PDAssess: A Privacy-preserving Free-speech based Parkinson's Disease Daily Assessment System

Baichen Yang[1], Qingyong Hu[1], Wentao Xie[1], Xinchen Wang[2], Wei Luo[2]*, Qian Zhang[1]*
{byangak,qhuag,wxieaj,qianzh}@cse.ust.hk
{wang_xc,luoweirock}@zju.edu.cn
[1]Department of Computer Science and Engineering, Hong Kong University of Science and Technology, Hong Kong
[2]Department of Neurology, The Second Affiliated Hospital, Zhejiang University School of Medicine, Zhejiang, China

ABSTRACT

In-time disease assessment is essential to better customize the medication scheme and improve the quality of life for chronic diseases like Parkinson's disease (PD). Toward the inconvenience problem in current clinical assessment practice, mobile sensing solutions based on detecting Parkinson's vocal changes are proposed. However, current solutions either can only achieve binary disease detection task or require patients to perform specific speaking tasks, which is not effective and practical for disease stage assessment in daily scenario. Moreover, most of existing solutions do not take speech privacy into consideration. In this work, we present PDAssess, a free speech-based daily assessment system that can perform 4-stage Parkinson's disease assessment in a privacy-preserving manner. We observe that current solutions did not fully leverage the rich information embedded in free speech due to the *linguistic content variations*, and therefore leverage a pre-trained automatic speech recognition (ASR) model to achieve a content variation-aware feature-extraction. In order to distinguish subtle stage-wise differences, we design a novel attention-based neural network architecture with a customized loss function for disease assessment task. Towards the potential privacy leakage problem, we design a Split Learning-based framework with pseudo-labeling and local domain adversarial training to better preserve speech content privacy. We collaborate with a medical center and evaluate the performance of PDAssess on real-world speech data collected from 50 PD subjects and 50 healthy subjects. The evaluation result shows that PDAssess can perform 4-stage PD assessment with an average person-wise F1 score of 89.1% and voice sample-wise F1 score of 75.1%.

CCS CONCEPTS

• **Human-centered computing** → **Ubiquitous and mobile computing**.

KEYWORDS

Parkinson's disease, Healthcare, Mobile Sensing

*Corresponding authors.

SenSys '23, November 12–17, 2023, Istanbul, Turkiye

ACM ISBN 979-8-4007-0414-7/23/11…$15.00
https://doi.org/10.1145/3625687.3625805

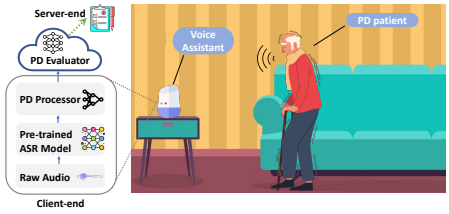


Figure 1: The application scenario of PDAssess system.

ACM Reference Format:
Baichen Yang[1], Qingyong Hu[1], Wentao Xie[1], Xinchen Wang[2], Wei Luo[2]*, Qian Zhang[1]*. 2023. PDAssess: A Privacy-preserving Free-speech based Parkinson's Disease Daily Assessment System. In *The 21st ACM Conference on Embedded Networked Sensor Systems (SenSys '23), November 12–17, 2023, Istanbul, Turkiye.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3625687.3625805

1 INTRODUCTION

Parkinson's disease (PD) is currently the second-most common progressive nervous system disorder that has already affected more than 6 million people worldwide and is expected to rise to 14.2 million by 2040 [16, 60]. To receive more effective and personalized treatment, PD patients need to constantly measure the disease severity by seeking medical consultation [1, 12, 15].

Currently clinicians use well-designed scale systems to measure PD severity, including the Hoehn and Yahr (H&Y) scale system [32] and the unified Parkinson's disease rating scale (UPDRS) system [53]. These scale systems require the patients to perform a series of tasks or fill in multiple questionnaires for the clinicians' evaluation. It is rather laborious for patients to visit the hospital for assessment constantly. And the shortage of professional clinicians increases assessment inconvenience. Therefore, to achieve better and more convenient Parkinson's disease management, an in-home daily Parkinson's disease assessment system is desirable.

With the help of mobile sensing technologies, research efforts have been put into automating PD assessment in daily life. Research works [4, 7, 37, 41, 76] have been proposed to enable motor-based daily PD assessment. However, these approaches either require patients to wear or use specially designed hardware [4, 7, 37], or require performing specific tasks [41, 76], leading to inconvenience in daily assessment. In addition, most motor symptoms will only be developed after more than 70% dopaminergic neurons are damaged [23], making it unsuitable for prodromal-stage PD assessment. More recently, there are other assessment methods leveraging non-motor biomarkers, such as breathing [78]. Though inspiring as it sounds, specially designed hardware device is still required. In contrast,

voice-based PD assessment is more suitable for daily scenarios and all-stage evaluation, as vocal impairment of PD appears early in the disease progression [67], and voice-based solutions can be achieved by using commercial off-the-shelf (COTS) microphone only.

Recently, many research efforts have been put into voice-based PD assessment [2, 3, 8, 10, 22, 36, 52, 63, 81]. However, current voice-based solutions are not effective enough for practical daily assessment owing to the following two limitations. Firstly, most of the recent works [2, 10, 22, 36, 52, 81] target PD detection instead of fine-grained assessment, which cannot have long-term benefit for patients. Moreover, most of the works [2, 3, 8, 52, 63] require patients to perform specific speaking tasks, such as sustained vowel, diadochokinetic task or read a specific paragraph. On the contrary, disease stage assessment based on free speech (voice recordings with arbitrary linguistic content) is more favorable than task-specific methods due to following reasons. Firstly, task-specific vocal tests might impact the assessment accuracy in a long-term perspective as the patients will get familiar with the task content [9]. Such an issue will be relieved in free-speech scenarios. In addition, free-speech based system can operate in a purely passive manner while task-specific solutions will however lead to a burden to patients in the long round. However, comparing to task-specific methods, which can already achieve around 90% staging accuracy, existing free speech-based solutions fail to achieve a good performance in severity assessment [10, 72], and have not considered content privacy which is sensitive under free speech setting.

We observe that the performance bottleneck of current free-speech based PD assessment is the mixing of linguistic content information and disease-related features. We therefore try to obtain a *high-fidelity audio representation*, from which disease-related information can be better disentangled. We will further elaborate our observation in Sec. 3 in detail. We leverage a pre-trained ASR model, HuBERT [33], to obtain such a audio representation. The design rationale is that, deep learning-empowered ASR models have the ability to extract high-dimensional audio embedding, which will contain richer information for disease analysis. HuBERT is one typical ASR model which can not only extract linguistic features, but also capture non-linguistic information like speaker-related details, which is beneficial for PD assessment [33].

With such an observation, however, designing a free speech-based PD assessment system still faces three main challenges: (i) *Subtle stage-wise difference.* Even if we have high-fidelity speech representation, it is still challenging to disentangle disease-related information and perform subtle stage assessment, especially with stage label supervision only. (ii) *Skewed disease distribution.* Real-world distribution of disease stage is highly skewed [46]. Specifically, there will be fewer patients in severe stages owing to disease progression. And many mild-stage PD patients is undiagnosed, leading to a smaller population as well. Such a skewed distribution will increase assessment difficulty. (iii) *Potential privacy leakage.* Daily free speech may contain sensitive information that the user does not want to expose to the system. It is difficult to achieve a good assessment result without sacrificing user speech privacy.

In this paper, we propose PDAssess, a privacy-preserving free speech-based Parkinson's disease daily assessment system, whose application scenario is demonstrated in Fig. 1. PDAssess addresses the above three challenges with the following designs. Towards the first challenge, we design a squeeze-and-excitation (SE) attention-based model, to extract and analyze disease-related features for better disease assessment. We additionally adopt momentum contrastive loss [28] to train the model in a more effective manner with person-wise and stage-wise consistency awareness, as the extracted features from the same patient or different patients in the same stage should be similar. Towards the second challenge, we incorporate multi-class focal loss [44], forcing the model to focus more on the classes of a small population. Towards the third challenge, we utilize Split Learning (SL) scheme [73] and further customize a local adversarial-training technique [20] to prevent direct uploading of audio representations and remove sensitive linguistic information.

H&Y	Definition
Stage 1	Unilateral involvement only
Stage 2	Bilateral involvement without impairment of balance
Stage 3	Mild to moderate involvement; Some postural instability but physically independent Needs assistance to recover from pull test
Stage 4	Severe disability; Still able to walk or stand unassisted
Stage 5	Wheelchair bound or bedridden unless aided

Table 1: The stages of the H&Y scale. [32]

We enroll PD patients and healthy controls to evaluate our system in a home-like scenario. Specifically, we enroll 50 PD patients and 50 healthy people for evaluation. We utilize a COTS microphone device [59] to collect speech in a home setting. Results show that we can achieve a 90.4% person-wise F1 score and a 76.5% sample-wise F1 score for stage prediction with centralized training scheme, and an 89.1% person-wise F1 score and a 75.1% sample-wise F1 score with our privacy-preserving training scheme.

We highlight our contributions in the following three folds:

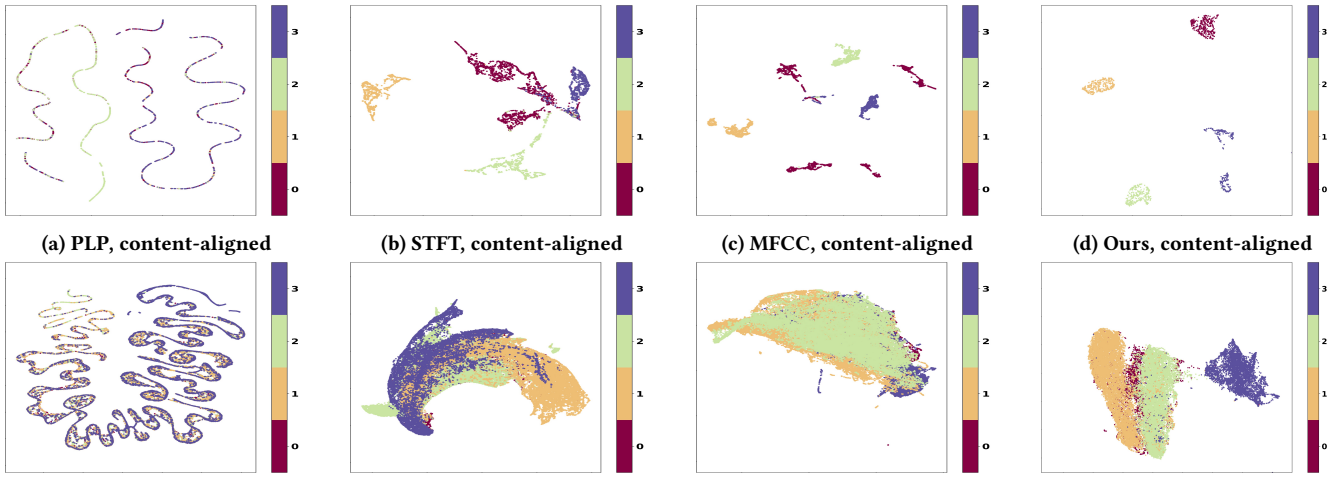
(i) We propose a novel hybrid neural network architecture for Parkinson's disease stage assessment based on free speech. We incorporate a pre-trained ASR model as a non-linear signal preprocessing method to extract high-fidelity audio representation. We further design a neural network powered by the attention mechanism to eliminate linguistic content variations and model the relationship between PD severity and human voice features.

(ii) We customize a privacy-preserving Split Learning framework for our task. We remove the sensitive information in the voice signals by embedding the adversarial training strategy. With such a design, our scheme can offload computation burdens to the cloud servers while protecting users' privacy.

(iii) We collaborate with a medical center to collect a large dataset with 50 patients and 50 healthy subjects and conduct extensive evaluations in real-world scenarios. The evaluation results validate that our system can not only achieve a high performance of around 90% person-wise F1 score and over 75% sample-wise F1 score on 4-stage PD assessment but also work in a privacy-preserving manner.

2 BACKGROUND

In current PD management, clinicians need to constantly and accurately evaluate PD patient's symptom severity before selecting the appropriate treatment. To track the disease progression, the clinicians mainly utilize the H&Y Scale [32] as the standard rating scale owing to the conciseness and powerful clinimetric performance for assessment. In this work we therefore adopt H&Y Scale as the



(a) PLP, content-aligned (b) STFT, content-aligned (c) MFCC, content-aligned (d) Ours, content-aligned

(e) PLP, free speech (f) STFT, free speech (g) MFCC, free speech (h) Ours, free speech

Figure 2: Sample UMAP visualizations of preprocessed audio data under content-aligned or free-speech setting with conventional techniques as well as our scheme with pre-trained ASR and feature extraction. The color legend on the right represents different PD stages (0 refers to healthy people).

disease progression metric. In H&Y Stage system, PD patients are classified into five stages (see in Table. 1), which can be categorized into mild (Stage 1-2), moderate (Stage 3-4) and severe (Stage 5). During the disease progression, the watershed is Stage 3, after which advanced interventions like surgeries will be needed [12].

Vocal impairment is an important biomarker for PD [65], which appears in early disease stages and becomes more severe when the disease progresses [67]. Recent clinical research [77] has demonstrated that some vocal impairment features, such as fundamental frequencies, have a strong correlation with H&Y scale. In this paper, we try to leverage such correlation to analyze PD severity.

3 PRELIMINARY

Toward the assessment performance issue in free speech-based solutions, we raise a research question: *how can we design an effective free speech-based PD assessment solution?* To answer this question, we collect real-world speech samples from PD patients and conduct analysis. We observe that one important factor limiting free speech-based solutions' effectiveness is the *linguistic content variations*. That is, under free speech condition, we cannot control the voice content to be the same. This is significantly different from the above-mentioned vocal tasks including sustained vowel and reading one specific sentence, whose voice or linguistic content are aligned. Compared to the content-aligned setting, vocal impairment features of PD will be more indistinguishable in free speech recordings, as these subtle features will be drowned in content variations. It will be difficult to use traditional acoustic analysis to mine these PD-related acoustic features out of free speech.

As a demonstration, Fig. 2a-2c and 2e-2g show a sample Uniform Manifold Approximation and Projection (UMAP) [47] visualization of content-aligned voice recordings and free speech from different PD patients, which are processed by conventional acoustic analyses [24, 30, 66]. We can see that, conventional techniques have a good ability to differentiate disease stages on content-aligned audio. However, the data of different PD stages highly overlap with each other in the feature space. This is owing to the fact that conventional methods can only extract low-fidelity features from audio, which will be more concentrated on content variations in free speech. Therefore, to achieve better performance in free speech-based PD assessment, we need a *high-fidelity audio representation* to capture and further disentangle hidden disease-related information from content variations.

With the aforementioned observation, we therefore propose to use one pre-trained ASR model, HuBERT, to obtain audio representation with rich information. The benefit of using HuBERT [33] is demonstrated in Fig. 2d and 2h. We can see that after the pre-trained HuBERT and feature extraction, the audio samples with different severity stages can be distinguished more easily.

4 SYSTEM OVERVIEW

In response to the aforementioned challenges in Sec. 1, we design and implement PDAssess. The overall training and inference architecture of the system are illustrated in Fig 3. The system comprises three main components:

(i) **Pre-trained Model-based Pre-processing.** In this module, the system collects the users' speech recording and performs voice activity detection, and then leverages one pre-trained ASR model, HuBERT [33], to extract high-fidelity audio representation for effective disentanglement of disease-related information. The detailed design of this module is presented in Sec. 5.1.

(ii) **Parkinson's disease Assessment Model.** In this module, an MLP module is utilized on the client side to weigh and extract disease-related features out. Another SE-Attention-based neural network with special loss design is proposed on the server side to proceed with the fine-grained assessment of Parkinson's disease. The module details are presented in Sec. 5.2.

(iii) **Privacy-preserving Training Mechanism.** In this module, we utilize SL framework [73] to provide architecture-level privacy preservation. In addition, we utilize a local domain adversarial training technique combined with K-means clustering [27] based

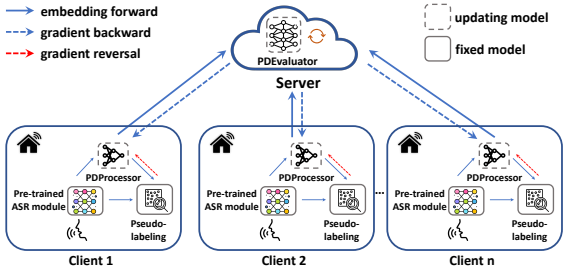


Figure 3: The overall architecture of the system.

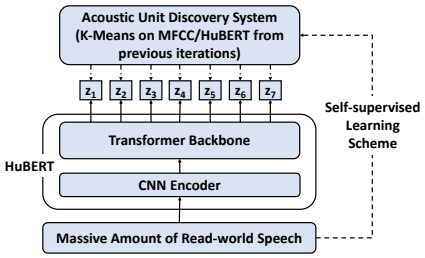


Figure 4: HuBERT model illustration.

pseudo-labeling to further protect the embedding privacy. The design details of this module are presented in Sec. 5.3.

5 SYSTEM DESIGN

5.1 Pre-trained Model-based Pre-processing

As illustrated in Fig. 2, traditional acoustic analysis techniques fail to present the disease severity-related information well due to the fact that linguistic content variation overshadowed the speaker-related disease information. One naive solution will be removing the linguistic variations from the representation by averaging features for different linguistic contents. However, this approach is ineffective as the vocal impairment symptoms in PD vary in different linguistic contents, such as presenting more phonatory changes in vowels and more articulatory changes in consonants [21]. Therefore, instead of processing the features in a content-agnostic manner, we need to extract high-fidelity audio representations that can preserve both linguistic content-wise information as well as disease-wise information to further enable disease-aware analysis.

To empower such a representation extraction, we leverage a self-supervised learning (SSL)-based pre-trained ASR model, *HuBERT* [33]. The overall architecture and training scheme of HuBERT is illustrated in Fig. 4. The SSL scheme adopted by HuBERT supervises the model with pseudo-labeled acoustic units. The pseudo-labeling is first performed with K-means [27] clustering on MFCC [66], and later with trained HuBERT model in the previous iteration.

The reasons for using HuBERT are as follows: Firstly, compared to the pre-defined operators or the filters in conventional methods like STFT and MFCC, HuBERT adopts deep learning-based pre-processing, which can extract complex characteristics from audio signals by learnable kernels and non-linear activation functions. Specifically, it adopts a transformer-based backbone, which can extract features from sequential data like audio more effectively. Secondly, HuBERT utilizes SSL scheme, which is based on clustering on pseudo acoustic labels instead of specific linguistic content labels. Such a design can empower better extraction of not only linguistic information, but also speaker-related characteristics [33]. Note that the speaker-related information is highly correlated with the vocal system dynamics [57], and therefore may benefit PD's vocal impairment feature extraction. Thirdly, the audio data for HuBERT pre-training is of a massive amount and with variations in content, speaker and acoustic conditions, which makes the learned representation more generalizable in real-world collected audio data. The pre-trained HuBERT can provide audio representation with a higher fidelity compared to conventional audio processing methods, which can lay a better foundation for further disease-relevant features extraction.

To be specific, we choose the large version of the HuBERT to capture enough acoustic features. The model starts with a 7-layer CNN encoder, followed by 24 Transformer layers with 1024-dim feature embedding. We use the final output of the Transformer backbone as our representation. Due to the fact that our enrolled subjects are Mandarin speakers, we use HuBERT model pre-trained on the WenetSpeech dataset [80] with more than 10000 hours Chinese speech. We give a demonstration of HuBERT's ability to distinguish PD patients in different stages from free speech in Fig. 2d, 2h. Compared to other pre-processing techniques like MFCC and STFT, our scheme provides better audio representation that can better cluster recordings from different stages based on free speech.

The detailed pre-processing procedure is the following: we first extract the voiced parts of the participating subject out of the collected conversational data; next we utilize rVAD, a robust voice activity detection method proposed in [69], to remove the silences between voiced segments and concatenate all voiced segments together for further processing; we then feed the voiced segments directly into the pre-trained HuBERT model for audio representation extraction. After such pre-processing, we can obtain audio feature embeddings that contain both vocal system-related and linguistic-related information with high fidelity.

5.2 Parkinson's disease Assessment Model

Though we have obtained high-fidelity audio representation from the SSL-based pre-trained ASR model, it remains challenging to perform an accurate PD severity assessment. Specifically, there are two main challenges towards disease assessment. Firstly, it is difficult to effectively disentangle disease-related information from high-fidelity audio representation. Secondly, given that we successfully extract disease-related features, the fine-grained disease assessment task is still hard in terms of subtle feature-wise relation modeling and disease-stage distribution skewness.

To tackle the above challenges, we design two neural network modules, namely *PDProcessor* and *PDEvaluator*, and the corresponding loss function for our task. Moreover, as our target is to make person-wise disease assessment, and there might be not enough information encoded in single input segment, we design a majority voting-based module to analyze segment sequence, which can provide more reliable results. The following will illustrate the detailed design of these modules.

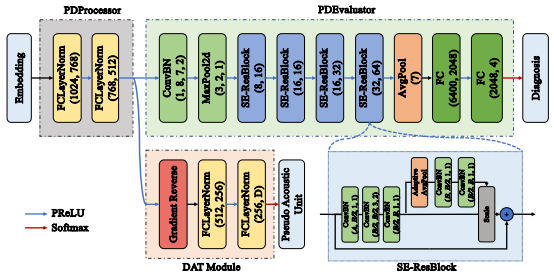


Figure 5: The architecture of our assessment model.

5.2.1 PDProcessor: Client-end Feature Extraction Module.
The client-end feature extractor adopts a multi-layer perceptron (MLP) architecture to perform disease-related information disentanglement. The architecture is shown in Fig. 5. As we previously mentioned, content-wise variation also contains disease information. So we adopt MLP to extract such disease-related information by dynamically weighing different input features. We set the hidden sizes of MLP layers to 768 and 512. As for the normalization usage, we utilize LayerNorm [75] instead of BatchNorm [35] to perform feature normalization. The reason is that HuBERT performs LayerNorm to normalize the feature space. In terms of the activation function, we use the Parametric ReLU (PReLU) unit [29].

5.2.2 PDEvaluator: Server-end Disease Assessment Module.
Next we illustrate our disease-assessment module at the server end. Owing to the lack of H&Y stage-5 patient and only 1 H&Y stage-4 patient (see Fig. 7), we combine the stage 3 and stage 4 as one class and formulate the PD assessment task as a 4-class classification problem, where the classes include healthy control (referred as stage-0), H&Y stage-1, stage-2 and stage-3&4.

As some of the PD vocal impairments, such as articulation dysfunction, are revealed more in several consecutive acoustic units instead of one single acoustic unit [51], we choose to bundle 512 samples together as one segment (correspond to 10.24 seconds) for analysis instead of analyzing in a sample-wise manner. We treat the segment as an image-like input with a width of 512 and a height of 512 to preserve the sequential locality, and train deep neural network module to perform the assessment.

Note that our PDEvaluator incorporates a relatively deep neural network design, and a large number of feature maps, which is essential for effective feature extraction as the PD vocal impairment is quite subtle. However, our task supervision is relatively sparse compared to our input in guiding such a complex neural network, the model will be easily affected by noisy feature channels. Therefore, to better empower the neural network's training, we integrate the design of squeeze-and-excitation (SE) attention into our model [34]. The SE-attention module contains one global average pooling layer and two fully connected (FC) layers on the feature dimension , which explore channel-wise significance through attention mechanism. We add the SE-attention module on top of the residual block as SE-ResBlock. Based on the SE-ResBlock, our final design of the *PDEvaluator* is presented in Fig. 5, which contains one convolutional layer with BatchNorm, four consecutive SE-ResBlocks and two FC layers to perform disease stage classification.

5.2.3 Loss Function Design.
As we mentioned in Sec. 1, the distribution of disease stages is highly skewed, which may lead to statistical imbalance for model training and lead to bad performance. Moreover, even though we have leveraged SE-attention, the disease assessment task is still difficult as we only have stage label supervision. Towards these two design challenges, we leverage two types of loss design to increase the model robustness under real-world scenarios: *Multi-class Focal Loss* and *Momentum Contrastive Loss*.

(i) **Multi-class Focal Loss.** The category distribution of our collected dataset is rather imbalanced, with a relatively small population of stage-1 and stage-3&4 patients (see Fig. 7). Such an uneven distribution will lead to skewed performance, that is, higher accuracy on categories with a large population, which is unfavourable for disease assessment tasks. Therefore, to improve the performance of minority classes, we utilize a modified version of Focal Loss [44] for our multi-class classification problem. The loss is defined as follows:

$$L_f^i = -\alpha_{y_i}(1-p_i)^\gamma \log(p_i),$$

where the p_i is the predicted probability of sample i, whose label is y_i, α_t controls the weight for different classes and γ controls the loss based on the classification difficulty. We set α to be $[0.1, 0.4, 0.1, 0.4]$ and γ to 2 in our experiment.

(ii) **Momentum Contrastive Loss.** Note that due to the linguistics content difference, the audio samples from PD patients might vary a lot, which leads to difficulty in neural network training. As stage classification labels are relatively weak to guide the complex neural network, we propose to consider person-wise and stage-wise consistency in neural network supervision. Specifically, due to the nature of Parkinson's disease, within a nearby time period, the prediction result should be the same. And the recordings from patients of the same stage should have similar features.

However, simply pulling feature embeddings from the same user or users in the same stage is not effective enough as it will be influenced by outliers. That is, some of the audio samples may not reveal enough vocal impairment as others depending on the contents [21]. To extract features using person-wise and stage-wise consistency and avoid the learned feature embeddings from being affected by outliers, we adopt momentum contrastive loss [28] to perform class-wise and person-wise momentum prototype learning.

The basic idea is to store and update a collection of user-wise and class-wise embedding prototypes, and make the model learn an embedding space where samples from the same disease stage and the same user will gather around its embedding prototype, while separating embedding prototypes from other classes or users. Specifically, we introduce the following two contrastive losses on the 2048-dim embedding before the last FC layer, class-wise contrastive loss L_{class} and person-wise contrastive loss L_{person}:

$$L_{class}^i = -\log \frac{\exp\left(\frac{\mathbf{z}_i \cdot \mathbf{c}_{y_i}}{\tau}\right)}{\sum_{k=1}^{K} \exp\left(\frac{\mathbf{z}_i \cdot \mathbf{c}_k}{\tau}\right)}, L_{person}^i = -\log \frac{\exp\left(\frac{\mathbf{z}_i \cdot \mathbf{m}_{u_i}}{\tau}\right)}{\sum_{r=1}^{R} \exp\left(\frac{\mathbf{z}_i \cdot \mathbf{m}_r}{\tau}\right)}$$

where z_i is the embedding of sample i, y_i is its corresponding class label and u_i is its correspondling user id. And c_{y_i} is the embedding prototype of the corresponding class and m_{u_i} is the embedding prototype of the corresponding user. τ is the temperature parameter. We set τ to be 0.1 during the experiments.

During the training of the neural network, we adopt a momentum-based approach to update the class and user prototype. Specifically,

$$\mathbf{c}_k = \alpha_c \mathbf{c}_k + (1-\alpha_c)\mathbf{z}_i, \forall i \in \{i|y_i = k\}$$

$$\mathbf{m}_r = \alpha_m \mathbf{m}_r + (1-\alpha_c)\mathbf{z}_i, \forall i \in \{i|u_i = r\}$$

where the notations are consistent with the above loss functions.

In addition, we leverage *weight decay* to prevent overfitting, which is implemented as an L2 regularization strategy that penalizes over large model parameters to prevent overfitting. To conclude, our overall loss function on the server side will be the following:

$$\mathcal{L} = \sum_{i=1}^{n} (L_f^i + \lambda_{class} L_{class}^i + \lambda_{person} L_{person}^i) + \lambda_{decay} \|\mathbf{w}\|_2,$$

here we set $\lambda_{class} = 1$, $\lambda_{person} = 0.8$ and λ_{decay} to 10^{-5} in our experiment.

5.2.4 *Majority Voting Module.*

As we have discussed in Sec. 5.1, PD's vocal impairment symptoms will vary with respect to the linguistic content difference. Though our model has incorporated HuBERT pre-processing and model design to tackle this problem, the sample-wise prediction may still be inaccurate as the amount of disease-related information encoded in different samples will be different, for instance, the disease feature might be more evident in vowels but less evident in some consonants. However, as our final target is to make a more reliable person-wise prediction, it is possible to cumulatively analyze a sequence of audio samples to offer a person-wise assessment result, easing the sample-wise prediction variation problem.

According to this idea, we design an extra majority voting module for audio sample sequence analysis, which can provide a more reliable person-wise disease assessment result. Specifically, for n continuous audio samples, the system will vote to give the following result y according to the majority of the sample predictions:

$$y = \begin{cases} c_i & \text{if } c_i > c_j, \forall j \neq i, \\ \text{refuse to predict} & otherwise \text{ (if equally distributed)}, \end{cases}$$

where c_i is the number of samples belong to class i, and $i \in [1, 2, 3, 4]$. In the evaluation, if the voting result is refuse to predict, then we treat such prediction as wrong.

5.3 Privacy-preserving Training Mechanism

To deploy such a PD assessment system in real-world scenarios, one more design challenge we need to tackle is the privacy issue. Since our training data is free speech, whose content is rather sensitive, directly uploading these data onto the server for processing will lead to privacy leakage. Therefore, we design a privacy-preserving training mechanism to solve this problem. Our training scheme is based on Split Learning architecture [73]. Towards better privacy preservation, we incorporate domain adversarial training (DAT) technique [20] to reduce the speech content information in the uploaded embedding, limiting the leakage of sensitive information. We don't adopt Federated Learning (FL) scheme here because FL will have a rather poor convergence due to distribution skewness, which is demonstrated in our evaluation Fig.10a. In this section, we will first explain our threat model, and elaborate on our privacy-preserving training mechanism design in detail.

Algorithm 1: PDAssess Server training procedure for step t.

params: B: mini-batchs; E: number of epochs; η_t: learning rate; $\mathbf{W_t}$: weight of *PDEvaluator*; $\mathbf{A_t^k}$: uploaded embedding; ∇L: gradient

```
function PDServer(step t)
    for each client k ∈ S_t in parallel do
        A_t^k ← PDClientUpdate(k, t)
        W_t ← W_t − η_t ∇L(W_t; A_t^k)
        PDClientBackProp(k, t, ∇L(A_t^k))
        Sync PDProcessor's weight
    end for
end function
```

5.3.1 *Threat Model.*

Split Learning (SL) is a novel distributed machine learning framework that can provide better client data privacy preservation compared to the traditional centralized training scheme. The basic idea is to split the neural network model into two parts, one for the client and one for the server. Instead of uploading the raw data to the server, SL asks the clients to feed their data into the client-side model and upload the embedding to the server for further processing. SL then claims that the privacy-preserving objective can be achieved since the raw data cannot be effectively inferred from the uploaded embedding.

However, one recent work proposes a potential attack named Feature-space Hijacking Attack (FSHA) towards SL [55]. The threat model is described as follows: consider the computation server is malicious and curious about the user's speech content. The attacking process is (i) the malicious server collects a speech dataset and pre-processes it with pre-trained HuBERT as well. (ii) the malicious server trains an attack model to convert the uploaded embedding back to the raw pre-processed representation instead of the training assessment model. (iii) the malicious server collects uploaded embeddings from some clients and uses the trained attack model to invert its post-HuBERT representation. (iv) the malicious server can train a decoder on pre-trained HuBERT and use it to obtain the raw input data from the hacked post-HuBERT representation.

Towards such threat model, we customize the split learning framework under our scenario for privacy-preserving.

5.3.2 *Server-side Training Procedure.*

On the server side, we adopt similar architecture as Split Learning. The server side mainly executes the training process of *PDEvaluator*, which is described in Algorithm. 1. For each training round, the server will sequentially query each client to perform feature extraction with *PDProcessor*, and calculate and backpropagate the loss based on the output of *PDEvaluator*. Afterward, the server will send the gradient back to the clients for client model update. At the end of each round, the server will inform the clients to synchronize the weights of their models for further training.

5.3.3 *Client-side Training Procedure.*

As the server might be malicious according to our threat model, we need to adjust our client training procedure from the vanilla SL design. In addition to using *PDProcessor* for embedding generation, we incorporate a DAT module with pseudo-labeling on the client

Algorithm 2: PDAssess Client training procedure for step t

params: B: mini-batchs; E: number of epochs; η_1, η_2, η_3: learning rate; $\mathbf{H}_t^k$, $\mathbf{U}_t^k$: weight of *PDProcessor* and DAT module f; $\mathbf{A_t^k}$: uploaded embedding; $\mathbf{D}_k$: HuBERT embedding; ∇L, ∇L_{adv}: normal and DAT gradient

function PDClientUpdate(client k, step t)
  if $t == 0$ **then**
    PseudoLabeling(k, p)
  end if
  $\mathbf{A}_t^k = \phi$
  for batch $b \in B$ in each local epoch E **do**
    $\mathbf{A}_t^k \leftarrow$ Concatenation of $\mathbf{A}_t^k$ and $f(b; \mathbf{H}_t^k)$
  end for
  return $\mathbf{A}_t^k$ to server
end function
function PDClientBackprop(client k, step t, $\nabla L(\mathbf{A}_t^k)$)
  for batch $b \in B$ **do**
    $\mathbf{H}_t^k \leftarrow \mathbf{H}_t^k - \eta_1 \nabla L(\mathbf{A}_t^k; \mathbf{H}_t^k; b) + \eta_2 \nabla L_{adv}(g(\mathbf{A}_t^k); \mathbf{H}_t^k; b)$
    $\mathbf{U}_t^k \leftarrow \mathbf{U}_t^k + \eta_3 \nabla L_{adv}(P_b, g(\mathbf{A}_t^k); \mathbf{U}_t^k; b)$
  end for
end function
function PseudoLabeling(client k, cluster num p)
  Randomly sample p embeddings out of $\mathbf{D}_k$ as the centroids of clusters C
  while not converge **do**
    Assign all embeddings to the closest cluster $c \in C$
    Recompute all the cluster centroids
  end while
  for all $d \in \mathbf{D}_k$ **do**
    Pseudo Label $P_d \leftarrow$ cluster index of d
  end for
end function

side for privacy-preserving. The architecture of the DAT module is shown in Fig. 5.

We notice that the linguistic contents of the speech are rather sensitive as private information like financial details can be inferred, and thus try to reduce linguistic information in order to prevent the content inversion from the server. The basic units of linguistic contents are defined as acoustic units or phonemes [5]. As labeling acoustic units can be labor-intensive in large dataset, we here adopt the pseudo-labeling technique [33, 45] to obtain these units for further removal from continuous speech. It is mentioned that by clustering on the HuBERT embeddings, the phone purity can achieve around 70% [33]. Hence, in our system design, we adopt K-means clustering [27] with K=100 on post-HuBERT representations to generate pseudo-labels. Normally the value of K should be set slightly greater than the number of basic acoustic units in a certain language. We here choose the value based on its usage in [33]. The detailed procedure of pseudo-labeling is illustrated in Algorithm. 2.

After pseudo-labelling these units, we need to remove their information from the calculated embedding. Thus we introduce the DAT module based on the pseudo-labels, which is a two-layer MLP module with Gradient Inverse Layer. The design of the DAT module is shown in Fig. 5. The basic idea of DAT is to inverse the gradient

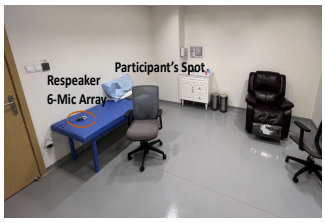


Figure 6: Experiment setup.

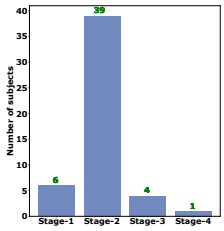


Figure 7: Patient H&Y stage distribution.

Stats	PD	Healthy
Population	50	50
Age (years)	63.0(10.6)	65.6(8.1)
# Female	24	25
# Smoker	2	17
Disease Duration (years)	4.92(4.03)	/

Age & duration form: mean (standard deviation).

Table 2: Participants statistics.

value through backpropagation, so that the neural network will try to reduce its emphasis on the DAT label differences. The detailed backpropagation process is illustrated in Alg. 2.

Based on such a training scheme, we can progressively remove information related to pseudo acoustic units as the training proceed, and therefore encode less linguistic content information in the embedding. Moreover, such a client training design can confront the malicious server's training scheme, which can push the embedding away from the objective of the potential malicious server and defense against the FSHA attack. Though the DAT scheme will lead to some performance degradation as the linguistic contents are also useful as our discussed in Sec. 5.2.2, the degradation is comparatively small as the linguistic content is not crucial in the assessment, which is shown in our evaluation. Moreover, such a trade-off is acceptable as we preserved the sensitive speech data privacy, which is important in our scenario.

6 EVALUATION

6.1 Evaluation Setup

6.1.1 *Dataset Collection.*

To evaluate the real-world performance of our system, we collaborate with the Second Affiliated Hospital of Zhejiang University School of Medicine to enroll 50 PD patients and 50 healthy control on the basis of research protocols approved by institutional review board. Written informed consent is obtained from all subjects. Medical professionals ensure that enrolled PD patients are idiopathic PD patients, excluding other related diseases such as PSP (progressive supranuclear palsy) and MSA (multiple system atrophy) [50]. Each patient's disease severity is evaluated with the H&Y scale system by professionals. For each participant, we conduct a demographic survey to record related information such as gender, age and smoking. The overall participants' statisitcs are listed in Table. 2.

Data collection is conducted in a room setup with around 20 m^2 with a noise of 40-50 dB, simulating a daily-life environment. The setup is shown in Fig. 6. We collect all the voice data using a combined device with a Seeed ReSpeaker 6-Mic Circular Array Kit [82] and a Raspberry Pi 4B development board [59], mimicking hardware configurations of most commodity smart speakers. The

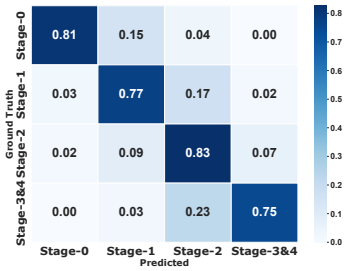


Figure 8: Sample-wise confusion matrix.

Model	Metric	Stage 0	Stage 1	Stage 2	Stage 3&4	Avg	Binary
RF	Acc	0.3728	0.2399	**0.9002**	0.0242	0.3843	0.6623
	F1	0.4583	0.3137	0.7604	0.0449	0.3943	0.4594
X-vector	Acc	0.6991	0.3599	0.8445	0.2766	0.5450	0.8047
	F1	0.6461	0.4596	0.7953	0.3297	0.5576	0.6475
PDVocal	Acc	0.7589	0.6149	0.8464	0.3420	0.6406	0.8380
	F1	0.7499	0.6377	0.8296	0.3879	0.6488	0.7416
PDAssess (w/o SL)	Acc	0.8077	**0.7721**	0.8265	**0.7460**	**0.7880**	0.9594
	F1	0.8300	**0.7007**	**0.8621**	0.6667	**0.7648**	0.9770
PDAssess (w/ SL)	Acc	**0.8462**	0.7051	0.7965	0.7354	0.7708	**0.9622**
	F1	**0.8748**	0.6262	0.8327	**0.6715**	0.7513	**0.9785**

Table 3: Sample-wise benchmark comparison. Best results are in Bold.

ReSpeaker is placed around 1 meter away from participants, and records speech with a 44.1 kHz sampling rate. We select the channel with the most average energy as the data source. Participants are free to talk in any direction to simulate a real-world conversational environment. They are asked to have some random conversation with the examiner for 1 to 2 minutes. The detailed conversation protocol is listed in Appendix. A. We collect 2.59 hours of audio recordings in total.

6.1.2 *Experimental Setting.*

For the data pre-processing, we resample the audio recordings from 44.1 kHz to 16 kHz, and segment the audio recordings into pieces with 10.2-second length with a sliding window of 4-second, and therefore have 2119 voice samples overall for training and evaluation. Our neural networks are implemented in Python 3.8 and Pytorch 1.12 [56]. The distributed training scheme is implemented using OpenMPI [19] and the mpi4py library [14]. The optimizer we adopted is Adam [38] with an initial learning rate of 10^{-5}. We load minibatch data from the whole dataset and set the batch size to 64. We conduct experiments on a cluster with NVIDIA Tesla V100 SXM2. For the evaluation setting, we mainly utilize a random 5-fold cross-validation with respect to users instead of samples. In order to check the performance of each individual, we conduct leave-one-subject-out validation in the person-wise evaluation sections.

6.1.3 *Performance Metrics.*

We compare our design according to the following two main performance metrics:

Person-wise Performance: After majority voting on the recorded 1-2 minute audio data, the system will output a person-wise result with respect to the majority of sample predictions. For the leave-one-subject-out experiments, we use macro-version person-wise F1 score as our system's performance metric. That is, we report one prediction result for each person and calculate the F1 score according to these results. Moreover, in the demographic results we also use person-wise F1 score as our metric.

Sample-wise Performance: Though the final prediction is provided by the majority voting module, it is still important to look at sample-wise performance for neural network model evaluation. Towards such performance evaluation, we select the macro version of F1 score and accuracy as our metrics. Note that in terms of multi-class classification, accuracy is equivalent to recall. Moreover, due to the previously mentioned imbalanced data distribution, we mainly use F1 score as our benchmark and ablation study metrics. And for the privacy-related discussion, we choose the weighted average of accuracy for phoneme recognition as there is no much data imbalance as in the disease prediction.

6.1.4 *Benchmark Models.*

We compare our system's performance with the following state-of-the-art voice-based Parkinson's disease assessment methods:

Random Forest [11] is one of the famous traditional machine learning algorithms that has been examined to be effective in many classification problems. Existing works [71, 81] in Parkinson's disease detection utilized this model with MFCC as baseline; we therefore use it together with 57-dim MFCC (19-dim and the first and second-order difference) as one of our benchmarks.

X-vector [36, 68] is a widely adopted neural network architecture for speaker identification and speaker-related information extraction with MFCC. Recently it has been proven to be an effective digital biomarker for Parkinson's disease [36]; we therefore utilize this together with 57-dim MFCC as one of our benchmarks.

PDVocal [81] uses a ResNet-like architecture to detect Parkinson's disease from the unvoiced breathing sound, and also has been examined to be effective with speech signal using the MFCC feature extraction method. With ResNet structure, it is able to extract more hidden patterns from the signal, we therefore use this together with 57-dim MFCC as one of our benchmarks.

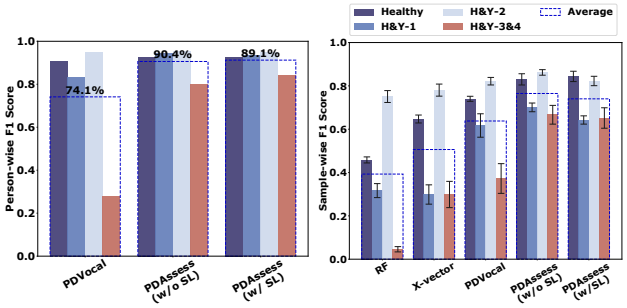
6.2 Evaluation Results

6.2.1 *Overall Performance.*

To evaluate our system's performance, we perform leave-one-out validation to see the effectiveness on different subjects. Fig. 8 demonstrates the overall confusion matrix. From the confusion matrix, we can see that our system can achieve around 75-85% accuracy over all the disease stages prediction, demonstrating the ability to accurately predict the disease progression. Fig. 9a shows the person-wise F1 score. We can see that our system can achieve averagely 90.4% F1 score without SL, and 89.1% F1 score with SL, both with a balanced prediction performance. In comparison, we also performed leave-one-subject-out evaluation on our implementation of PDVocal's system, which achieves an around 75% average F1 score. Moreover, PDVocal shows a performance degradation on the severe stage classification (H&Y-3&4), which is of a significance in disease management such as taking surgeries. Also, we can see that our system's performance stays relatively consistent after introducing SL and DAT scheme, showing that a good performance can be achieved while preserving privacy. From such a person-wise evaluation, we can see that our system can achieve a superior result in assessing PD severity, especially in classifying more serious stages.

6.2.2 *Sample-wise Benchmark Performance.*

To further validate the performance of the PDAssess, we perform the following sample-wise benchmark evaluation.



(a) Person-wise **(b) Sample-wise**

Figure 9: F1 score Comparison.

We conduct sample-wise benchmark evaluations compared to other state-of-the-art voice-based Parkinson's disease assessment systems. The results are shown in Fig. 9b and Table. 3. As the benchmarks are mainly designed for binary classification tasks, we provide the binary classification result in the last column as well to prove the correctness of our implementation. In the Fig. 9b, different color bars represent the F1 score of different stages, and the dashed outline represents the average F1 score of all four stages. We can see that our model achieves the best average F1 score and accuracy with a minimal stage-wise variation. Moreover, for almost all stages our model achieves the best performance, with just minor inter-run variation shown by error bars. Note that for H&Y-2, due to the large data skewness, we can see that RF achieves the best accuracy, which is due to the prediction extremity of the RF model responding to the dataset. That is, the insufficient modeling ability leads the RF model to an overfitting situation, where all the samples are likely to be classified as H&Y-2, leading to a high accuracy (recall) of this stage. However, such overfitting will lead to a decrease in the overall system performance as shown in the averaged results. Therefore, through this benchmark we can see the prediction effectiveness of our model against different disease stages. Though in the SL version of our system, we get a little performance degradation due to a trade-off between accuracy and privacy, the overall accuracy is still acceptable and we suppose it can be further improved with more data being collected in the future.

6.2.3 Ablation Study.

In the following section, we will discuss the ablation study results of our system.

Impact of Pre-processing. We evaluate our proposed pre-processing technique against PLP [30], STFT [24], and MFCC [66] and Wav2Vec2 [6] on top of our model. The first three conventional feature extraction methods have widely been used in human voice processing as well as speech-based PD assessment tasks. The last one is another widely-used ASR model. For the PLP implementation, we use Rasta-PLP [31] to extract the initial 13-dim feature, and perform first and second-order differences to get the final 39-dim features. For the STFT Spectrogram, we resample the audio recordings to 16kHz, and set the number of frequency bins to 320 with a Hanning sliding window of length 160 to align the segment length with HuBERT, so that the feature dimension of the output spectrogram is 161. For MFCC, we use the same setting in STFT to obtain the basic Spectrogram, and extract 19-dim MFCC initially, and perform first and second-order differences to get the final 57-dim feature. For Wav2Vec2, we also use its pre-trained version on Wenetspeech, who will also extract out 1024-dim audio embedding.

We report the macro-version F1 score as our metric here, and the results are shown in Fig. 10b. We can observe that, in terms of average F1 score, our pre-trained model-based method achieves the best among all five methods, with an increase of around 10% F1 score compared to MFCC, the previous state-of-the-art method on related tasks. Also we can see a 5% F1 score increase compared to Wav2Vec2, another ASR model that widely used in audio tasks. Moreover, we can see a decrease in prediction variation among different disease stages. These demonstrate that our proposed pre-processing method improves both the F1 score and the robustness of the prediction. The main reason for such improvement is that the pre-trained HuBERT model enables a more sophisticated mapping between the raw audio space and the disease-oriented feature space comparing to conventional audio preprocessing. And compared to other existing ASR models like Wav2Vec2, HuBERT can extract more non-linguistic information, which is helpful in our assessment.

Impact of Distributed Training Scheme. We conduct evaluations on our distributed training scheme against two other privacy-preserving distributed training schemes, FedAvg [48], which is one of the standard Federated Learning (FL) frameworks, and FedProx [42], an FL framework that is improved towards imbalanced data distribution. Our distributed training simulation is performed with 100 clients, each corresponding to one subject. The evaluation results are shown in Fig. 10a. We witness a huge F1 score degradation when training our model on top of the FedAvg architecture, where the final F1 Score is only around 20% with a huge class-wise variation. This is mainly caused by the heavy imbalance in our data distribution. Firstly, the number of subjects in H&Y-1 and H&Y-3&4 is relatively smaller than those of H&Y-2 and healthy control. The subject-wise imbalance factor is around 10, which will lead to heavy performance degradation. Secondly, in our problem setting, each subject can only access one kind of disease label, making the weights on other disease classes not randomly generalized, and generating a bad weight aggregation result. And we can see a performance improvement in FedProx-based training, which achieves around 30% F1 score and a decrease in the variation. However, the performance is still not satisfying. Comparatively, our proposed usage of SL architecture can effectively stabilize the training and improve performance. We obtain around 70% F1 score with a maximum class-wise F1 Score variation of 5%, which is a direct result of the weight synchronization and sequential training in SL.

Impact of SE-Attention. We first evaluate the effectiveness of introducing the SE-attention into our system. We compare the performance of this module to see if inter-channel relations exist for our feature dimension. The result is shown in Fig. 10c. By integrating the SE-attention into the system, we witness an around 5% overall F1 score improvement compared to the original residual architecture. This demonstrates the effectiveness of leveraging feature-wise relation in the acoustic embedding generated by the pre-trained HuBERT model. What's more, such performance improvement is demonstrated in prediction results for all stages, which shows that the benefit brought by the SE-attention design is generally applicable to all stages' predictions.

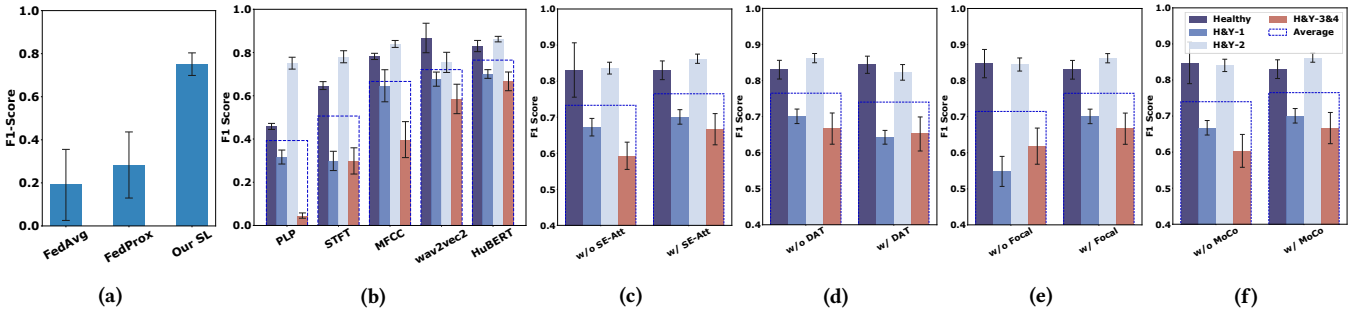


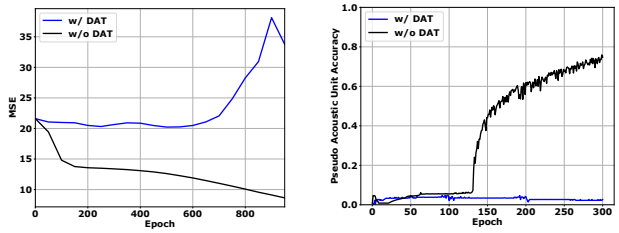
Figure 10: Ablation Study Results (a) Impact of our distributed training scheme. (b) Impact of our pre-processing scheme. (c) Impact of SE-attention. (d) Impact of domain adversarial training. (e) Impact of multi-class focal loss. (f) Impact of momentum contrastive loss.

Impact of Domain Adversarial Training. Next, we evaluate the impact of introducing the domain adversarial training module into our system. By leveraging pseudo acoustic unit targeted domain adversarial training, the neural network will try to remove the phoneme-related information from the uploaded embedding. Though the speech content privacy can be preserved better, which we will discuss in Sec. 6.2.4, the evaluation results in Fig. 10d show that a performance degradation will occur due to such adversarial training. From the result we can see that, after using the domain adversarial training scheme, the overall F1 score is decreased by 1.3%. Moreover, the performance instability increases among different stages, including 8% F1 score decrease for stage-1 prediction, and a slight increase of F1 score in stage-0 prediction due to the training instability. This demonstrates the privacy-performance trade-off in our system, even though the F1 score degradation is not much. By involving more training data in the system through collaborative training, it is possible to improve the overall performance and compensate for such negative impact.

Impact of Loss Design. We here evaluate the impact on the results of our two loss designs. Our ablation study involving loss functions are conducted by replacing the special loss part with normal cross entropy loss function.. First is the multi-class focal loss, the result of which is shown in Fig. 10e. Focal loss will penalize the model for those classes of greater difficulty to classify well, and thus dampen the prediction F1 Score variations between different classes. We can see that the introduction of focal loss improves the system's performance in H&Y-1 and H&Y-3&4, especially the latter one. Without the focal loss design, the prediction F1 score of H&Y-1 decreases to less than 60%, and the prediction F1 score of H&Y-3&4 also decreases down to 60%, which shows the effectiveness of focal loss in our system to improve the stage-wise prediction imbalance. We also evaluate the impact of momentum contrastive loss. The result is shown in Fig. 10f. The contrastive loss will generally improve the overall performance of the system, which can be witnessed in the figure. We can see an improvement of around 5% in terms of overall F1 score. Especially, the introduction of contrastive loss improves the performance of minor class H&Y-3&4, contributing to the robustness of our system.

6.2.4 Privacy Analysis.

We next evaluate and discuss our privacy-preserving scheme. In our system design we mainly adopt the SL architecture, therefore we have the original privacy-related advantages of SL. Firstly, the client can use its own model to process the data before uploading, and the server has no access to the client model, preventing both direct data access and potential model inversion attacks. Moreover, the server-side model will not publish to clients, which protects the model's privacy and the model owner's intellectual property.



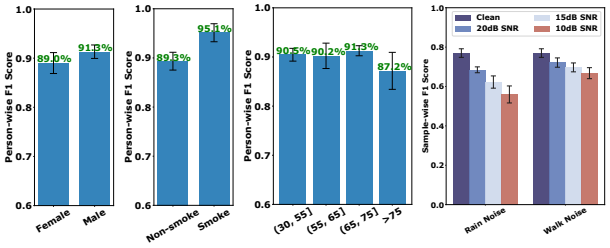
(a) Reconstruction Error under FSHA attack. **(b) Pseudo phonetic unit recognition accuracy.**

Figure 11: Privacy Analysis.

In addition to the above advantages, we incorporate the domain adversarial training technique into our design, so that in the uploaded embedding, there will be less speech-content-wise information as we cannot distinguish linguistic units from the embedding. Therefore, to examine our design's benefit, we evaluate our system with the previously mentioned FSHA attack [55]. The result is shown in Fig. 11a. We can see that with our domain-adversarial-training (DAT) enhancement, the reconstruction error in terms of Mean Square Error(MSE) is comparatively higher than that without DAT after 1000 epoch training. Moreover, we can see from the result in Fig. 11b when we train the whole system using DAT, the pseudo-acoustic unit's prediction accuracy is restricted to less than 5%, compared to an accuracy of around 80% without using DAT. Therefore, we can see that we achieve a high level of privacy preservation on the speech-content information with our DAT-enabled Split Learning architecture.

6.2.5 Robustness Analysis.

In this section, we discuss our system's robustness under different noise settings. We evaluated our system on two types of real world noises, including raining and walking. The noise data is adopted



(a) Gender **(b) Smoke** **(c) Age** **(d) Noise**

Figure 12: Demographic and Robustness Analysis.

from a noise dataset [18]. We examined the performance under three different SNR level, 20dB, 15dB and 10dB. The results are shown in Fig. 12d. We can see that our system achieves a robust performance under walking conditions with 70-80% F1 score. However, the system's performance is a bit unsatisfactory under raining situation, will drop to around 60% when SNR is small. This might due to the fact that the frequency components of rain sound falls more into human voice frequencies than walking. In the future, we will adopt data augmentation and other techniques to enhance the system's robustness under noisy conditions.

6.2.6 Demographic Analysis.
In this section, we discuss our system's performance on different demographic groups. We analyze our system's performance on gender, smoking and age differences. The results are shown in Fig. 12a-12c. In terms of gender, we can see that our system performs consistently on male and female population, both with an F1 Score around 90% . As for smoking, we can can witness a 5.8% F1 score difference between smokers and non-smokers. This might indicate that smoking can be a factor in impacting the system's performance, possibly due that smoking makes the vocal symptoms more distinct for discovery. And for different age groups, we can also see that the performance is relatively stable while dropping a bit in population aged over 75-year-old. This might because the age distribution of our dataset is also skewed, that is, we only get 15 objects greater than 75-year old, making the reported person-wise F1 score not that representative. In the future, we would like to collect more data to analyze the system's performance on different age groups to provide more reliable and holistic analysis.

6.2.7 Complexity Analysis.
In this section, we evaluate and discuss our system's computational complexity. Table. 4 shows the model's training complexity measurement. As our system follows the client-server architecture, we should mainly focus on the client side model of PDAssess. From the comparison we can see that the PDAssess (Client) has comparatively less parameters and FLOPs than the same-size version of MobileNet, i.e. MobileNet-512 [64], which is the state-of-the-art lightweight machine learning model for edge processing and has also been utilized in PDVocal [81]. This can effectively demonstrate our system's feasibility in real-world deployment. Note that the utilized version of MobileNet in PDVocal is 0.25-MobileNet-224, whose complexity is small. However, it only accept small input, 1 second sample in its case, which cannot be fairly compared with our model as the training intervals are different. Moreover, we might adopt some model pruning techniques in the future to further reduce the client-side model complexity, making it more suitable for edge processing.

Model Name	Params (M)	FLOPs (M)
0.25-MobileNet-224 [81]	0.47	41
MobileNet-224 [64]	4.2	569
MobileNet-512	4.2	2276
PDAssess (Client)	1.3	688
PDAssess	14.9	850

Table 4: Model Complexity.

7 DISCUSSION

In addition to the existing functionalities, in this section, we will discuss some potential future directions of our system.

Other Languages. Our current collected dataset only involves Mandarin speakers. Based on this fact we utilize the Chinese pre-trained HuBERT model for audio pre-processing. However, it has been validated that ASR models can perform well on different languages with corresponding training data [58]. Though we haven't examined our system's performance on Parkinsonism speech recordings in other languages, we believe our system will still be applicable in these scenarios as the pre-processing module can be adjusted correspondingly, the assessment analysis is content-invariant and the vocal feature of PD patients should be irrelevant to languages. Evaluations of our system with different languages will be left as future work.

Other Related Diseases. Currently our dataset collection protocol ensures that only idiopathic PD patients are enrolled in our evaluation. However, note that other similar neurological system diseases will also lead to vocal impairments, such as Alzheimer's Disease [49], non-idiopathic PD (PSP, MSA) [50]. Moreover, inflammation diseases such as laryngitis might also lead to dysphonia symptoms [17, 62]. Though we haven't investigated the impact of these related diseases on our system in this work, we suppose it is possible to distinguish these diseases from PD as the vocal impairment patterns vary in different diseases [61]. In the future, we will continue to collect data from patients with these related diseases and refine our neural network design by transfer learning techniques [13, 74] to design a more robust system for PD assessment, and possibly extend to other related diseases assessment.

Other Environmental Factors. Currently our dataset collection is performed in a daily-life room setup with a changing environmental noise of 40-50 dB, and the participants can freely choose speaking directions or volumes. Benchmark evaluation and robustness study can show good system performance under real and noisy scenarios. However, owing to the time and resource constraint we are not able to collect more data for more holistic system evaluations like performance on changing room sizes and distances, or recordings under cross-talk conditions. Though these environmental changes might impact the system performance, we believe it can be solved with domain adaptation and generalization techniques [20, 79]. In the future, we will conduct more experiments to explore our system's performance under different recording conditions.

Other Related Attacks. In our current privacy-preserving training scheme, we only discuss the threat model with a malicious server. However, due to the client model's weight synchronization

process, there is also a possibility that the attack can be launched on the client side when the malicious user wants to steal the other user's information. Recent research has revealed the potential of such an attack [55]. Nevertheless, existing defensive solutions on SL for such attack [70] should also be applicable given that our system is adopted from SL. Exploring unique defensive mechanisms towards such client-wise SL attacks under our scenario will be explored in our future work.

8 RELATED WORK

8.1 Voice-based Parkinson's Disease Assessment

Recently, research efforts have been put into voice-based PD assessment. Most of the current research focuses on Parkinson's disease detection based on vocal tasks or speech, which is a coarse-grained assessment and cannot continuously benefit PD patients. [2, 22] proposed to use conventional speech analysis methods on PD detection. However, they require the user to read content-aligned tasks, which is not practical for real-world scenarios. [10] explored the use of glottal source information in the speech-based PD assessment. Though the classification result is good, the method suffers from noisy conditions. [36, 52] both used the X-vector speaker identification technique to detect PD. The solution outperforms previous solutions for text-independent free speech, but does not consider privacy issues in free speech setting. Moreover, to leverage the power of deep learning and avoid privacy leakage of speech data, PDVocal [81] proposed a system to extract non-speech body sound and design a neural network for PD detection. Though such a solution can solve the privacy issue to some extent, environmental noises can influence the system performance as the non-speech body sound can hardly be sensed in a noisy environment. Moreover, the above-mentioned works are all limited to PD detection task.

There are also researches that exploit the possibility of severity assessment based on speech. [3, 8] are two early works on fine-grained voice-based PD assessment targeting UPDRS-based PD severity prediction. They both used specific speaking tasks as speech materials, and therefore may lead to a lack of user adherence. Alternatively, [25, 54] conducted assessment based on monologue materials. However, they mainly utilized conventional feature extraction methods which did not fully exploit the hidden characteristics in free speech, leading to unsatisfactory assessment results. Therefore, we try to leverage high-fidelity audio representations from free speech for more practical fine-grained assessment.

8.2 Privacy-preserving Distributed Machine Learning

To empower collaborative learning from different participants without sacrificing data privacy, several distributed machine learning frameworks have been proposed.

Federated Learning (FL) is one of the most popular privacy-preserving distributed learning frameworks. In each round of the algorithm, each client receives the whole global model and trains locally with their own data, and sends back the model to the server, where the updated weights are averaged and distributed again. McMahan et al. [48] presented the FedAvg algorithm, which is regarded as the baseline of FL algorithms. However, FL suffers from skewed distribution problem in real-world conditions. To tackle such problem, many algorithms have been proposed such as FedProx [42] and FedBN [43]. However, these solutions still suffer in heavy imbalance scenarios.

Split Learning (SL) is another framework proposed recently to tackle the convergence problem of FL [26, 73]. In SL, a neural network will be split into two parts, which are located separately on the client side and the server side. And instead of uploading raw data or model weight, SL only requires the user to upload embedding to the server. Recently many works have leveraged SL to obtain a privacy-preserving solution which achieves better performance [39, 40]. However, recent research reveals that SL may have potential security issues due to the uploading of embeddings [55]. In this work, we adopt and modify SL architecture in our problem and improve the privacy-preserving ability.

9 CONCLUSION

In this paper, we present PDAssess, a privacy-preserving free speech-based Parkinson's disease daily assessment system. The system will passively record the daily speech of the user and automatically analyze the disease severity. PDAssess is able to perform a 4-stage PD assessment with robustness and accuracy while preserving speech content privacy. We leverage several techniques to achieve the assessment objectives: We utilize a pre-trained ASR model, HuBERT, for speech preprocessing to extract high-fidelity speech representation. We design a hybrid neural network architecture with SE-attention and utilize special loss designs to achieve an assessment of high accuracy. We adopt and customize the Split Learning framework with a local adversarial training mechanism with pseudo labels to better preserve speech content privacy. We collect real-world speech data from PD patients and conduct comprehensive experiments to evaluate the system. The evaluation results show that our system can achieve high accuracy on 4-stage assessment around 90% person-wise F1 score and over 75% sample-wise F1 score among 100 subjects while preserving client data privacy.

ACKNOWLEDGEMENTS

We thank all anonymous reviewers and shepherd for their insightful comments on this paper. This research is supported in part by RGC under Contract CERG 16204820, 16206122, R6021-20, AoE/E-601/22-R, Contract R8015, and 3030_006.

A CONVERSATION PROTOCOL

Our conversation protocol is as follows: we will randomly select 3-4 daily questions out of 5, regarding weather and city description as well as personal interests on book, food, seasons:

- Can you describe today's weather?
- Can you recommend some places to visit? Why?
- Can you recommend something good to eat and give some description?
- Can you recommend a book and give some description?
- What's your favorite season and why?

We choose to design insensitive open questions and adopt random question selection schemes to ensure the diversity of speech data within and across subjects for the system's generalizability in daily conditions.

REFERENCES

[1] Giovanni Abbruzzese, Roberta Marchese, Laura Avanzino, and Elisa Pelosin. 2016. Rehabilitation for Parkinson's disease: Current outlook and future challenges. *Parkinsonism & related disorders* 22 (2016), S60–S64.

[2] Federica Amato, Luigi Borzì, Gabriella Olmo, and Juan Rafael Orozco-Arroyave. 2021. An algorithm for Parkinson's disease speech classification based on isolated words analysis. *Health Information Science and Systems* 9, 1 (2021), 1–15.

[3] Meysam Asgari and Izhak Shafran. 2010. Predicting severity of Parkinson's disease from speech. In *2010 Annual International Conference of the IEEE Engineering in Medicine and Biology*. IEEE, 5201–5204.

[4] Tunç Aşuroğlu, Koray Açıcı, Çağatay Berke Erdaş, Münire Kılınç Toprak, Hamit Erdem, and Hasan Oğul. 2018. Parkinson's disease monitoring from gait analysis via foot-worn sensors. *Biocybernetics and Biomedical Engineering* 38, 3 (2018), 760–772.

[5] Alexei Baevski, Steffen Schneider, and Michael Auli. 2019. vq-wav2vec: Self-supervised learning of discrete speech representations. *arXiv preprint arXiv:1910.05453* (2019).

[6] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems* 33 (2020), 12449–12460.

[7] E Balaji, D Brindha, and R Balakrishnan. 2020. Supervised machine learning based gait classification system for early detection and stage classification of Parkinson's disease. *Applied Soft Computing* 94 (2020), 106494.

[8] Alireza Bayestehtashk, Meysam Asgari, Izhak Shafran, and James McNames. 2015. Fully automated assessment of the severity of Parkinson's disease from speech. *Computer speech & language* 29, 1 (2015), 172–185.

[9] David R Beukelman and Kathryn M Yorkston. 1980. Influence of passage familiarity on intelligibility estimates of dysarthric speech. *Journal of Communication Disorders* 13, 1 (1980), 33–41.

[10] Tanuka Bhattacharjee, Jhansi Mallela, Yamini Belur, Atchayaram Nalini, Ravi Yadav, Pradeep Reddy, Dipanjan Gope, and Prasanta Kumar Ghosh. 2021. Source and Vocal Tract Cues for Speech-Based Classification of Patients with Parkinson's Disease and Healthy Subjects.. In *Interspeech*. 2961–2965.

[11] Leo Breiman. 2001. Random forests. *Machine learning* 45, 1 (2001), 5–32.

[12] Jeff M Bronstein, Michele Tagliati, Ron L Alterman, Andres M Lozano, Jens Volkmann, Alessandro Stefani, Fay B Horak, Michael S Okun, Kelly D Foote, Paul Krack, et al. 2011. Deep brain stimulation for Parkinson disease: an expert consensus and review of key issues. *Archives of neurology* 68, 2 (2011), 165–165.

[13] Xingyuan Bu, Junran Peng, Junjie Yan, Tieniu Tan, and Zhaoxiang Zhang. 2021. GAIA: A Transfer Learning System of Object Detection That Fits Your Needs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 274–283.

[14] Lisandro Dalcin and Yao-Lung L Fang. 2021. mpi4py: Status update after 12 years of development. *Computing in Science & Engineering* 23, 4 (2021), 47–54.

[15] Rohit Dhall and David L Kreitzman. 2016. Advances in levodopa therapy for Parkinson disease: review of RYTARY (carbidopa and levodopa) clinical efficacy and safety. *Neurology* 86, 14 Supplement 1 (2016), S13–S24.

[16] E Ray Dorsey and Bastiaan R Bloem. 2018. The Parkinson pandemic—a call to action. *JAMA neurology* 75, 1 (2018), 9–10.

[17] James Paul Dworkin. 2008. Laryngitis: types, causes, and treatments. *Otolaryngologic Clinics of North America* 41, 2 (2008), 419–436.

[18] Eduardo Fonseca, Manoj Plakal, Daniel PW Ellis, Frederic Font, Xavier Favory, and Xavier Serra. 2019. Learning sound event classifiers from web audio with noisy labels. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 21–25.

[19] Edgar Gabriel, Graham E Fagg, George Bosilca, Thara Angskun, Jack J Dongarra, Jeffrey M Squyres, Vishal Sahay, Prabhanjan Kambadur, Brian Barrett, Andrew Lumsdaine, et al. 2004. Open MPI: Goals, concept, and design of a next generation MPI implementation. In *European Parallel Virtual Machine/Message Passing Interface Users' Group Meeting*. Springer, 97–104.

[20] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. 2016. Domain-adversarial training of neural networks. *The journal of machine learning research* 17, 1 (2016), 2096–2030.

[21] Alexander M Goberman and Carl Coelho. 2002. Acoustic analysis of Parkinsonian speech I: Speech characteristics and L-Dopa therapy. *NeuroRehabilitation* 17, 3 (2002), 237–246.

[22] JI Godino-Llorente, S Shattuck-Hufnagel, JY Choi, L Moro-Velázquez, and JA Gómez-García. 2017. Towards the identification of Idiopathic Parkinson's Disease from the speech. New articulatory kinetic biomarkers. *PloS one* 12, 12 (2017), e0189583.

[23] Marvin M Goldenberg. 2008. Medical management of Parkinson's disease. *Pharmacy and Therapeutics* 33, 10 (2008), 590.

[24] Daniel Griffin and Jae Lim. 1984. Signal estimation from modified short-time Fourier transform. *IEEE Transactions on acoustics, speech, and signal processing* 32, 2 (1984), 236–243.

[25] Tamás Grósz, Róbert Busa-Fekete, Gábor Gosztolya, and László Tóth. 2015. Assessing the degree of nativeness and Parkinson's condition using Gaussian processes and deep rectifier neural networks. (2015).

[26] Otkrist Gupta and Ramesh Raskar. 2018. Distributed learning of deep neural network over multiple agents. *Journal of Network and Computer Applications* 116 (2018), 1–8.

[27] John A Hartigan and Manchek A Wong. 1979. Algorithm AS 136: A k-means clustering algorithm. *Journal of the royal statistical society. series c (applied statistics)* 28, 1 (1979), 100–108.

[28] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 9729–9738.

[29] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*. 1026–1034.

[30] Hynek Hermansky. 1990. Perceptual linear predictive (PLP) analysis of speech. *the Journal of the Acoustical Society of America* 87, 4 (1990), 1738–1752.

[31] Hynek Hermansky, Nelson Morgan, Aruna Bayya, and Phil Kohn. 1991. RASTA-PLP speech analysis. In *Proc. IEEE Int'l Conf. Acoustics, speech and signal processing*, Vol. 1. 121–124.

[32] Margaret M Hoehn, Melvin D Yahr, et al. 1998. Parkinsonism: onset, progression, and mortality. *Neurology* 50, 2 (1998), 318–318.

[33] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 29 (2021), 3451–3460.

[34] Jie Hu, Li Shen, and Gang Sun. 2018. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 7132–7141.

[35] Sergey Ioffe and Christian Szegedy. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*. PMLR, 448–456.

[36] Laetitia Jeancolas, Dijana Petrovska-Delacrétaz, Graziella Mangone, Badr-Eddine Benkelfat, Jean-Christophe Corvol, Marie Vidailhet, Stéphane Lehéricy, and Habib Benali. 2021. X-vectors: New quantitative biomarkers for early Parkinson's disease detection from speech. *Frontiers in Neuroinformatics* 15 (2021), 578369.

[37] Zachary Kabelac, Christopher G Tarolli, Christopher Snyder, Blake Feldman, Alistair Glidden, Chen-Yu Hsu, Rumen Hristov, E Ray Dorsey, and Dina Katabi. 2019. Passive monitoring at home: a pilot study in Parkinson disease. *Digital biomarkers* 3, 1 (2019), 22–30.

[38] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[39] Yusuke Koda, Jihong Park, Mehdi Bennis, Koji Yamamoto, Takayuki Nishio, and Masahiro Morikura. 2019. One pixel image and RF signal based split learning for mmWave received power prediction. In *Proceedings of the 15th International Conference on emerging Networking EXperiments and Technologies*. 54–56.

[40] Yusuke Koda, Jihong Park, Mehdi Bennis, Koji Yamamoto, Takayuki Nishio, Masahiro Morikura, and Kota Nakashima. 2020. Communication-efficient multimodal split learning for mmWave received power prediction. *IEEE Communications Letters* 24, 6 (2020), 1284–1288.

[41] Elina Kuosmanen, Valerii Kan, Aku Visuri, Assam Boudjelthia, Lokmane Krizou, and Denzil Ferreira. 2019. Measuring Parkinson's disease motor symptoms with smartphone-based drawing tasks. In *Adjunct Proceedings of the 2019 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2019 ACM International Symposium on Wearable Computers*. 1182–1185.

[42] Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. 2020. Federated optimization in heterogeneous networks. *Proceedings of Machine Learning and Systems* 2 (2020), 429–450.

[43] Xiaoxiao Li, Meirui Jiang, Xiaofei Zhang, Michael Kamp, and Qi Dou. 2021. Fedbn: Federated learning on non-iid features via local batch normalization. *arXiv preprint arXiv:2102.07623* (2021).

[44] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*. 2980–2988.

[45] Loren Lugosch, Tatiana Likhomanenko, Gabriel Synnaeve, and Ronan Collobert. 2021. Pseudo-Labeling for Massively Multilingual Speech Recognition. *arXiv preprint arXiv:2111.00161* (2021).

[46] Juan Carlos Martínez-Castrillo, Pablo Martínez-Martín, Ángel Burgos, Gloria Arroyo, Natalia García, María Rosario Luquín, and José Matías Arbelo. 2021. Prevalence of Advanced Parkinson's Disease in Patients Treated in the Hospitals of the Spanish National Healthcare System: The PARADISE Study. *Brain Sciences* 11, 12 (2021), 1557.

[47] Leland McInnes, John Healy, and James Melville. 2018. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426* (2018).

[48] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. 2017. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*. PMLR,

1273–1282.

[49] Juan JG Meilán, Francisco Martínez-Sánchez, Juan Carro, José A Sánchez, and Enrique Pérez. 2012. Acoustic markers associated with impairment in language processing in Alzheimer's disease. *The Spanish journal of psychology* 15, 2 (2012), 487–494.

[50] Nick Miller, Uma Nath, Emma Noble, and David Burn. 2017. Utility and accuracy of perceptual voice and speech distinctions in the diagnosis of Parkinson's disease, PSP and MSA-P. *Neurodegenerative disease management* 7, 3 (2017), 191–203.

[51] Laureano Moro-Velazquez, Jorge A Gomez-Garcia, Julian D Arias-Londoño, Najim Dehak, and Juan I Godino-Llorente. 2021. Advances in Parkinson's disease detection and assessment using voice and speech: A review of the articulatory and phonatory aspects. *Biomedical Signal Processing and Control* 66 (2021), 102418.

[52] Laureano Moro-Velazquez, Jesus Villalba, and Najim Dehak. 2020. Using x-vectors to automatically detect parkinson's disease from speech. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 1155–1159.

[53] Movement Disorder Society Task Force on Rating Scales for Parkinson's Disease. 2003. The unified Parkinson's disease rating scale (UPDRS): status and recommendations. *Movement Disorders* 18, 7 (2003), 738–750.

[54] Juan Rafael Orozco-Arroyave, JC Vdsquez-Correa, Florian Hönig, Julián D Arias-Londono, Jesús Francisco Vargas-Bonilla, Sabine Skodda, Jan Rusz, and E Noth. 2016. Towards an automatic monitoring of the neurological state of Parkinson's patients from speech. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 6490–6494.

[55] Dario Pasquini, Giuseppe Ateniese, and Massimo Bernaschi. 2021. Unleashing the tiger: Inference attacks on split learning. In *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*. 2113–2129.

[56] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems* 32 (2019).

[57] Adriano Petry and Dante Augusto Couto Barone. 2002. Speaker identification using nonlinear dynamical features. *Chaos, Solitons & Fractals* 13, 2 (2002), 221–231.

[58] Ngoc-Quan Pham, Alex Waibel, and Jan Niehues. 2022. Adaptive multilingual speech recognition with pretrained models. *arXiv preprint arXiv:2205.12304* (2022).

[59] Raspberry Pi. 2022. Raspberry pi 4 model B. https://www.raspberrypi.com/products/raspberry-pi-4-model-b

[60] Shige Qi, Peng Yin, Linhong Wang, Ming Qu, Ge Lin Kan, Han Zhang, Qingjun Zhang, Yize Xiao, Ying Deng, Zhong Dong, et al. 2021. Prevalence of Parkinson's disease: A community-based study in China. *Movement Disorders* 36, 12 (2021), 2940–2944.

[61] K Uma Rani and Mallikarjun S Holi. 2012. Analysis of speech characteristics of neurological diseases and their classification. In *2012 Third International Conference on Computing, Communication and Networking Technologies (ICCCNT'12)*. IEEE, 1–6.

[62] Douglas Roth and Berrylin J Ferguson. 2010. Vocal allergy: recent advances in understanding the role of allergy in dysphonia. *Current Opinion in Otolaryngology & Head and Neck Surgery* 18, 3 (2010), 176–181.

[63] Betul Erdogdu Sakar, M Erdem Isenkul, C Okan Sakar, Ahmet Sertbas, Fikret Gurgen, Sakir Delil, Hulya Apaydin, and Olcay Kursun. 2013. Collection and analysis of a Parkinson speech dataset with multiple types of sound recordings. *IEEE Journal of Biomedical and Health Informatics* 17, 4 (2013), 828–834.

[64] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. 2018. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4510–4520.

[65] Anthony HV Schapira, K Chaudhuri, and Peter Jenner. 2017. Non-motor features of Parkinson disease. *Nature Reviews Neuroscience* 18, 7 (2017), 435–450.

[66] Sigurdur Sigurdsson, Kaare Brandt Petersen, and Tue Lehn-Schiøler. 2006. Mel Frequency Cepstral Coefficients: An Evaluation of Robustness of MP3 Encoded Music.. In *ISMIR*. 286–289.

[67] Sabine Skodda, W Grönheit, N Mancinelli, and U Schlegel. 2013. Progression of voice and speech impairment in the course of Parkinson's disease: a longitudinal study. *Parkinson's disease* 2013 (2013).

[68] David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur. 2018. X-vectors: Robust dnn embeddings for speaker recognition. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 5329–5333.

[69] Zheng-Hua Tan, Najim Dehak, et al. 2020. rVAD: An unsupervised segment-based robust voice activity detection method. *Computer speech & language* 59 (2020), 1–21.

[70] Zhiyi Tian, Lei Cui, Jie Liang, and Shui Yu. 2022. A Comprehensive Survey on Poisoning Attacks and Countermeasures in Machine Learning. *ACM Computing Surveys (CSUR)* (2022).

[71] Evaldas Vaiciukynas, Antanas Verikas, Adas Gelzinis, and Marija Bacauskiene. 2017. Detecting Parkinson's disease from sustained phonation and speech signals. *PloS one* 12, 10 (2017), e0185613.

[72] Juan Camilo Vasquez-Correa, Tomas Arias-Vergara, Philipp Klumpp, Paula Andrea Pérez-Toro, Juan Rafael Orozco-Arroyave, and Elmar Nöth. 2021. End-2-End Modeling of Speech and Gait from Patients with Parkinson's Disease: Comparison Between High Quality Vs. Smartphone Data. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 7298–7302.

[73] Praneeth Vepakomma, Otkrist Gupta, Tristan Swedish, and Ramesh Raskar. 2018. Split learning for health: Distributed deep learning without sharing raw patient data. *arXiv preprint arXiv:1812.00564* (2018).

[74] Matthew Wallingford, Hao Li, Alessandro Achille, Avinash Ravichandran, Charless Fowlkes, Rahul Bhotika, and Stefano Soatto. 2022. Task Adaptive Parameter Sharing for Multi-Task Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 7561–7570.

[75] Jingjing Xu, Xu Sun, Zhiyuan Zhang, Guangxiang Zhao, and Junyang Lin. 2019. Understanding and improving layer normalization. *Advances in Neural Information Processing Systems* 32 (2019).

[76] Shoujiang Xu and Zhigeng Pan. 2020. A novel ensemble of random forest for assisting diagnosis of Parkinson's disease on small handwritten dynamics dataset. *International Journal of Medical Informatics* 144 (2020), 104283.

[77] Shu Yang, Fengbo Wang, Liqiong Yang, Fan Xu, Man Luo, Xiaqing Chen, Xixi Feng, and Xianwei Zou. 2020. The physical significance of acoustic parameters and its clinical significance of dysarthria in Parkinson's disease. *Scientific Reports* 10, 1 (2020), 11776.

[78] Yuzhe Yang, Yuan Yuan, Guo Zhang, Hao Wang, Ying-Cong Chen, Yingcheng Liu, Christopher G Tarolli, Daniel Crepeau, Jan Bukartyk, Mithri R Junna, et al. 2022. Artificial intelligence-enabled detection and assessment of Parkinson's disease using nocturnal breathing signals. *Nature medicine* (2022), 1–9.

[79] Kaichao You, Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I. Jordan. 2019. Universal Domain Adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[80] Binbin Zhang, Hang Lv, Pengcheng Guo, Qijie Shao, Chao Yang, Lei Xie, Xin Xu, Hui Bu, Xiaoyu Chen, Chenchen Zeng, et al. 2022. Wenetspeech: A 10000+ hours multi-domain mandarin corpus for speech recognition. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 6182–6186.

[81] Hanbin Zhang, Chen Song, Aosen Wang, Chenhan Xu, Dongmei Li, and Wenyao Xu. 2019. Pdvocal: Towards privacy-preserving parkinson's disease detection using non-speech body sounds. In *The 25th Annual International Conference on Mobile Computing and Networking*. 1–16.

[82] Baozhu Zuo. 2022. Respeaker 6-mic circular array kit for Raspberry Pi. https://wiki.seeedstudio.com/ReSpeaker_6-Mic_Circular_Array_kit_for_Raspberry_Pi