, and HuBERT models, respectively. The numbers on the *x*-axis indicate the index of the corresponding layer.

The results of the detection experiments (in accuracy) for the TORGO database are shown in Fig. 7 for the three baseline features and HuBERT features (as they have shown better detection performance in UA-Speech) for the SVM and CNN classifiers. The results in the figure show that the HuBERT features derived from a few final layers gave higher detection accuracies compared to openSMILE (i.e., the best-performing feature among the baseline features).

Table 2 presents the detection results in all five metrics for the three baseline features and the best-performing HuBERT features for both the SVM and CNN classifiers. From the results in the table, it can be seen that the HuBERT features outperformed the other three baseline features in all the metrics (except for sensitivity, where openSMILE showed a slightly higher performance) for both SVM and CNN. Compared to the best-performing baseline feature (openSMILE), the SVM classifier with HuBERT-24 and the CNN classifier with HuBERT-25 showed an absolute improvement of 1.33% and 2.32% in detection accuracy, respectively. Confusion matrices of the detection experiments for the TORGO database are shown in Fig. 8 for openSMILE (the best baseline feature) and for the best-performing HuBERT features for the SVM and CNN classifiers. It can be observed that the HuBERT features outperformed openSMILE in the detection of healthy samples in both classifiers. On the other hand, openSMILE showed better detection accuracy for dysarthric samples.

4.2. Results for severity level classification of dysarthric speech

The severity level classification performance (in accuracy) obtained using the speech data of UA-Speech is shown in Fig. 9 for the three baseline features and the HuBERT features. Among the baseline features, eGeMAPS outperformed the other two features in both classifiers. From the results obtained for the HuBERT features, it can be clearly seen that all the features (except for HuBERT-1 and HuBERT-2) showed higher accuracy compared to eGeMAPS in both classifiers. The overall classification accuracy and the class-wise accuracies obtained for UA-Speech using the SVM and CNN classifiers are presented in Table 3 for

Table 1
Dysarthria detection results for the three baseline features (MFCCs, openSMILE, and eGeMAPS) along with the best-performing wav2vec2-BASE, wav2vec2-LARGE, and HuBERT features for the SVM and CNN classifiers using the UA-Speech database. Here ACC refers to accuracy, SE refers to sensitivity, SP refers to specificity, and F1 refers to F1-score.

Classifier	Feature	ACC [%]	SE	SP	F1	EER
SVM	MFCCs	89.22	0.86	0.93	0.90	0.102
	OpenSMILE	92.64	0.88	0.97	0.92	0.074
	eGeMAPS	88.22	0.87	0.90	0.89	0.116
	wav2vec2-BASE-1	93.96	0.92	0.95	0.94	0.059
	wav2vec2-LARGE-4	95.17	0.92	0.97	0.95	0.046
	Hubert-12	**95.50**	**0.93**	**0.98**	**0.96**	**0.038**
CNN	MFCCs	86.58	0.84	0.88	0.87	0.132
	OpenSMILE	92.52	0.90	0.97	0.92	0.078
	eGeMAPS	87.33	0.87	0.88	0.88	0.126
	wav2vec2-BASE-1	94.00	0.93	0.94	0.94	0.060
	wav2vec2-LARGE-2	**95.79**	**0.94**	**0.98**	**0.96**	**0.040**
	Hubert-8	95.25	0.93	**0.98**	0.95	0.043

Table 2
Dysarthria detection results for the three baseline features (MFCCs, openSMILE, and eGeMAPS) along with the best performing HuBERT features for the SVM and CNN classifiers using the TORGO database. Here ACC refers to accuracy, SE refers to sensitivity, SP refers to specificity, and F1 refers to F1-score.

Classifier	Feature	ACC [%]	SE	SP	F1	EER
SVM	MFCCs	63.13	0.44	0.73	0.46	0.417
	OpenSMILE	74.79	**0.59**	0.69	0.54	0.351
	eGeMAPS	59.42	0.43	0.62	0.40	0.425
	Hubert-24	**76.12**	0.55	**0.81**	**0.57**	**0.340**
CNN	MFCCs	65.24	0.55	0.66	0.50	0.407
	OpenSMILE	72.80	**0.62**	0.60	0.52	0.382
	eGeMAPS	68.74	0.57	0.61	0.50	0.400
	Hubert-25	**75.12**	0.57	**0.77**	**0.55**	**0.356**

Table 3
Severity level classification accuracies and class-wise accuracies obtained using the SVM and CNN classifiers for the three baseline features (MFCCs, openSMILE, and eGeMAPS) along with the best HuBERT features for the UA-Speech database. Here ACC refers to accuracy and C refers to class.

Classifier	Feature	ACC [%]	ACC [%]			
			$C_{very\text{-}low}$	C_{low}	C_{medium}	C_{high}
SVM	MFCCs	33.95	50.63	7.32	21.02	56.78
	OpenSMILE	35.63	55.36	12.13	4.58	70.48
	eGeMAPS	35.99	66.90	19.80	6.16	51.09
	Hubert-12	**46.45**	67.00	19.25	13.50	86.03
CNN	MFCCs	37.67	51.75	11.22	25.89	61.85
	OpenSMILE	36.61	61.38	7.80	9.90	67.39
	eGeMAPS	39.87	67.63	18.50	16.70	56.65
	Hubert-7	**48.01**	70.60	10.18	33.71	77.53

the best-performing baseline feature (eGeMAPS) and the best HuBERT features. From the table, it can be seen in the results obtained for the SVM classifier that HuBERT-12 gave the best accuracy corresponding to an absolute improvement of 10.46% compared to the baseline eGeMAPS feature. In the case of CNN, HuBERT-7 showed an absolute improvement of 8.14% in accuracy compared to eGeMAPS.

Fig. 10 shows the confusion matrices for the severity level classification experiments for eGeMAPS (the best baseline feature) and the best-performing HuBERT feature for the UA-Speech database. It can be seen that the two extreme ends of the severity level scale (“very low” and “high”) were recognized better than the two other severity levels (“low” and “medium”). By comparing the results between the HuBERT and eGeMAPS features, HuBERT showed higher accuracy for all severity levels (except in “low” for which eGeMAPS achieved a better accuracy). The results also showed that there are more confusions between “low” and “medium” and vice versa, and also between “low” and “very low”, and “medium” and “high”.

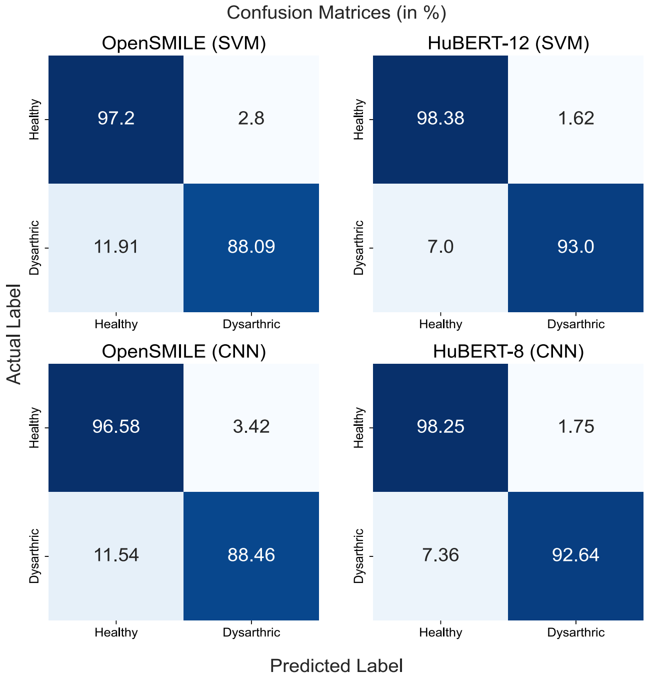


Fig. 6. Confusion matrices of dysarthria detection given by the SVM and CNN classifiers for the best-performing baseline feature (openSMILE) and the best HuBERT features for the UA-Speech database.

For the TORGO database, the performance of severity level classification (in accuracy) is shown in Fig. 11 for the three baseline features and the HuBERT features using the SVM and CNN classifiers. From the figure, it can be observed that almost all the HuBERT features (except for a few features derived from the first layers) clearly performed better than eGeMAPS (i.e., the best-performing feature among the baseline features) by showing higher accuracy for both SVM and CNN. The overall classification accuracy and the class-wise accuracies obtained for TORGO using the SVM and CNN classifiers are presented in Table 4 for the eGeMAPS feature and the best HuBERT features. From the table, it can be seen in the results obtained for the SVM classifier that HuBERT-13 gave the best accuracy, corresponding to an absolute improvement of 6.54% compared to the eGeMAPS feature. In the case of CNN, HuBERT-10 showed an absolute improvement of 8.62% in accuracy compared to eGeMAPS.

Finally, the authors would like to point out that in addition to the detection and severity level classification experiments that were reported above in Sections 4.1 and 4.2, extra experiments were conducted in which a feature selection method was used to reduce the dimensions of the openSMILE and eGeMAPS sets. As a feature selection method, the popular random forest importance (RFI) algorithm (Kursa and Rudnicki, 2011) was used. However, the use of feature selection both in the detection and severity level classification tasks gave results that were inferior to those that were reported above using the original feature dimensions of openSMILE and eGeMAPS. Therefore, the results of these additional experiments are not reported.

5. Discussion

This section discusses the main observations of the study by first discussing the results of the detection experiments that were reported in Section 4.1 and then discussing the results of the severity classification experiments that were reported in Section 4.2 .

5.1. Discussion of the detection results

From the results shown in Figs. 4 and 5, two main observations can be made. First, there is a falling trend in accuracy when moving

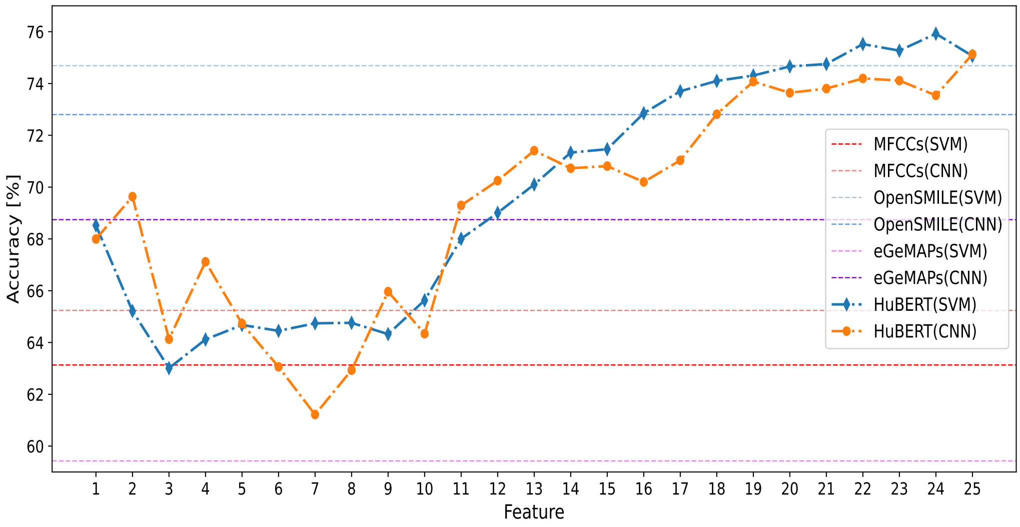


Fig. 7. Dysarthria detection accuracy given by the SVM and CNN classifiers for the three baseline features (MFCCs, openSMILE, and eGeMAPS) and all 25 HuBERT features for the TORGO database. The dashed lines represent the mean accuracy for the three baseline features using SVM and CNN. The bold blue and orange dots represent the mean accuracy for the 25 HuBERT features. The numbers on the *x*-axis indicate the index of the corresponding layer.

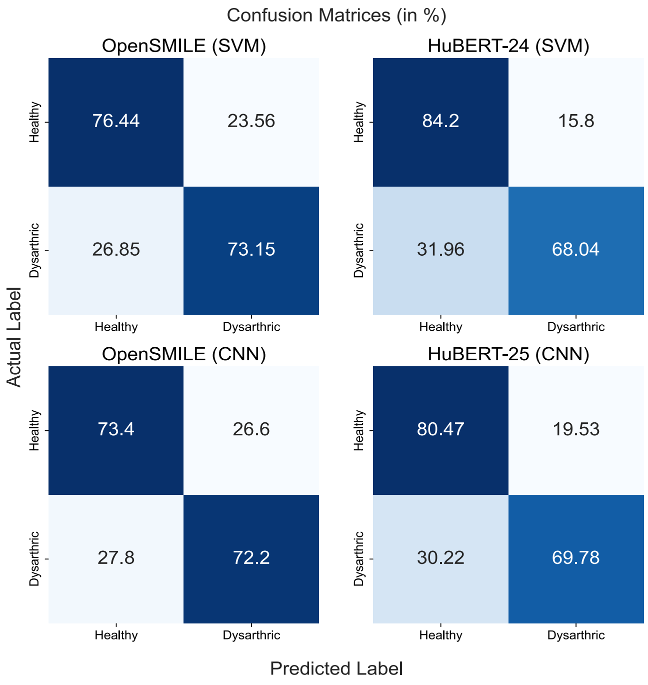


Fig. 8. Confusion matrices of dysarthria detection given by the SVM and CNN classifiers for the best-performing baseline feature (openSMILE) and the best HuBERT features for the TORGO database.

from the initial layer towards the final layer for the wav2vec2-BASE, wav2vec2-LARGE and HuBERT features. Originally, these models were pre-trained on a large amount of unlabeled speech and fine-tuned using a small set to perform automatic speech recognition (ASR) tasks. Therefore, the initial layers of these models tend to capture more generic speech information, which is fundamental for ASR. Such information includes, for example, basic phonetic elements and acoustic properties like pitch and timbre. These elements are crucial for identifying dysarthria, which often manifests itself as atypical articulation and acoustic variation. This can be the reason why the initial layers of these models showed better performance in dysarthria detection. As we progress from the initial layers to the final layers in these models, there is a notable shift in focus from basic acoustic features to more complex linguistic information. The final layers are designed to capture more abstract linguistic aspects such as syntax and semantics. These aspects are typically less linked to the specific characteristics of dysarthric speech, which often involves more distinct phonetic and acoustic irregularities. Since these final layers are adept at processing typical linguistic patterns, they may be less effective in recognizing atypical speech patterns characteristic of dysarthria. Therefore, this shift in focus could contribute to the observed falling trend in dysarthria detection accuracy in the final layers.

Table 4
Severity level classification accuracies and class-wise accuracies obtained using the SVM and CNN classifiers for the three baseline features (MFCCs, openSMILE, and eGeMAPS) along with the best HuBERT features for the TORGO database. Here ACC refers to accuracy and C refers to class.

Classifier	Feature	ACC [%]	ACC [%]		
			$C_{very\text{-}low}$	C_{low}	C_{medium}
SVM	MFCCs	31.25	63.71	7.29	22.74
	OpenSMILE	37.84	73.44	0.00	40.08
	eGeMAPS	42.12	68.37	0.00	57.98
	Hubert-13	**48.66**	86.00	0.22	59.76
CNN	MFCCs	33.29	48.18	12.23	39.47
	OpenSMILE	40.41	73.30	0.05	47.89
	eGeMAPS	41.21	67.30	0.00	56.35
	Hubert-10	**49.83**	79.52	0.05	69.91

Second, it can be observed that the accuracy difference between the 25 wav2vec2-LARGE features and 25 HuBERT features are relatively smaller compared to the 13 wav2vec2-BASE features. This variation can be attributed to the differences in model parameters and the amount of training data used. Both wav2vec2-LARGE and HuBERT have larger model sizes and were trained with extensive training data which results in capturing a wide range of speech variances, including those relevant to dysarthria. The wav2vec2-BASE model with fewer parameters trained on less diverse data showed a more pronounced decrease in accuracy across layers. This indicates that the capacity of a model to learn and adapt to diverse speech patterns, including pathological speech, is significantly influenced by its complexity and the comprehensiveness of its training data. Moreover, using vast amounts of normal speech in the training of larger models raises the possibility that some of the speech samples may originate from speakers with undiagnosed pathologies. This means that the models, especially HuBERT and wav2vec2-LARGE, might have been unintentionally exposed to the characteristics of pathological speech during training. Such exposure

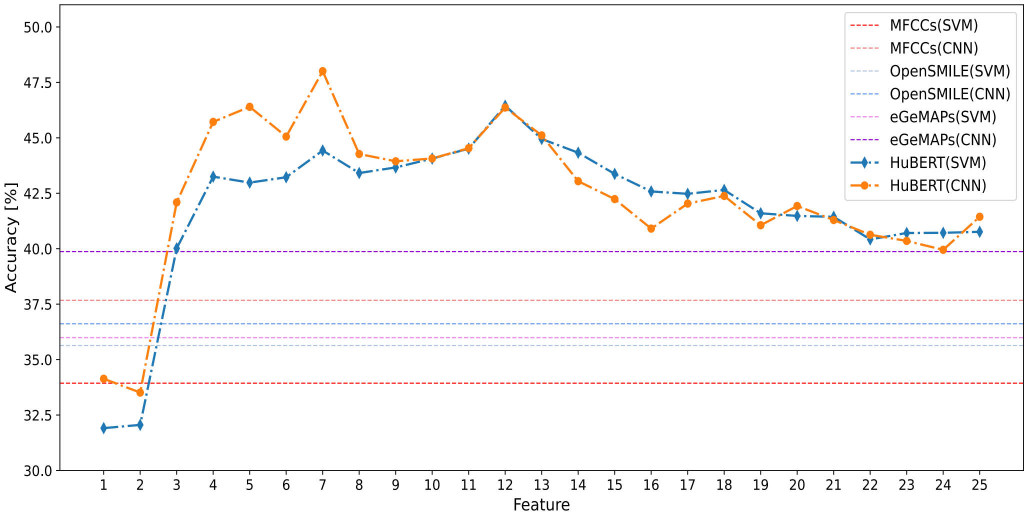


Fig. 9. Severity level classification accuracies given by the SVM and CNN classifiers for the three baseline features (MFCCs, openSMILE, and eGeMAPS) and all 25 HuBERT features for the UA-Speech database. The dashed lines represent the mean accuracy for the three baseline features using SVM and CNN. The bold blue and orange dots represent the mean accuracy for the 25 HuBERT features. The numbers on the *x*-axis indicate the index of the corresponding layer.

(even if incidental) might enhance the models' ability to recognize and classify pathology-related speech characteristics, and therefore it helps to improve the performance in detecting dysarthric speech compared to the wav2vec2-BASE model trained with a much smaller amount of normal speech.

From the results of confusion matrices reported in Fig. 6, it can be seen that there is a tendency for the classifiers to misclassify dysarthric speech as healthy speech more frequently compared to misclassifying healthy speech as dysarthric speech. Table 5 shows the details of the misclassified dysarthric samples. The majority of dysarthric samples that were wrongly classified as healthy samples were from only 4 dysarthric speakers (three males and a female). The three male speakers were M10, M14 and M16, and the female speaker was F05. From these four speakers, three (M10, M14, F05) had high severity level of dysarthria and one speaker (M16) had low severity level of dysarthria. Misclassifying dysarthric samples of low severity level as healthy can be expected because speakers of low severity level typically show high speech intelligibility making their speech similar to healthy speech. On the other hand, misclassification of dysarthric speech of high severity level as healthy speech is an unexpected outcome, particularly given the conventional understanding that higher severity level of dysarthria typically correlates with lower speech intelligibility, as opposed to clear articulation observed in healthy speech. This is a surprising finding especially in the context of utilizing the openSMILE features. Given its comprehensive nature, the openSMILE set is designed to capture a wide range of acoustic features, including those particularly relevant to dysarthric speech such as articulatory, phonatory and prosody features. This robustness in feature extraction should enhance the classifier's ability to distinguish between dysarthric and healthy speech across varying degrees of severity. Therefore, the misclassification of dysarthric samples of high level of severity as healthy samples calls for deeper exploration to understand whether these misclassifications are due to the complex and variable nature of dysarthric speech, the influence of speaker-specific characteristics, or the adaptive capabilities of the classification algorithms employed.

By comparing the results obtained with the HuBERT features for UA-Speech and TORGO, it can be clearly observed that the HuBERT features show a rising trend in accuracy for TORGO (Fig. 7), whereas the HuBERT features show a falling trend in accuracy for UA-Speech (Figs. 4 and 5). This difference may be attributed to the distinct characteristics of the two datasets. The UA-Speech database contains only one speaking task (i.e., word pronunciation), whereas the TORGO database includes three speaking tasks (pronunciation of non-words, words, and sentences). This variation in the speaking task might influence the effectiveness of feature extraction at different layers of the HuBERT model. In addition to the speaking tasks, some factors such as the amount of data, the diversity of speech impairments, and background noise levels in each database could further account for the observed trends. For instance, a richer or more complex database (TORGO) might benefit more from the deeper, more abstract representations in HuBERT's later layers. On the other hand, a simpler database (UA-Speech) might be adequately represented by the initial layers. Moreover, the training regime and data diversity of HuBERT might influence its adaptability to different speech tasks. HuBERT was originally trained on a diverse set of speaking tasks. Therefore, this might generalize better to a complex database (TORGO) and could explain why deeper layers perform better for TORGO than for UA-Speech.

One more observation that was found in the detection experiments for the UA-Speech and TORGO databases is that the openSMILE features performed better among the baseline features. We argue that the better performance of the openSMILE features can be due to the richer, more comprehensive nature of its features compared to the other two feature sets. The MFCCs are effective in capturing particularly vocal tract information of speech, and eGeMAPS extends this by including additional features that capture prosodic and voice quality information. The openSMILE features, consisting of all the eGeMAPS features and of many other features, provide a broader and more detailed representation of speech characteristics. This extensive acoustic profiling by the openSMILE features likely contributes to the enhanced ability of openSMILE to capture dysarthric speech, leading to better detection performance than using MFCCs or eGeMAPS alone.

5.2. Discussion of the severity level classification results

The results of the severity level classification experiments showed that the eGeMAPS features outperformed the MFCCs and openSMILE features for both UA-Speech and TORGO databases (which was in contrast to the detection experiments, where openSMILE was the best baseline feature set). We argue that the performance differences between openSMILE and eGeMAPS in the current study can be attributed to the nature and composition of these feature sets. The openSMILE features covers a comprehensive range of acoustic features. This feature set is well suited for dysarthria detection, which is a task where identi-

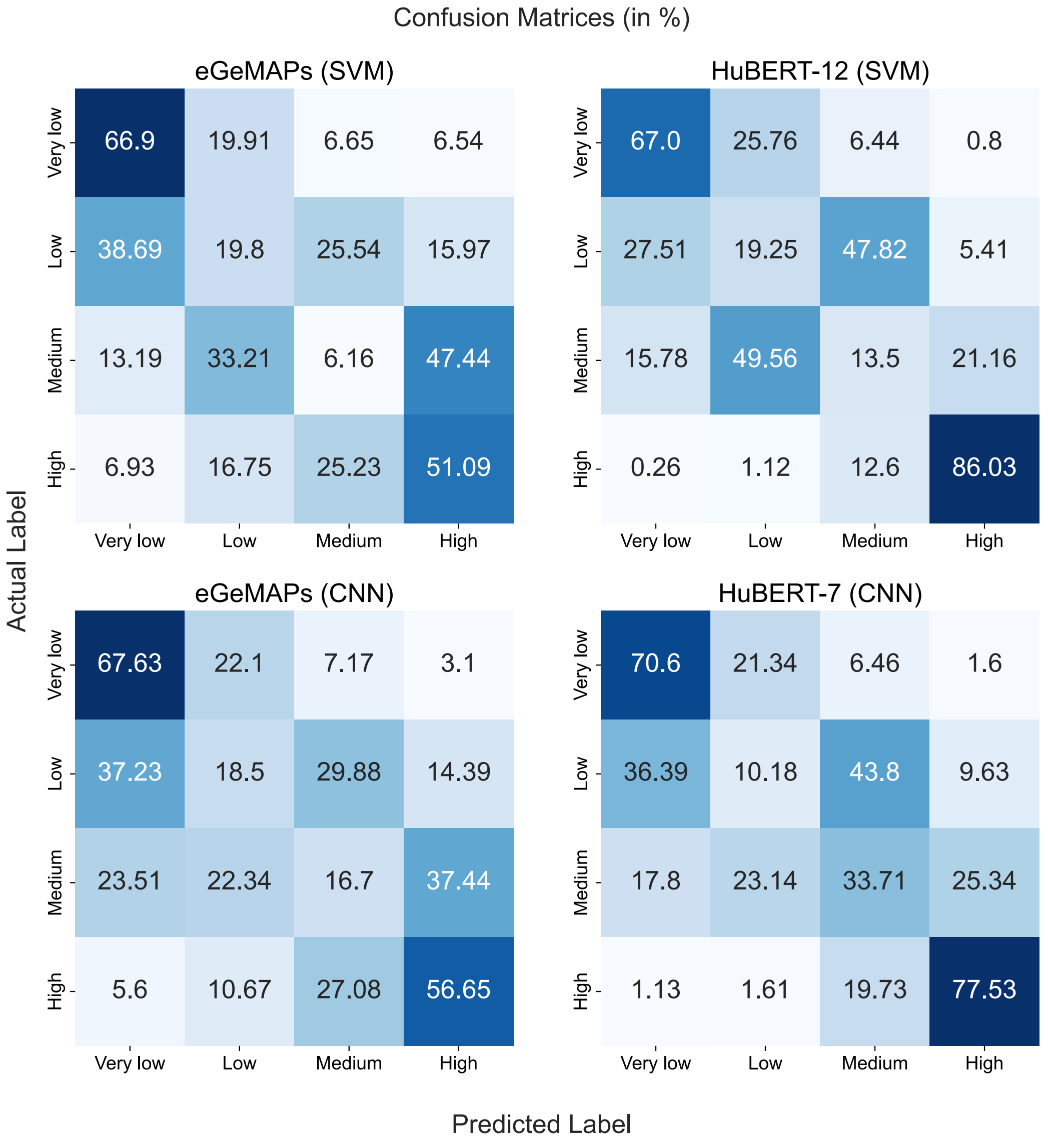


Fig. 10. Confusion matrices of severity level classification given by the SVM and CNN classifiers for the best-performing baseline feature (eGeMAPS) and the best HuBERT features for the UA-Speech database.

Table 5
Details of the misclassified dysarthric samples..

Speaker ID	Severity level	No. of misclassified samples			
		openSMILE (SVM)	openSMILE (CNN)	HuBERT-12 (SVM)	HuBERT-8 (CNN)
M10	High	90 (0.78%)	98 (0.85%)	165 (1.44%)	136 (1.19%)
M14	High	517 (4.5%)	421 (3.67%)	310 (2.7%)	158 (1.38%)
M16	Low	555 (4.84%)	410 (3.57%)	125 (1.09%)	233 (2.03%)
F05	High	203 (1.77%)	390 (3.4%)	180 (1.57%)	300 (2.62%)
All other	–	2 (0.02%)	5 (0.04%)	23 (0.2%)	17 (0.15%)
Total	–	1367 (11.91%)	1324 (11.54%)	803 (7%)	844 (7.36%)

fying the presence of speech abnormalities is key. The large openSMILE feature set with its 6373 dimensions allows for detailed analysis of speech characteristics which is crucial for detecting the presence of dysarthria. On the other hand, eGeMAPS is a more compact collection of 88 acoustic features, specifically chosen for their effectiveness in past studies for detecting physiological changes in voice production. These features focus on key speech elements such as pitch, jitter, formant frequencies, shimmer, and loudness, which are particularly relevant for assessing voice quality and prosodic variations which are critical for classifying the severity of dysarthria.

Table 6 summarizes the main characteristics of a few recent studies (including the current work) on severity level classification of dysarthria that used the data of UA-Speech and TORGO. It can observed that the results reported in Joshy and Rajan (2021) are notably higher than those in other studies, including the current investigation. More specifically, the results published in Joshy and Rajan (2021) showed the best classification accuracy of 93.24% for UA-Speech and 96.18% for TORGO, while our study reported accuracies of 48.01% for UA-Speech and 49.83% for TORGO. These large differences can be attributed to two main disparities between (Joshy and Rajan, 2021) and the current

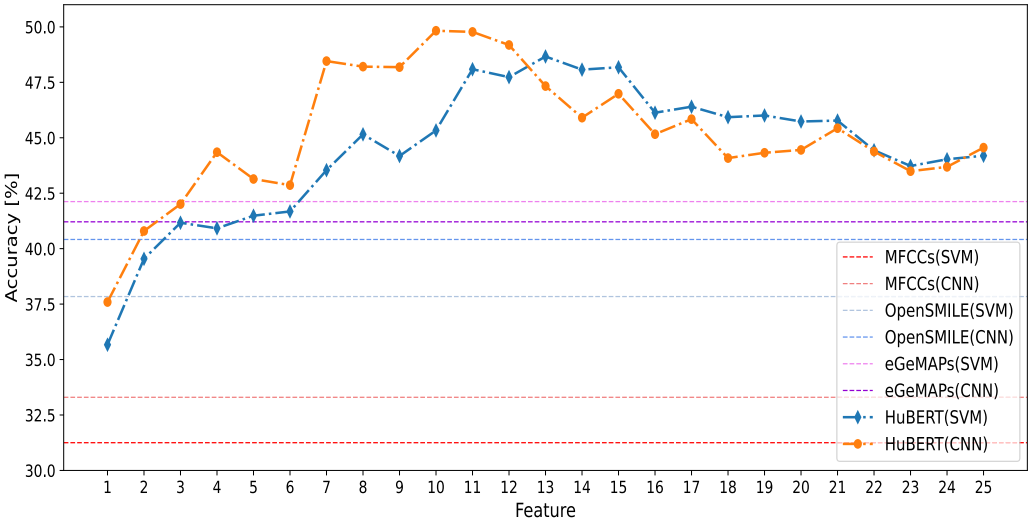


Fig. 11. Severity level classification accuracies given by the SVM and CNN classifiers for the three baseline features (MFCCs, openSMILE, and eGeMAPS) and all 25 HuBERT features for the TORGO database. The dashed lines represent the mean accuracy for the three baseline features using SVM and CNN. The bold blue and orange dots represent the mean accuracy for the 25 HuBERT features. The numbers on the x-axis indicate the index of the corresponding layer.

study. First, the difference in the amount of training and testing data. Second, the difference in the strategy implemented in the training and testing of the classifier. In Joshy and Rajan (2021), the classifier was trained using only the common words of UA-Speech and tested with the uncommon words (i.e., testing the classifier on seen speakers with unseen utterances). For TORGO, only word utterances were used in which 60% of the data was used for training the classifier, 20% for validation and another 20% for testing. These methodological choices have most likely contributed to the high accuracy rates reported in Joshy and Rajan (2021).

In contrast, the results of our study are closer to the results reported in Joshy and Rajan (2022) and Chandrashekar et al. (2019). However, there are still differences between the results published in Joshy and Rajan (2022) and in Chandrashekar et al. (2019) and in the current study, and these differences are most likely also due to the two issues mentioned above. In Joshy and Rajan (2022), the Leave-One-Speaker-Out (LOSO) cross-validation strategy was utilized by using the common words of 14 speakers of UA-Speech for training, and the uncommon words of the left-out speaker for testing in each round (i.e., testing the classifier on unseen speaker with unseen utterances). In Chandrashekar et al. (2019), only 455 samples of UA-Speech were used with the LOSO cross-validation strategy (i.e., testing the classifier on unseen speaker with seen utterances). For TORGO, the authors of Chandrashekar et al. (2019) only used word utterances of 6 dysarthric speakers, of which 1358 words were used for training and 339 words for testing the classifier. The details of data usage and the strategy of the training and testing used in the current study were explained in Section 3.3.

6. Summary and conclusions

In the current study, we have investigated the effectiveness of pre-trained models as feature extractor in the detection of dysarthric speech as well as in the severity level classification of dysarthria. The experiments were carried out by building different detection and classification systems using three popular pre-trained models (wav2vec2-BASE, wav2vec2-LARGE, and HuBERT) to extract features, and two classifiers (SVM and CNN) for predicting the output labels. The speech signals of two publicly available dysarthria databases (UA-Speech and TORGO) were used in the study. In order to evaluate the performance of the features derived from the pre-trained models, a comparison with the three popularly used feature sets (MFCCs, openSMILE, and eGeMAPS) was conducted in the detection and severity level classification.

The experimental results for the detection systems indicated that the features derived from the pre-trained models outperformed the three baseline features. This suggests that the unsupervised learning process of pre-trained models, where the models have seen large amounts of healthy speech of different acoustical characteristics, is actually a more effective basis for feature extraction of pathological voice than using individual hand-crafted features that have been designed to describe specific phenomena that are expected to be present in pathological voice (e.g., jitter for quantifying speech aperiodicity). The results also showed that the features extracted from the HuBERT model performed better than the wav2vec2-BASE and wav2vec2-LARGE features. The higher performance of the HuBERT features may be due to re-using embeddings from the intermediate layer of the context network to generate the hidden units (targets), while the wav2vec2 model uses the output of the CNN encoder for quantization to create the targets. The authors of the HuBERT paper (Hsu et al., 2021) showed that the process of re-using the embeddings from the context networks leads to improved target quality. This improvement in target quality could potentially be a reason for achieving better results compared to the wav2vec2 models in our experiments on detection and severity level classification of dysarthria. Overall, the HuBERT features improved the detection accuracy with absolute accuracy improvements in the range from 1.33% (the SVM classifier, the TORGO database) to 2.86% (the SVM classifier, the UA-Speech database) compared to the baseline features. The results of severity level classification also revealed that the features extracted from the HuBERT model performed better than the baseline features by showing absolute accuracy improvements in the range from 6.54% (the SVM classifier, the TORGO database) to 10.46% (the SVM classifier, the UA-Speech database).

In conclusion, our experiments conducted on two dysarthria databases indicated that the features extracted from pre-trained models are generalizable and useful in both detection and severity level classification problems. Moreover, fine-tuning using dysarthric speech can be explored to further improve the performance of the detection and severity level classification systems in individual and cross-database scenarios. In addition, further studies are needed to understand the effectiveness of features derived from pre-trained models (with and without fine-tuning) for other disorders.

Table 6
A brief comparison of recent works in the severity level classification of dysarthria from speech..

Ref.	Feature & Classifier	Database	Data Usage	Accuracy	Remarks
Joshy and Rajan (2021)	MFCCs with CNN	UA-Speech TORGO	UA-Speech: common words used for training and uncommon words used for testing TORGO: 60% of the data used for training, 20% for validation and 20% for testing	UA-Speech: 93.24% TORGO: 96.18%	– MFCCs in form of 2-D were used – For TORGO, all the samples were not used
Joshy and Rajan (2022)	MFCC-based i-vectors with DNN	UA-Speech	Common words used for training and uncommon words used for testing with LOSO cross-validation	UA-Speech: 49.22%	– TORGO was not used
Chandrashekar et al. (2019)	Mel-spectrogram with CNN	UA-Speech TORGO	UA-Speech: only 455 words of each speaker used with LOSO cross-validation TORGO: 1358 words used for training and 339 words used for testing	UA-Speech: 54.10% TORGO: 54.00%	– For UA-Speech and TORGO, all the samples were not used – LOSO cross-validation was not used for TORGO
Current study	HuBERT features with CNN	UA-Speech TORGO	UA-Speech: all the samples from only 12 dysarthric speakers used TORGO: all the samples from all 8 dysarthric speakers used	UA-Speech: 48.01% TORGO: 49.83%	– UA-Speech: the accuracy represents the average accuracy of 81 rounds (four speakers (one per class) for testing and the 8 remaining speakers used for training in each round) – TORGO: the accuracy represents the average accuracy of 18 rounds

CRediT authorship contribution statement

Farhad Javanmardi: Methodology, Writing – original draft, Writing – review & editing, Conceptualization, Data curation, Formal analysis, Investigation, Resources, Software, Validation, Visualization. **Sudarsana Reddy Kadiri:** Supervision, Writing – review & editing, Conceptualization, Data curation, Methodology, Resources, Validation. **Paavo Alku:** Supervision, Writing – review & editing, Funding acquisition, Validation.

Declaration of competing interest

We declare that we have no conflict of interest in this submission.

Data availability

The authors do not have permission to share data.

Acknowledgments

This study was supported by the Academy of Finland (Project No. 330139).

References

Al-Ali, A., Al-Maadeed, S., Saleh, M., Naidu, R.C., Alex, Z.C., Ramachandran, P., Khoodeeram, R., et al., 2023. Classification of dysarthria based on the levels of severity. a systematic review. arXiv preprint arXiv:2310.07264.

Baevski, A., Zhou, Y., Mohamed, A., Auli, M., 2020. Wav2vec 2.0: A framework for self-supervised learning of speech representations. In: Advances in Neural Information Processing Systems. Curran Associates, Inc., pp. 12449–12460.

Chandrashekar, H., Karjigi, V., Sreedevi, N., 2019. Spectro-temporal representation of speech for intelligibility assessment of dysarthria. IEEE J. Sel. Top. Sign. Proces. 14 (2), 390–399.

Chandrashekar, H.M., Karjigi, V., Sreedevi, N., 2020. Investigation of different time-frequency representations for intelligibility assessment of dysarthric speech. IEEE Trans. Neural Syst. Rehabil. Eng. 28 (12), 2880–2889.

De Bodt, M.S., Hernández-Díaz Huici, M.E., Van De Heyning, P.H., 2002a. Intelligibility as a linear combination of dimensions in dysarthric speech. J. Commun. Disorders 35 (3), 283–292.

De Bodt, M.S., Huici, M.E.H.-D., Van De Heyning, P.H., 2002b. Intelligibility as a linear combination of dimensions in dysarthric speech. J. Commun. Disorders 35 (3), 283–292.

Doyle, P.C., Leeper, H.A., Kotler, A.-L., Thomas-Stonell, N., O'Neill, C., Dylke, M.-C., Rolls, K., 1997. Dysarthric speech: A comparison of computerized speech recognition and listener intelligibility. J. Rehabil. Res. Develop. 34 (3), 309–316.

Duffy, J.R., 2019. Motor Speech Disorders E-Book: Substrates, Differential Diagnosis, and Management. Elsevier Health Sciences.

Enderby, P., 1980. Frenchay dysarthria assessment. Br. J. Disord. Commun. 15 (3), 165–173.

Eyben, F., Scherer, K.R., Schuller, B.W., Sundberg, J., André, E., Busso, C., Devillers, L.Y., Epps, J., Laukka, P., Narayanan, S.S., Truong, K.P., 2016. The geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. IEEE Trans. Affect. Comput. 7 (2), 190–202.

Eyben, F., Weninger, F., Gross, F., Schuller, B., 2013. Recent developments in opensmile, the munich open-source multimedia feature extractor. In: Proceedings of the 21st ACM International Conference on Multimedia. MM '13, Association for Computing Machinery, New York, NY, USA, pp. 835–838.

Eyben, F., Wöllmer, M., Schuller, B., 2010. Opensmile: The munich versatile and fast open-source audio feature extractor. In: Proceedings of the 18th ACM International Conference on Multimedia. MM '10, Association for Computing Machinery, pp. 1459–1462.

Falk, T.H., Chan, W.-Y., Shein, F., 2012. Characterization of atypical vocal source excitation, temporal dynamics and prosody for objective measurement of dysarthric word intelligibility. Speech Commun. 54 (5), 622–631, Advanced Voice Function Assessment.

Fan, Z., Li, M., Zhou, S., Xu, B., 2021. Exploring wav2vec 2.0 on Speaker Verification and Language Identification. In: Proc. Interspeech. pp. 1509–1513.

Fernández-Díaz, M., Gallardo-Antolín, A., 2020. An attention long short-term memory based system for automatic classification of speech intelligibility. Eng. Appl. Artif. Intell. 96, 103976.

Gauder, L., Pepino, L., Ferrer, L., Riera, P., 2021. Alzheimer disease recognition using speech-based embeddings from pre-trained models.. In: Proc. Interspeech. pp. 3795–3799.

Grósz, T., Porjazovski, D., Getman, Y., Kadiri, S., Kurimo, M., 2022. Wav2vec2-based paralinguistic systems to recognise vocalised emotions and stuttering. In: Proceedings of the 30th ACM International Conference on Multimedia. pp. 7026–7029.

Gupta, S., Patil, A.T., Purohit, M., Parmar, M., Patel, M., Patil, H.A., Guido, R.C., 2021. Residual neural network precisely quantifies dysarthria severity-level based on short-duration speech segments. Neural Netw. 139, 105–117.

Gurugubelli, K., Vuppala, A.K., 2019. Perceptually enhanced single frequency filtering for dysarthric speech detection and intelligibility assessment. In: IEEE International Conference on Acoustics, Speech and Signal Processing. ICASSP, pp. 6410–6414.

Gurugubelli, K., Vuppala, A.K., 2020. Analytic phase features for dysarthric speech detection and intelligibility assessment. Speech Commun. 121, 1–15.

Hernandez, A., Pérez-Toro, P.A., Nöth, E., Orozco-Arroyave, J.R., Maier, A., Yang, S.H., 2022. Cross-lingual self-supervised speech representations for improved dysarthric speech recognition. arXiv preprint arXiv:2204.01670.

Hsu, W.-N., Bolte, B., Tsai, Y.-H.H., Lakhotia, K., Salakhutdinov, R., Mohamed, A., 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. IEEE/ACM Trans. Audio Speech Lang. Process. 29, 3451–3460.

Javanmardi, F., Kadiri, S.R., Kodali, M., Alku, P., 2022. Comparing 1-dimensional and 2-dimensional spectral feature representations in voice pathology detection using machine learning and deep learning classifiers. In: Proc. Interspeech. pp. 2173–2177.

Javanmardi, F., Tirronen, S., Kodali, M., Kadiri, S.R., Alku, P., 2023. Wav2vec-based detection and severity level classification of dysarthria from speech. In: IEEE International Conference on Acoustics, Speech and Signal Processing. ICASSP, pp. 1–5.

Joshy, A.A., Rajan, R., 2021. Automated dysarthria severity classification using deep learning frameworks. In: 28th European Signal Processing Conference. EUSIPCO, pp. 116–120.

Joshy, A.A., Rajan, R., 2022. Automated dysarthria severity classification: A study on acoustic features and deep learning techniques. IEEE Trans. Neural Syst. Rehabil. Eng. 30, 1147–1157.

Joshy, A.A., Rajan, R., 2023a. Dysarthria severity assessment using squeeze-and-excitation networks. Biomed. Signal Process. Control 82, 104606.

Joshy, A.A., Rajan, R., 2023b. Dysarthria severity classification using multi-head attention and multi-task learning. Speech Commun. 147, 1–11.

Kadi, K.L., Selouani, S.A., Boudraa, B., Boudraa, M., 2016. Fully automated speaker identification and intelligibility assessment in dysarthria disease using auditory knowledge. Biocybern. Biomed. Eng. 36 (1), 233–247.

Kain, A.B., Hosom, J.-P., Niu, X., van Santen, J.P., Fried-Oken, M., Staehely, J., 2007. Improving the intelligibility of dysarthric speech. Speech Commun. 49 (9), 743–759.

Kent, R.D., Weismer, G., Kent, J.F., Rosenbek, J.C., 1989. Toward phonetic intelligibility testing in dysarthria. J. Speech Hearing Disorders 54 (4), 482–499.

Kim, H., Hasegawa-Johnson, M., Perlman, A., Gunderson, J., Huang, T., Watkin, K., Frame, S., 2008. Dysarthric speech database for universal access research. In: Proc. Interspeech. pp. 1741–1744.

Kim, J., Kumar, N., Tsiartas, A., Li, M., Narayanan, S.S., 2015. Automatic intelligibility classification of sentence-level pathological speech. Comput. Speech Lang. 29 (1), 132–144.

Kursa, M.B., Rudnicki, W.R., 2011. The all relevant feature selection using random forest. arXiv preprint arXiv:1106.5112.

McFee, B., Raffel, C., Liang, D., Ellis, D.P., McVicar, M., Battenberg, E., Nieto, O., 2015. Librosa: Audio and music signal analysis in python. In: Proceedings of the 14th Python in Science Conference, Vol. 8. Citeseer, pp. 18–25.

Menendez-Pidal, X., Polikoff, J.B., Peters, S.M., Leonzio, J.E., Bunnell, H.T., 1996. The nemours database of dysarthric speech. In: Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP'96, IEEE, pp. 1962–1965.

Mohamed, O., Aly, S.A., 2021. Arabic speech emotion recognition employing wav2vec2. 0 and hubert based on BAVED dataset. arXiv preprint arXiv:2110.04425.

Narendra, N., Alku, P., 2018. Dysarthric speech classification using glottal features computed from non-words, words and sentences.. In: Proc. Interspeech. pp. 3403–3407.

Narendra, N., Alku, P., 2019. Dysarthric speech classification from coded telephone speech using glottal features. Speech Commun. 110, 47–55.

Narendra, N., Alku, P., 2020. Automatic intelligibility assessment of dysarthric speech using glottal parameters. Speech Commun. 123, 1–9.

Narendra, N., Alku, P., 2021. Automatic assessment of intelligibility in speakers with dysarthria from coded telephone speech using glottal features. Comput. Speech Lang. 65, 101117.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al., 2019. Pytorch: An imperative style, high-performance deep learning library. Adv. Neural Inf. Process. Syst. 32, 8024–8035.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E., 2011. Scikit-learn: Machine learning in python. J. Mach. Learn. Res. 12, 2825–2830.

Rong, P., Yunusova, Y., Wang, J., Zinman, L., Pattee, G.L., Berry, J.D., Perry, B., Green, J.R., 2016. Predicting speech intelligibility decline in amyotrophic lateral sclerosis based on the deterioration of individual speech subsystems. PLoS One 11 (5), e0154971.

Rudzicz, F., Namasivayam, A.K., Wolff, T., 2012. The TORGO database of acoustic and articulatory speech from speakers with dysarthria. Lang. Resour. Evaluat. 46 (4), 523–541.

Sheikh, S.A., Sahidullah, M., Hirsch, F., Ouni, S., 2022. Introducing ECAPA-TDNN and Wav2Vec2. 0 embeddings to stuttering detection. arXiv preprint arXiv:2204.01564.

Tirronen, S., Javanmardi, F., Kodali, M., Reddy Kadiri, S., Alku, P., 2023a. Utilizing Wav2Vec in database-independent voice disorder detection. In: IEEE International Conference on Acoustics, Speech and Signal Processing. ICASSP, pp. 1–5.

Tirronen, S., Kadiri, S.R., Alku, P., 2023b. Hierarchical multi-class classification of voice disorders using self-supervised models and glottal features. IEEE Open J. Signal Process. 4, 80–88.

Vaessen, N., Van Leeuwen, D.A., 2022. Fine-tuning wav2vec2 for speaker recognition. In: IEEE International Conference on Acoustics, Speech and Signal Processing. ICASSP, pp. 7967–7971.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., et al., 2019. Huggingface's transformers: State-of-the-art natural language processing. arXiv preprint arXiv:1910.03771.

Wrench, A., 1999. The MOCHA-TIMIT articulatory database. URL http://www.cstr.ed.ac.uk/research/projects/artic/mocha.html.

Xue, W., Cucchiarini, C., van Hout, R., Strik, H., 2019. Acoustic correlates of speech intelligibility: the usability of the eGeMAPS feature set for atypical speech. In: Proc. SLaTE 2019: 8th ISCA Workshop on Speech and Language Technology in Education. pp. 48–52.

Yorkston, K.M., Beukelman, D.R., Traynor, C., 1984. Assessment of Intelligibility of Dysarthric Speech. Pro-ed Austin, TX.