applied sciences

MDPI

Article

A Study on Model Training Strategies for Speaker-Independent and Vocabulary-Mismatched Dysarthric Speech Recognition

Jinzi Qi and Hugo Van hamme *

Department Electrical Engineering-ESAT-PSI, Katholieke Universiteit Leuven (KU Leuven) Kasteelpark Arenberg 10, 3001 Leuven, Belgium; jinzi.qi@kuleuven.be
* Correspondence: hugo.vanhamme@kuleuven.be; Tel.: +32-16321842

Abstract: Automatic speech recognition (ASR) systems often struggle to recognize speech from individuals with dysarthria, a speech disorder with neuromuscular causes, with accuracy declining further for unseen speakers and content. Achieving robustness for such situations requires ASR systems to address speaker-independent and vocabulary-mismatched scenarios, minimizing user adaptation effort. This study focuses on comprehensive training strategies and methods to tackle these challenges, leveraging the transformer-based Wav2Vec2.0 model. Unlike prior research, which often focuses on limited datasets, we systematically explore training data selection strategies across diverse source types (languages, canonical vs. dysarthric, and generic vs. in-domain) in a speaker-independent setting. For the under-explored vocabulary-mismatched scenarios, we evaluate conventional methods, identify their limitations, and propose a solution that uses phonological features as intermediate representations for phone recognition to address these gaps. Experimental results demonstrate that this approach enhances recognition across dysarthric datasets in both speaker-independent and vocabulary-mismatched settings. By integrating advanced transfer learning techniques with the innovative use of phonological features, this study addresses key challenges for dysarthric speech recognition, setting a new benchmark for robustness and adaptability in the field.

Keywords: dysarthric speech recognition; phonological features; speaker-independent; vocabulary-mismatched

check for updates

Academic Editor: Douglas O'Shaughnessy

Received: 16 December 2024
Revised: 2 February 2025
Accepted: 10 February 2025
Published: 14 February 2025

Citation: Qi, J.; Van hamme, H. A Study on Model Training Strategies for Speaker-Independent and Vocabulary-Mismatched Dysarthric Speech Recognition. *Appl. Sci.* **2025**, *15*, 2006. https://doi.org/10.3390/app15042006

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Dysarthria is a speech disorder caused by neuromuscular diseases or conditions, characterized by poor pronunciation. It leads to reduced intelligibility, which significantly hinders daily communication for individuals with dysarthria [1]. Automatic speech recognition (ASR) systems, which have demonstrated high effectiveness for canonical speech, hold great potential for improving communication in this population [2]. In addition, speakers with diseases that cause dysarthria often have limited motor control and need assistive devices to compensate for that. In this context, ASR can be instrumental in developing voice-controlled assistive devices, offering enhanced independence and accessibility.

State-of-the-art dysarthric speech recognition (DSR) models, typically leveraging deep neural networks, have achieved satisfactory recognition performance [2–8]. An ideal DSR model, however, must go beyond high accuracy to exhibit robust generalization capabilities and perform well on speech from unseen users (target speakers) excluded from the training set. This requirement, referred to as the speaker-independent setting, is essential for reducing the need for additional data collection and model retraining, thereby minimizing user effort and enabling the model to adapt more effectively to user variability.

Beyond generalization to new speakers, an ideal DSR model should also handle words not present in the training data. Current dysarthric speech databases typically contain less than ten hours of recordings and a vocabulary limited to only a few hundred words. These limited training data do not cover the diverse vocabulary needed for everyday communication. Consequently, an ASR model that recognizes only the vocabulary seen during training would have limited usability in the real world [9]. This challenge is framed as the vocabulary-mismatched setting, in which the DSR model must perform reliably even when there is a mismatch between the vocabulary used during training and testing.

In this work, we focus on improving dysarthric speech recognition under two critical scenarios: speaker-independent and vocabulary-mismatched settings. Instead of developing new model architectures, our emphasis lies on identifying effective training strategies and identifying appropriate training datasets and approaches that enable the model to acquire and transfer relevant knowledge for accurate recognition of target dysarthric speaker speech.

In the speaker-independent setting, the model does not have direct access to the specific speech patterns of the target speaker. However, it can still benefit from leveraging shared characteristics among dysarthric speakers, such as similarities in speech patterns that arise due to shared medical diagnoses, gender, or age. Additionally, linguistic commonalities can be exploited: for example, dysarthric and canonical speakers of the same language share overlapping vocabulary, and even across different languages, phonetic units often exhibit similarities. By training on data from other dysarthric speakers, canonical speakers of the same language, or languages with shared phonetic inventories, the model learns knowledge that can improve its recognition performance for unseen dysarthric speakers, which has been validated in previous works [10–13]. However, previous research has largely focused on specific domains of source data (details are discussed in Section 2.1), leaving the integration of diverse resources unexamined. To address this gap, we systematically investigate comprehensive training strategies that integrate diverse resources, including multilingual canonical speech data, in-domain canonical speech data that align with the content of the target dysarthric speech, and dysarthric speech data within or outside the target speaker's dataset. Using the widely adopted transformer-based Wav2Vec model [7,14–16] as the base model, we evaluate the recognition performance of target speaker speech across various combinations of these data sources. This investigation aims to determine the training strategy that maximizes knowledge transfer and achieves optimal performance in speaker-independent scenarios.

Furthermore, we consider the challenge of a vocabulary-mismatched setting, where the model should recognize out-of-vocabulary (OOV) words that are absent from the training data. Vocabulary mismatches often arise in two situations: very-low-resource settings, where only a limited number of words are recorded, and their phones do not cover the full phone vocabulary of the language; and cross-lingual settings, where no databases exist in the target dysarthric speaker's language. Recognition tasks for target speakers under such mismatched conditions remain under-researched. To examine this issue, we design experiments simulating these scenarios and evaluate model performance. Our results indicate that models perform poorly after conventional training in vocabulary-mismatched cases, particularly when no pretrained model with a sufficiently extensive phonetic inventory is available.

To address these limitations, we propose using phonological features (PFs) as intermediate representations in phone recognition tasks. Phonological features [17,18], also known as articulatory attributes or distinctive features, are bundles of characteristics used to describe speech sounds. PFs are shared among phones and across languages, enabling the recognition of previously unseen phones when PFs are accurately detected and prior

knowledge of their composition is available. Previous works have explored the benefit of PFs brought by their independence from phones in canonical speech recognition [19] (discussed further in Section 2.2). However, the potential of PFs in solving dysarthric speech recognition under vocabulary-mismatched settings has not been studied. In this work, we apply PFs to phone recognition for dysarthric speech in vocabulary-mismatched scenarios. Specifically, PFs are extracted from speech features and transformed into phones based on prior phonetic knowledge [19]. This approach ensures that even if certain phones are missing from the training data, all PF variations may be observed, allowing the PF model to be well trained for recognition tasks. Experimental results demonstrate that employing PFs in very-low-resource and cross-lingual settings significantly improves recognition performance, mitigating mismatches between training and testing vocabularies and enhancing model accuracy and robustness.

In summary, this work makes the following contributions:

1. We systematically explore optimal model training strategies that improve DSR performance in speaker-independent scenarios. Unlike previous studies, our work incorporates the most diverse range of resources, including multilingual canonical speech data and multiple dysarthric speech datasets, to maximize knowledge transfer to the target dysarthric speaker.
2. We investigate the under-explored issue of vocabulary mismatch in dysarthric speech recognition by designing experiments that evaluate model performance in very-low-resource and cross-lingual settings. Our findings highlight the limitations of conventional recognition approaches in these scenarios.
3. We propose using phonological features as intermediate representations as the solution for enhancing phone recognition for dysarthric speech in vocabulary-mismatched cases. Experimental results confirm the effectiveness of this method in mitigating vocabulary mismatches between training and test data, significantly improving recognition performance and robustness.

The paper is organized as follows: In Section 2, we introduce related work on knowledge transfer in dysarthric speech-related tasks and discuss the use of phonological features in speech technology. Section 3 explores model training strategies based on transfer learning, which enhance recognition performance in the speaker-independent setting. Section 4 examines model performance in the vocabulary-mismatched setting. Section 5 presents the improved recognition achieved through phonological features in these settings. Finally, Section 6 concludes the paper and outlines directions for future research.

2. Related Works

2.1. Knowledge Transfer for Dysarthric Speech Processing

In dysarthric speech technology, transferring knowledge from training data to enhance model performance on a test set or to facilitate model adaptation to target data domains is a common strategy [10,12,20–22], as dysarthric data resources are too scarce for training a state-of-the-art model from scratch. Therefore, in speaker-independent settings, a model pretrained on canonical speech will be fine-tuned or adapted with data that follow a distribution closer to a target speaker's. For example, feature extractors and acoustic models pretrained or fine-tuned on *canonical* and *out-of-domain* dysarthric speech datasets have demonstrated improved ASR accuracy for target dysarthric datasets [23]. During pretraining, base knowledge for speech recognition is acquired from canonical speech, while fine-tuning enables the model to learn dysarthric speech patterns. Similarly, ref. [12] utilizes data from *different languages* or *diseases* to enhance shared knowledge across diseases or languages, improving classification accuracy for dysarthric speakers. Another key data

source is in-domain canonical or dysarthric speech, which shares the same speech content as the target data. Studies such as [13,20,24,25] have shown that utilizing *in-domain* canonical or dysarthric datasets, which introduce the possible content of the target speech to models in addition to knowledge of normal or dysarthric speech patterns, improves DSR performance for unseen dysarthric speakers. In [26], in-domain control and dysarthric speech data are used to create word-level prototypes for target speaker speech recognition, leading to significant improvements in speech recognition for unseen speakers. However, this work does not address the challenges posed by vocabulary-mismatched scenarios.

Given the high variability among dysarthric speakers, speaker-independent settings pose significant challenges, leading to limited research in this area. To provide a more comprehensive overview, we also discuss *speaker-dependent* cases, which have been studied more extensively. In speaker-dependent scenarios, models are typically pretrained on source domain data and then fine-tuned using data from target speakers. One commonly used source domain is *in-domain* dysarthric speech data from source speakers [11,27,28], which provide an excellent initialization for recognition models, facilitating more effective adaptation to target speakers. Another frequently employed source domain is *canonical* speech data [20,21], which serve as a rich resource and can be used to train large-scale, up-to-date models, ultimately resulting in improved recognition performance. Combining canonical and dysarthric speech data for model pretraining has also proven to enhance recognition performance [5,20,29], as this approach leverages the complementary strengths of both data types. In addition to using source domains in the same language as the target speaker, researchers have explored using *multilingual* datasets to expand data diversity and capture more general speech patterns, which improves model robustness and recognition performance for dysarthric speech in the target language. For example, refs. [10,30] investigate knowledge transfer between English and Japanese in the Listener model of the 'Listen, Attend and Spell' model. Similarly, ref. [31] examines bilingual model training followed by language-specific adaptation, resulting in further performance gains. In another study [32], an acoustic-to-articulatory model pretrained on typical English speech is adapted to disordered English and Cantonese dysarthric speech, demonstrating the potential of incorporating knowledge beyond conventional speech recognition features.

Overall, existing research demonstrates the advantages of leveraging multiple data sources for improving recognition performance in both speaker-independent and speaker-dependent cases, including rich-resource and in-domain canonical speech, as well as out-of-domain and in-domain dysarthric speech. Each resource has been shown to contribute distinct and valuable knowledge that can be transferred for recognition of target speech. However, no prior work has systematically investigated the optimal utilization of all available training data types, particularly in speaker-independent scenarios. To address this gap, our study focuses on developing a comprehensive model training strategy that integrates all accessible data resources to optimize performance in speaker-independent settings. This systematic approach aims to advance the field by identifying effective strategies for leveraging diverse knowledge to enhance dysarthric speech recognition.

2.2. Phonological Features Used in Speech Technology

In speech technology, PFs help to systematically analyze speech and contribute to speech classification and recognition within and across languages. In canonical speech recognition, PFs are commonly used in multilingual and cross-lingual settings. Studies like [33–35] have trained PF extractors demonstrating general advantages of integrating detected PFs with traditional acoustic features for phone recognition tasks across various model architectures and languages. Beyond direct integration at the feature level, other approaches have explored the transformation of detected PFs into phone tokens using prior

phonetic knowledge. For instance, refs. [19,36–38] achieved better recognition of unseen phones or in cross-lingual scenarios by leveraging such transformations on multilingual datasets. In [39], the focus was on better initializing the output layer for phone recognition. This work utilized detected PFs of unseen phones to determine their closest seen phones and initialized the weights of the unseen phones using those of the closest seen phones, significantly enhancing recognition performance for unseen phones. Additionally, ref. [40] proposed a more compact transformation between articulatory attributes and speech transcriptions, achieving significantly better recognition performance compared to conventional word- or character-based recognition models.

In the domain of dysarthric speech, research on PFs spans various tasks. One prevalent task involves using neural networks to detect PFs from dysarthric speeches in different languages. Studies such as [8,41–44] have demonstrated that PF detectors pretrained on canonical speech generalize effectively, enabling reliable PF extraction from dysarthric speech across different language datasets and base model architectures. Accurate PF detection also enables the automatic assessment and analysis of dysarthric speakers based on their pronunciation profile. Research in this field [43,45–47] uses PFs either directly obtained from their speech or transformed from recognized phones to provide insights into individual speech characteristics. For speech recognition tasks, detected PFs have been employed to enhance recognition performance by combining them with traditional acoustic features or model representations [8,48–51]. Additionally, PFs have been employed as a form of regularization during model training, as shown in [52], to further improve the adequacy and robustness of the information extracted during model training.

Across all approaches explored in previous studies, PFs have consistently served an auxiliary role in achieving more accurate characteristic modeling of speech or speakers, ultimately benefiting primary tasks such as speech recognition. For dysarthric speech recognition, their effectiveness has been demonstrated primarily when used as supplementary information alongside acoustic features or model representations. The unique advantage of PFs, their independence from specific phones and languages, enables recognition of unseen phones in vocabulary-mismatched settings, a capability that has been validated in canonical speech recognition. However, this potential remains under-explored in dysarthric speech tasks. This study addresses this gap by employing PFs as intermediate representations to enhance model performance on DSR tasks in vocabulary-mismatched scenarios.

3. Exploring Speaker-Independent Dysarthric Speech Recognition

3.1. Motivation

To achieve reasonable recognition performance on target speaker speech without any data from the speaker, training the model with relevant datasets is the most straightforward solution. However, current research has yet to systematically examine how to select and combine training data to maximize effectiveness. In this work, we investigate optimal training strategies by considering all available datasets and evaluating how knowledge transfer from diverse sources impacts target dysarthric speech recognition in a speaker-independent setting.

Our exploration covers three types of knowledge resources: rich-resource (multilingual) canonical speech, in-domain canonical speech (sharing vocabulary with the target data), and dysarthric speech (from speakers either within or outside the target speaker's dataset). The base model is trained using one or more types of source data and tested on target dysarthric speech. We evaluate the effectiveness of these model training strategies (different combinations of source data) using the target speech recognition performance as the metric.

3.2. Datasets and Data Split

In this subsection, we outline the datasets utilized for each resource type and detail the data split for training, validation, and testing.

3.2.1. Rich-Resource Canonical Speech

For rich-resource canonical speech, we use the **Common Voice** dataset [53] (version 13.0), which is a well-known dataset consisting of speech recordings of sentences, noted for its diversity in accents, languages, and demographics, making it ideal for training robust and inclusive speech recognition models. To ensure a comprehensive knowledge base and better alignment with target speech languages, we selected data from five languages: English, Spanish, Italian, Dutch, and Tamil.

For each language, we limited the training set to a maximum of 100 h and the validation set to 1 h. This resulted in 100 h of training data for English, Spanish, and Italian, 75 h for Tamil, and 38 h for Dutch. Training data that combines all five languages are referred to as **C_rich** while data using only the target language's data are referred to as **C_rich_tgt**.

3.2.2. In-Domain Canonical Speech

We utilized seven dysarthric speech datasets covering the five languages mentioned above to provide both in-domain canonical and dysarthric speech data. Table 1 provides the statistics of these datasets, including the length as well as the number of speakers and utterances for the canonical and dysarthric parts.

Table 1. Statistics of dysarthric datasets used in the experiment. (N/U: Not used in this work; N/A: not available; #Spk., #Utt.: number of speakers, number of utterances.)

Name	Language	Content	Dysarthric Part			Canonical Part		
			Length	#Spk.	#Utt.	Length	#Spk.	#Utt.
Torgo [54]	English	Word, Sentence	3.0 h	8	3.1 k	4.6 h	7	5.9 k
UASPEECH [55]	English	Word	10.2 h	16	12.0 k	4.9 h	13	9.9 k
Domotica [56]	Dutch	Sentence	8.8 h	17	4.2 k	N/A	N/A	N/A
COPAS [57]	Dutch	Word list, Sentence	N/U	N/U	N/U	4.5 h	131	1.8 k
PC-GITA [58]	Spanish	Word, Sentence	1.0 h	50	1.8 k	1.0 h	50	1.8 k
EasyCall [59]	Italian	Word, Phrase	8.2 h	31	11.3 k	4.0 h	24	10.1 k
SSNCE [60]	Tamil	Word, Phrase	5.5 h	20	7.3 k	2.9 h	10	3.7 k

Six of the dysarthric datasets include recordings of the same content as the dysarthric speech within the dataset spoken by canonical speakers. These recordings are used as the in-domain canonical speech for target speakers within the respective datasets, referred to as **C_indm**. For the Dutch dataset **Domotica**, which lacks canonical speakers, we use the canonical part of the **COPAS** dataset as its in-domain canonical data. The **COPAS** dataset covers 21% of the words and 100% of the phones in **Domotica**. The dysarthric part of **COPAS** is excluded in this work due to its limited number of recordings per speaker.

In model training involving **C_indm** data, 90% of the data from each dataset are used for training, with the remaining 10% for validation. For canonical speakers in the **UASPEECH** dataset, which provides a predefined data split, data from block 1 and 3 are used for training, and block 2 is used for validation.

3.2.3. Dysarthric Speech

Dysarthric speech from the six dysarthric datasets serves as the dysarthric training data. A five-fold speaker cross-validation is implemented in the training, dividing dysarthric speakers in each dataset into five subsets. For each fold, speakers from four subsets are

used as source speakers for training and validation, whereas the remaining subset serves as the target speakers used for evaluation (testing). Training data that included all source speakers across the six datasets are labeled as **D**. If the training data only included source speakers from the same dataset as the target speakers, they are referred to as **D_indm**.

For the **Torgo** dataset, which contains only eight dysarthric speakers, we conducted a four-fold cross-validation, dividing speakers into four subsets and using three subsets for training and one for evaluation in each fold. When using the training data from multiple dysarthric datasets (**D**), we maintained five folds, but in the fifth fold, the same training speakers as the fourth fold were used for **Torgo**. To avoid evaluation overlap, the remaining **Torgo** speakers were excluded from evaluation in that fold.

For each source dysarthric speaker, 90% of the recordings were randomly selected for training, while the remaining 10% were used for validation. In the **UASPEECH** dataset, block 1 and 3 were used for training, and block 2 was for validation.

3.2.4. Evaluation Set

For performance evaluation, all data from each target speaker were used in the recognition task. (For **UASPEECH**, data from block 2 were used for recognition.) The final recognition result for each dataset was calculated as the average performance across all speakers. Similarly, the overall result for each trained model was determined by the average results across all six datasets, unless mentioned.

3.2.5. Summary

The various training data used in the experiments are summarized below:

- **C_rich**: Rich-resource canonical speech from five languages in **Common Voice**;
- **C_rich_tgt**: Rich-resource canonical speech for target language in **Common Voice**;
- **C_indm**: In-domain canonical speech from the dataset of target dysarthric speakers;
- **D**: Dysarthric speech from source speakers across all dysarthric datasets, including in-domain dysarthric data;
- **D_indm**: In-domain dysarthric speech from source speakers within the dataset of target dysarthric speakers.

All datasets used in this work are publicly available and have been extensively utilized in numerous published studies. During preprocessing, mislabeled utterances (as identified in related works) and recordings shorter than 0.2 s were excluded. These preprocessing steps, combined with the datasets' widespread use, ensure robust annotation accuracy and reliable sample representativeness.

We utilized multiple canonical and dysarthric speech datasets for training and testing to increase variability in our test sets, aiming to ensure a representative sample for practical applications. However, potential biases may still exist in the data, particularly related to age, disorder type, language, and accent. To mitigate these effects, we have consistently reported recognition results averaged across six dysarthric datasets, covering a broader range of variations compared to most recent studies. Consequently, the potential negative impact of this variability has been effectively minimized, ensuring a more robust and generalized evaluation of our approach.

3.3. Experiment Setting

3.3.1. Experiment Design

We use the Wav2Vec2.0 model [14] as the base model, which has been commonly used in recent research. The model consists of a feature extraction network (six convolutional layers), a feature projection layer (one dense layer), twenty-four transformer encoder layers, and a final output layer dedicated to downstream tasks (see Figure 1). Specifically, we

utilize the Wav2Vec2.0 model that is self-supervised, pretrained on multilingual data, and fine-tuned for phone recognition tasks across 26 languages [61]. This base model contains approximately 315 million parameters.

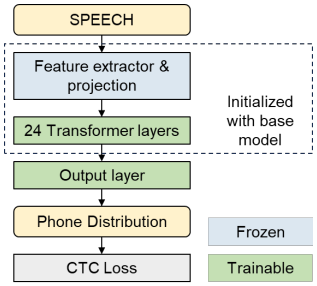


Figure 1. The Wav2Vec2-based phone recognition model used in this work.

In our experiments, we kept phones as the recognition target and used Phone Error Rate (PER) as the metric for recognition performance. The base model's vocabulary includes 392 phones, which is redundant for our five-language scenario. Thus, we adopt a smaller vocabulary of 82 phones, including a <*blk*> (blank) token for loss calculation. This reduced vocabulary was derived from phonetic transcriptions of all datasets we used. These transcriptions are expressed in the International Phonetic Alphabet (IPA) and were obtained using the *Epitran* toolkit [62]. For the **SSNCE** database, we utilized the provided phonetic transcriptions. The final vocabulary was refined by removing rarely occurring phones, splitting diphthongs into two parts, and merging phones with identical phonological features. Greedy decoding was used during evaluation without the application of a language model. The Connectionist Temporal Classification (CTC) loss [63] was employed for model training.

In this study, we investigate model training strategies using combinations of available data. Each strategy involves one or multiple training steps, each using only one type of source data. Models are trained sequentially on the data. In the training step using **C_rich**, only one model is trained, and it is tested with all dysarthric datasets. When the training step involves target-related data, including using language-related data **C_rich_tgt** and in-domain data **C_indm**, **D_indm**, we train one model for each language or each dataset and evaluate the model with the corresponding dysarthric dataset(s). If the training involves dysarthric speech **D** and **D_indm**, we conduct speaker-cross-validation, train one model for each fold, and test it on the target speakers in each fold. This means that for in-domain dysarthric data **D_indm**), there will be one model for each fold of each dataset.

3.3.2. Experiment Details

During training, we used a batch size of 8 and implemented a gradient accumulation of 4 steps, resulting in an effective batch size of 32. The optimizer used was *AdamW*. The training procedure begins with training only the output layer for 2 epochs, using a linearly increasing learning rate that ramps up to 10^{-4}. Subsequently, we proceed with joint training of the full model, which starts with a learning rate warm-up period of 2 epochs up to 10^{-4}, followed by a linear decrease ending at 0 after the last epoch. The feature extraction network and projection layer are always frozen. The number of epochs for full model training and the checkpointing frequency vary across training data, as detailed in Table 2. For training with dysarthric data (**D, D_indm**), we incorporate early stopping with the patience of 10 checkpoints, utilizing the PER on the validation set as the metric. After training, we select the checkpoint with the best PER on the validation set as the final model.

The selection of hyperparameters and training parameters in this work was largely carried out according to the corresponding reference works, as these parameters were

previously optimized by various original authors [61]. Other parameters, particularly training parameters, were chosen to balance model convergence and efficiency in terms of training time and storage space or to align with our hardware constraints (e.g., GPU memory size) to ensure successful training.

Table 2. Training settings for each training step. (#Epoch: Number of full model training epochs; CKPT Freq: the checkpointing frequency.)

Training Step	#Epoch	CKPT Freq.
C_rich **C_rich_tgt**	20	1/2 Epoch
C_indm	40	1 Epoch (2 for **PC-GITA**)
D **D_indm**	200	1/2 Epoch 2 Epochs

3.4. Results

In this subsection, we present the PER results of models trained with various strategies. The objective is to identify the optimal combination of training data and steps that maximize relevant knowledge extraction and recognition performance for target speech in the speaker-independent setting.

Table 3 provides the characteristics of each training dataset at the top, while the middle and bottom sections detail the data selection and corresponding PER results for each experiment. Specifically, rows **E1**–**E7** display results when dysarthric speech is excluded from training, while the subsequent rows show results when dysarthric speech is included. For experiments involving multiple datasets, training begins with the dataset listed first in the table and proceeds sequentially from left to right.

Table 3. Top: Characteristics of training data. **Middle**: PER results averaged across all dysarthric datasets using models trained on only canonical data. **Bottom**: Idem, making use of dysarthric training data. (Target Dataset: training vocabulary equals testing vocabulary; All Dataset: training vocabulary covers words of all six datasets.)

Data Name	**C_rich**	**C_rich_tgt**	**C_indm**	**D**	**D_indm**	
Type	Canonical			Dysarthric		
Size	Rich-resource			Low-resource		
Train Vocabulary	5 Languages	Target Language	Target Dataset	All Datasets	Target Dataset	
EXP	**Training Strategy**					**PER**
E1	**Base model**					0.620
E2	✓*					0.554
E3	✓					0.536
E4		✓				0.456
E5			✓			0.340
E6	✓		✓			**0.315**
E7		✓	✓			**0.322**
E8				✓		0.232
E9					✓	0.205

Table 3. *Cont.*

EXP	Training Strategy					PER
E10				✓	✓	0.194
E11	✓				✓	0.185
E12		✓			✓	0.182
E13			✓		✓	**0.181**
E14	✓		✓		✓	**0.181**
E15		✓	✓		✓	**0.180**

✓: The data is used for model training in the corresponding experiment; **Bold numbers**: Best PER(s) within the statistical significance level of the middle or bottom table; *: train only the output layer.

In the result tables throughout this work, the statistical significance of differences is represented using bold numbers, unless mentioned otherwise. Bolded values represent the best PER in the scenario and any PER values that are not significantly different from the best result (p-value > 0.05). Un-bold numbers indicate results that are significantly different from the best (p-value < 0.05). In our discussion, claims are made only when results show significant differences. The significance level was calculated using the Wilcoxon signed rank test [64] across 142 dysarthric speakers from all dysarthric datasets.

3.4.1. Training with Canonical Speech Only

The evaluation of the base Wav2Vec2.0 model [14] is provided in the **Base model** scenario in Table 3, which yields high-PER results. This is due to the different phone vocabulary and phonetic transcriptions used during the base model's training, which were derived from a different toolkit than the one we used in this work. In scenario **E2**, simply adapting the output layer of the base model to our new phone vocabulary improves the PER, as it aligns the transcription target with our setup. Further fine-tuning the model jointly with the output layer (**E3**) led to additional gains, as the multilingual base model was adapted to the narrower five-language phone set. Utilizing only the canonical speech of the target language (**E4**) further reduced the PER, as the model was trained with data closely relevant to the target dysarthric speech.

Incorporating in-domain canonical speech (**C_indm**) into the model training (**E5**–**E7**) significantly enhanced performance, showing that the knowledge of vocabulary is well learned and relevant for the test speech, despite pronunciation differences between canonical and dysarthric speech. Pretraining the model on rich-resource Common Voice data before training on **C_indm** (**E6**–**E7**) resulted in the best recognition accuracy on target dysarthric speech, as the rich-resource data provides a wide vocabulary base. These training strategies are optimal in the absence of dysarthric speech.

3.4.2. Training Involving Dysarthric Speech

The PER results for models trained with dysarthric speech are presented in Table 3, rows **E8**–**E15**. The significant overall improvement compared to models trained solely on in-domain canonical data underscores the critical role of dysarthric-specific knowledge transfer in enhancing DSR performance. While using only in-domain dysarthric data (**E9**) achieves reasonable accuracy, incorporating other source data in the initial training phase and subsequently adapting the model to in-domain dysarthric data (**E10**–**E15**) yields even better performance.

The highest recognition accuracy (**E13**–**E15**) reflects the optimal training strategy. This approach builds on the best-performing models (**E6**–**E7**) from the previous scenario.

This optimal strategy uses rich-resource and in-domain canonical data to provide a broad and relevant vocabulary base while employing in-domain dysarthric speech to enhance vocabulary knowledge and integrate crucial dysarthric-specific knowledge.

Notably, similar recognition results are observed between models trained with **C_rich** and **C_rich_tgt**, such as in **E6**–**E7**, **E11**–**E12**, and **E13**–**E15**. This similarity arises because the base model itself is pretrained on a rich-resource multilingual dataset, diminishing the unique impact of **C_rich** as a fundamental factor in achieving high recognition accuracy.

4. Exploring Vocabulary-Mismatched Dysarthric Speech Recognition

4.1. Motivation

The previous section examined model training strategies aimed at improving target speech recognition in a speaker-independent setting. The findings show that incorporating in-domain datasets, whether from canonical or dysarthric speakers, significantly enhances recognition accuracy, indicating the feasibility of accurate unseen speakers' speech recognition if we have prior knowledge of the words they are likely to use.

However, a critical limitation of current dysarthric datasets is their restricted vocabulary, which falls far short of covering the daily vocabulary needs of users. The challenge becomes even more pronounced in extreme cases, such as when dysarthric recordings of very few words are available, or when the target speakers use a language that lacks corresponding dysarthric or even in-domain canonical speech data. In such scenarios, the model has to be trained with vocabulary that has minimal or no overlap with the target vocabulary, a situation known as a vocabulary-mismatched scenario.

In this section, we design experiments to simulate the vocabulary-mismatched settings to investigate how the training vocabulary impacts model performance after conventional training. For simplicity, we only use in-domain dysarthric data (**D_indm**) for training. We examine two types of scenarios: the very-low-resource setting and the cross-lingual setting. In the very-low-resource setting, we use data of limited vocabulary from each in-domain source speaker. As the speech content among speakers within the same dataset is typically similar, having fewer words in the training data increases the mismatch between the training and test data. For the cross-lingual setting, we train the model with dysarthric speech of languages different from the target dysarthric dataset.

4.2. Experiment Design

4.2.1. Model Initialization

In previous experiments, we initialized the model with the base model (**E1**) and a newly initialized output layer, as the phone vocabulary in our training data was smaller than that of the base model. This approach discarded the mapping from speech features to phone tokens learned during base model training on rich-resource data. While this prevented leveraging the pretrained phone distribution knowledge, it ensured a stricter mismatch between the vocabularies of the training and testing datasets.

However, given that the base model is publicly available and its phone vocabulary may align well with the test data, discarding this learned knowledge may not be necessary. To address this, we consider two different model initializations in the experiments. In the first one, we continue using the base model and the newly initialized output layer, maintaining strict vocabulary limitations. In the second one, we use model **E3** to simulate a new base model tailored to our phone inventory, referred to as a vocabulary-informed base model. This setup retains the pretrained wide vocabulary knowledge in both the output layer and the overall model.

For the cross-lingual scenario, instead of using **E3**, we train the vocabulary-informed base models with the **C_rich_extgt** dataset, which excludes target language data from the

C_rich dataset. This ensures that the vocabulary-informed base models draw knowledge exclusively from non-target languages.

4.2.2. Training Data Selection

For the very-low-resource setting, we selected training data with limited vocabulary by sorting each source speaker's text transcriptions alphabetically and then using a fraction (10%, 30%, or 90%) of the data for training, reserving the last 10% for validation (**Ordered**). The fewer data we select, the smaller the training vocabulary. However, the models trained with **Ordered** data are influenced by both limited vocabulary and limited data size, complicating our goal of studying the impact of vocabulary alone. To address this, we implemented a second method, **Random**, where a 10%, 30%, or 90% fraction of the data are randomly selected for training, with 10% reserved for validation. With **Random** selection, even a small amount of data may cover the full vocabulary, as utterances from different speakers are randomly selected and contain different vocabulary. With this method, the trained model will suffer more from the limited data size rather than the vocabulary limitation. By comparing models trained on data from both methods, we can better isolate the impact of vocabulary mismatch.

For the cross-lingual experiments, we use two specific datasets:

- **C_rich_extgt**: The **C_rich** data with the target language speech excluded, used to train vocabulary-informed base models;
- **D_extgt**: The **D** data with the target language dataset(s) excluded, used as the model training data in the relevant setting.

In the cross-lingual setting, we do not implement cross-valuation and we train one model for each (target) language. All speakers in **D_extgt** are used as source speakers, while all dysarthric speakers from the target dataset(s) are used for evaluation. A total of 90% of each source speaker's data are randomly selected for training, with the remaining 10% reserved for validation.

To quantify the extent of vocabulary mismatch between the training and testing data in our experiment design, we calculate the fraction of words and phones in the test data that appear in the training data, referred to as the Overlap. Table 4 presents the Overlap averaged across datasets in the very-low-resource setting and across languages in the cross-lingual setting.

Table 4. Overlap between the word/phone vocabulary of training and testing set, averaged over all datasets/languages.

	Averaged Overlap in % (Word/Phone)			
	Very-Low-Resource			
Data Selection	**Training Data Amount**			**Cross-Lingual (D_extgt)**
	10%	**30%**	**90%**	
Random	74.7 / 96.4	86.2 / 99.7	90.0 / 99.9	5.1 / 84.9
Ordered	20.5 / 91.0	33.8 / 99.5	82.8 / 99.8	

The table shows that, in the very-low-resource setting, vocabulary mismatch is more pronounced when using a smaller amount of training data, but this mismatch decreases as the amount of training data increases. The phone vocabulary nearly disappears when training data are sufficient (90%). As we expected, the **Random** selection method covers significantly more test words than **Ordered** under the same amount of training data. Compared to the very-low-resource setting, the cross-lingual setting exhibits a much more severe word mismatch, with only 5.1% of test words seen in the training set. The Overlap of phones is even smaller than with minimal training data in the very-low-resource setting,

indicating a more challenging task for the model. Overall, the designed experiments for vocabulary-mismatched settings evaluate the model's performance on both new words and new phones but place greater emphasis on word vocabulary mismatch.

4.2.3. Experiment Details

During training, when using 10% and 30% of the training data in the very-low-resource setting, we keep the other settings the same as in the training step **D_indm** but adjust the checkpointing frequency to every four epochs and set the early stopping patience to 20 checkpoints. For the cross-lingual setting, the training settings for **C_rich_extgt** and **D_extgt** are the same as those for **C_rich** and **D**, respectively.

4.3. Results

This section presents the model performance under the very-low-resource setting and the cross-lingual setting, both characterized by a mismatch between the training and test vocabulary. We aim to examine the model's generalization ability to target speech containing unseen words and phones after conventional training.

4.3.1. Very-Low-Resource Setting

Table 5 provides the PER results of target speaker speech recognition in the very-low-resource setting. It compares results obtained with different data selection methods, amounts of training data, and model initializations. Specifically, **E9** and **E16** use the base model with a new output layer, whereas **E11** and **E17** are based on the vocabulary-informed base model **E3**.

Table 5. Averaged PER results across all dysarthric datasets of models trained with different data selections, amounts of training data, and model initializations. (A → B: train model A with data B.)

EXP	Data Selection	Training Strategy	Training Data Amount		
			10%	30%	90%
E9	**Random**	**D_indm**	0.326	0.247	0.205
E11		**E3 → D_indm**	**0.251**	**0.213**	**0.185**
E16	**Ordered**	**D_indm**	0.690	0.478	**0.258**
E17		**E3 → D_indm**	**0.361**	**0.320**	**0.241**

Bold numbers: Best PER(s) within the statistical significance level in each setting.

The results for the **Random** selection method indicate that increasing the amount of training data consistently improves PER, highlighting the importance of training data quantity. Comparing the **Ordered** and **Random** methods, we observe a consistent performance advantage for the **Random** method, but the performance gap between the two cases narrows as the amount of training data increases. This suggests the significant model performance enhancement brought by the improved vocabulary coverage in training data. Notably, when using only 10% of the training data with the **Ordered** method, the model shows a high PER of 0.690, underscoring the limitations of conventional model training approaches in very-low-resource scenarios.

When comparing the results of different model initializations, it is evident that employing the vocabulary-informed base model **E3** provides a consistent advantage across all data selection methods and training data amounts. This advantage is weakened as the amount of training data increases, indicating that increased training data help the model to adapt better to limited vocabulary training data while exacerbating forgetting of the knowledge acquired during pretraining. Comparing results from **E9/11** and **E16/17**, the benefit of using the vocabulary-informed base model is particularly pronounced in the

Ordered method. This is likely because the extensive pretrained knowledge in **E3** serves as a crucial complement to the limited vocabulary in the **Ordered** experiment. By contrast, with the **Random** selection method, which already provides more comprehensive training vocabulary coverage, the additional benefit from **E3** is reduced. Another factor is the larger overlap between the vocabularies of the training and validation sets in the **Random** method, enabling the model to adapt more effectively to the training data and more readily forget the pretrained knowledge embedded in **E3**.

4.3.2. Cross-Lingual Setting

Table 6 presents the averaged PER results across all dysarthric datasets in a cross-lingual setting. **E18** represents the result of vocabulary-informed base models trained on **C_rich_extgt**, which contain canonical speech knowledge from non-target languages. **E19** and **E20** denote models trained with cross-lingual dysarthric data **D_extgt** using different model initializations.

Table 6. Averaged PER results across all dysarthric datasets using vocabulary-informed base model and using models trained with **D_extgt** with different model initializations. (A → B: train model A with data B.)

EXP	Training Strategy	PER
E18	**C_rich_extgt**	**0.694**
E19	**D_extgt**	0.840
E20	**E18 → D_extgt**	0.818

Bold numbers: Best PER(s) within the statistical significance level in each case.

Compared to model **E3** in Table 3, which utilizes rich-resource training data including the target language, **E18** exhibits lower recognition accuracy due to the lack of target language knowledge. When adapting the base models to dysarthric speech using cross-lingual dysarthric data **D_extgt** (**E19-20**), the PER is even higher compared to the results of base models (**E1** in Table 3 and **E18**). This is attributed to the significantly smaller size of **D_extgt** in both data volume and vocabulary, exacerbating overfitting to the training data and leading to poor recognition in word and phone vocabulary mismatched cases. Furthermore, employing the vocabulary-informed base model **E18**, which contains wide phone knowledge from multiple source languages, potentially mitigates overfitting as evidenced by slightly improved PER results (**E20**). Nevertheless, the retraining on **D_extgt** leads to forgetting previously acquired vocabulary knowledge, outweighing these benefits.

In conclusion, this section examined model recognition performance in scenarios with a vocabulary mismatch, where access to a sufficient training vocabulary is limited. The findings reveal that conventional training approaches result in suboptimal recognition performance in very-low-resource and cross-lingual settings, especially when a vocabulary-informed base model is unavailable for model initialization.

5. Improved DSR Using Phonological Features in Vocabulary-Mismatched Setting

5.1. Motivation

Previous research on dysarthric speech recognition explored the significant impact of training vocabulary on enhancing the accuracy of recognizing target dysarthric speech in both speaker-independent and vocabulary-mismatched contexts. When using very-low-resource data for model training or working in a cross-lingual setting, models frequently encounter speech with previously unseen words and even phones, leading to poor recognition accuracy. While using a vocabulary-informed base model with extensive vocabulary

knowledge in model initialization can enhance performance, such base models are not always available or suitable in practical scenarios.

In this section, we explore methods to improve model performance in the vocabulary-mismatched case without relying on vocabulary-informed base models. We propose using phonological features (PFs) as the intermediate representations in the phone recognition task.

5.2. The Implementation of Phonological Features

Phonological features, as described in previous works [17,18], are bundles of characteristics used to describe phones. These features encompass various attributes, including major sound class, place and manner of articulation, and glottal states [36,65]. As PFs are the basic units of the phonological structure shared among phones, their independence from phones and languages allows the recognition of unseen phones in test sets or new languages, provided that the PF representations of the target phones are detected accurately and the PF compositions of the phones are known.

To implement PFs in phone recognition, we first discretely represent the PFs of each phone. We utilize the Panphon toolkit [65], which transforms phones in IPA into vector representations of their corresponding PFs. Initially, the PF vector consists of 24 dimensions, each representing a phonological feature as *syllabic*, *sonorant*, *consonantal*, *continuant*, *delayed release*, *lateral*, *nasal*, *strident*, *voiced*, *spread glottis*, *constricted glottis*, *anterior*, *coronal*, *distributed labial*, *labial*, *high*, *low*, *back*, *rounded*, *velaric*, *tense*, *long*, *hitone*, and *hireg* [65]. For phone recognition loss calculation, we augment the vector with an additional dimension dedicated to a <*blk*> token, resulting in a 25-dimensional vector representation for each phone.

Each feature is encoded with a value in the set $\{-1,0,1\}$, indicating the absence, irrelevance, and presence of the feature. In this representation, the <*blk*> token is the only one with a non-zero element in the first dimension, whereas all other dimensions are zero. For example, the vector representation of the PFs for the phone *'[a]'* is [0, 1, 1, −1, 1, −1, −1, −1, 0, 1, −1, −1, 0, −1, 0, −1, −1, 1, −1, −1, −1, 1, −1, 0, 0]. The first element is 0, indicating that it is not the <*blk*> token. The second and third elements, both 1, signify that the phone is syllabic and sonorant, respectively. The remaining elements adhere to a similar pattern.

When using PFs as the intermediate representation for phone recognition, the recognition involves two steps, as shown in Figure 2. First, the PFs are extracted from the transformer layer output $\boldsymbol{x}$ using PF layer $\boldsymbol{F}_{PFs}(\cdot)$, resulting in PF logits $\boldsymbol{x}_{PFs} = \tanh\left(\boldsymbol{F}_{PFs}(\boldsymbol{x})\right)$ after applying *tanh* activation to limit their range between -1 and 1. These vectors are then transformed into phone logits (distribution) $\boldsymbol{y}_{PFs}$ via a 'Signature Matrix' [19]. The Signature Matrix $\boldsymbol{A}_{sig}$ contains the predefined PFs vector $\boldsymbol{a}_i$ for each phone i in the vocabulary, based on prior phonetic knowledge of phone composition. By computing the dot product between the extracted PF vector $\boldsymbol{x}_{PFs}$ and the predefined PF vector $\boldsymbol{a}_i$ for phone i, we obtain the logits of phone i given the current speech features. When calculating the dot product between $\boldsymbol{x}_{PFs}$ and the entire Signature Matrix $\boldsymbol{A}_{sig}$, we derive the phone logits across the vocabulary: $\boldsymbol{y}_{PFs} = \boldsymbol{A}_{sig} \times \tanh\left(\boldsymbol{F}_{PFs}(\boldsymbol{x})\right)$.

Hence, we do not train directly for PF targets but indirectly via the Signature matrix. This strategy imposes fewer constraints on PF prediction, since zero values in the Signature matrix now really express that we do not care about the predicted PF value, rather than forcing the model to output a zero value. Also, PFs are not always realized as the context-independent PFs would suggest so the actual PF values can adjust accordingly.

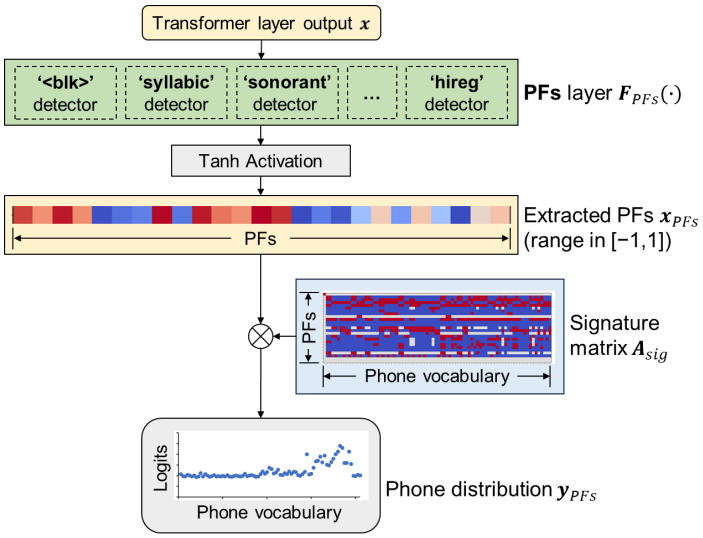


Figure 2. Proposed phone recognition using PFs as intermediate representation.

In the experiments, we also investigate the mutual complementation of conventional phone recognition and the PF method by combining the phone logits obtained from both the phone (**PHN**) and Phonological Features (**PFs**) layers: $\boldsymbol{y} = \boldsymbol{y}_{PHN} + \boldsymbol{y}_{PFs}$. (Logits from both methods are weighted equally, as a trainable weight ends up with a value close to unity without performance improvement.) Figure 3 provides flow charts for using only the **PF** layer and for using both layers.

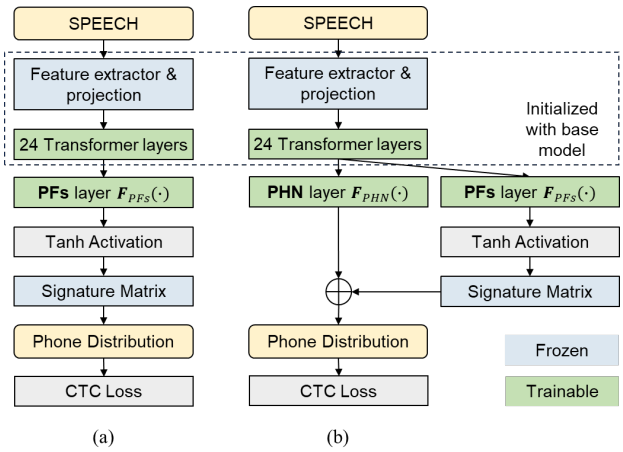


Figure 3. Wav2Vec-based phone recognition models using different output layers. (**a**) Phonological Features (**PFs**) method; (**b**) combination of **PFs** and phone (**PHN**) method.

5.3. Experiment Details

In the experiment, we reconsider the 'PFs composition' of the <*blk*> token $\boldsymbol{a}_{<blk>}$, which has a value of '1' in the first element and zeros in all other positions. Thus, the maximum dot product achievable for this token is '1', provided the extracted PF vector $\boldsymbol{x}_{PFs}$ is accurate. However, for other phones in the vocabulary, there is more than one non-zero element. If $\boldsymbol{x}_{PFs}$ for the <*blk*> is not sufficiently accurate, the resulting phone distribution may assign higher probabilities to other phones. To address this, in our experiment, we set the first element of $\boldsymbol{a}_{<blk>}$ to '8', which was optimal on the validation set. This adjustment aims to improve the recognition of the <*blk*> token. Moreover, the other phones in the phone vocabulary also have varying numbers of non-zero elements, leading to different maximum dot product values when PFs are accurately inferred. However, these differences are small enough to not cause confusion. Normalization of the phone logits based on

the number of non-zero elements turned out to have hardly any impact on recognition accuracy.

For the training procedure, we follow the same settings as Section 4, using **D_indm** and **D_extgt** as the training data and conducting experiences in the very-low-resource and cross-lingual settings. Both the **PHN** and **PFs** layers are single dense layers, while the **PFs** method has fewer trainable parameters as it needs to output only 25-dimensional PF vectors. However, this reduction in parameters is minimal since we jointly train the entire model.

5.4. Results

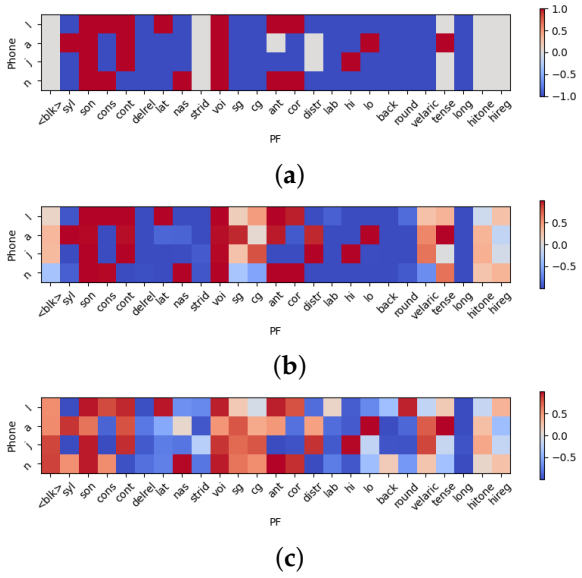
In this section, we aim to show the benefits of using PFs in recognizing dysarthric speech in very-low-resource and cross-lingual scenarios via experimental results, revealing their advantages in unseen word and phone recognition.

5.4.1. Qualitative Evaluation of PFs

As introduced before, PFs are used as an intermediate representation to achieve phone recognition. If the PFs of a speech frame are accurately inferred, the dot product of the corresponding phone for this frame will be the largest among the phone vocabulary, leading to successful recognition.

To illustrate how well PFs are inferred and utilized during phone recognition, we plot the PFs of a canonical and a dysarthric speech sample extracted from the PF models trained with canonical and dysarthric data, respectively. The chosen canonical speech sample is from the validation set of a canonical speaker *CF02* of **UASPEECH**, while the dysarthric speech sample is from the test set of a dysarthric speaker *M10* of **UASPEECH**. Both samples contain the word '*line*', which has phonetic transcription '[*l, a, j, n*]' (via *Epitran*), and the phone sequences are accurately recognized using the corresponding PFs model.

Figure 4 shows plots of the extracted PFs from the canonical (Figure 4b) and dysarthric (Figure 4c) speech samples, alongside the standard PFs for reference (Figure 4a). In the plots, the *x*-axis represents each PF, while the *y*-axis is the phone sequence.



(**a**)

(**b**)

(**c**)

Figure 4. Plots of standard and extracted phonological features of the phone sequence '*[l, a, j, n]*'. (**a**) Standard PFs of the phone sequence; (**b**) extracted PFs of a canonical speech sample inferred by a PFs model trained with canonical data; (**c**) extracted PFs of a dysarthric speech sample inferred by a PFs model trained with dysarthric data.

The plots show that the extracted PFs match the ideal values for both types of speech, with smaller absolute values for dysarthric speech. This less confident estimation reflects the poor pronunciation in dysarthric speech.

Most misestimations happen when a zero value is expected, but this will not lead to inaccurate recognition as explained above. Non-zero PF elements can also be extracted incorrectly. The PF '*sg*' (spread glottis) should be '−1' for all four phones but are often estimated as positive in both canonical and dysarthric examples. These wrong estimations will also not lead to a higher dot product for other phones as these PFs are always '−1' across the whole phone vocabulary. Another mistake in the dysarthric example is the '*round*' PF of phone '*l*' being extracted as positive instead of negative. Despite this, phone recognition remains accurate as no other phones in the vocabulary have the same PFs as '*l*' but differ in the '*round*' PF. Such conditions illustrate that while accurate estimation of the entire 25-dimensional PF vector is expected for phone recognition, certain inaccuracies do not necessarily lead to errors. This robustness may contribute to the **PFs** method's ability to mitigate overfitting.

5.4.2. Very-Low-Resource Setting

To evaluate the **PFs** method in the very-low-resource setting, we used selected training data from **D_indm**. Table 7 presents the averaged PER results across all dysarthric databases, comparing different recognition methods, data selection methods, and amounts of training data. In the table, **PHN** and **PFs** denote the utilization of the respective layer alone, whereas **Combi.** represents using a combination of both layers.

Table 7. Averaged PER across all dysarthric databases, using different recognition methods, data selection methods, and amounts.

EXP	Data Selection	Method	Training Data Amount		
			10%	**30%**	**90%**
E9	**Random**	**PHN**	0.326 †	0.247	0.205
E21		**PFs**	0.269 *	0.225	0.196
E22		**Combi.**	**0.263** *†	**0.216** *†	**0.184** *†
E16	**Ordered**	**PHN**	0.690 †	0.478 †	0.258
E23		**PFs**	**0.445** *	**0.355** *	0.248
E24		**Combi.**	0.459 *†	**0.352** *	**0.240** *†

Bold numbers: Best PER(s) within statistical significance level in each case; *: p-value ≤ 0.05 compared to corresponding **PHN** method; †: p-value ≤ 0.05 compared to corresponding **PFs** method.

Across all data selections and training data amounts, the **PFs** method consistently outperforms the **PHN** method, highlighting its effectiveness in capturing word- and phone-independent information and its ability to generalize to unseen words and phones in very-low-resource scenarios. However, it is noteworthy that as training data increase, the advantage of the **PFs** layer diminishes. With 90% training data, the improvement in PER brought by the **PFs** layer becomes statistically insignificant.

This trend may stem from the fact that when sufficient training data are available, the unique strengths of **PFs** in recognizing unseen phones or overcoming vocabulary over-fitting vanish as vocabulary mismatch is resolved. The reduced impact of the **PFs** layer is also evident with the **Random** selection method, where the training data better covers the vocabulary. Additionally, with sufficient training data, the limitations of PF-based recognition show up. It requires precise detection of each phonological feature for accurate phone recognition, which could be challenging, particularly for dysarthric speakers. Alternatively, while the PF detection may still be accurate, the Signature matrix

may not accurately model phone realization for dysarthric speech. Due to the lack of PF ground truth, it is difficult to discriminate between these hypotheses.

When the **PFs** layer is combined with the **PHN** layer for speech recognition, the combined layer performs better than using only **PFs** layer in most cases. Especially when sufficient training data are available, the performance advantage proves the supplementary effect between the two recognition layers. However, the **Combi.** layer gives higher PER results or non-significantly lower PER results with limited training vocabulary (**Ordered** selection and 10%/30% data). In this scenario, the advantage of the **PFs** layer may be negatively influenced by the **PHN** layer, which is more sensitive to the mismatch between the training and test vocabulary. Moreover, compared to the results **E11/17** using the **PHN** layer and the vocabulary-informed base model in Table 5, the **Combi.** layer achieves similar recognition performance without needing a suitable vocabulary-informed base model.

To better understand the impact of using PFs in recognizing dysarthric speech, we categorize all dysarthric speakers based on their intelligibility levels (High/Medium/Low) using the information provided in the databases or estimated when such information is unavailable. The PER results averaged across all speakers in each intelligibility level are presented in Table 8, focusing exclusively on experiments of **Ordered** selection method for simplicity (**E16,23,24**). The statistical results here are calculated across speakers from each intelligibility level.

Table 8. Averaged PER across speakers from each intelligibility level of different recognition methods using training data selected by **Ordered** method and for different data amounts.

Intelligibility Level	Method	Training Data Amount		
		10%	**30%**	**90%**
High	**PHN**	0.633 †	0.404 †	0.116
	PFs	**0.359** *	**0.241** *	0.108
	Combi.	0.387 *†	**0.242** *	**0.102** *†
Medium	**PHN**	0.622 †	0.447†	0.191
	PFs	**0.378** *	**0.298** *	0.183
	Combi.	0.406 *†	**0.300** *	**0.174** *†
Low	**PHN**	0.764 †	0.631 †	0.485
	PFs	**0.642** *	**0.570** *	0.483
	Combi.	**0.648** *	**0.565** *	**0.469** *†

Bold numbers: Best PER(s) within statistical significance level in each case; *: p-value ≤ 0.05 compared to corresponding **PHN** method; †: p-value ≤ 0.05 compared to corresponding **PFs** method.

The table reveals a consistent benefit of the **PFs** method, in line with previous findings. However, the advantages are most significant among speakers with high and medium intelligibility levels, nearly diminishing in the low intelligibility groups. And this disparity in performance becomes more pronounced with limited training data. These observations suggest that either accurately recognizing PFs from speakers with low intelligibility levels is challenging, or the mapping of PFs to phones becomes inaccurate, both resulting in inaccurate phone recognition. Conversely, for speakers with good intelligibility levels, precise PF recognition and mapping to phones via the Signature matrix appears to be achievable, thereby yielding superior performance compared to other methods. This observation confirms our previous hypothesis about the limitations of the **PFs** method.

5.4.3. Cross-Lingual Setting

We also demonstrate the effectiveness of PFs in phone recognition in the cross-lingual setting where the model is trained with **D_extgt** data and tested with the target language

dataset(s). In Table 9, we present the PER results averaged across all datasets and across speakers at each intelligibility level.

Table 9. PER averaged across all datasets and across speakers at each intelligibility level, using different recognition methods.

EXP	Method	Averaged PER Across			
		Datasets	Intelligibility Level		
			High	Medium	Low
E19	**PHN (E19)**	0.840 †	0.754 †	0.774 †	0.968 †
E25	**PFs**	**0.801** *	**0.709** *	**0.733** *	**0.913** *
E26	**Combi.**	0.852 *†	0.769 *†	0.786 *†	0.981 †

Bold numbers: Best PER(s) within statistical significance level in each case; *: p-value ≤ 0.05 compared to corresponding **PHN** method; †: p-value ≤ 0.05 compared to corresponding **PFs** method.

The **PFs** layer demonstrates a clear advantage in the cross-lingual scenario across all speaker intelligibility levels and in the average results across all datasets when compared to the **PHN** layer. However, combining the two layers diminishes the benefits provided by PF detection, resulting in the highest recognition error. The superior recognition results of the **PFs** layer underscore its effectiveness in recognizing unseen phones and mitigating vocabulary fitting, particularly when contrasted with the phone target. This advantage is consistent across different speaker groups. Nonetheless, despite these improvements, the recognition method still fails to provide practical phone transcriptions due to the relatively high-PER results.

5.5. Discussion

In our experiments, PFs have shown their advantages in enhancing vocabulary-mismatched dysarthric speech recognition without using a suitable vocabulary-informed base model. However, it is important to note that the base model [61] used for training in this section was originally pretrained with phones as the recognition targets. The speech features it outputs should contain well-trained phone knowledge. As a result, there exists the possibility that the extracted PFs are inferred from this phone information in the speech features rather than from PFs-related information directly contained in the base model. If this is the case, the trained model would lack the ability to generalize to unseen phones, as it would not have the relevant information to infer their PFs. Although we observe improved performance with the PFs method using mismatched data, where there is both word and phone vocabulary mismatch between training and test data, it is challenging to demonstrate its effectiveness in recognizing unseen phones due to the lack of reliable aligned phone transcription in the datasets used.

Previous research [19] has demonstrated the effectiveness of using PFs for recognizing unseen phones, achieving higher accuracy compared to baseline models. This work did not involve a pretrained model with phones as the target, thereby avoiding certain concerns associated with such models. To further verify how PFs are extracted from a base model pretrained with phones as the recognition target, we plan to conduct additional experiments using reliable phone-aligned data to analyze the performance of unseen-phone recognition in future work.

6. Conclusions and Future Work

Dysarthric speech recognition systems significantly enhance the quality of life for individuals with dysarthria. For practical use, these systems must generalize effectively to unseen speakers and handle words or phones not encountered during training, ad-

dressing limitations of small-vocabulary datasets. These requirements necessitate robust performance in speaker-independent and vocabulary-mismatched scenarios.

This work focuses on improving model performance under these two scenarios through optimized training strategies. For the speaker-independent case, we systematically explored training methods to maximize knowledge transfer from available data to unseen dysarthric speech. Unlike prior research relying on limited datasets, we utilized all available data, categorized as rich-resource canonical speech, in-domain canonical speech, and dysarthric speech. Experimental results showed the optimal strategy involves training the model on rich-resource canonical speech, followed by retraining on in-domain canonical speech, and finally adapting to in-domain dysarthric speech when available.

For vocabulary-mismatched scenarios, an under-explored area, we examined model performance in very-low-resource and cross-lingual settings. Results revealed the limitations of conventional training approaches, especially when models lacked sufficient vocabulary knowledge during initialization.

To address these challenges, we proposed using phonological features (PFs) as intermediate representations for phone recognition. Experimental findings demonstrated that PFs effectively mitigate vocabulary mismatches and overfitting while enabling recognition of unseen words and phones due to their independence from specific phones. Combining PFs with traditional methods achieved superior recognition performance in very-low-resource scenarios when adequate training data was available.

Future work will focus on enhancing the use of PFs and improving the Signature matrix. Currently derived from canonical speech, the Signature matrix does not fully account for mispronounced dysarthric phones. Experiments showed limited benefits from adapting the matrix to dysarthric data during training. To address this, we aim to develop advanced adaptation methods and create personalized Signature matrices tailored to individual speakers.

Another focus is advancing cross-lingual dysarthric speech recognition to support speakers of rare languages lacking adequate training data. While our experiments in cross-lingual settings highlighted challenges, including poor recognition accuracy and vocabulary overfitting, future work will focus on mitigating these issues. This includes analyzing phone-phonological feature confusion and leveraging data augmentation techniques to improve performance in these under-explored scenarios.

Author Contributions: Conceptualization, J.Q. and H.V.h.; methodology, J.Q.; software, J.Q.; validation, J.Q.; formal analysis, J.Q.; investigation, J.Q.; resources, H.V.h.; data curation, J.Q.; writing—original draft preparation, J.Q.; writing—review and editing, J.Q. and H.V.h.; visualization, J.Q.; supervision, H.V.h.; project administration, H.V.h.; funding acquisition, H.V.h. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by KU Leuven Special Research Fund grant number C24M/22/025 and the Flemish Government under the "Onderzoeksprogramma Artificiële Intelligentie (AI) Vlaanderen" program.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: No new data were created or analyzed in this study. Data sharing is not applicable to this article.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

ASR	Automatic Speech Recognition
DSR	Dysarthric Speech Recognition
OOV	Out-of-Vocabulary
PF	Phonological Feature
CTC	Connectionist Temporal Classification
PER	Phone Error Rate
IPA	International Phonetic Alphabet

References

1. Enderby, P. Disorders of communication: Dysarthria. *Handb. Clin. Neurol.* **2013**, *110*, 273–281. [PubMed]
2. Qian, Z.; Xiao, K.; Yu, C. A survey of technologies for automatic Dysarthric speech recognition. *EURASIP J. Audio Speech Music. Process.* **2023**, *2023*, 48. [CrossRef]
3. Dhanjal, A.S.; Singh, W. A comprehensive survey on automatic speech recognition using neural networks. *Multimed. Tools Appl.* **2024**, *83*, 23367–23412. [CrossRef]
4. Almadhor, A.; Irfan, R.; Gao, J.; Saleem, N.; Rauf, H.T.; Kadry, S. E2E-DASR: End-to-end deep learning-based dysarthric automatic speech recognition. *Expert Syst. Appl.* **2023**, *222*, 119797. [CrossRef]
5. Shahamiri, S.R.; Lal, V.; Shah, D. Dysarthric speech transformer: A sequence-to-sequence dysarthric speech recognition system. *IEEE Trans. Neural Syst. Rehabil. Eng.* **2023**, *31*, 407–3416. [CrossRef]
6. Mahum, R.; El-Sherbeeny, A.M.; Alkhaledi, K.; Hassan, H. Tran-DSR: A hybrid model for dysarthric speech recognition using transformer encoder and ensemble learning. *Appl. Acoust.* **2024**, *222*, 110019. [CrossRef]
7. Hu, S.; Xie, X.; Geng, M.; Jin, Z.; Deng, J.; Li, G.; Wang, Y.; Cui, M.; Wang, T.; Meng, H.; et al. Self-supervised ASR Models and Features For Dysarthric and Elderly Speech Recognition. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2024**, *32*, 3561–3575. [CrossRef]
8. Lin, Y.; Wang, L.; Dang, J.; Minematsu, N. Exploring Pre-trained Speech Model for Articulatory Feature Extraction in Dysarthric Speech Using ASR. In Proceedings of the Interspeech, Kos, Greece, 1–5 September 2024; pp. 4598–4602.
9. Jaddoh, A.; Loizides, F.; Rana, O. Interaction between people with dysarthria and speech recognition systems: A review. *Assist. Technol.* **2023**, *35*, 330–338. [CrossRef]
10. Takashima, Y.; Takashima, R.; Takiguchi, T.; Ariki, Y. Knowledge transferability between the speech data of persons with dysarthria speaking different languages for dysarthric speech recognition. *IEEE Access* **2019**, *7*, 164320–164326. [CrossRef]
11. Takashima, R.; Takiguchi, T.; Ariki, Y. Two-step acoustic model adaptation for dysarthric speech recognition. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 6104–6108.
12. Vásquez-Correa, J.C.; Rios-Urrego, C.D.; Arias-Vergara, T.; Schuster, M.; Rusz, J.; Nöth, E.; Orozco-Arroyave, J.R. Transfer learning helps to improve the accuracy to classify patients with different speech disorders in different languages. *Pattern Recognit. Lett.* **2021**, *150*, 272–279. [CrossRef]
13. Leivaditi, S.; Matsushima, T.; Coler, M.; Nayak, S.; Verkhodanova, V. Fine-Tuning Strategies for Dutch Dysarthric Speech Recognition: Evaluating the Impact of Healthy, Disease-Specific, and Speaker-Specific Data. In Proceedings of the Interspeech, Kos, Greece, 1–5 September 2024; pp. 1295–1299.
14. Baevski, A.; Zhou, Y.; Mohamed, A.; Auli, M. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 12449–12460.
15. Hu, S.; Xie, X.; Jin, Z.; Geng, M.; Wang, Y.; Cui, M.; Deng, J.; Liu, X.; Meng, H. Exploring self-supervised pre-trained asr models for dysarthric and elderly speech recognition. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, 4–10 June 2023; IEEE: Piscataway, NJ, USA, 2023; pp. 1–5.
16. Javanmardi, F.; Kadiri, S.R.; Alku, P. Exploring the impact of fine-tuning the wav2vec2 model in database-independent detection of dysarthric speech. *IEEE J. Biomed. Health Inform.* **2024**, *28*, 4951–4962. [CrossRef]
17. Chomsky, N.; Halle, M. *The Sound Pattern of English*; Harper & Row: Manhattan, NY, USA, 1968.
18. King, S.; Taylor, P. Detection of phonological features in continuous speech using neural networks. *Comput. Speech Lang.* **2000**, *14*, 333–353. [CrossRef]
19. Li, X.; Dalmia, S.; Mortensen, D.; Li, J.; Black, A.; Metze, F. Towards zero-shot learning for automatic phonemic transcription. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 9–11 February 2020; Volume 34, pp. 8261–8268.

20. Xiong, F.; Barker, J.; Yue, Z.; Christensen, H. Source domain data selection for improved transfer learning targeting dysarthric speech recognition. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 7424–7428.
21. Mariya Celin, T.; Vijayalakshmi, P.; Nagarajan, T. Data Augmentation Techniques for Transfer Learning-Based Continuous Dysarthric Speech Recognition. *Circuits Syst. Signal Process.* **2023**, *42*, 601–622. [CrossRef]
22. Rathod, S.; Charola, M.; Patil, H.A. Transfer learning using whisper for dysarthric automatic speech recognition. In *Speech and Computer, Proceedings of the 25th International Conference, SPECOM 2023, Dharwad, India, 29 November–2 December 2023*; Springer: Cham, Switzerland, 2023; pp. 579–589.
23. Yılmaz, E.; Mitra, V.; Sivaraman, G.; Franco, H. Articulatory and bottleneck features for speaker-independent ASR of dysarthric speech. *Comput. Speech Lang.* **2019**, *58*, 319–334. [CrossRef]
24. Lin, Y.; Wang, L.; Li, S.; Dang, J.; Ding, C. Staged Knowledge Distillation for End-to-End Dysarthric Speech Recognition and Speech Attribute Transcription. In Proceedings of the Interspeech 2020, Shanghai, China, 25–29 October 2020; pp. 4791–4795.
25. Wang, S.; Zhao, J.; Sun, S. Effective Domain Adaptation for Robust Dysarthric Speech Recognition. In *Neural Information Processing, Proceedings of the 30th International Conference, ICONIP 2023, Changsha, China, 20–23 November 2023*; Springer: Singapore, 2023; pp. 62–73.
26. Wang, S.; Zhao, S.; Zhou, J.; Kong, A.; Qin, Y. Enhancing Dysarthric Speech Recognition for Unseen Speakers via Prototype-Based Adaptation. In Proceedings of the Interspeech 2024, Kos, Greece, 1–5 September 2024; pp. 1305–1309.
27. Wang, D.; Yu, J.; Wu, X.; Sun, L.; Liu, X.; Meng, H. Improved end-to-end dysarthric speech recognition via meta-learning based model re-initialization. In Proceedings of the 12th International Symposium on Chinese Spoken Language Processing (ISCSLP), Hong Kong, 24–27 January 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 1–5.
28. Qi, J.; Van hamme, H. Parameter-efficient dysarthric speech recognition using adapter fusion and householder transformation. In Proceedings of the Interspeech 2023, Dublin, Ireland, 20–24 August 2023; pp. 151–155.
29. Hernandez, A.; Pérez-Toro, P.A.; Nöth, E.; Orozco-Arroyave, J.R.; Maier, A.; Yang, S.H. Cross-lingual self-supervised speech representations for improved dysarthric speech recognition. In Proceedings of the Interspeech 2022, Incheon, Republic of Korea, 18–22 September 2022; pp. 51–55.
30. Takashima, Y.; Takiguchi, T.; Ariki, Y. End-to-end dysarthric speech recognition using multiple databases. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 2–17 May 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 6395–6399.
31. Baskar, M.K.; Herzig, T.; Nguyen, D.; Diez, M.; Polzehl, T.; Burget, L.; Černockỳ, J. Speaker adaptation for Wav2vec2 based dysarthric ASR. In Proceedings of the Interspeech 2022, Incheon, Republic of Korea, 18–22 September 2022; pp. 3403–3407.
32. Hu, S.; Xie, X.; Geng, M.; Cui, M.; Deng, J.; Li, G.; Wang, T.; Liu, X.; Meng, H. Exploiting cross-domain and cross-lingual ultrasound tongue imaging features for elderly and dysarthric speech recognition. In Proceedings of the Interspeech 2023, Dublin, Ireland, 20–24 August 2023; pp. 2313–2317.
33. Stuker, S.; Metze, F.; Schultz, T.; Waibel, A. Integrating multilingual articulatory features into speech recognition. In Proceedings of the Eighth European Conference on Speech Communication and Technology, Geneva, Switzerland, 1–4 September 2003.
34. Zhan, Q.; Motlicek, P.; Du, S.; Shan, Y.; Ma, S.; Xie, X. Cross-lingual Automatic Speech Recognition Exploiting Articulatory Features. In Proceedings of the Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Lanzhou, China, 8–21 November 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 1912–1916.
35. Zhan, Q.; Xie, X.; Hu, C.; Zuluaga-Gomez, J.; Wang, J.; Cheng, H. Domain-Adversarial Based Model with Phonological Knowledge for Cross-Lingual Speech Recognition. *Electronics* **2021**, *10*, 3172. [CrossRef]
36. Zhu, C.; An, K.; Zheng, H.; Ou, Z. Multilingual and crosslingual speech recognition using phonological-vector based phone embeddings. In Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), Cartagena, Colombia, 13–17 December 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 1034–1041.
37. Lee, J.; Mimura, M.; Kawahara, T. Embedding Articulatory Constraints for Low-resource Speech Recognition Based on Large Pre-trained Model. In Proceedings of the Interspeech 2023, Dublin, Ireland, 20–24 August 2023; pp. 1394–1398.
38. Yen, H.; Siniscalchi, S.M.; Lee, C.H. Boosting End-to-End Multilingual Phoneme Recognition Through Exploiting Universal Speech Attributes Constraints. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Seoul, Republic of Korea, 14–19 April 2024; IEEE: Piscataway, NJ, USA, 2024; pp. 11876–11880.
39. Tong, S.; Garner, P.N.; Bourlard, H. Fast Language Adaptation Using Phonological Information. In Proceedings of the Interspeech 2018, Hyderabad, India, 2–6 September 2018; pp. 2459–2463.
40. Li, S.; Ding, C.; Lu, X.; Shen, P.; Kawahara, T.; Kawai, H. End-to-End Articulatory Attribute Modeling for Low-Resource Multilingual Speech Recognition. In Proceedings of the Interspeech 2019, Graz, Austria, 15–19 September 2019; pp. 2145–2149.
41. Wong, K.H.; Yeung, Y.T.; Wong, P.C.; Levow, G.A.; Meng, H. Analysis of dysarthric speech using distinctive feature recognition. In Proceedings of the 6th Workshop on Speech and Language Processing for Assistive Technologies (SLPAT), Dresden, Germany, 11 September 2015; pp. 86–90.

42. Wong, K.H.; Yeung, W.S.; Yeung, Y.T.; Meng, H. Exploring articulatory characteristics of Cantonese dysarthric speech using distinctive features. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Shanghai, China, 20–25 March 2016; IEEE: Piscataway, NJ, USA, 2016; pp. 6495–6499.
43. Jiao, Y.; Berisha, V.; Liss, J. Interpretable phonological features for clinical applications. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, USA, 5–9 March 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 5045–5049.
44. Lin, Y.; Wang, L.; Dang, J.; Li, S.; Ding, C. End-to-End articulatory modeling for dysarthric articulatory attribute detection. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Virtual, 4–9 May 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 7349–7353.
45. Kim, M.J.; Kim, Y.; Kim, H. Automatic intelligibility assessment of dysarthric speech using phonologically-structured sparse linear model. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2015**, *23*, 694–704. [CrossRef]
46. Liu, Y.; Penttilä, N.; Ihalainen, T.; Lintula, J.; Convey, R.; Räsänen, O. Language-independent approach for automatic computation of vowel articulation features in dysarthric speech assessment. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2021**, *29*, 2228–2243. [CrossRef]
47. Wong, K.H.; Meng, H.M.L. Automatic Analyses of Dysarthric Speech based on Distinctive Features. *APSIPA Trans. Signal Inf. Process.* **2023**, *12*, e18. [CrossRef]
48. Rudzicz, F. Phonological features in discriminative classification of dysarthric speech. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Taipei, Taiwan, 19–24 April 2009; IEEE: Piscataway, NJ, USA, 2009; pp. 4605–4608.
49. Rudzicz, F. Applying discretized articulatory knowledge to dysarthric speech. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Taipei, Taiwan, 19–24 April 2009; IEEE: Piscataway, NJ, USA, 2009; pp. 4501–4504.
50. Xiong, F.; Barker, J.; Christensen, H. Deep learning of articulatory-based representations and applications for improving dysarthric speech recognition. In Proceedings of the Speech Communication; 13th ITG-Symposium, Oldenburg, Germany, 10–12 October 2018; VDE: Berlin, German, 2018; pp. 1–5.
51. Hsieh, I.; Wu, C.H. Dysarthric Speech Recognition Using Curriculum Learning and Articulatory Feature Embedding. In Proceedings of the Interspeech 2024, Kos, Greece, 1–5 September 2024; pp. 1300–1304.
52. Lin, Y.; Wang, L.; Dang, J.; Li, S.; Ding, C. Disordered speech recognition considering low resources and abnormal articulation. *Speech Commun.* **2023**, *155*, 103002. [CrossRef]
53. Ardila, R.; Branson, M.; Davis, K.; Kohler, M.; Meyer, J.; Henretty, M.; Morais, R.; Saunders, L.; Tyers, F.; Weber, G. Common Voice: A Massively-Multilingual Speech Corpus. In Proceedings of the Twelfth Language Resources and Evaluation Conference, Marseille, France, 11–16 May 2020; pp. 4218–4222.
54. Rudzicz, F.; Namasivayam, A.K.; Wolff, T. The TORGO database of acoustic and articulatory speech from speakers with dysarthria. *Lang. Resour. Eval.* **2012**, *46*, 523–541. [CrossRef]
55. Kim, H.; Hasegawa-Johnson, M.; Perlman, A.; Gunderson, J.; Huang, T.S.; Watkin, K.; Frame, S. Dysarthric speech database for universal access research. In Proceedings of the Ninth Annual Conference of the International Speech Communication Association, Brisbane, Australia, 22–26 September 2008.
56. Ons, B.; Gemmeke, J.F.; Van hamme, H. The self-taught vocal interface. *EURASIP J. Audio, Speech, Music. Process.* **2014**, *2014*, 43. [CrossRef]
57. Van Nuffelen, G.; De Bodt, M.; Middag, C.; Martens, J.P. *Dutch Corpus of Pathological and Normal Speech (Copas)*; Technical Report; Antwerp University Hospital and Ghent University: Gent, Belgium, 2009.
58. Orozco-Arroyave, J.R.; Arias-Londoño, J.D.; Vargas-Bonilla, J.F.; Gonzalez-Rátiva, M.C.; Nöth, E. New Spanish speech corpus database for the analysis of people suffering from Parkinson's disease. In Proceedings of the International Conference on Language Resources and Evaluation (LREC), Reykjavik, Iceland, 26–31 May 2014; pp. 342–347.
59. Turrisi, R.; Braccia, A.; Emanuele, M.; Giulietti, S.; Pugliatti, M.; Sensi, M.; Fadiga, L.; Badino, L. EasyCall corpus: A dysarthric speech dataset. In Proceedings of the Interspeech, Brno, Czechia, 30 August–3 September 2021; pp. 41–45.
60. TA, M.C.; Nagarajan, T.; Vijayalakshmi, P. Dysarthric speech corpus in tamil for rehabilitation research. In Proceedings of the IEEE Region 10 Conference (TENCON), Singapore, 22–25 November 2016; IEEE: Piscataway, NJ, USA, 2016; pp. 2610–2613.
61. Xu, Q.; Baevski, A.; Auli, M. Simple and effective zero-shot cross-lingual phoneme recognition. In Proceedings of the Interspeech 2022, Incheon, Republic of Korea, 18–22 September 2022; pp. 2113–2117.
62. Mortensen, D.R.; Dalmia, S.; Littell, P. Epitran: Precision G2P for many languages. In Proceedings of the International Conference on Language Resources and Evaluation (LREC), Miyazaki, Japan, 7–12 May 2018.
63. Graves, A.; Fernández, S.; Gomez, F.; Schmidhuber, J. Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. In Proceedings of the 23rd International Conference on Machine Learning (ICML), Pittsburgh, PA, USA, 25–29 June 2006; pp. 369–376.

64. Wilcoxon, F. Individual comparisons by ranking methods. In *Breakthroughs in Statistics: Methodology and Distribution*; Springer: New York, NY, USA, 1992; pp. 196–202.
65. Mortensen, D.R.; Littell, P.; Bharadwaj, A.; Goyal, K.; Dyer, C.; Levin, L. Panphon: A resource for mapping IPA segments to articulatory feature vectors. In Proceedings of the 26th International Conference on Computational Linguistics: Technical Papers (COLING), Osaka, Japan, 11–16 December 2016; pp. 3475–3484.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.