# Exploring the Impact of Fine-Tuning the Wav2vec2 Model in Database-Independent Detection of Dysarthric Speech

Farhad Javanmardi, *Student Member, IEEE*, Sudarsana Reddy Kadiri, *Member, IEEE*, and Paavo Alku, *Fellow, IEEE*

***Abstract*—Many acoustic features and machine learning models have been studied to build automatic detection systems to distinguish dysarthric speech from healthy speech. These systems can help to improve the reliability of diagnosis. However, speech recorded for diagnosis in real-life clinical conditions can differ from the training data of the detection system in terms of, for example, recording conditions, speaker identity, and language. These mismatches may lead to a reduction in detection performance in practical applications. In this study, we investigate the use of the wav2vec2 model as a feature extractor together with a support vector machine (SVM) classifier to build automatic detection systems for dysarthric speech. The performance of the wav2vec2 features is evaluated in two cross-database scenarios, language-dependent and language-independent, to study their generalizability to unseen speakers, recording conditions, and languages before and after fine-tuning the wav2vec2 model. The results revealed that the fine-tuned wav2vec2 features showed better generalization in both scenarios and gave an absolute accuracy improvement of 1.46%–8.65% compared to the non-fine-tuned wav2vec2 features.**

***Index Terms*—Dysarthria, fine-tuning, self-supervised learning, wav2vec 2.0.**

## I. Introduction

DYSARTHRIA occurs due to various neurodegenerative conditions and diseases, including stroke, cerebral palsy, Parkinson's disease, and amyotrophic lateral sclerosis. These conditions affect muscle control in organs (the lips, tongue, and throat) involved in speech production [1]. Therefore, individuals with dysarthria often produce speech characterized by abnormalities in phonatory, resonatory, articulatory, and prosodic aspects of speech [2]. Automatic detection of dysarthria from acoustic speech signals has become a widely-studied research topic due to progress in signal processing and machine learning. Automatic detection systems can be used as effective tools to facilitate the clinical diagnosis and treatment of dysarthria. The techniques studied in dysarthric speech detection are mainly based on the popular two-stage architecture consisting of separate feature extraction and classification stages. The detection system is built by training a machine learning algorithm based on supervised learning using a set of collected speech samples and their labels (healthy vs. dysarthric).

Research on automatic, speech-based detection of dysarthria has mainly been conducted by using speech samples from just one database and the popular cross-validation (CV) approach in which the system is trained and tested using samples of the corresponding database. The databases used in the study area have mainly been recorded in controlled laboratory environments using professional equipment. However, when these detection systems are applied in medical diagnosis using realistic test speech samples recorded in clinical environments, the system performance may decrease because several factors (e.g. the level of environment noise, recording equipment, speaker identity, language) can be different between training and inference. In order to enhance the generalization ability of detection systems to data from unseen speakers and recording conditions, a large amount of training data is needed. In the area of dysarthria detection, the amount of available training data is, however, much smaller compared to areas such as speech synthesis and automatic speech recognition (ASR). In order to artificially increase the volume of training data, a widely used approach is to employ data augmentation (DA) [3], [4]. Ideally, the new data generated by DA should enrich the training data by introducing natural variability generated by different speakers and recording conditions [4], [5].

Another approach to enrich the training data is to use self-supervised deep learning models such as the wav2vec2 [6] that have become highly popular in the past few years. Such models have been pre-trained in an unsupervised manner on large speech datasets to be used in automatic speech recognition (ASR) tasks. Several studies have shown that the so-called context embeddings extracted from the transformer layers of the pre-trained wav2vec2 model contain information that is useful for a wide variety of speech-related tasks, including dysarthric speech detection [7], [8], [9], [10], [11], [12]. Therefore, the pre-trained

Manuscript received 7 October 2023; revised 29 February 2024; accepted 20 April 2024. Date of publication 26 April 2024; date of current version 7 August 2024. This work was supported by the Academy of Finland under Grant 330139. *(Corresponding author: Farhad Javanmardi.)*

Farhad Javanmardi and Paavo Alku are with the Department of Information and Communications Engineering, Aalto University, FI-00076 Espoo, Finland (e-mail: farhad.javanmardi@aalto.fi; paavo.alku@aalto.fi).

Sudarsana Reddy Kadiri was with the Department of Information and Communications Engineering, Aalto University, FI-00076 Espoo, Finland. He is now with the Signal Analysis and Interpretation Laboratory (SAIL), University of Southern California, Los Angeles 90089, USA. (e-mail: skadiri@usc.edu).

Digital Object Identifier 10.1109/JBHI.2024.3392829

models can serve as powerful feature extractors in detection systems based on the two-stage pipeline architecture [11], [13]. The main advantage of pre-trained models is that they can be easily fine-tuned using small amounts of labeled data to achieve state-of-art results in the required task [14], [15], [16], [17]. When the wav2vec2 model is fine-tuned on a specific task, the model is capable of using its knowledge of general characteristics of speech that it has learned by seeing a large amount of speech data in the pre-training phase. This learned knowledge allows the wav2vec model to adapt to the nuances of a new task. In fact, the pre-trained wav2vec2 model simply refines its pre-learned representations to align with the unique characteristics of the new task, without the need to learn these representations again. This capability dramatically reduces the necessity to use large volumes of labeled data for fine-tuning and enables building better classification systems with small data. This adaptability and efficiency of the pre-trained wav2vec2 model in learning from small amount of labeled data is a considerable advantage in fields such as pathological speech detection where most datasets are relatively small. In addition, the pre-trained model is beneficial for the task in which out-of-domain data is used [18], because pre-training the models on extensive datasets inherently develops robustness with respect to variations and noise. This robustness translates into a strong generalization ability when applied to out-of-domain data.

Previous studies on dysarthria detection have mainly focused on designing handcrafted acoustic features that characterize phonation, articulation, and prosodic aspects of speech production [19], [20], [21], [22], [23], [24], [25], [26]. In [27], [28], the single frequency filtering-based features were investigated for dysarthric speech detection. In [29], automatic feature extraction from raw speech waveforms was studied using a fully-learnable audio frontend. Due to success of deep learning (DL) in several speech processing tasks, many studies have recently explored various DL-based techniques in the detection of dysarthric speech [30], [31], [32], [33], [34], [35], [36]. DL-based techniques for dysarthria detection include mapping from handcrafted acoustic features to output labels, as well as modern end-to-end systems in which the raw speech signal or time—frequency spectrogram is directly used by a DL model to compute the output labels. However, even though the modern end-to-end DL techniques have shown significant progress, they can still be criticized for the following two major issues: 1) large amounts of speech data are needed in the system training [37], and 2) interpretability of results provided by DL approaches is difficult and therefore the clinical relevance of the technology might be questioned by specialists [38].

While the aforementioned studies have demonstrated promising performance in dysarthric speech detection through both the development of handcrafted acoustic features and the application of advanced DL techniques, the necessity for further verification in cross-database scenarios (i.e., training and testing the system using different databases) remains an important research topic. Studying cross-database scenarios enables assessing the effect of different mismatches (e.g., recording environment, equipment, and language) between the system training and testing. In the field of pathological speech detection, several studies have investigated cross-database scenarios in one language or across different languages for voice disorder detection [10], [39], [40], Parkinson's disease detection [41], [42], and dementia detection [43]. To the best of our knowledge, however, only three studies have investigated cross-database scenarios in dysarthric speech detection [44], [45], [46]. In [44], the usage of spectral and prosodic features was studied in cross-database experiments. The experiments of [45] explored a specific disease (hypokinetic dysarthria for Parkinson's disease) in cross-language experiments using mel frequency cepstral coefficients (MFCCs) along with prosodic features. In [46], domain-adversarial training and mutual information minimization were proposed to extract domain-invariant biomarker embeddings from acoustic features (e.g., mel-spectrogram) in cross-database dysarthric speech detection. Importantly, pre-trained models have not been studied before as feature extractors in cross-database scenarios in detection of dysarthria.

In this paper, we investigate the effectiveness of fine-tuning the wav2vec2 model as a feature extractor in dysarthric speech detection. More specifically, the features extracted using the fine-tuned wav2vec2 model are compared in two scenarios where training and testing of the systems were first conducted with two different English databases, and then training and testing of the systems were carried out using two databases representing two different languages (English and Italian). In these experiments, we used wav2vec2-BASE [6] as an English-based wav2vec2 model and wav2vec2-XLSR [47] as a multilingual wav2vec2 model.

The main contributions of this study are:

- Conducting a layer-by-layer comparison between features extracted by the fine-tuned wav2vec2 model and by the non-fine-tuned wav2vec2 model in the detection of dysarthria (healthy vs. dysarthric).
- Studying the detection of dysarthria in a language-dependent and language-independent scenario using a wav2vec2 model trained in English (wav2vec2-BASE) and a multilingual wav2vec2 model (wav2vec2-XLSR).
- Presenting new results on speech-based biomarking of dysarthria showing that the fine-tuned wav2vec2 features improved the performance in the detection of the disease.

## II. Databases

The current study uses three publicly available dysarthria databases. Two of the databases include speech spoken in English, and one of the databases includes speech spoken in Italian. The two English databases are the Universal Access Speech (UA-Speech) database [19] and the TORGO database [48]. The Italian database used is EasyCall [49].

### A. UA-Speech

This database comprises 15 dysarthric speakers (4 females, 11 males) with cerebral palsy and 13 healthy controls (4 females, 9 males), aged between 18 and 58 years [19]. Each speaker produced 765 isolated words in three blocks, each containing 255 words. Among these, 155 words are common to all blocks, encompassing 19 computer commands, 26 radio alphabet letters, 10 digits, and the 100 most common English words. The remaining 100 words in each block were chosen from Project Gutenberg novels. Speech was recorded using an eight-microphone array using a sampling frequency of 16 kHz, with microphones spaced 1.5 inches apart. This study utilized speech utterances from all three blocks, captured by microphone number six.

## B. TORGO

This database comprises recordings from 8 patients (3 females, 5 males) diagnosed with cerebral palsy or amyotrophic lateral sclerosis (ALS) and 7 healthy control speakers (3 females, 4 males), aged between 16 and 50 years. TORGO includes three categories of speech signals: non-words, words, and sentences. The non-words category features repetitions of */iy – p – ah/*, */ah – p – iy/*, */p – ah – t – ah – k – ah/*, and vowels at high and low pitches, each lasting 5 seconds. The words category contains 50 words from the Frenchay Dysarthria Assessment [50] and 360 words from the Yorkston-Beukelman Assessment of Intelligibility of Dysarthric Speech. The sentences are drawn from various sources, totaling 162 sentences from the Yorkston-Beukelman Assessment, 460 sentences from the MOCHA database, the Grandfather passage from the Nemours database [51], and spontaneously elicited descriptive texts. Recordings were made using a head-mounted microphone and an array microphone, sampled at 16 kHz. This study used all three categories of speech signals (i.e., non-words, words and sentences) from the array microphone [48].

## C. EasyCall

The EasyCall database encompasses recordings from a total of 24 healthy speakers (10 females and 14 males) and 31 dysarthric speakers (11 females and 20 males) [49]. Various underlying conditions contributing to dysarthria, such as Parkinson's disease, Huntington's disease, amyotrophic lateral sclerosis, peripheral neuropathy, and myopathic or myasthenic lesions, are represented by the database. The degree of speech impairment of the dysarthric speakers was assessed by neurologists through the therapy outcome measure (score ranges from 1–5 corresponding to mild, mild-moderate, moderate, moderate-severe, and severe). The data consists of 37 commands, encompassing words and sentences pertinent to the specific task at hand, as well as 30 non-command utterances. Each participant performed between 2 to 8 sessions and in each recording session, the speaker repeated one utterance. Consequently, the dataset comprises 21,386 recordings, with 10,077 from healthy speakers and 11,309 from dysarthric speakers. The speech was recorded using a sampling frequency of 8 kHz. In this study, speech utterances including words and sentences from all recording sessions were used. Fig. 3 shows the details of the UA-Speech, TORGO and EasyCall databases.

# III. Detection System

This study explores the effect of fine-tuning the wav2vec2 model as a feature extractor in binary classification problems to distinguish dysarthric speech from healthy speech automatically (i.e., a detection problem) in the following two scenarios: (1) in a cross-database language-dependent scenario where the training and testing of the detection system are carried out using two different dysarthric databases, sharing the same language (i.e., English), and (2) in a cross-database language-independent scenario where the detection experiments are conducted using two different dysarthric databases with different languages (i.e., English and Italian). Fig. 1 shows the systems built using the popular two-stage pipeline approach (consisting of a feature extraction stage and a classifier stage) for the two scenarios mentioned above. In the feature extraction stage, the feature vectors are derived from raw speech waveforms using two popular pre-trained models (wav2vec2-BASE [6] and wav2vec2-XLSR [47]) that were fine-tuned using the dysarthric databases as described in Section III-B). The classifier stage uses a support vector machine (SVM) to predict the output labels (healthy vs. dysarthric). The feature extraction and classifier are explained in the following sub-sections.

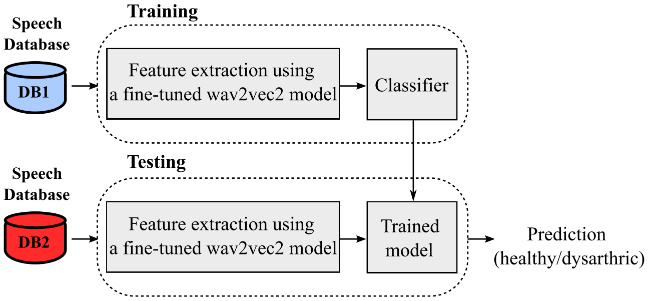
Training
Speech Database
DB1
Feature extraction using a fine-tuned wav2vec2 model
Classifier
Testing
Speech Database
DB2
Feature extraction using a fine-tuned wav2vec2 model
Trained model
Prediction (healthy/dysarthric)

Fig. 1. Proposed system for database-independent detection of dysarthric speech using features derived from the fine-tuned wav2vec2 model and using SVM as a classifier.

## A. Pre-Training of the wav2vec2 Model

The selected pre-trained wav2vec2 models used as feature extractors for the detection problem are the English-based wav2vec2-BASE model and the multilingual wav2vec2-XLSR model. In our initial study [11], we found that the features extracted from the starting layers of the wav2vec2-BASE model, which was pre-trained on 960 hours of speech from the English Librispeech corpus [6], showed a better capability to distinguish between healthy and dysarthric speech. Therefore, in the present investigation, we decided to study the fine-tuning of the wav2vec2-BASE model for cross-database scenarios involving the English language. For cross-database scenarios involving different languages, we considered the use of the wav2vec2-XLSR model, which was originally pre-trained on a combination of three ASR databases, encompassing 56,000 hours of speech representing 53 different languages. The wav2vec2 model uses a multi-layer convolutional feature encoder, a context network, and a quantization module. The context network contains 12 transformer blocks with a model dimension of 768 for the wav2vec2-BASE model and 12 transformer blocks with a model dimension of 1024 for the wav2vec2-XLSR model.

TABLE I
DETAILS OF THE DATABASES USED IN THE CURRENT STUDY

| Database | Language | Speaking Task | Amount of data (≈) | Speaker Class | No. of Speakers | | |
|---|---|---|---|---|---|---|---|
| | | | | | Male | Female | Total |
| UA-Speech | English | Words | 15 hours | Healthy | 9 | 4 | 13 |
| | | | | Dysarthric | 11 | 4 | 15 |
| TORGO | English | Non-words, Words, and Sentences | 4 hours | Healthy | 4 | 3 | 7 |
| | | | | Dysarthric | 5 | 3 | 8 |
| EasyCall | Italian | Words and Sentences | 13 hours | Healthy | 14 | 10 | 24 |
| | | | | Dysarthric | 20 | 11 | 31 |

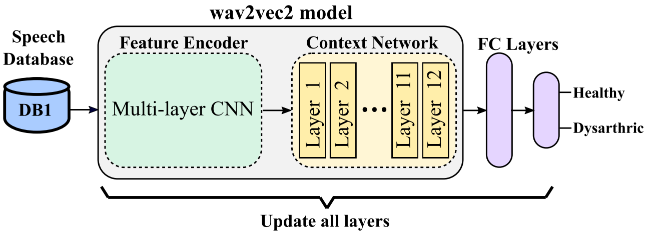

Fig. 2. Overview of the fine-tuning of the wav2vec2 model.

## B. Fine-Tuning of the wav2vec2 Model

Fig. 2 shows the fine-tuning process for the wav2vec2 model. The feature encoder and context network of wav2vec2 are fine-tuned on the labeled healthy and dysarthric speech data by adding two new fully-connected layers, where the second fully-connected layer performs the binary detection (i.e., classifying the input as healthy speech or dysarthric speech). The two added fully-connected layers are randomly initialized and the wav2vec2 models (wav2vec2-BASE and wav2vec2-XLSR) are initialized with the original models released in [6], [47]. The models are optimized using the cross-entropy loss function. As hyper-parameters used for fine-tuning the wav2vec2 models, a batch size of 8 and the Adam optimizer with a learning rate of 3e-5 are used. Because the wav2vec2 models were originally trained with speech spoken by a large number of healthy speakers, the fine-tuning of the wav2vec2 models was conducted using all the dysarthric samples and 20% of the healthy samples of the speech database. It should be noted that the three speech databases (UA-Speech, TORGO and EasyCall) were individually used to fine-tune the wav2vec2 models. This process resulted in 6 fine-tuned models (3 wav2vec2-BASE and 3 wav2vec2-XLSR models that were fine-tuned using the UA-Speech, TORGO and EasyCall databases).

## C. Feature Extraction Using the Fine-Tuned wav2vec2 Models

After fine-tuning the wav2vec2 models, the outputs from each of the transformer layers of the context network are utilized as features in dysarthric speech detection. More specifically, thirteen 768-dimensional feature vectors (i.e., the temporal average of the inputs to the first transformer layer and the outputs of all 12 transformer layers) are derived for each speech signal using the fine-tuned wav2vec2-BASE model. For the fine-tuned wav2vec2-XLSR model, the context network contains 24 transformer blocks with a model dimension of 1024. Therefore, twenty-five 1024-dimensional feature vectors are extracted for each speech signal. The first feature vector is the temporal average of the inputs to the first transformer layer and the remaining feature vectors are the outputs of all 24 transformer layers.

In the following sections, we use the term "FT-wav2vec2-BASE features" when referring to the feature vectors extracted using the fine-tuned wav2vec2-BASE model. Similarly, we use the term "NO-FT-wav2vec2-BASE features" when referring to the features extracted by the non-fine-tuned wav2vec2-BASE model. Likewise, we denote the feature vectors derived from the fine-tuned wav2vec2-XLSR model as the "FT-wav2vec2-XLSR features," and the features from the non-fine-tuned wav2vec2-XLSR model as the "NO-FT-wav2vec2-XLSR features." Additionally, when referring to the features from the N-th transformer layer, we use the "FT-wav2vec2-BASE-**N**" notation.

## D. Classifiers and Evaluation

In order to distinguish between healthy and dysarthric samples, the SVM classifier was chosen in the current study as it is a very popular ML classifier in the detection and classification of speech disorders [10], [11], [24]. The parameters used for the SVM classifiers are as follows: radial basis function as kernel, a regularization parameter value of 1, and the following scaling $\gamma = 1/(D \cdot Var(X))$ as gamma parameter, where $D$ is the dimensionality of the feature vectors and $Var(X)$ is the variance of the training data. Balanced detection accuracy (ACC) serves as the primary metric for evaluation, and the results reported in Section IV are discussed based on this metric. In addition, four other evaluation metrics (sensitivity (SE), specificity (SP), F1-score (F1), and equal error rate (EER)) are reported in order to get a comprehensive overview of the results.

## E. Experiments

The goal of the current study is to assess whether the features extracted from the fine-tuned wav2vec2 models can reduce the training-testing mismatch caused by different recording conditions, recording equipment, and language in the detection of dysarthria. Therefore, the evaluation of the FT-wav2vec2 features was conducted in the following two scenarios: (1) the cross-database language-dependent scenario in which English databases were used for training (UA-Speech) and testing (TORGO) and vice versa and (2) the cross-database language-independent scenario in which English databases were used for training (UA-Speech/TORGO) and the Italian database (EasyCall) was used for testing and vice versa. These two evaluation scenarios are described in more detail below.

*1) The Cross-Database Language-Dependent Scenario:* The evaluation of the wav2vec2 features was conducted in three parts consisting of a single-database evaluation (using the NO-FT-wav2vec2 features) and a cross-database evaluation (using the NO-FT-wav2vec2 and FT-wav2vec2 features). In the single-database evaluation, only the data from an individual database was used to train and evaluate the SVM classifiers (with the leave-one-speaker-out (LOSO) cross-validation strategy). This process was repeated for UA-Speech, TORGO and EasyCall. The single-database evaluation enables us to observe the performance of the NO-FT-wav2vec2 features in a scenario with the same recording environment, equipment, and language in the training and testing phases, where the speaker identity is the only difference between the two phases. Hence, the single-database evaluation assesses the ability of the NO-FT-wav2vec2 features to generalize to unseen speakers. It can also be considered as baseline for comparison with the cross-database scenarios using the NO-FT-wav2vec2 and FT-wav2vec2 features.

The cross-database evaluation was first conducted using the NO-FT-wav2vec2 features and then using the FT-wav2vec2 features. In the cross-database evaluation, the training and testing were carried out using data from different databases. The detection systems were first trained using the UA-Speech data and then evaluated using the TORGO data. This process was repeated using the TORGO samples as training data and using the UA-Speech samples as testing data. The cross-database scenario enables studying the generalizability of the NO-FT-wav2vec2 features in conditions when there is a mismatch due to recording environment and equipment between training and testing but no mismatch due to language. In addition, the sensitivity of the NO-FT-wav2vec2 features to the mismatch between training and testing can be observed when compared with the single-database scenario.

In the cross-database evaluation using the FT-wav2vec2 features, we first trained the detection systems using the FT-wav2vec2 features extracted from the UA-Speech data (i.e., the same data used for fine-tuning the wav2vec2 models). Then we evaluated it using the FT-wav2vec2 features extracted from the TORGO data. Similarly, the system was trained using the TORGO data (i.e., the same data used for fine-tuning the wav2vec2 model) and then evaluated using the UA-Speech data. This scenario enables us to assess the effectiveness of features extracted using the fine-tuned models. Furthermore, comparing this scenario to the one where the NO-FT-wav2vec2 features were used enables investigating whether fine-tuning can reduce the training-testing mismatch caused by environmental factors.

*2) The Cross-Database Language-Independent Scenario:* In this scenario, the evaluation of the wav2vec2 features was conducted in a similar manner as explained in Section III-E1 except for the cross-database evaluation, in which the training and testing were carried out by using data from different databases with different languages. We first trained the detection systems using English speech samples of UA-Speech and then evaluated them using Italian speech samples of EasyCall, and then trained the systems using Italian samples of EasyCall and evaluated them using English samples of UA-Speech. This process was repeated using TORGO (for English samples) and EasyCall (for Italian samples). The cross-database evaluation allows us to study the generalizability of the FT-wav2vec2 features in a scenario with different recording environments, equipment, and language between training and testing. In addition, the sensitivity of the FT-wav2vec2 features to the mismatch between training and testing caused by language can be assessed when compared with the NO-FT-wav2vec2 features. Table II gives a brief summary of all experiments conducted in this study.

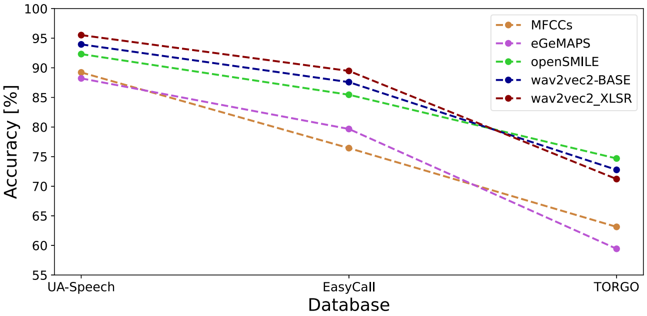

Fig. 3. Detection accuracy for the three baseline features (MFCCs, openSMILE, and eGeMAPS) and for the best-performing wav2vec2-BASE and wav2vec2-XLSR features for the UA-Speech, EasyCall, and TORGO databases.

## IV. Results

This section reports the results obtained using the features derived from the non-fine-tuned and fine-tuned wav2vec2-BASE and wav2vec2-XLSR models. First, the results of the cross-database language-dependent experiments are presented in Section IV-A. Then the results of the cross-database language-independent experiments are presented in Section IV-B. Before reporting the main results of the current study, we briefly present the results of the experiments where the features obtained from the wav2vec2-BASE and wav2vec2-XLSR models were compared with three baseline features that represent conventional widely-used acoustical features (MFCCs [11], openSMILE [52], and eGeMAPS [53]). In these experiments, the system training and testing was based on the same database.

Fig. 3 shows the detection accuracies for the baseline features and for the best-performing wav2vec2-BASE and wav2vec2-XLSR features (see Figs. 4, 6, and 7) for the UA-Speech, EasyCall, and TORGO databases. It can be seen that the wav2vec2-BASE and wav2vec2-XLSR features outperformed all three baseline features for all databases (except for the TORGO database, in which openSMILE showed a slightly higher detection accuracy compared to the two wav2vec2 features). As the wav2vec2 features performed better than the other features,

TABLE II
SUMMARY OF ALL EXPERIMENTS CONDUCTED IN THIS STUDY FOR THE LANGUAGE-DEPENDENT AND LANGUAGE-INDEPENDENT SCENARIOS

| **Language-dependent scenario** | | | | |
|---|---|---|---|---|
| **Experiment** | **Train data** | **Test data** | **Feature** | **Evaluation strategy** |
| 1. Single-database | UA-Speech | UA-Speech | NO-FT-wav2vec2 (no Fine-tuning) | LOSO cross-validation |
| 2. Single-database | TORGO | TORGO | NO-FT-wav2vec2 (no Fine-tuning) | LOSO cross-validation |
| 3. Single-database | EasyCall | EasyCall | NO-FT-wav2vec2 (no Fine-tuning) | LOSO cross-validation |
| 4. Cross-database | UA-Speech | TORGO | NO-FT-wav2vec2 (no Fine-tuning) | Whole test data |
| 5. Cross-database | TORGO | UA-Speech | NO-FT-wav2vec2 (no Fine-tuning) | Whole test data |
| 6. Cross-database | UA-Speech | TORGO | FT-wav2vec2 (Fine-tuned with train data) | Whole test data |
| 7. Cross-database | TORGO | UA-Speech | FT-wav2vec2 (Fine-tuned with train data) | Whole test data |
| **Language-independent scenario** | | | | |
| 8. Cross-database | UA-Speech | EasyCall | NO-FT-wav2vec2 (no Fine-tuning) | Whole test data |
| 9. Cross-database | EasyCall | UA-Speech | NO-FT-wav2vec2 (no Fine-tuning) | Whole test data |
| 10. Cross-database | TORGO | EasyCall | NO-FT-wav2vec2 (no Fine-tuning) | Whole test data |
| 11. Cross-database | EasyCall | TORGO | NO-FT-wav2vec2 (no Fine-tuning) | Whole test data |
| 12. Cross-database | UA-Speech | EasyCall | FT-wav2vec2 (Fine-tuned with train data) | Whole test data |
| 13. Cross-database | EasyCall | UA-Speech | FT-wav2vec2 (Fine-tuned with train data) | Whole test data |
| 14. Cross-database | TORGO | EasyCall | FT-wav2vec2 (Fine-tuned with train data) | Whole test data |
| 15. Cross-database | EasyCall | TORGO | FT-wav2vec2 (Fine-tuned with train data) | Whole test data |

The term “NO-FT-wav2vec2” refers to the feature vectors extracted using the non-fine-tuned wav2vec2 models. Similarly, “FT-wav2vec2” refers to the feature vectors extracted using the fine-tuned wav2vec2 models. LOSO refers to leave-one-speaker-out cross-validation strategy.

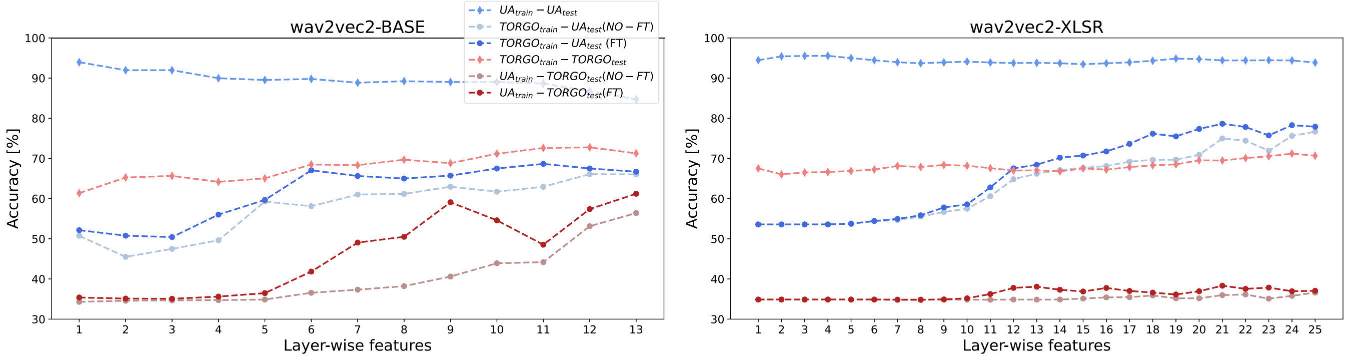

Fig. 4. Detection accuracy for the wav2vec2-BASE features (left panel) and for the wav2vec2-XLSR features (right panel). The following conditions are shown: (1) The single-database scenario (training and testing with the same data, i.e., $UA_{train}$–$UA_{test}$ and $TORGO_{train}$–$TORGO_{test}$), (2) the cross-database scenario with two English databases (training with one data and testing with another data). In (2), testing was done using both non-fine-tuned wav2vec2 features (NO-FT) and fine-tuned wav2vec2 features (FT).

only the wav2vec2 features are considered in the remainder of this study.

## A. Detection Results in the Cross-Database Language-Dependent Scenarios

Fig. 4 displays the detection accuracies for the features derived from the wav2vec2 models (wav2vec2-BASE and wav2vec2-XLSR) in the single-database scenario, and also for the features derived from the non-fine-tuned and fine-tuned models in the cross-database scenarios using two English databases (UA-Speech and TORGO). The results in Fig. 4 show for both models (wav2vec2-BASE and wav2vec2-XLSR) that in the single-database scenarios (training and testing with the same data, i.e., $UA_{train}$–$UA_{test}$ and $TORGO_{train}$–$TORGO_{test}$), all the layer-wise features performed better than in the cross-database scenarios (i.e., training with one dataset and testing with another dataset, e.g., $TORGO_{train}$–$UA_{test}$ and vice versa), regardless of whether the models were fine-tuned (FT) or not (No-FT). In the cross-database scenarios, it is evident that the fine-tuned (FT) model features outperformed the non-fine-tuned (No-FT) model features.

The performance of the best-performing features in the single-database and cross-database scenarios for the UA-speech and TORGO databases is given in Table III. From the results of the wav2vec2-BASE model (left side of the table), it can be observed that when the system was trained with TORGO and tested with

TABLE III

Performance Metrics of the Single-Database and Cross-Database Scenarios for the Best Performing Features (wav2vec2-BASE (Left Side of the Table) and wav2vec2-XLSR (Right Side of the Table)) Using the UA-Speech and TORGO Databases

| wav2vec2-BASE model | | | | | |
|---|---|---|---|---|---|
| **Feature** | **ACC** | **SE** | **SP** | **F1** | **ERR** |
| **UA-Speech (trained also on UA-Speech)** | | | | | |
| wav2vec2-BASE-1 | 93.96 | 0.93 | 0.95 | 0.94 | 0.059 |
| **UA-Speech (trained on TORGO)** | | | | | |
| NO-FT-wav2vec2-BASE-12 | 66.09 | 0.61 | 0.72 | 0.66 | 0.341 |
| FT-wav2vec2-BASE-11 | 68.65 | 0.66 | 0.72 | 0.69 | 0.311 |
| **TORGO (trained also on TORGO)** | | | | | |
| wav2vec2-BASE-12 | 72.77 | 0.61 | 0.84 | 0.64 | 0.385 |
| **TORGO (trained on UA-Speech)** | | | | | |
| NO-FT-wav2vec2-BASE-13 | 56.40 | 0.62 | 0.50 | 0.49 | 0.405 |
| FT-wav2vec2-BASE-13 | 61.20 | 0.67 | 0.55 | 0.54 | 0.367 |

| wav2vec2-XLSR model | | | | | |
|---|---|---|---|---|---|
| **Feature** | **ACC** | **SE** | **SP** | **F1** | **ERR** |
| **UA-Speech (trained also on UA-Speech)** | | | | | |
| wav2vec2-XLSR-3 | 95.53 | 0.93 | 0.98 | 0.96 | 0.044 |
| **UA-Speech (trained on TORGO)** | | | | | |
| NO-FT-wav2vec2-XLSR-25 | 76.67 | 0.81 | 0.72 | 0.79 | 0.226 |
| FT-wav2vec2-XLSR-21 | 78.67 | 0.85 | 0.71 | 0.81 | 0.201 |
| **TORGO (trained also on TORGO)** | | | | | |
| wav2vec2-XLSR-24 | 71.21 | 0.59 | 0.83 | 0.62 | 0.383 |
| **TORGO (trained on UA-Speech)** | | | | | |
| NO-FT-wav2vec2-XLSR-25 | 36.61 | 0.62 | 0.11 | 0.37 | 0.422 |
| FT-wav2vec2-XLSR-13 | 38.07 | 0.58 | 0.18 | 0.37 | 0.396 |

The number at the end of each feature refers to the number of the corresponding layer of the pre-trained model. ACC, SE, SP, F1, and EER refer to accuracy, sensitivity, specificity, F1-score, and equal error rate, respectively.

UA-Speech with no fine-tuning (NO-FT-wav2vec2-BASE-12), the performance dropped drastically in comparison to when the system was both trained and tested using UA-Speech, decreasing from 93.96% to 66.09% (i.e., an absolute drop of 27.87%). Similarly, when the system was trained using UA-Speech and tested using TORGO, the performance dropped in comparison to when it was trained and tested using TORGO, decreasing from 72.77% to 56.40% (i.e., an absolute drop of 16.37%). When the models were fine-tuned (FT-wav2vec2-BASE), the performance improved in both scenarios. The absolute accuracy improvement was 2.5% (from 66.09% to 68.65%) for UA-Speech and 4.8% (from 56.40% to 61.20%) for TORGO.

From the results of the wav2vec2-XLSR model (right side of the table), it can be observed that when the system was trained using TORGO and tested using UA-Speech with no fine-tuning (NO-FT-wav2vec2-XLSR-25), the performance dropped in comparison to when it was trained and tested using UA-Speech, decreasing from 95.53% to 76.67% (i.e., an absolute drop of 18.86%). Similarly, when the system was trained using UA-Speech and tested using TORGO, the performance dropped drastically in comparison to when it was trained and tested using the TORGO samples alone, decreasing from 71.21% to 36.61% (i.e., an absolute drop of 34.6%). When the models were fine-tuned (FT-wav2vec2-XLSR), the performance improved in both scenarios. The absolute accuracy improvement was 2% (from 76.67% to 78.67%) for UA-Speech and 1.46% (from 36.61% to 38.07%) for TORGO. Between the two models (wav2vec2-BASE and wav2vec2-XLSR), wav2vec2-XLSR performed better, especially when the system was trained using TORGO and tested using UA-Speech. However, further investigation is required to understand the poorer performance found for wav2vec2-XLSR when the system was trained with UA-Speech and tested with TORGO.

Confusion matrices are shown in Fig. 5 for the best performing non-fine-tuned and fine-tuned wav2vec2-BASE features for the system trained using UA-Speech and tested using TORGO. It can be seen that there are less confusions between healthy and dysarthric speech for the fine-tuned wav2vec2-BASE feature (FT-wav2vec2-BASE-13) compared to the non-fine-tuned wav2vec2-BASE feature (NO-FT-wav2vec2-base-13).

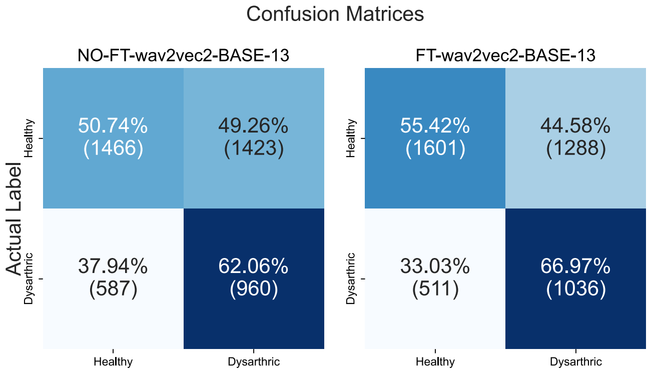

Fig. 5. Confusion matrices of dysarthria detection in the cross-database language-dependent scenario for NO-FT-wav2vec2-BASE-13 (the best performing non-fine-tuned wav2vec2-BASE feature) and for FT-wav2vec2-BASE-13 (the best performing fine-tuned wav2vec2-BASE feature).

### B. Detection Results in the Cross-Database Language-Independent Scenarios

Fig. 6 displays the detection accuracies for the features derived from the wav2vec2 models (wav2vec2-BASE and wav2vec2-XLSR) in the single-database scenario, and also for the features derived from the non-fine-tuned and fine-tuned models in the cross-database scenarios using one English database (UA-Speech) and one Italian database (EasyCall). Other evaluation metrics are provided in Table IV for the best-performing features from both scenarios (single-database and cross-database) for the UA-Speech and EasyCall databases.

From the results reported in Table IV for the wav2vec2-BASE model (left side of the table), it can be observed that when the system was trained using EasyCall and tested with UA-Speech with no fine-tuning (NO-FT-wav2vec2-BASE-4), the performance dropped from 93.96% to 70.52% in comparison to when it was trained and tested with UA-Speech (i.e., an absolute drop of 23.44%). Similarly, when the system was trained with UA-Speech and tested with EasyCall, the performance dropped in comparison to when it was trained and tested with EasyCall, decreasing from 87.58% to 68.46% (i.e., an absolute drop of

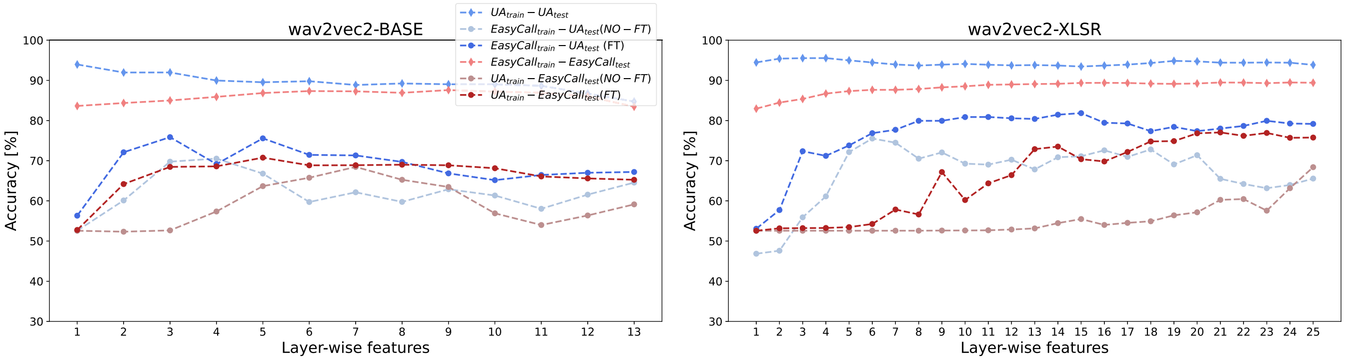

Fig. 6. Detection accuracy for the wav2vec2-BASE features (left panel) and for the wav2vec2-XLSR features (right panel). The following conditions are shown: (1) The single-database scenario (training and testing with the same data, i.e., $UA_{train}$–$UA_{test}$ and $EasyCall_{train}$–$EasyCall_{test}$), (2) the cross-database scenario using an English database and an Italian database (training with one data and testing with another data). In (2), testing was done using both non-fine-tuned wav2vec2 features (NO-FT) and fine-tuned wav2vec2 features (FT).

TABLE IV

Performance Metrics of the Single-Database and Cross-Database Scenarios for the Best Performing Features (wav2vec2-BASE (Left Side of the Table) and wav2vec2-XLSR (Right Side of the Table)) Using the UA-Speech and EasyCall Databases

| wav2vec2-BASE model | | | | | |
|---|---|---|---|---|---|
| **Feature** | **ACC** | **SE** | **SP** | **F1** | **ERR** |
| UA-Speech (trained also on UA-Speech) | | | | | |
| wav2vec2-BASE-1 | 93.96 | 0.93 | 0.95 | 0.94 | 0.059 |
| UA-Speech (trained on EasyCall) | | | | | |
| NO-FT-wav2vec2-BASE-4 | 70.52 | 0.82 | 0.57 | 0.75 | 0.287 |
| FT-wav2vec2-BASE-3 | 75.87 | 0.76 | 0.76 | 0.77 | 0.242 |
| EasyCall (trained also on EasyCall) | | | | | |
| wav2vec2-BASE-9 | 87.58 | 0.87 | 0.88 | 0.88 | 0.130 |
| EasyCall (trained on UA-Speech) | | | | | |
| NO-FT-wav2vec2-BASE-7 | 68.46 | 0.83 | 0.54 | 0.74 | 0.284 |
| FT-wav2vec2-BASE-5 | 70.76 | 0.69 | 0.72 | 0.71 | 0.287 |

| wav2vec2-XLSR model | | | | | |
|---|---|---|---|---|---|
| **Feature** | **ACC** | **SE** | **SP** | **F1** | **ERR** |
| UA-Speech (trained also on UA-Speech) | | | | | |
| wav2vec2-XLSR-3 | 95.53 | 0.93 | 0.98 | 0.96 | 0.044 |
| UA-Speech (trained on EasyCall) | | | | | |
| NO-FT-wav2vec2-XLSR-6 | 75.53 | 0.58 | 0.96 | 0.72 | 0.232 |
| FT-wav2vec2-XLSR-15 | 81.84 | 0.73 | 0.92 | 0.81 | 0.181 |
| EasyCall (trained also on EasyCall) | | | | | |
| wav2vec2-XLSR-24 | 89.48 | 0.90 | 0.89 | 0.90 | 0.113 |
| EasyCall (trained on UA-Speech) | | | | | |
| NO-FT-wav2vec2-XLSR-25 | 68.39 | 0.79 | 0.58 | 0.73 | 0.247 |
| FT-wav2vec2-XLSR-21 | 77.04 | 0.84 | 0.70 | 0.79 | 0.227 |

The number at the end of each feature refers to the number of the corresponding layer of the pre-trained model. ACC, SE, SP, F1, and EER refer to accuracy, sensitivity, specificity, F1-score, and equal error rate, respectively.

19.12%). When the models were fine-tuned (FT-wav2vec2-BASE), the performance improved in both scenarios. An absolute improvement of 5.35% and 2.3% in accuracy was obtained for UA-Speech and EasyCall, respectively.

From the results of the wav2vec2-XLSR model (right side of the table), it can be observed that when the system was trained with EasyCall and tested with UA-Speech with no fine-tuning (NO-FT-wav2vec2-XLSR-6), the performance dropped in comparison to when it was trained and tested with UA-Speech, decreasing from 95.53% to 75.53% (i.e., a drop of 20%). Similarly, when the system was trained with UA-Speech and tested with EasyCall, the performance dropped drastically in comparison to when it was trained and tested with EasyCall, decreasing from 89.48% to 68.39% (i.e., a drop of 21.09%). When the models were fine-tuned (FT-wav2vec2-XLSR), the performance improved in both scenarios. The absolute accuracy improvement was 6.31% for UA-Speech and 8.65% for EasyCall. Between the two models (wav2vec2-BASE and wav2vec2-XLSR), wav2vec2-XLSR performed better, especially when the system was trained with EasyCall speech and tested with UA-Speech.

The detection accuracies using the TORGO and EasyCall databases for the features derived from the wav2vec2 models (wav2vec2-BASE and wav2vec2-XLSR) in the single-database scenario and also for the non-fine-tuned and fine-tuned wav2vec2 features in the cross-database scenarios are shown in Fig. 7. Moreover, the performances of the best-performing features from both the single-database and cross-database scenarios are reported in Table V. From the results in Table V, it can be seen that fine-tuning the model improved the detection performance in terms of accuracy compared to the non-fine-tuned model, and absolute improvements of 3.18% for TORGO and 3.81% for EasyCall were achieved using the wav2vec2-BASE model. For the wav2vec2-XLSR model, this improvement was 4.02% for TORGO and 3.95% for EasyCall. Confusion matrices are shown in Fig. 8 for the best performing non-fine-tuned and fine-tuned wav2vec2-BASE features for the system trained using UA-Speech and tested using EasyCall. It can be seen that there are less confusions between healthy and dysarthric speech for the fine-tuned wav2vec2-XLSR feature (FT-wav2vec2-XLSR-21) compared to the non-fine-tuned wav2vec2-XLSR feature (NO-FT-wav2vec2-XLSR-25).

Finally, the authors would like to point out that in addition to the detection experiments that were reported above in Sections IV-A and IV-B, preliminary experiments were conducted to find out how the dysarthria detection using the wav2vec2 features is affected when a system trained with clean speech and tested with noisy speech (i.e., another type of mismatch). For this case, we used one type of additive noise (babble, SNR of 5 dB)

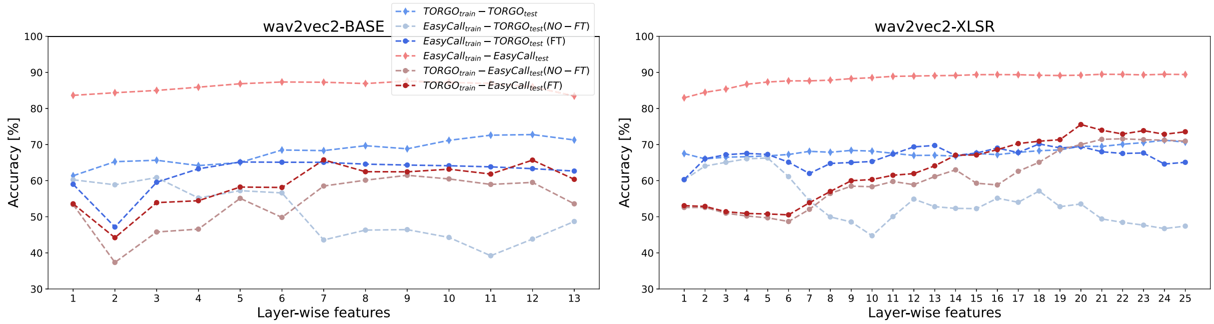

Fig. 7. Detection accuracy for the wav2vec2-BASE features (left panel) and for the wav2vec2-XLSR features (right panel). The following conditions are shown: (1) The single-database scenario (training and testing with the same data, i.e., $TORGO_{train}$–$TORGO_{test}$ and $EasyCall_{train}$–$EasyCall_{test}$), (2) the cross-database scenario using an English database and an Italian database (training with one data and testing with another data). In (2), testing was done using both non-fine-tuned wav2vec2 features (NO-FT) and fine-tuned wav2vec2 features (FT).

TABLE V
PERFORMANCE METRICS OF THE SINGLE-DATABASE AND CROSS-DATABASE SCENARIOS FOR THE BEST PERFORMING FEATURES (WAV2VEC2-BASE (LEFT SIDE OF THE TABLE) AND WAV2VEC2-XLSR (RIGHT SIDE OF THE TABLE)) USING THE TORGO AND EASYCALL DATABASES

| wav2vec2-BASE model | | | | | |
|---|---|---|---|---|---|
| **Feature** | **ACC** | **SE** | **SP** | **F1** | **ERR** |
| **TORGO (trained also on TORGO)** | | | | | |
| wav2vec2-BASE-12 | 72.77 | 0.61 | 0.84 | 0.64 | 0.385 |
| **TORGO (trained on EasyCall)** | | | | | |
| NO-FT-wav2vec2-BASE-3 | 60.87 | 0.59 | 0.62 | 0.51 | 0.408 |
| FT-wav2vec2-BASE-5 | 64.05 | 0.53 | 0.75 | 0.53 | 0.402 |
| **EasyCall (trained also on EasyCall)** | | | | | |
| wav2vec2-BASE-9 | 87.58 | 0.87 | 0.88 | 0.88 | 0.130 |
| **EasyCall (trained TORGO)** | | | | | |
| NO-FT-wav2vec2-BASE-9 | 62.05 | 0.32 | 0.92 | 0.46 | 0.340 |
| FT-wav2vec2-BASE-7 | 65.86 | 0.47 | 0.85 | 0.58 | 0.335 |

| wav2vec2-XLSR model | | | | | |
|---|---|---|---|---|---|
| **Feature** | **ACC** | **SE** | **SP** | **F1** | **ERR** |
| **TORGO (trained also on TORGO)** | | | | | |
| wav2vec2-XLSR-24 | 71.21 | 0.59 | 0.83 | 0.62 | 0.383 |
| **TORGO (trained on EasyCall)** | | | | | |
| NO-FT-wav2vec2-XLSR-5 | 66.28 | 0.40 | 0.93 | 0.51 | 0.426 |
| FT-wav2vec2-XLSR-18 | 70.30 | 0.59 | 0.81 | 0.61 | 0.398 |
| **EasyCall (trained also on EasyCall)** | | | | | |
| wav2vec2-XLSR-24 | 89.48 | 0.90 | 0.89 | 0.90 | 0.113 |
| **EasyCall (trained on TORGO)** | | | | | |
| NO-FT-wav2vec2-XLSR-22 | 71.60 | 0.81 | 0.62 | 0.75 | 0.263 |
| FT-wav2vec2-XLSR-20 | 75.55 | 0.86 | 0.65 | 0.79 | 0.244 |

The number at the end of each feature refers to the number of the corresponding layer of the pre-trained model. ACC, SE, SP, F1, and EER refer to accuracy, sensitivity, specificity, F1-score, and equal error rate, respectively.

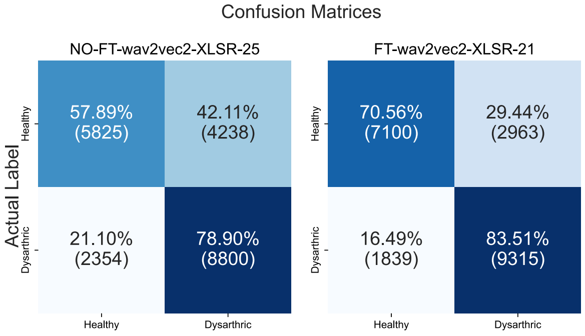

Fig. 8. Confusion matrices of dysarthria detection in the cross-database language-independent scenario for NO-FT-wav2vec2-XLSR-21 (the best performing non-fine-tuned wav2vec2-XLSR feature) and for FT-wav2vec2-XLSR-25 (the best performing fine-tuned wav2vec2-XLSR feature).

and conducted our experiments only for the TORGO database. Moreover, the conventional MFCCs features was used for comparison. From the results, it was found that the wav2vec2 features outperformed the MFCCs. This experiment suggests that the wav2vec2 features are more generalizable not only to different databases and languages as shown in the current study but they also show better robustness to noise conditions compared to the conventional features. Robustness of the wav2vec2 features to mismatch caused by different noise conditions is, however, outside the scope of this article and it calls for further studies.

## V. SUMMARY AND CONCLUSION

In this paper, we studied the automatic detection of dysarthria from speech signals by comparing features derived from fine-tuned and non-fine-tuned wav2vec2 models (specifically wav2vec2-BASE as a model trained with only English samples and wav2vec2-XLSR as a multilingual model trained with samples representing many languages). This comparison involved studying cross-database training and testing in two scenarios: language-dependent and language-independent. In the language-dependent scenario, we performed training and testing of the systems using two English databases, namely UA-Speech and TORGO. For the language-independent scenario, the training and testing of the systems were conducted using both an English database (UA-Speech/TORGO) and an Italian database (EasyCall), and vice versa. In both scenarios, a comparison with a single-database scenario (training and testing the system with the same database) was conducted. The primary goal of this study was to examine the effectiveness of fine-tuned wav2vec features in generalizing to unseen speakers, recording environments, and languages.

The results showed that the performance of the non-fine-tuned wav2vec2 features remarkably decreased when the system trained with one English database and tested with another English database was compared to the system trained and tested with the same English database. Similarly, a decrease in performance was found for a scenario, where databases of different languages (English and Italian) were used for training and testing. When the models (wav2vec2-BASE and wav2vec2-XLSR) were fine-tuned, the performance was improved and the improvements were between 1.46% and 4.8% (absolute) in accuracy for the language-dependent scenario. For the language-independent scenario, the fine-tuned wav2vec2 features showed absolute accuracy improvements between 2.3% and 8.65%. The reason why the fine-tuned wav2vec2 features improved detection performance is because the pre-trained wav2vec2 model has learned general speech representations from a large amount of data, and when fine-tuning is carried out on dysarthric speech data, the model can transfer this general knowledge to better represent dysarthric speech characteristics, including atypical articulation, prosody, and phonation patterns. In other words, fine-tuning allows the model to adapt its existing representations to the specific traits of dysarthric speech, and this adaptation enhances the model's ability to discriminate between healthy and dysarthric speech patterns.

A comparison between the two wav2vec2 models (wav2vec2-BASE and wav2vec2-XLSR) indicates that wav2vec2-XLSR performed better in both cross-database scenarios by showing higher detection accuracy compared to the wav2vec-BASE model. The improvement shown by wav2vec2-XLSR may be due to fact that the model was pre-trained on a diverse set of languages and designed to learn shared representations across languages. When this multilingual model is fine-tuned on dysarthric speech data from different languages, the model can leverage these shared representations to capture cross-linguistic dysarthric speech characteristics which results in enhanced generalization ability to unseen speakers and language. In addition, this improvement can also be attributed to the differences in the amount of training data and the complexity of the model parameters. The wav2vec2-XLSR model benefits from being trained on a vast corpus comprising 56,000 hours of audio data of 56 different languages. This extensive training allows the model to capture a broader spectrum of speech variations, including those characteristic to dysarthric speech. The wav2vec2-BASE model with fewer parameters trained on a much smaller dataset (960 hours of audio) showed a more pronounced decrease in detection accuracy. This indicates that the wav2vec2-BASE model has a narrower scope for learning such varied speech patterns. As a result, the model's complexity together with the larger training dataset can influence the capacity of the model to learn and adapt to diverse speech patterns, including pathological speech.

Two more observations that are worth mentioning can be made from the results reported in Section IV. First, it was found that the detection performance for the TORGO database in the cross-database language-independent scenario (i.e., training with EasyCall and testing with TORGO) was better compared to the detection performance in the cross-database language-dependent scenario (i.e., training with UA-Speech and testing with TORGO). This improvement can be attributed to several factors related to the distinct characteristics and composition of the database involved. The UA-Speech database includes only one speaking task (word pronunciation), whereas the number of speaking tasks in EasyCall is two (pronunciation of words and sentences) and in TORGO it is three (pronunciation of non-words, words, and sentences). This overlap in the speaking task between TORGO and EasyCall potentially enhances the wav2vec2 model's capability to extract more generalizable and robust features, thereby improving performance in detecting dysarthric speech for the TORGO database. Moreover, fine-tuning the wav2vec2 model with the EasyCall database, which is solely in Italian, introduces the model to a different linguistic context with its unique phonetic and prosodic characteristics. This exposure to the Italian language could still enrich the model's acoustic diversity. It may help the model in developing refined sensitivity to variations in speech that transcends language barriers, aiding in the identification of dysarthric speech characteristics that are less language-dependent and more universally present across different speech disorders. This aspect of cross-lingual training, even with a single non-English language, might contribute to the model's enhanced ability to generalize across varied expressions of dysarthric speech. This observation can also be seen from the results of the detection experiment in which the system was trained with EasyCall and tested with UA-Speech.

Second, the wav2vec2 features applied to TORGO showed a different trend in results compared to UA-Speech (i.e., a raising trend in the accuracy when moving from the first layer towards the final layer). This implies that the features extracted from final layers showed a better discriminability between healthy and dysarthric speech. In contrast, our initial study [11] (which focused exclusively on the wav2vec2-BASE model and utilized only the UA-Speech database) demonstrated that the features extracted from the starting layers of the wav2vec2-BASE model showed a better capability to distinguish between healthy and dysarthric speech, because these models were originally pre-trained on a large amount of unlabeled data and fine-tuned using a small set to perform automatic speech recognition (ASR). Therefore, the early layers of these models primarily tend to capture generic speech information. This information includes acoustic properties such as pitch, formants, and timbre. These features are vital for detecting dysarthria, which typically presents with unusual articulation and acoustic fluctuations. The observed discrepancies in detection accuracy between UA-Speech and TORGO may be attributed to the unique attributes of each database such as speaking task, the amount of data, diversity of speech impairments, and varying levels of background noise. Specifically, a richer and more complex database like TORGO (with a broader range of speaking tasks) possibly benefits more from the nuanced representations captured by the wav2vec2 model's later layers. Conversely, the comparatively straightforward nature of the UA-Speech database might be sufficiently addressed by the features extracted from the initial layers.

Taken together, the experimental findings of the study indicate that fine-tuning the wave2vec2 models allows to extract features that are more generalizable to different speakers, recording environments, and language in the cross-database scenarios. Even though the multilingual wav2vec2-XLSR showed better generalizability, there still exists a gap in its performance when compared to the single-database scenario. Therefore, further research is required to study the generalizability of the fine-tuned models by investigating, for example, how much the severity level of the disease (e.g., mildly dysarthric vs. healthy) affects the detection performance. In addition, other popular pre-trained models such as WavLM [54] and HuBERT [55] can be explored in cross-database scenarios. Another potential future topic is to continue the current work by studying the features of pre-trained models in automatic multi-class classification of various diseases from speech signals.

## References

[1] Y. Yunusova, G. Weismer, J. Westbury, and M. Lindstrom, "Articulatory movements during vowels in speakers with dysarthria and healthy controls," *J. Speech, Lang., Hear. Res.*, vol. 51, pp. 596–611, 2008.

[2] J. R. Duffy, *Motor Speech Disorders E-Book: Substrates, Differential Diagnosis, and Management*. Amsterdam, The Netherlands: Elsevier, 2019.

[3] F. Javanmardi, S. R. Kadiri, and P. Alku, "A comparison of data augmentation methods in voice pathology detection," *Comput. Speech Lang.*, vol. 83, 2024, Art. no. 101552. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0885230823000712

[4] Y. Jiao, M. Tu, V. Berisha, and J. M. Liss, "Simulating dysarthric speech for training data augmentation in clinical speech applications," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2018, pp. 6009–6013.

[5] M. Hireš, M. Gazda, L. Vavrek, and P. Drotár, "Voice-specific augmentations for Parkinson's disease detection using deep convolutional neural network," in *Proc. 20th Jubilee World Symp. Appl. Mach. Intell. Inform.*, 2022, pp. 000213–000218.

[6] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, 2020, pp. 12449–12460.

[7] A. Hernandez, P. A. Pérez-Toro, E. Nöth, J. R. Orozco-Arroyave, A. Maier, and S. H. Yang, "Cross-lingual self-supervised speech representations for improved dysarthric speech recognition," in *Proc. Interspeech*, 2022, pp. 51–55.

[8] S. A. Aly, "Arabic speech emotion recognition employing wav2vec2. 0 and hubert based on baved dataset," *Trans. Eng. Comput. Sci.*, vol. 9, no. 6, pp. 1–8, 2021.

[9] L. Pepino, P. Riera, and L. Ferrer, "Emotion recognition from speech using wav2vec 2.0 embeddings," in *Proc. Interspeech*, 2021, pp. 3400–3404.

[10] S. Tirronen, F. Javanmardi, M. Kodali, S. R. Kadiri, and P. Alku, "Utilizing wav2vec in database-independent voice disorder detection," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2023, pp. 1–5.

[11] F. Javanmardi, S. Tirronen, M. Kodali, S. R. Kadiri, and P. Alku, "Wav2vec-based detection and severity level classification of dysarthria from speech," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2023, pp. 1–5.

[12] S. Hu et al., "Exploring self-supervised pre-trained ASR models for dysarthric and elderly speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2023, pp. 1–5.

[13] F. Javanmardi, S. R. Kadiri, and P. Alku, "Pre-trained Models for Detection and Severity Level Classification of Dysarthria from Speech," *Speech Commun.*, vol. 158, Mar. 2024, Art. no. 103047.

[14] Z. Fan, M. Li, S. Zhou, and B. Xu, "Exploring wav2vec 2.0 on speaker verification and language identification," in *Proc. Interspeech*, 2021, pp. 1509–1513.

[15] Y. Wang, A. Boumadane, and A. Heba, "A fine-tuned wav2vec 2.0/hubert benchmark for speech emotion recognition, speaker verification and spoken language understanding," 2021, *arXiv:2111.02735*.

[16] N. Vaessen and D. A. V. Leeuwen, "Fine-tuning wav2vec2 for speaker recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2022, pp. 7967–7971.

[17] T. Grósz, D. Porjazovski, Y. Getman, S. Kadiri, and M. Kurimo, "Wav2vec2-based paralinguistic systems to recognise vocalised emotions and stuttering," in *Proc. ACM Int. Conf. Multimedia*, 2022, pp. 7026–7029.

[18] J. P. Zuluaga et al., "How does pre-trained wav2vec 2.0 perform on domain-shifted ASR? An extensive benchmark on air traffic control communications," in *Proc. IEEE Spoken Lang. Technol. Workshop*, 2023, pp. 205–212.

[19] H. Kim et al., "Dysarthric speech database for universal access research," in *Proc. Ninth Annu. Conf. Int. Speech Commun. Assoc.*, 2008, pp. 1741–1744.

[20] T. H. Falk, W.-Y. Chan, and F. Shein, "Characterization of atypical vocal source excitation, temporal dynamics and prosody for objective measurement of dysarthric word intelligibility," *Speech Commun.*, vol. 54, no. 5, pp. 622–631, 2012.

[21] A. B. Kain, J.-P. Hosom, X. Niu, J. P. v. Santen, M. Fried-Oken, and J. Staehely, "Improving the intelligibility of dysarthric speech," *Speech Commun.*, vol. 49, no. 9, pp. 743–759, 2007.

[22] J. Kim, N. Kumar, A. Tsiartas, M. Li, and S. S. Narayanan, "Automatic intelligibility classification of sentence-level pathological speech," *Comput. Speech Lang.*, vol. 29, no. 1, pp. 132–144, 2015.

[23] P. Rong et al., "Predicting speech intelligibility decline in amyotrophic lateral sclerosis based on the deterioration of individual speech subsystems," *PLoS One*, vol. 11, no. 5, 2016, Art. no. e0154971.

[24] N. Narendra and P. Alku, "Dysarthric speech classification from coded telephone speech using glottal features," *Speech Commun.*, vol. 110, pp. 47–55, 2019.

[25] I. Kodrasi and H. Bourlard, "Super-gaussianity of speech spectral coefficients as a potential biomarker for dysarthric speech detection," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2019, pp. 6400–6404.

[26] A. Mayle, Z. Mou, R. Bunescu, S. Mirshekarian, L. Xu, and C. Liu, "Diagnosing dysarthria with long short-term memory networks," in *Proc. Interspeech*, 2019, pp. 4514–4518.

[27] K. Gurugubelli and A. K. Vuppala, "Perceptually enhanced single frequency filtering for dysarthric speech detection and intelligibility assessment," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2019, pp. 6410–6414.

[28] K. Gurugubelli and A. K. Vuppala, "Analytic phase features for dysarthric speech detection and intelligibility assessment," *Speech Commun.*, vol. 121, pp. 1–15, 2020.

[29] J. Millet and N. Zeghidour, "Learning to detect dysarthria from raw speech," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2019, pp. 5831–5835.

[30] H. Chandrashekar, V. Karjigi, and N. Sreedevi, "Spectro-temporal representation of speech for intelligibility assessment of dysarthria," *IEEE J. Sel. Topics Signal Process.*, vol. 14, no. 2, pp. 390–399, Feb. 2020.

[31] H. M. Chandrashekar, V. Karjigi, and N. Sreedevi, "Investigation of different time-frequency representations for intelligibility assessment of dysarthric speech," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 28, no. 12, pp. 2880–2889, Dec. 2020.

[32] M. Fernández-Díaz and A. Gallardo-Antolín, "An attention long short-term memory based system for automatic classification of speech intelligibility," *Eng. Appl. Artif. Intell.*, vol. 96, 2020, Art. no. 103976.

[33] S. Gupta et al., "Residual neural network precisely quantifies dysarthria severity-level based on short-duration speech segments," *Neural Netw.*, vol. 139, pp. 105–117, 2021.

[34] A. A. Joshy and R. Rajan, "Automated dysarthria severity classification using deep learning frameworks," in *Proc. 28th Eur. Signal Process. Conf.*, 2021, pp. 116–120.

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

[36] A. A. Joshy and R. Rajan, "Dysarthria severity assessment using squeeze-and-excitation networks," *Biomed. Signal Process. Control*, vol. 82, 2023, Art. no. 104606.

[37] J. Hestness et al., "Deep learning scaling is predictable, empirically," 2017, *arXiv:1712.00409*.

[38] P. Gómez-Vilda, A. Gómez-Rodellar, D. Palacios-Alonso, V. Rodellar-Biarge, and A. Á. Marquina, "The role of data analytics in the assessment of pathological speech—A critical appraisal," *Appl. Sci.*, vol. 12, no. 21, 2022, Art. no. 11095.

[39] M. Markaki and Y. Stylianou, "Normalized modulation spectral features for cross-database voice pathology detection," in *Proc. Tenth Annu. Conf. Int. Speech Commun. Assoc.*, 2009, pp. 935–938.

[40] J. A. Gómez-García, L. Moro-Velázquez, and J. I. Godino-Llorente, "On the design of automatic voice condition analysis systems. Part II: Review of speaker recognition techniques and study on the effects of different variability factors," *Biomed. Signal Process. Control*, vol. 48, pp. 128–143, 2019.

[41] A. Favaro et al., “Do phonatory features display robustness to characterize parkinsonian speech across corpora?,” in *Proc. INTERSPEECH*, 2023, pp. 2388–2392.

[42] E. J. Ibarra, J. D. Arias-Londoño, M. Zañartu, and J. I. Godino-Llorente, “Towards a corpus (and language)-independent screening of Parkinson’s disease from voice and speech through domain adaptation,” *Bioengineering*, vol. 10, no. 11, 2023, Art. no. 1316.

[43] F. Braun et al., “Classifying dementia in the presence of depression: A cross-corpus study,” in *Proc. INTERSPEECH*, 2023, pp. 2308–2312.

[44] S. Gillespie, Y.-Y. Logan, E. Moore, J. Laures-Gore, S. Russell, and R. Patel, “Cross-database models for the classification of dysarthria presence,” in *Proc. Interspeech*, 2017, pp. 3127–3131.

[45] J. R. Orozco-Arroyave et al., “Automatic detection of Parkinson’s disease in running speech spoken in three different languages,” *J. Acoustical Soc. Amer.*, vol. 139, no. 1, pp. 481–500, 2016.

[46] D. Wang, L. Deng, Y. T. Yeung, X. Chen, X. Liu, and H. Meng, “Unsupervised domain adaptation for dysarthric speech detection via domain adversarial training and mutual information minimization,” in *Proc. Interspeech*, 2021, pp. 2956–2960.

[47] A. Conneau, A. Baevski, R. Collobert, A. Mohamed, and M. Auli, “Unsupervised cross-lingual representation learning for speech recognition,” in *Proc. Interspeech*, 2021, pp. 2426–2430.

[48] F. Rudzicz, A. K. Namasivayam, and T. Wolff, “The torgo database of acoustic and articulatory speech from speakers with dysarthria,” *Lang. Resour. Eval.*, vol. 46, no. 4, pp. 523–541, 2012.

[49] R. Turrisi et al., “Easycall corpus: A dysarthric speech dataset,” in *Proc. Interspeech*, 2021, pp. 41–44.

[50] P. Enderby, “Frenchay dysarthria assessment,” *Brit. J. Disord. Commun.*, vol. 15, no. 3, pp. 165–173, 1980.

[51] X. Menendez-Pidal, J. B. Polikoff, S. M. Peters, J. E. Leonzio, and H. T. Bunnell, “The nemours database of dysarthric speech,” in *Proc. 4th Int. Conf. Spoken Lang. Process.*, 1996, pp. 1962–1965.

[52] F. Eyben, M. Wöllmer, and B. Schuller, “Opensmile: The munich versatile and fast open-source audio feature extractor,” in *Proc. 18th ACM Int. Conf. Multimedia*, 2010, pp. 1459–1462, doi: 10.1145/1873951.1874246.

[53] F. Eyben et al., “The geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing,” *IEEE Trans. Affect. Comput.*, vol. 7, no. 2, pp. 190–202, Apr.–Jun. 2016.

[54] S. Chen et al., “WavLM: Large-scale self-supervised pre-training for full stack speech processing,” *IEEE J. Sel. Topics Signal Process.*, vol. 16, no. 6, pp. 1505–1518, Oct. 2022.

[55] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, “HuBERT: Self-supervised speech representation learning by masked prediction of hidden units,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 3451–3460, 2021.