The phonetic footprint of Parkinson's disease

Philipp Klumpp, Tomás Arias-Vergara, Juan Camilo Vásquez-Correa, Paula Andrea Pérez-Toro, Juan Rafael Orozco-Arroyave, Anton Batliner, Elmar Nöth

arXiv:2112.11514v1 [eess.AS] 21 Dec 2021

The Phonetic Footprint of Parkinson's Disease

Philipp Klumpp[a,*], Tomás Arias-Vergara[a,b], Juan Camilo Vásquez-Correa[a,b], Paula Andrea Pérez-Toro[a,b], Juan Rafael Orozco-Arroyave[b], Anton Batliner[a,c], Elmar Nöth[a]

[a] *Friedrich-Alexander-University Erlangen-Nuremberg, Pattern Recognition Lab, Martensstrasse 3, 91058 Erlangen, Germany*
[b] *Universidad de Antioquia, Medellín, Colombia*
[c] *Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Germany*

Abstract

As one of the most prevalent neurodegenerative disorders, Parkinson's disease (PD) has a significant impact on the fine motor skills of patients. The complex interplay of different articulators during speech production and realization of required muscle tension become increasingly difficult, thus leading to a dysarthric speech. Characteristic patterns such as vowel instability, slurred pronunciation and slow speech can often be observed in the affected individuals and were analyzed in previous studies to determine the presence and progression of PD. In this work, we used a phonetic recognizer trained exclusively on healthy speech data to investigate how PD affected the phonetic footprint of patients. We rediscovered numerous patterns that had been described in previous contributions although our system had never seen any pathological speech previously. Furthermore, we could show that intermediate activations from the neural network could serve as feature vectors encoding information related to the disease state of individuals. We were also able to directly correlate the expert-rated intelligibility of a speaker with the mean confidence of phonetic predictions. Our results support the assumption that pathological data is not necessarily required to train systems that are capable of analyzing PD speech.

*Corresponding author
Email address: philipp.klumpp@fau.de (Philipp Klumpp)
URL: https://lme.tf.fau.de/person/klumpp/ (Philipp Klumpp)

1. Introduction

Parkinson's disease (PD) is one of the most common neurological conditions [1] in aging societies. This neurodegenerative disorder is strongly characterized by its progressive motor symptoms which can be classified into four types: bradykinesia (slowness of movement), tremor (involuntary muscle activity), rigidity (freezing of gait) and postural instability [2]. Whilst all these indications are increasingly visible in an individual as the degeneration of brain cells progresses, the ongoing deterioration has long become audible. The production of healthy speech requires a well-coordinated fine-motor interaction between the different articulators. Throughout the different stages of PD, fine-motor deficits lead to a number of speech impairments that are often referred to as Parkinsonian Dysarthria. It comprises various articulatory (slurred consonants), prosodic (monotonous pitch), phonatory (breathy voice) and respiratory (decreased loudness) symptoms [3]. A computational assessment of speech signals can be performed to either classify the presence of PD or to monitor the progression of the ongoing neurological disorder. Previous studies used phonation and articulation features [4], voice onset time [5], autoencoder representations [6] and other representations to detect the condition. Voice onset and offset features have also proven informative to track the progression of PD [7]. In a very recent study, the speech signal was utilized as a surrogate to estimate the medication state of patients [8]. Acoustic models trained on large datasets from healthy speakers had already been used in previous studies [9] to predict the intelligibility of dysarthric speakers. Unlike the approach presented in this article, the described system required a phonetic reference to perform automatic speech alignment. The system described in [10] was also trained exclusively with healthy speakers (seven male subjects). Despite the very small sample size, a correlation between the speech recognition performance and Frenchay

Dysarthria Assessment scores could be reported. Training background models from larger datasets that only included healthy contributors was also performed for gait analysis [11]. Robust features could be learned from the signals of healthy participants that were then used to classify PD. In a multi-modal study setup, the progression of PD was predicted with universal background models estimated from speech, handwriting and gait features of healthy subjects [12].

A major weakness of most classification and regression (monitoring in the case of PD) methods is their lack of generalization when parameters have to be estimated from a small dataset. For modern deep architectures, the problem of small PD datasets can be dealt with by applying transfer learning methods using data from a secondary domain [13]. However, this solution is not optimal. The parameter-intensive feature extraction part could be trained with a large dataset from a different domain. Afterwards, these parameters are frozen and only the final layers would have to be estimated for the low-resource classification problem. Whilst this allows for a robust feature extraction, the following feature interpretation could be strongly affected by acoustic conditions, utilized hardware, signal pre-processing or study design, to mention a few.

Another important factor that naturally hinders generalization of machine learning models is the varying progression of the disease among patients. The presence and further development of symptoms is different for every individual [14]. This implies that any PD monitoring solution must be able to adapt to a patient-specific baseline instead of the one from a large study cohort.

With respect to study designs, most related works perform speech analysis on signals acquired from dedicated speech tasks, such as sustained vowels or diadochokinesis (DDK) tasks (rapid repetition of consonant-vowel clusters) [15, 16, 17, 18]. The major advantage of this setup is the combination of exercise and data collection. The main disadvantage is quite obvious as well: The trained model is bound to such speech exercises, hence unable to interpret free speech. To overcome all the presented drawbacks, we propose the computation of phonetic footprints based on a fundamental acoustic analysis of speech signals. A phonetic footprint resembles the distribution of production probabilities among

different phonemes or phonological classes for an individual speaker. The core idea behind the design of our recognizer was that it is trained only with speech samples from healthy speakers. This enabled us to include large general-purpose automatic speech recognition (ASR) datasets into the training procedure. To improve generalization even more and ensure that we are not learning strong language dependent patterns, we trained our model with independent datasets from three different languages. We used the final model to compute acoustic class probability densities for healthy subjects as well as PD patients to show how the phonetic footprint of speech signals changed between groups. Step by step we demonstrate how alterations in articulation are conserved from speech exercises over read text all the way to free speech. Furthermore, we tried not only to identify changes in speech productions, but also to link these changes to the involved articulators.

The last hypothesis of this work arose from the nature of our training data. Our acoustic model has barely seen any dysarthric speech sample before. For a human listener, we know that the more dysarthric a speech sample sounds, the more unintelligible it becomes. We demonstrate that this perceived decrease of intelligibility was very well conserved in the results of our recognition model.

Phonetic footprints were not designed to serve as a basis for classifying PD. In fact, in a study about the evolution of PD diagnosis, only 11 % of patients reported speech problems as one of the initial symptoms [19]. On the other side, more than three out of four patients reported tremor at the beginning of the disease. To mitigate PD symptoms, patients are being administered a pre-stage of the neurotransmitter dopamine. Only this pre-stage is able to pass the blood-brain barrier and make up for the lack of dopamine in the basal ganglia which was caused by the loss of dopaminergic brain cells [20]. The therapy success with levodopa has to be controlled on a regular basis, because at some point, the drug itself would cause strong motor symptoms which render the therapy useless [20]. An automatic and unobtrusive speech analysis could provide such supervision of therapy success on a daily basis and without requiring an expert clinician. Symptoms are developed differently by individual patients and their

response to treatment with levodopa varies strongly [21]. A phonetic footprint would have to be estimated individually for each patient to serve as a unique baseline. After the initial calibration, the footprint could be updated regularly with speech samples collected from different voice-user interfaces, such as smartphones or smart speakers. A larger amount of speech samples could also help to alleviate the influence of recognition errors on the final footprint. With the knowledge about how PD would affect the average footprint of a person as it progresses, it would be possible to search for distinct patterns, deviations from their individual baseline, to track if a patient responded well to therapy, and if not, whether this deviation was only "a bad day" or manifested over weeks. The study of phonetic footprints could ultimately improve PD monitoring in terms of availability, closer coverage and better adaption to the individual patient.

After a brief introduction into the different datasets used for training and PD analysis, we want to outline the multilingual concept of PHONE unions employed in this study. We then explain our acoustic model and how it was used to extract knowledge from unseen data. In the results section, we embed our findings with respect to different degrees of freedom in speech production, ranging from constricted speech exercises to free speech. We will then interpret the outcome of our experiments both with respect to the detected characteristics of PD speech as well as their progression throughout disease stages in the discussion. Finally, we want to highlight the key strengths of our method and we will point out remaining weak spots.

2. Materials and Methods

2.1. Training data

Our acoustic model was trained with German, English and Spanish ASR datasets. They were carefully selected due to their high quality in annotation and acoustic properties. The high annotation quality would result in fewer phonetic label errors, thus giving us a more reliable acoustic model. The PD dataset was recorded in a noise-free environment with high-quality microphones. We

5

Table 1: Summary of training data distribution before and after augmentation.

Language	Gender [f / m]	Hours before Aug.	Hours after Aug.
German	307 / 286	29.1	58.2
English	56 / 112	5.3	10.6
Mexican Spanish	51 / 49	6.2	12.4
Total	**414 / 447**	**40.6**	**81.2**

therefore tried to train our acoustic models with data of similar acoustic properties. We used a subset of the German Verbmobil corpus [22] containing around 29 hours of dialogue speech recordings from 593 speakers (307 female, 286 male). This was the largest of the three datasets. For our purpose we downsampled the data to 16 kHz using 16 bits/sample and mono-channel configuration. We distributed the speakers randomly into training (26.1 h), development (1.5 h) and test (1.5 h) sets. Phonetic segmentations were created by forced-alignment with the Kaldi [23] ASR toolkit.

We also incorporated the TIMIT Acoustic-Phonetic Continuous Speech Corpus [24]. It is composed of American English speech samples collected from a total of 630 speakers, 438 (70%) male and 192 (30%) female, from eight major dialect regions. Every speaker contributed ten recordings of phonetically rich text reading tasks. The PCM audio files had been recorded with a resolution of 16 bits at a sampling rate of 16 kHz and mono-channel configuration. The TIMIT corpus was hand-labeled to provide time-aligned segmentation at the orthographic, phonetic and word level. It is distributed with a predefined split into training (3.9 h) and test (1.4 h) sets. 168 of the original 630 speakers were part of the test subset, 112 (67%) male and 56 (33%) female. During training, we used a portion of 5% of all speakers from the training split for validation.

Our PD dataset, which is described in subsection 2.2, was collected from native Colombian Spanish speakers. Therefore, we decided to include a Spanish corpus into the training as well. The DIMEx100 corpus [25] was recorded with 100 native Mexican Spanish speakers (49% male, 51% female). Every individual

contributed 50 unique and 10 reference (identical for all speakers) phrases. Audio files were recorded in raw format at 44.1 kHz (once more we downsampled to 16 kHz), 16 bit resolution for a single channel. All speakers were distributed into training (5 h), development (0.6 h) and test set (0.6 h).

Every audio sample was used in its clean version and also with added Gaussian noise at different SNR levels (randomly chosen from 10 dB, 15 dB, 20 dB). The overall distribution of training data is summarized in Table 1.

2.2. Parkinson's disease dataset

In this study, we used the extended version of the PC-GITA corpus [26] collected from native Colombian Spanish speakers. It consists of 68 pathological and 50 healthy control (HC) speakers. We reduced this distribution to those 58 (PD) and 48 (HC) who were recorded in a (portable) soundproof booth and had been assessed with respect to the modified Frenchay Dysarthria Assessment (mFDA) [26], an adapted version of the original FDA [27] for assessments from speech recordings. FDA is a common procedure to evaluate speech symptoms of PD patients. Unlike the very popular Unified Parkinson Disease Rating Scale (UPDRS) [28], which mainly focuses on motor abilities, FDA only evaluates the oromotor abilities of a subject, thus providing a better impression of an individuals ability to articulate. Only including recordings from a noise-free environment was important to ensure equal acoustic conditions throughout our experiments. Every individual contributed samples of six diadochokinetic exercises, ten isolated sentences, one read text and one monologue. Audio signals were recorded in raw PCM, 16 bits resolution and mono-channel configuration at 16 kHz. Every participant was assessed by an expert therapist with respect to mFDA. The score aims to estimate an individual's articulatory abilities with respect to various items, namely the lips [score range: 0 - 8], palate [0 - 8], larynx [0 - 12], respiration [0 - 8], intelligibility [0 - 4], monotonicity [0 - 4] and tongue [0 - 8]. This allows for a more precise localization of the morphological origin of a speech impairment inside the vocal tract. Throughout the whole mFDA value range [0 - 52], larger values always indicate a stronger impairment.

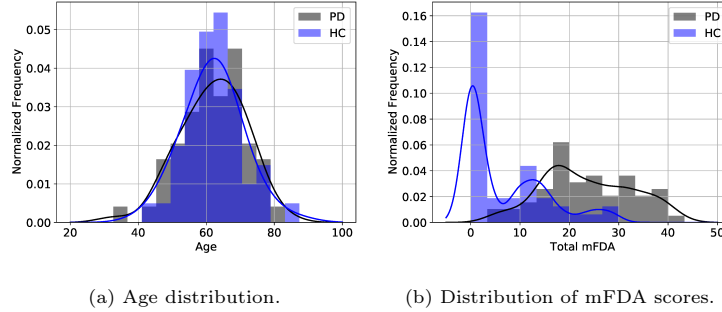(a) Age distribution.   (b) Distribution of mFDA scores.

Figure 1: Distribution of age and mFDA assessment scores for the groups of healthy speakers and Parkinson's patients.

The gender distribution among the HC group was 25 females and 23 males, among the PD patients it was 27 females, 31 males. The mean age was 62.0 (8.0 standard deviation) and 61.5 (9.3) for the HC and PD groups, respectively. The total mFDA score can be computed as a sum of all items described previously and was on average 6.6 (8.3) for the HC group and 22.7 (8.5) for the PD group. Figure 1 provides an additional insight into the age and mFDA score distribution among healthy subjects and PD patients.

2.3. Multilingual PHONE *concept*

For any given language, there is a phoneme inventory (consonant system and vowel system) whose members represent those units that are both perceptually distinct and distinguish words in this language. One or more phones can be a realisation of the same phoneme; they are then called allophones or combinatorial variants.

For our multilingual acoustic model, we had to revise this concept as the term *phoneme* is easily misused [29]. We instead decided to define a set of 35 PHONE unions (including one for silence), where every such PHONE would describe the set of elementary phones that could be used in German, English or Spanish to produce said PHONE class.

For some experiments, we also grouped PHONES into phonetic categories. This helped us interpret the overall capabilities of a subject with respect to certain

Table 2: List of all 34 PHONES (excluding silence) used in this work, according to the international phonetic alphabet (IPA).

PHONES	Coarse phonetic class	Fine phonetic class
l / j / w	Approximants	Approximants
u / i / y	Vowels	Closed Vowels
e / o / œ / ɛ / ɜ	Vowels	Mid-open Vowels
a	Vowels	Open Vowels
s / ʃ / z / ʒ	Fricatives	Sibilant Fricatives
θ / ð / h / v / x / f / ɣ	Fricatives	Non-sibilant Fricatives
n / m / ŋ / ɲ	Nasals	Nasals
r	Rhotics	Rhotics
p / t / k	Stops	Unvoiced Stops
b / d / g	Stops	Voiced Stops

classes of PHONES. We used a coarse and a fine set of phonetic classes. A full list of all phonetic categories is provided in Table 2.

2.4. Audio processing

Every audio recording was first resampled to 16 kHz when necessary, followed by a root mean square normalization to a level of −10 dB and removal of any DC offset. We then computed amplitude and phase spectrograms (2048 FFT points) over a window of 25 ms which was shifted by 5 ms. We found that including the phase information slightly improved the PHONE recognition results. Both spectrograms were converted to logarithmic scale of base 10 and afterwards filtered with a triangular Mel-bank [30] with 128 frequency bands to resemble the human perception of speech with a high resolution in the lower frequencies and a coarse resolution for high frequencies. The resulting dual-channel spectrograms were used throughout all experiments of this study.

2.5. Deep recurrent PHONE *recognition*

Our PHONE recognition model was comprised of two major components, a convolutional feature extraction part as well as a recurrent sequence analysis part. The whole neural network was trained in two separate steps to improve the final sequence prediction result. First, the network was optimized to perform framewise PHONE classification. At every time step t, the network would distribute a probability mass over 44 PHONE targets. Notice that for the first training stage, we had 9 additional PHONE targets. This was necessary because some of the PHONE labels of the described datasets were simply composites of other PHONE classes. For example, the ground truth class [ʧ] included in TIMIT is a composite of the PHONES [t] and [ʃ] that had already been defined. However, we could not just replace [ʧ] by the two underlying PHONES because we were missing the corresponding alignment information about where [t] ended and [ʃ] started. It would have been possible to compute new alignments for TIMIT, but then the exceptional quality of the dataset's manual phonetic annotation would have been lost.

A detailed description of the PHONE recognition model's architecture is provided in Appendix A. After we had successfully trained the framewise PHONE classifier, we slightly increased the parameter complexity of our recurrent network and retrained the recognizer, this time without alignment information. In other words, instead of telling the network about the explicit target PHONE at every time step, we used the connectionist temporal classification (CTC) loss function [31] to predict an alignment-free PHONE sequence. This allowed us to disassemble the composite PHONES into their elementary components. Furthermore, framewise phonetic annotations are usually prone to label imprecision due to poor forced alignments or unclear PHONE boundaries. With the CTC loss, our model learned to identify inherent PHONE boundaries by itself.

All proposed architectures were trained with Adam optimizer [32] with initial learning rate of 0.001. The learning rate decayed by a factor of 0.5 if the validation loss did not improve for at least five subsequent epochs. One training batch contained 20 samples and each of the samples comprised a subsequence

10

of 6 seconds duration chosen randomly over a recorded utterance. The core idea during pre-training was to learn a meaningful initialization of the convolutional feature extraction layers. After reaching convergence on this alignment task, we removed the two recurrent layers (200 hidden units) as well as the final classification layer for 44 PHONE targets and replaced them with two new recurrent layers (240 hidden units) and a classification layer projecting to the 35 sequential PHONES. During the framewise classification in the first step, we used cross entropy loss function. In the second stage, we used CTC loss function as described in [31]. To improve overall model generalization, parameters were penalized by applying $L2$ regularization [33] with a rate of 0.0001. During early epochs, we observed rather noisy validation loss values. One way of fixing this was to reduce the initial learning rate, but this would result in an overall slower training. Instead, we used gradient clipping to limit the $L2$ norm of gradients to a maximum value of 1.0. This approach has proven to be beneficial for training other RNN architectures as well [34].

A model trained with CTC loss still performs a framewise prediction, estimating a probility density function (PDF) over the total number of classes (35 in our case) and an additional helper class, commonly referred to as the blank label. A blank label indicates preservation of the previous state, meaning that the network's last emitted non-blank class is still present. This holds until a new non-blank class is predicted. For a given sequential input X of length T, one can decode the sequence of predicted states s of length $L \leq T$. The probability of observing path p can be defined as

$$P(p|X) = \prod_{t=1}^{T} y_c^t \tag{1}$$

where y_c^t denotes the probability of observing class c at time t. The probability of any sequence s can then be formulated as the sum of probabilities of all paths $p \in P_s$ that decode to s:

$$P(s|X) = \sum_{p \in P_s} P(p|X) \tag{2}$$

11

Consequently, the most probable path p might not always decode to the most probable sequence s. With a beam search decoding, it is possible to get a reliable estimate of s. To keep computational costs low, we decided to perform beam search with a beam width of 20, thus only keeping track of the 20 most probable PHONE sequences during decoding. It is common practice to extend equation 2 with an additional term representing a phonetic language model. We decided not to include any such language dependent transition probabilities because they are likely to conflict with certain patterns of dedicated speech exercises that are uncommon in spoken language.

For our experiments, we used the trained PHONE recognition model to predict three types of information. The most important one was the PHONE sequence. It is alignment-free and contains a single PHONE symbol for every PHONE realization, independent of its duration. For every such sequence, we estimated each PHONE's confidence by taking its posterior probability y_s^t from the most probable path $p \in P_s$. Lastly, we also extracted the concatenated hidden states of the forward and backward passes from the last BiLSTM at the particular timestep t as an intermediate, high-dimensional representation of this point in time.

2.6. Methodology

Although healthy speakers had also been assessed with respect to mFDA, they were never split into different sets according to their respective scores. The motivation was to compare pathological speech not only to unimpaired speech, but to speech from healthy subjects in general, among which pronunciation deficits could also occur for many other reasons.

2.6.1. Speech exercises

We first investigated speech productions from healthy and PD speakers collected during dedicated speech exercises. The exercise was split into three isolated tasks where an individual was asked to fluently repeat one of the syllables [pa], [ta] or [ka]. The group of PD patients was split into three sets according

to their mFDA scores for lips (syllable [pa]) and tongue ([ta] and [ka]) such that the distribution among sets was as equal as possible. All sets used in the following were created with the goal to create the most equal distributions. We then investigated the phonetic capabilities of all groups by estimating the production probability of different PHONES or phonetic classes. Our hypothesis is that the decreasing muscle tension in PD patients and the resulting slurred pronunciation of stop sounds would leave a noticeable footprint in their phonetic profile for each task. Imprecisions in the production of PHONES [p], [t] and [k] would indicate how severely the involved articulators (lips and tongue, respectively) were affected. In all histograms showing relative frequencies of PHONES or phonetic groups, error bars indicate the standard deviation from the respective mean value.

2.6.2. Intermediate network activations

For the dedicated speech tasks, we also computed a 1-dimensional decomposition of intermediate RNN hidden states that were associated with one of the three stop sounds by performing a principal component analysis (PCA). A very similar analysis of intermediate neural network activations was successfully applied in a previous study as well [35]. For the sake of readability, we only split the PD group into two sets in this case. In the first step, all hidden states from the last RNN layer that were classified as one of the three unvoiced stop PHONES [p], [t] or [k], were collected from all 106 participant's [pa], [ta] and [ka] syllable repetitions. We then computed three one-dimensional PCAs, one for each collection of hidden states. After applying this PCA, we sorted the results according to mFDA scores to identify differences between hidden states of HC, mildly and severely affected patients. Lastly, we computed the mean confidence of PHONE predictions over all DDK exercises as a measure of perceived intelligibility of the PHONE recognition model, sorted according to the total mFDA score. This evaluation was also done for all following experiments.

2.6.3. Isolated sentences

In the next step, we performed a phonetic analysis of isolated sentences. Here, we did not evaluate with respect to each individual PHONE anymore, but rather evaluated the coarse and fine phonetic classes, because unlike for the speech tasks, we could not define any target PHONES anymore. The group of PD speakers was now divided into two sets, sorted according to their mFDA intelligibility rating. Notice that intelligibility resembles more the overall pronunciation capabilities of an individual, whereas the previously used items where dedicated to certain functional organs in the vocal tract that are involved in the production of a certain PHONE.

2.6.4. Text reading and monologues

The analyses of read texts (multiple sentences) and monologues (free speech) were done in the same way as for the isolated sentences. All explained differences were found to be statistically significant by performing t-tests and observing p-values $\ll 0.001$.

To provide a better interpretation of the presented results, we performed a statistical analysis. We first determined all fine phonetic classes that were found to be significantly affected (as defined earlier in this subsection). For every class, we computed Spearman's correlation with the mFDA intelligibility score as well as the effect size (Cohen's d) for every set as a measure of how strong each group deviated from the healthy reference. For the remaining three speech tasks, we chose the most affected coarse and fine phonetic class from the dedicated speech exercises and determined how correlation with intelligibility score and effect sizes changed. For the mean PHONE confidences, we computed Spearman's correlation with respect to total mFDA score and mFDA intelligibility. Additionally, we determined the effect sizes of the individual PD-sets and tasks with respect to the healthy reference. Whenever we computed Spearman's correlation, we always considered the assigned mFDA values of HC speakers as well to get an exact representation of how well a certain property (e.g. the mean prediction confidences) resembled this score. Statistical analysis was not performed on the
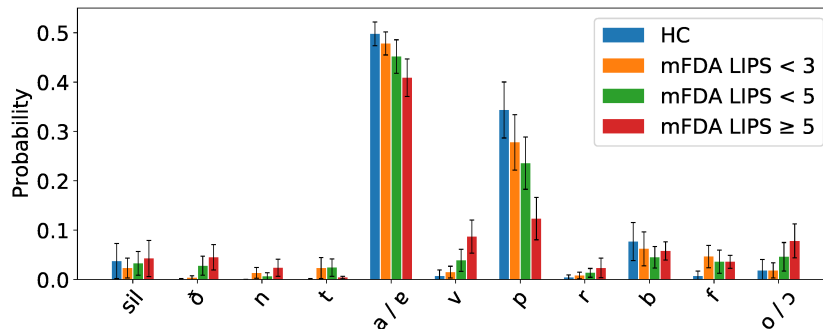
Figure 2: Mean posterior PHONE probabilities for the repetition of syllable [pa]. Error bars represent standard deviation from mean.

PCA results because the intention behind this experiment was only to show that an intermediate feature vector can encode PD-related information.

3. Results

3.1. PHONE recognition

Our framewise PHONE recognition model achieved a PHONE error rate (PER) of 18.3%. The final CTC model reached a PER of 16.6%. In this case, PER was computed as the ratio of deletions, insertions and replacements required to transform a predicted sequence to its ground truth. If we did not use the parameters from the framewise model's feature extraction layers as initialization (cold start) for the CTC model, we observed a slightly worse PER of 18.7%.

3.2. Phonetic analysis of Parkinson's Disease

3.2.1. Speech exercises

Figures 2 and 3 show the mean posterior probabilities (MPP) of different PHONES (only PHONES where at least one set had an MPP > 2%) and coarse phonetic classes for the repetition of syllable [pa], respectively. For the HC, the MPP for vowel [a] equals 49.8%, which is very close to the expected value of 50% if the syllable was repeated correctly. Notice that the expected value is not affected by the duration of PHONE productions, because the recognizer outputs
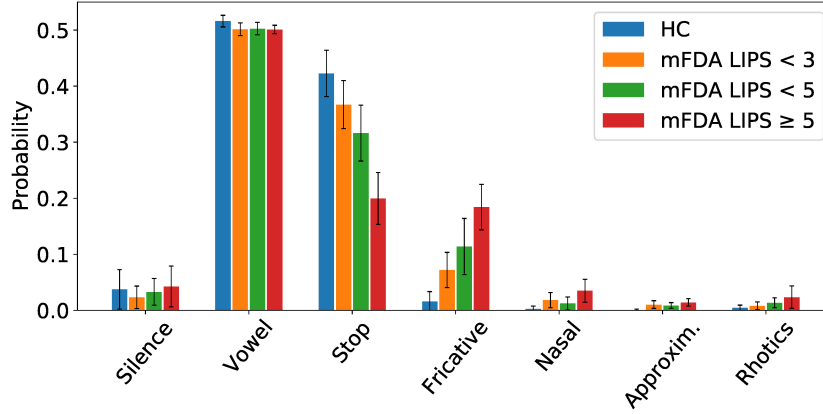
Figure 3: Mean posterior coarse phonetic class probabilities for the repetition of syllable [pa].

a single PHONE symbol per realization. PD patients with an mFDA lips score below 3 achieved [a] MPP of 47.8%. This value further decreased to 45.1% in the second set for mFDA lips scores ranging from 3 to below 5. The last group with scores greater or equal to 5 only reached 40.9%. Looking at the results in Figure 3, we observed that the MPP for vowels in general did not change much between the sets.

We found differences for the stop [p] to be even more pronounced. The MPP for HCs was 34.3%. The patients with low, intermediate and high scores for the lips item achieved 27.8%, 23.6% and 12.3% in MPP, respectively. Sometimes, instead of the unvoiced stop [p], we observed the voiced counterpart [b]. Such confusions could result from mispronunciations as well as from false recognition, because the two PHONES are very similar. For PD patients, we found that they have an increased probability of replacing the stop with fricatives such as [f], [v] or [ð]. While healthy participants had an MPP of 1.6% to produce a faulty fricative during the exercise, this value constantly increased along with the mFDA lips score, and reached a value of 18.4% for the strongly impaired group.

For the repetition of syllable [ta], the differences between sets appeared to be less pronounced compared to those of the first diadochokinesis (DDK) task. Figure 4 shows the MPPs for all relevant PHONES again. The set of mildly
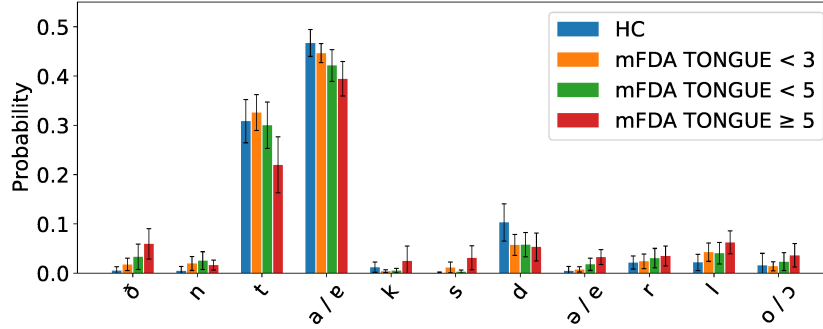
16

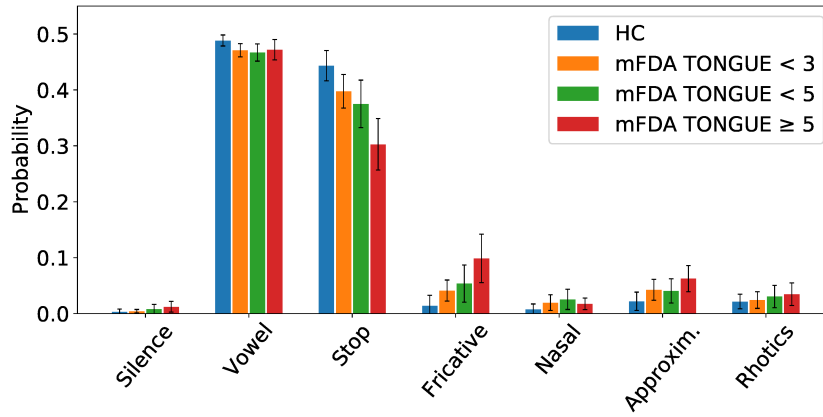Figure 4: Mean posterior PHONE probabilities for the repetition of syllable [ta].

Figure 5: Mean posterior coarse phonetic group probabilities for the repetition of syllable [ta].

affected PD patients according to the mFDA score for the tongue item even achieved a slightly higher MPP for the stop [t]. In contrast to this, the HC group produced far more [d] stops, the voiced counterpart of [t]. Looking at the phonetic classes (Figure 5), the MPP of a stop sound in general (voiced or unvoiced) was still higher for the HC. When the stop [t] was not produced correctly, patients produced more fricatives and approximants at the same time. For the vowel [a], we observed the same results as before: The highest value for the HC group was 46.7%, followed by a constant decline in MPP down to 39.4% for patients with strong deficits associated with their tongue.

Results for the last DDK task (syllable [ka]) are presented in Figures 6 and 7. The stability of vowel [a] showed the same characteristics as in the exercises
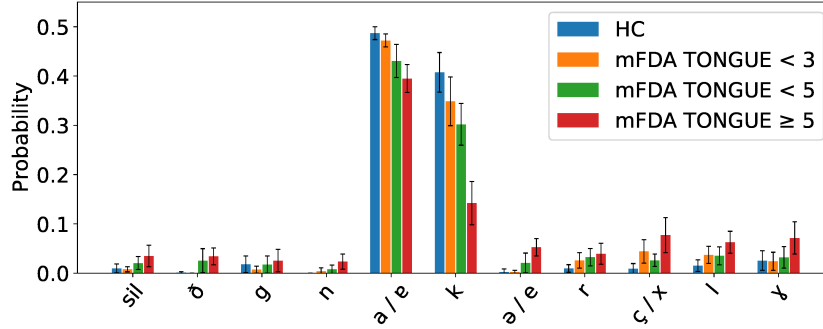
17

Figure 6: Mean posterior PHONE probabilities for the repetition of syllable [ka].
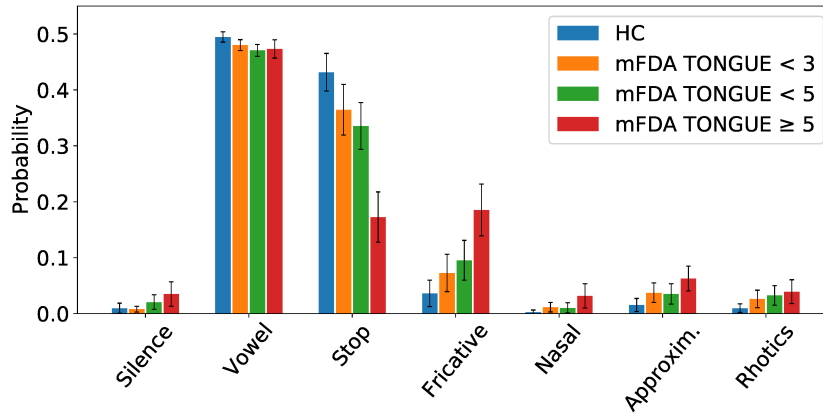

Figure 7: Mean posterior coarse phonetic group probabilities for the repetition of syllable [ka].

before. The stop sound [k] had an MPP of 40.7% for the HC. Mildly affected patients produced the correct stop with an MPP of 35.0%, intermediates reached 30.2%. The PD participants with an mFDA tongue score equal or greater than 5 only had an MPP of 14.2%. Among the replacing PHONES were fricatives ([ð], [x], [ɣ]), rhotics ([r]) and approximants ([l]). For all three consonant-vowel-cluster exercises, we found that the standard deviations for the respective unvoiced and voiced stop PHONES were mostly higher than for the vowel. A similar pattern was observed for the coarse phonetic groups of stops and fricatives when compared to the low standard deviation for the vowels.

To better interpret the overall results, we visualized MPPs for the fine phonetic classes in Figure 8. As we now looked at the complete set of exercises, we
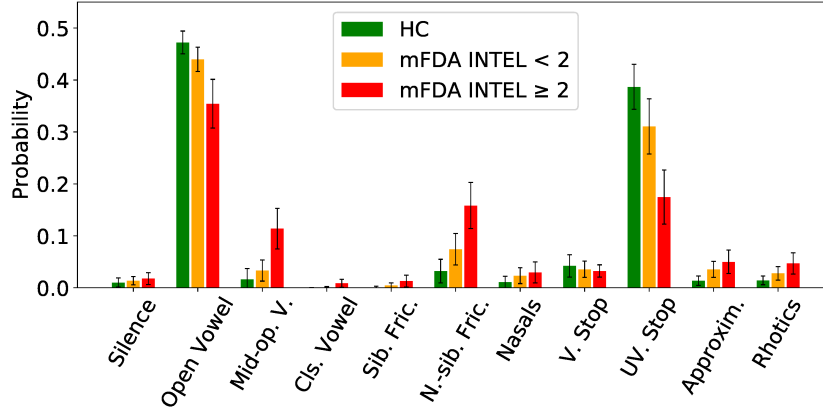
Figure 8: Mean posterior fine phonetic group probabilities for all DDK tasks.

did not sort by an mFDA item associated to one particular articulator (e.g. the lips). Instead, we split the PD group into two sets by their overall intelligibility score. Throughout all exercises, we observed that the MPP for open vowels decreased along with the severity of speech symptoms experienced by a patient. In that case, it was very common for individuals to replace [a] with schwa [ə] which was included in PHONE class [e]. The slight decline of MPP for voiced stops and the increasing amount of intermittent silence were both not considered significant with a t-test: for silence: p = 0.11, for voiced stop: p = 0.06. On the other side, the increasing probabilities of non-sibilant fricatives and the decline for unvoiced stops were found to be more expressive.

Table 3 shows the results for a statistical evaluation of the entire DDK results. We found a positive correlation with mFDA intelligibility scores for MPPs of mid-open vowels and non-sibilant fricatives and a negative correlation for the open vowels. A clear negative correlation was found for the unvoiced stop. Effect sizes were large for the set of PD patients with an intelligibility score below 2 for three out of the four tasks. For the set of poorly intelligible patients, we observed strong effect sizes compared to the healthy reference.

Table 3: Correlation between phonetic classes' mean posterior probability and mFDA intelligibility score. Columns 3 and 4 show effect sizes (Cohen's d) between healthy reference and the respective split according to mFDA intelligibility score.

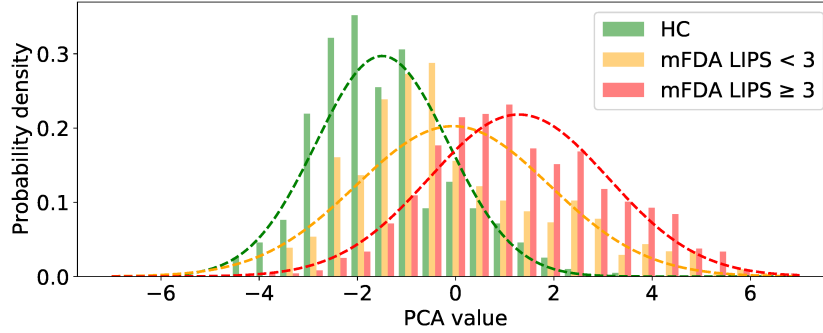| Phonetic class	Spearman's ρ	mFDA Intel. < 2	mFDA Intel. ≥ 2
Mid-open Vowel	0.57	0.41	1.81
Open Vowel	-0.59	0.72	1.92
Non-sib. Fricative	0.64	0.79	2.11
Unvoiced Stop	-0.71	0.79	2.32

Figure 9: Distribution of PCA values computed from RNN hidden states at stop [p] (syllable [pa]) for HC, mildly and severely affected PD patients according to their mFDA lips score. Dashed curves illustrate underlying Gaussian distributions.

3.2.2. Intermediate network activations

The PCA results for the three DDK tasks are depicted in Figures 9, 10 and 11. For the syllable [pa], the mean PCA value of the HC group was -1.5 ($\sigma^2 = 1.3$). Mean values of the PD patients with low and high mFDA lips score were -0.06 ($\sigma^2 = 2.0$) and 1.3 ($\sigma^2 = 1.8$), respectively. We observed a marked deviation in the PCA embeddings of hidden states between the three sets, although all hidden states were ultimately classified to the same PHONE, the unvoiced stop [p].

The results for syllable [ta] indicated that the production of the stop sound [t] would also change for PD patients. Here, the mean PCA value of the control group was -1.5 ($\sigma^2 = 1.6$). Patients with an mFDA tongue score of not more
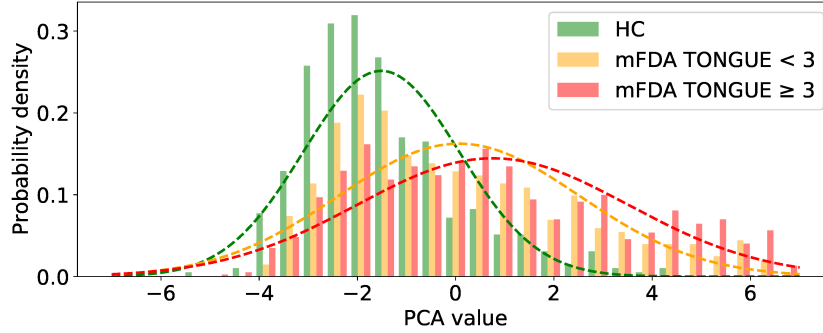
20

Figure 10: Distribution of PCA values computed from RNN hidden states at stop [t] (syllable [ta]) for HC, mildly and severely affected PD patients according to their mFDA tongue score. Dashed curves illustrate underlying Gaussian distributions.
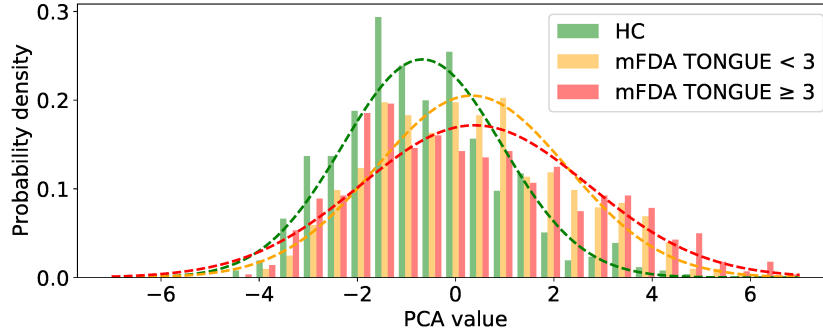

Figure 11: Distribution of PCA values computed from RNN hidden states at stop [k] (syllable [ka]) for HC, mildly and severely affected PD patients according to their mFDA tongue score. Dashed curves illustrate underlying Gaussian distributions.

than 2 had an average PCA value of 0.1 ($\sigma^2 = 2.5$), and those with scores above 2 were found to have $\mu = 0.8$ ($\sigma^2 = 2.8$). Compared to the results of the first DDK task, the PCA values of the two PD sets showed a greater overlap for syllable [ta].

The last DDK task to evaluate was the production of stop sound [k] in the syllable [ka]. Here, we found the smallest deviation between PCA values among the different sets when sorting them according to the mFDA tongue score. The mean value of the control group was -0.7 ($\sigma^2 = 1.6$). The PD patients with low and high mFDA scores showed mean values of 0.3 ($\sigma^2 = 1.9$) and 0.4 ($\sigma^2 = 2.3$).
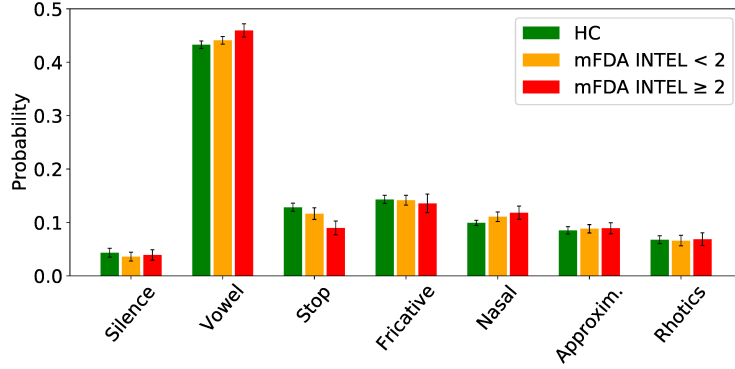
Figure 12: Mean posterior probabilities of coarse phonetic classes computed for isolated sentences.

Figure 13: Mean posterior probabilities of fine phonetic classes computed for isolated sentences.

We also sorted the PCA results by mFDA intelligibility score, which resulted in greater differences between the sets. This was particularly interesting for the stops [t] and [k], because we already found [p] stops to show a good separation when sorted according to the involved articulator's score (lips). The mean values for [t] of the three sets after arranging them by intelligibility were -1.5 ($\sigma^2 = 1.6$), 0.4 ($\sigma^2 = 2.7$) and 1.0 ($\sigma^2 = 2.0$). For the last DDK task and stop [k], we observed mean values of -0.7 ($\sigma^2 = 1.6$), 0.2 ($\sigma^2 = 2.1$) and 1.4 ($\sigma^2 = 2.2$).

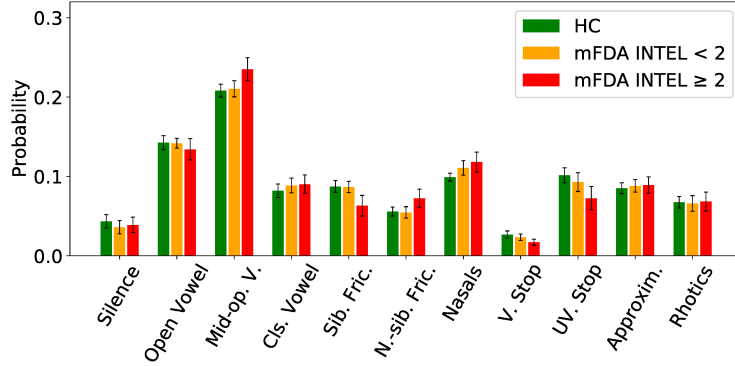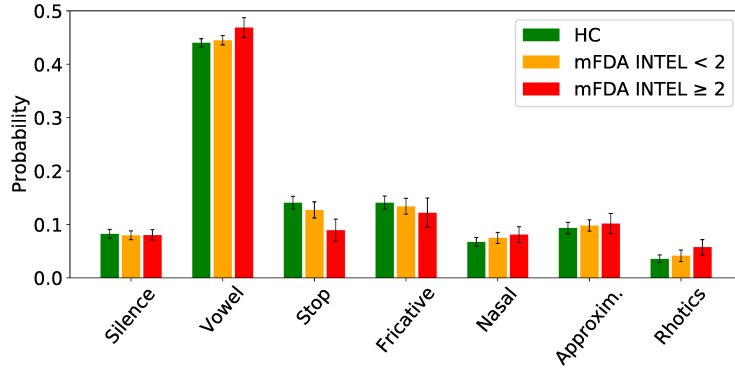Figure 14: Mean posterior probabilities of coarse phonetic classes computed for the text reading task.

3.2.3. Isolated sentences

In the results for isolated sentences, we rediscovered some of the patterns observed for the dedicated speech exercises already. Figures 12 and 13 show the MPPs of coarse and fine phonetic classes. The most pronounced change between sets sorted by mFDA intelligibility score was observed for the stop sounds of the coarse phonetic categories with values of 12.8% for the HC, 11.7% for mildly and 9.0% for severely affected PD patients. This pattern was identical to the trends observed for the DDK tasks, and once more we found the results to be significant (p-value ≪ 0.001). The increase of mean posterior probability of fricative sounds could not be found here. However, we found a pronounced shift of MPP from sibilant to non-sibilant fricatives for the poorly intelligible PD group in the fine phonetic categories. We also noticed an increasing MPP for nasal PHONES, decreasing open and increasing mid-open vowels.

3.2.4. Text reading

For the task of reading a short text, the decreasing MPP of stop sounds was found once more in the coarse phonetic classes (Figure 14). This time, the difference was a bit more pronounced with values of 14.1% for the HC, 12.7% for the mildly affected and 8.9% for the severely affected PD patients. We also checked the nasal sounds in this scenario and found differences with

23
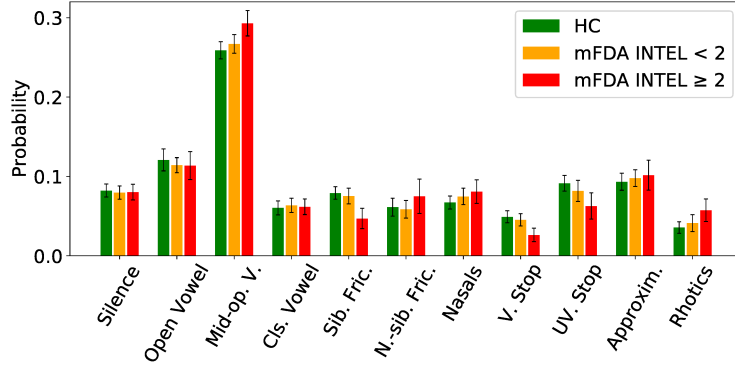
Figure 15: Mean posterior probabilities of fine phonetic classes computed for the text reading task.
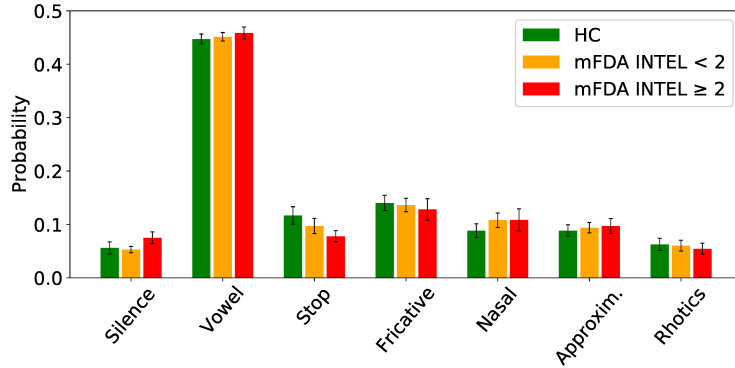
Figure 16: Mean posterior probabilities of coarse phonetic classes computed for the monologue.

values of 6.7%, 7.5% and 8.1%. The increasing MPP of the rhotics was also noticeable, but we did not find this pattern in the isolated sentences. Looking at the results for the fine phonetic categories in Figure 15, we rediscovered the decreasing amount of sibilant and increasing ratio of non-sibilant fricatives for the poorly intelligible PD participants that we already observed before. The increasing MPP of mid-open vowels was also preserved.

3.2.5. Monologues

The final phonetic analysis results were computed for the monologue where the patients produced free speech. Results are shown in Figures 16 and 17.
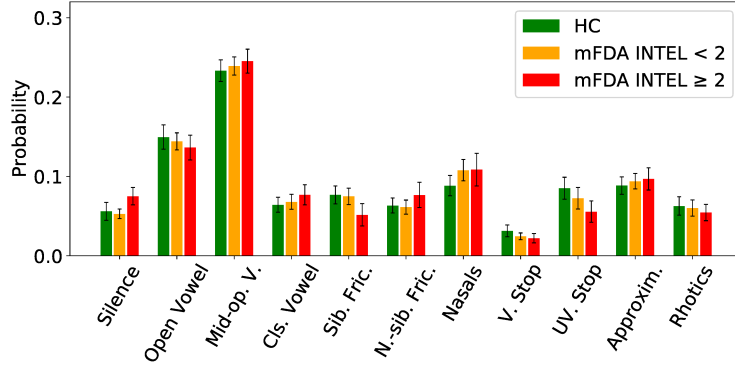
Figure 17: Mean posterior probabilities of fine phonetic classes computed for the monologue.

The most prominent differences in the coarse categories was again the decline of MPP for stop PHONES. The HC had an MPP of 11.7%, this decreased to 9.7% for the mild PD set and 7.8% for the severe set. The latter equaled to a relative reduction of 33% compared to the HC group. We also found an increased amount of nasal sounds for both PD sets, as we did for the previous tasks. An interesting result was the marked increase in MPP for silent segments (7.6% compared to 5.6% and 5.3%) found for the poorly intelligible PD patients. We did not observe any such pattern in the previous tasks. The increasing MPP of trill PHONES found for the text reading task was not corroborated in the monologues. In the fine PHONE categories, we found the substitution of open vowels with their mid-open counterparts as we did for the other tasks. Similarly, the increasing MPP of the poorly intelligible PD group with respect to non-sibilant fricatives was conserved.

3.2.6. Statistical interpretation

Tables 4 and 5 show the detailed evaluation of results for the most affect fine (Unvoiced stop) and coarse (Stop) phonetic classes. Throughout all tasks, we observed medium negative correlations between MPPs of the groups and the mFDA intelligibility score. This correlation was slightly stronger for the coarse groups. We observed large effect sizes for strongly affected patients. For the mildly affected group, these values are considerably lower.

Table 4: Correlation between unvoiced stop (fine groups) mean posterior probabilities for different speech tasks and mFDA intelligibility score. Columns 3 and 4 show effect sizes (Cohen's d) between healthy reference and the respective split according to mFDA intelligibility score.

Speech task	Spearman's ρ	mFDA Intel. < 2	mFDA Intel. ≥ 2
Sentences	-0.50	0.40	1.30
Text reading	-0.36	0.41	1.19
Monologue	-0.43	0.46	1.08

Table 5: Correlation between stop (coarse groups) mean posterior probabilities for different speech tasks and mFDA intelligibility score. Columns 3 and 4 show effect sizes (Cohen's d) between healthy reference and the respective split according to mFDA intelligibility score.

Speech task	Spearman's ρ	mFDA Intel. < 2	mFDA Intel. ≥ 2
Sentences	-0.58	0.65	2.07
Text reading	-0.45	0.50	1.74
Monologue	-0.48	0.63	1.29

3.2.7. Prediction confidence

We also evaluated the confidence of recognition results. To do so, we computed the mean posterior probability over all PHONE predictions for the different tasks. Results are summarized in Table 6. For each of the tasks, we observed a decrease in confidence as the total mFDA scores of patients increased. We found this decreasing confidence to be consistent throughout all tasks. Additionally, we found that the more freedom we had in our tasks (going from very restricted speech exercises to completely free speech), the overall confidence slightly decreased. It is important to notice that even for the free speech task (monologue), the mean posterior probability for the controls was higher than that of the mildly affected patients during the DDK exercises.

Tables 7 and 8 provide the results of an evaluation of findings for the confidence scores. Mean confidence values showed medium to strong correlations to the total mFDA score. Notice that the correlations and effect sizes increased along with increasingly complex speech tasks. Correlations were found to be

26

Table 6: Mean confidence (posterior probability) in percent associated with a recognized PHONE. PD sets were arranged according to total mFDA score.

	HC	mFDA < 20	mFDA < 30	mFDA ≥ 30
DDK tasks	87.3	84.8	83.6	79.0
Sentences	86.2	83.8	82.4	78.0
Read text	85.7	82.8	82.0	77.7
Monologue	85.1	82.3	81.0	76.7

Table 7: Correlation between mean *phone* confidences for different speech tasks and total mFDA score. Columns 3, 4 and 5 show effect sizes (Cohen's d) between healthy reference and the respective split according to total mFDA score.

Speech task	Spearman's ρ	mFDA < 20	mFDA < 30	mFDA ≥ 30
DDK	-0.53	0.49	0.68	1.75
Sentences	-0.64	0.63	1.01	2.12
Read text	-0.65	0.79	1.00	2.22
Monologue	-0.67	0.78	1.07	2.40

even more pronounced when compared to only the mFDA intelligibility item, with the highest (absolute) correlation of -0.76. Effect sizes showed clear differences between sets once again. As we already observed for the total mFDA, the lowest correlation of mean confidence with mFDA intelligibility was observed for the DDK task.

4. Discussion

4.1. Speech exercises

The results for all DDK speech exercises showed clear differences between the HC group and the PD patients. This difference was also observed to markedly increase along with the mFDA scores associated with the involved articulators. We also found the results to be well interpretable. A deficit in muscle tension lead to slurred pronunciation of stop PHONES, where it was increasingly challenging for patients to produce the characteristic closure-burst combination of

Table 8: Correlation between mean *phone* confidences for different speech tasks and mFDA intelligibility score. Columns 3 and 4 show effect sizes (Cohen's *d*) between healthy reference and the respective split according to mFDA intelligibility score.

Speech task	Spearman's ρ	mFDA Intel. < 2	mFDA Intel. ≥ 2
DDK	-0.63	0.58	1.95
Sentences	-0.74	0.74	2.07
Read text	-0.70	0.81	2.43
Monologue	-0.76	0.97	2.79

unvoiced stops. Without the full closure of the lips for example, a [p] would transition into an [f] or [v]. In this context, the transition to [v] is particularly important, because in Colombian Spanish there is no acoustic/phonetic difference between [p] and [v]. The difficulties related to lip muscle control observed from the phonetic profile for stop sound [p] are well-known [36]. Similar patterns were found for the other places of articulation and their respective PHONES [t] (substitution with [ð] (voiced English *th*) or [l]) and [k] (substitution with [r], [ɣ] or [x]). For syllable [ta], the observation that mildly affected PD patients were on average more likely to correctly produce [t] was surprising. The high MPP of healthy speakers for the stop [d] could give serve as an explanation that some speakers in the HC group tended to produce more voiced [d] than unvoiced stops [t]. Overall, they were still more likely to produce a stop than any PD set Notice that the position of the involved articulators (lips or tongue) and the manner of articulation of the different stop sounds were preserved in the transition PHONES, but the required muscle tension was lost, thus making it difficult to correctly produce the target PHONE. Throughout all DDK tasks, we found that the PD patients showed an increased instability for producing the vowel [a]. This instability was already found to be significant in other studies [37, 38]. Instead, patients produced more mid-open (centralized schwa-like) vowels and showed a reduced ability to fully open their vocal tract, possibly due to increasing rigidity [39]. In general, for all exercises of stop-vowel syllable repetitions, we found that PD patients showed clear difficulties to produce stop

sounds reliably and instead were more likely to replace them with fricatives.

The strong negative correlation of unvoiced stop *phone* MPPs and the mFDA intelligibility score clearly indicated that an increasing effect of PD on an individual's speech would result in a reduced MPP of unvoiced stop sounds. On the other hand, non-sibilant fricatives turned out to be positively correlated to the score. This could be caused for example by a transition from stop *phones* or sibilant fricatives over to the non-sibilants, but this is subject of further investigation.

The high standard deviations observed for speech exercises in stop sounds and fricatives were likely caused by confusions of unvoiced and voiced stops (both during production and recognition) as well as spontaneous transitions from stops to fricatives.

4.2. Intermediate network activations

With the PCA decomposition of hidden states that were classified as the correct stop PHONES [p], [t] or [k], we tried to find if there were noticeable differences in PHONE productions between HC and PD, even if the correct PHONE was detected. Not only did our results indicate these differences, but they also became more pronounced with increasing mFDA score of the involved articulator. Because the only objective of a PCA decomposition is to preserve the maximum amount of variation from a higher-dimensional feature space, we assume that the disease state introduced such a variation in the hidden states of stop sound productions and that it was sufficiently large to remain visible after the one-dimensional PCA. We noticed the largest discrepancy between sets for the stop [p] if we arranged sets by mFDA lips rating. This seemed plausible because the lips play the major role in creating the obstruction for the closure and release for the burst of the corresponding stop PHONE. Similar discoveries were made for the stops [t] and [k], but there, the difference between mildly and strongly affected patients was not as big. For PHONE [k], it is likely that the tongue item was also not perfectly resembling a patient's ability to produce velar PHONES. This assumption was further supported by the fact that the ob-

served differences between the sets were higher for [t] and [k] when we sorted by mFDA intelligibility instead of mFDA tongue rating. We chose a subset of the PD speech corpus described in section 2.2 to ensure equal acoustic recording conditions for HC and PD subjects. Therefore, we could attribute the observed differences in the hidden states to the effects of PD on an individual's speech production.

4.3. Isolated sentences and text reading

Whilst the differences in MPP for the stop sounds were very pronounced in the speech exercises, they were expected to lower when analysing spoken sentences or short texts. Nevertheless, we found that for both tasks, the MPP of unvoiced and voiced stops continued to decline along with the rated intelligibility. At the same time, we observed a new pattern, an increasing amount of nasal sounds. The development of a measurable hypernasality caused by PD has been reported previously [40]. It is likely caused by reduced control over the velopharyngeal tract due to loss of muscle tension, thereby allowing air to pass through the nasal cavity [41]. Our results indicate that this hypernasality could be found with the presented PHONE recognizer, but it is yet to be confirmed that the nasalization of vowels would result in an increased MPP of nasal PHONES. As for the speech exercises, PD patients were more likely to produce fewer open and more mid-open vowels as the disease progressed. An impaired vowel articulation was observed in related works [42, 43], which resulted in a smaller vowel space and would therefore lead to fewer realizations of open vowels. Our findings for isolated sentences and read texts among different sets supported the assumption that an impaired speech production and the linked decline in intelligibility would leave a detectable footprint in the phonetic profile of affected individuals. With respect to the correlation between MPPs and intelligibility scores, it did not make much difference if we arranged *phones* into fine (unvoiced stop) or coarse (stop) classes. In both cases, we observed medium correlations for spoken sentences and text reading exercises.

4.4. Monologues

One key research question was if it would be possible to translate any of the aforementioned findings to free speech. The results from speech recordings of monologues indicated that many of the patterns observed in the previous experiments were present in free speech as well. For example, the deterioration of stop sounds, the increasing amount of nasality and the transition from open to mid-open vowels were conserved. The previously described articulation deficits (reduced vowel space, impaired control over velopharyngeal port and overall decreased muscle tension) manifest in the phonetic footprints of PD patients during free speech. MPPs of (unvoiced) stop sounds once more showed a medium correlation to mFDA intelligibility scores in the monologue. Another important finding was the ratio of silent segments for the poorly intelligible PD patients. Relative to the mildly affected PD group and the HC (which showed only minor differences), they showed 39% more silent events. These intermittent pauses could potentially be related to a mild cognitive impairment (MCI) often developing in the later stages of PD. MCI in PD has already been described in previous studies [44, 45, 46]. Speech disfluencies in PD have been investigated in detail before [47] and they could be attributed to two major causes, stuttering and hesitations. Stuttering occurred more frequently within words and was provoked by mispronunciations. Hesitations were found more frequently between words and were far more likely to happen in free speech tasks. With our PHONE recognizer, we were able to find patterns of hesitations as well for poorly intelligible PD patients, and we could confirm their predominant occurrence in free speech.

4.5. Prediction confidence

As our PHONE recognition model was trained only with healthy speech, we also tried to investigate if impaired speech samples would result in an overall decreased confidence of network predictions. Throughout all tasks, we observed high confidences for the HC group (minimum of 85.1 % for monologues). To put this into more context, the mean confidence computed over the test set of all

three languages of our training data was almost 90%. The control group was very close to this value. With increasing mFDA scores of our PD sets, we found the recognizer mean confidence to decline equally for every task. The slight decline within every set going from dedicated speech tasks to free speech was expected due to the increasing variability and length of the signals. The fact that the mean confidence for monologues of HC speakers was still higher than that of any other PD task or set lead us to the conclusion that our model had a decent generalization of what healthy speech sounded like. Accordingly, predictions became markedly less confident as the severity of PD increased. Confidences were clearly correlated to both mFDA total and intelligibility scores. These correlations were particularly pronounced in the monologues and lowest for the dedicated speech exercises. A possible explanation is the rather short duration of an exercise (a couple of seconds) compared to that of a monologue (a few minutes). The longer a patient had to speak, the more likely they were to show signs of a declining articulation. Compared to the findings reported in [9], their models showed stronger correlations (0.793 to 0.943) with intelligibility scores. However, they employed dedicated regression models for score prediction and did not correlate to individual features such as prediction confidences, as it was done here. Furthermore, our models do not require any transcription and yield promising results even for free speech.

5. Conclusion

We presented a neural network for PHONE sequence recognition trained on datasets from three different languages. Pretraining the network with alignment information helped to increase the accuracy of the final CTC model. We then used the recognizer to analyze speech samples from a PD dataset and created phonetic profiles for different subgroups over various tasks. For the dedicated speech exercises, our PHONE recognition model was able to find clear differences in the mean posterior probabilities of stop sounds, which were also strongly correlated to mFDA intelligibility scores, and instabilities in vowel productions

of PD patients. Furthermore, it was possible to transfer these findings to more complex speech samples. Even in free speech recordings, the decreased amount of stop sounds was observed. We also found other patterns that have previously been reported in other studies, such as increased nasalization, reduced vowel articulation space or, for the severely affected PD patients, an increased amount of pauses. Particularly for the nasalization, it could be interesting to incorporate a French healthy speech dataset to include nasalized vowels in the PHONE space.

With the intermediate hidden state vectors from the RNN, we showed that even though all observed hidden state were classified to the same stop sound, they showed clear differences after applying a PCA decomposition that were directly correlated to the impairments of certain articulators.

The analysis of MPPs over all network predictions of a particular set showed that the network was most confident while predicting samples of the control group. Both the total mFDA score (and therefore the degree of impairment) as well as only the intelligibility item were clearly correlated to the mean confidences of PHONE predictions. This was explained by the network being trained exclusively on healthy speech. The less intelligible a speech signal sounded, the farther away it would be from the familiar hyperspace of healthy speech, thus resulting in a lower confidence.

In the field of pathological speech analysis, modern deep learning methods often lack the required amount of data to be applied to any such problem. Instead of training a model with pathological speech to identify characteristic patterns for classification (HC or PD) or regression (PD progression) tasks, it can be sufficient to train the entire model solely with large amounts of healthy speech. These models could serve as a reference of what healthy speech sounds like. At the same time, they enabled us to determine a phonetic footprint of a healthy cohort for a given language. Such footprints could also be computed for speaker groups of Parkinson's patients to compare their articulatory abilities to a healthy reference. Phonetic footprinting holds two major advantages over other methods. Models would not be prone to overfitting on a small domain-

specific dataset, simply because they would never see such data in the process of parameter optimization. Secondly, the footprints and their differences between speaker groups are highly interpretable. Particularly in the field of pathologic datasets, this property is extremely important as it helps to better understand a classifier's decision process. In our case, we confirmed a number of speech patterns of PD patients that had been described in other studies before, although our model was trained on PHONE recognition and had never received any information about Parkinson's disease.

6. Acknowledgements

References

[1] A. Lee, R. M. Gilbert, Epidemiology of Parkinson Disease, Neurologic Clinics 34 (4) (2016) 955–965. `doi:10.1016/j.ncl.2016.06.012`.

[2] J. Jankovic, Parkinson's disease: clinical features and diagnosis, Journal of neurology, neurosurgery & psychiatry 79 (4) (2008) 368–376.

[3] S. Pinto, C. Ozsancak, E. Tripoliti, S. Thobois, P. Limousin-Dowsey, P. Auzou, Treatments for dysarthria in Parkinson's disease, Lancet Neurology 3 (9) (2004) 547–556. `doi:10.1016/S1474-4422(04)00854-3`.

[4] J. S. Almeida, P. P. Rebouças Filho, T. Carneiro, W. Wei, R. Damaše-vičius, R. Maskeliūnas, V. H. C. de Albuquerque, Detecting Parkinson's disease with sustained phonation and speech signals using machine learning techniques, Pattern Recognition Letters 125 (2019) 55–62. `doi:10.1016/j.patrec.2019.04.005`.

[5] T. Arias-Vergara, P. Arguello-Velez, J. C. Vásquez-Correa, E. Nöth, M. Schuster, M. C. González-Rátiva, J. R. Orozco-Arroyave, Automatic detection of Voice Onset Time in voiceless plosives using gated recurrent units, Digital Signal Processing: A Review Journal 104 (2020) 102779. `doi:10.1016/j.dsp.2020.102779`.

[6] F. J. Martinez-Murcia, A. Ortiz, J. M. Gorriz, J. Ramirez, D. Castillo-Barnes, D. Salas-Gonzalez, F. Segovia, Deep Convolutional Autoencoders vs PCA in a Highly-Unbalanced Parkinson's Disease Dataset: A DaTSCAN Study, in: Advances in Intelligent Systems and Computing, Vol. 771, Springer, 2019, pp. 47–56. `doi:10.1007/978-3-319-94120-2_5`.

[7] J. R. Orozco-Arroyave, J. C. Vásquez-Correa, F. Hönig, J. D. Arias-Londono, J. F. Vargas-Bonilla, S. Skodda, J. Rusz, E. Nöth, Towards an automatic monitoring of the neurological state of Parkinson's patients from speech, in: ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, Vol. 2016-May, IEEE, 2016, pp. 6490–6494. `doi:10.1109/ICASSP.2016.7472927`.

[8] R. Norel, C. Agurto, S. Heisig, J. J. Rice, H. Zhang, R. Ostrand, P. W. Wacnik, B. K. Ho, V. L. Ramos, G. A. Cecchi, Speech-based characterization of dopamine replacement therapy in people with Parkinson's disease, npj Parkinson's Disease 6 (1) (2020) 1–8. `doi:10.1038/s41531-020-0113-5`.

[9] G. Van Nuffelen, C. Middag, M. De Bodt, J.-P. Martens, Speech technology-based assessment of phoneme intelligibility in dysarthria, International journal of language & communication disorders 44 (5) (2009) 716–730.

[10] S. L. Christina, P. Vijayalakshmi, T. Nagarajan, Hmm-based speech recognition system for the dysarthric speech evaluation of articulatory subsystem, in: 2012 International Conference on Recent Trends in Information Technology, IEEE, 2012, pp. 54–59.

[11] A. Som, N. Krishnamurthi, M. Buman, P. Turaga, Unsupervised pretrained models from healthy adls improve parkinson's disease classification of gait patterns, in: 2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), IEEE, 2020, pp. 784–788.

[12] J. C. Vásquez-Correa, T. Bocklet, J. R. Orozco-Arroyave, E. Nöth, Comparison of user models based on gmm-ubm and i-vectors for speech, handwriting, and gait assessment of parkinson's disease patients, in: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2020, pp. 6544–6548.

[13] A. Naseer, M. Rani, S. Naz, M. I. Razzak, M. Imran, G. Xu, Refining Parkinson's neurological disorder identification through deep transfer learning, Neural Computing and Applications 32 (3) (2020) 839–854. `doi:10.1007/s00521-019-04069-0`.

[14] A. Schrag, R. Dodel, A. Spottke, B. Bornschein, U. Siebert, N. P. Quinn, Rate of clinical progression in Parkinson's disease. A prospective study, Movement Disorders 22 (7) (2007) 938–945. `doi:10.1002/mds.21429`.

[15] J. C. Vásquez-Correa, C. D. Rios-Urrego, A. Rueda, J. R. Orozco-Arroyave, S. Krishnan, E. Nöth, Articulation and Empirical Mode Decomposition Features in Diadochokinetic Exercises for the Speech Assessment of Parkinson's Disease Patients, in: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), Vol. 11896 LNCS, Springer, 2019, pp. 688–696. `doi:10.1007/978-3-030-33904-3_65`.

[16] M. Chun, D. Wei, W. Qing, Speech Analysis for Wilson's Disease Using Genetic Algorithm and Support Vector Machine, Advances in Intelligent Systems and Computing 1017 (2020) 1286–1295. `doi:10.1007/978-3-030-25128-4_160`.

[17] S. Skodda, W. Grönheit, N. Mancinelli, U. Schlegel, Progression of voice and speech impairment in the course of Parkinson's disease: A longitudinal study, Parkinson's Disease 2013. `doi:10.1155/2013/389195`.

[18] J. Rusz, R. Čmejla, H. Růžičková, J. Klempíř, V. Majerová, J. Picmausová, J. Roth, E. Růžička, Evaluation of speech impairment in early stages of Parkinson's disease: A prospective study with the role of pharmacotherapy, Journal of Neural Transmission 120 (2) (2013) 319–329. `doi:10.1007/s00702-012-0853-4`.

[19] J. Jankovic, A. H. Rajput, M. P. McDermott, D. P. Perl, P. S. Group, et al., The evolution of diagnosis in early parkinson disease, Archives of neurology 57 (3) (2000) 369–372.

[20] W. Poewe, K. Seppi, C. M. Tanner, G. M. Halliday, P. Brundin, J. Volkmann, A.-E. Schrag, A. E. Lang, Parkinson disease, Nature reviews Disease primers 3 (1) (2017) 1–21.

[21] S. S. Hacisalihzade, M. Mansour, C. Albani, Optimization of symptomatic therapy in parkinson's disease, IEEE transactions on biomedical engineering 36 (3) (1989) 363–372.

[22] W. Wahlster, Verbmobil: foundations of speech-to-speech translation, Springer Science & Business Media, 2013.

[23] D. Povey, G. Boulianne, L. Burget, P. Motlicek, P. Schwarz, The Kaldi Speech Recognition, in: IEEE 2011 Workshop on Automatic Speech Recognition and Understanding, no. January, IEEE Signal Processing Society, 2011.
URL `http://kaldi.sf.net/`

[24] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1-1.1, NASA STI/Recon Technical Report N 93 (1993) 27403. URL `papers://e7d065ae-9998-4287-8af0-c9fa85af8e96/Paper/p44370`

[25] L. A. Pineda, L. V. Pineda, J. Cuétara, H. Castellanos, I. López, DIMEx100: A new phonetic and speech corpus for Mexican Spanish, in: Lecture Notes in Artificial Intelligence (Subseries of Lecture Notes in Computer Science), Vol. 3315, Springer, 2004, pp. 974–983. `doi:10.1007/978-3-540-30498-2_97`.

[26] J. C. Vásquez-Correa, J. R. Orozco-Arroyave, T. Bocklet, E. Nöth, Towards an automatic evaluation of the dysarthria level of patients with Parkinson's disease, Journal of Communication Disorders 76 (2018) 21–36. `doi:10.1016/j.jcomdis.2018.08.002`.

[27] P. ENDERBY, Frenchay Dysarthria Assessment, International Journal of Language & Communication Disorders 15 (3) (1980) 165–173. `doi:10.3109/13682828009112541`.

[28] C. G. Goetz, B. C. Tilley, S. R. Shaftman, G. T. Stebbins, S. Fahn, P. Martinez-Martin, W. Poewe, C. Sampaio, M. B. Stern, R. Dodel, et al., Movement disorder society-sponsored revision of the unified parkinson's disease rating scale (mds-updrs): scale presentation and clinimetric testing results, Movement disorders: official journal of the Movement Disorder Society 23 (15) (2008) 2129–2170.

[29] R. K. Moore, L. Skidmore, On the use/misuse of the term 'phoneme', Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH 2019-Septe (October) (2019) 2340–2344. `arXiv:1907.11640`, `doi:10.21437/Interspeech.2019-2711`.

[30] S. S. Stevens, J. Volkmann, E. B. Newman, A Scale for the Measurement of the Psychological Magnitude Pitch, Journal of the Acoustical Society of America 8 (3) (1937) 185–190. `doi:10.1121/1.1915893`.

[31] A. Graves, S. Fernández, F. Gomez, J. Schmidhuber, Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks, in: ACM International Conference Proceeding Series, Vol. 148, 2006, pp. 369–376. `doi:10.1145/1143844.1143891`.

[32] D. P. Kingma, J. L. Ba, Adam: A method for stochastic optimization, 3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings`arXiv:1412.6980`.

[33] I. Loshchilov, F. Hutter, Decoupled weight decay regularization, arXiv abs/1711.0. `arXiv:1711.05101`.

[34] A. Graves, Generating Sequences With Recurrent Neural Networks, arXiv preprint arXiv:1308.0850`arXiv:1308.0850`.
URL `http://arxiv.org/abs/1308.0850`

[35] P. Klumpp, T. Arias-Vergara, J. C. Vásquez-Correa, P. A. Pérez-Toro, F. Hönig, E. Nöth, J. R. Orozco-Arroyave, Surgical mask detection with deep recurrent phonetic models, Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH 2020-Octob (2020) 2057–2061. `doi:10.21437/Interspeech.2020-1723`.

[36] R. Leanderson, B. A. Meyerson, A. Persson, Lip muscle function in parkinsonian dysarthria, Acta Oto-Laryngologica 74 (1-6) (1972) 350–357. `doi:10.3109/00016487209128462`.

[37] B. E. Sakar, M. E. Isenkul, C. O. Sakar, A. Sertbas, F. Gurgen, S. Delil, H. Apaydin, O. Kursun, 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. `doi:10.1109/JBHI.2013.2245674`.

[38] A. Tsanas, M. A. Little, P. E. McSharry, J. Spielman, L. O. Ramig, Novel speech signal processing algorithms for high-accuracy classification

of Parkinsons disease, IEEE Transactions on Biomedical Engineering 59 (5) (2012) 1264–1271. `doi:10.1109/TBME.2012.2183367`.

[39] J. Proença, A. Veiga, S. Candeias, F. Perdigão, Acoustic, Phonetic and Prosodic Features of Parkinson's disease Speech, Tech. rep. (2013).
URL `http://www.aclweb.org/anthology/W/W13/W13-4827.pdf`

[40] M. Novotnỳ, J. Rusz, R. Čmejla, H. Růžičková, J. Klempíř, E. Růžička, Hypernasality associated with basal ganglia dysfunction: Evidence from Parkinson's disease and Huntington's disease, PeerJ 2016 (9). `doi:10.7717/peerj.2530`.

[41] J. I. Godino-Llorente, S. Shattuck-Hufnagel, J. Y. Choi, L. Moro-Velázquez, J. A. Gómez-García, Towards the identification of idiopathic parkinson's disease from the speech. new articulatory kinetic biomarkers, PloS one 12 (12) (2017) e0189583.

[42] S. Skodda, W. Grönheit, U. Schlegel, Impairment of vowel articulation as a possible marker of disease progression in parkinson's disease, PloS one 7 (2) (2012) e32132.

[43] J. A. Whitfield, A. M. Goberman, Articulatory–acoustic vowel space: Application to clear speech in individuals with parkinson's disease, Journal of communication disorders 51 (2014) 19–28.

[44] D. Verbaan, J. Marinus, M. Visser, S. M. v. Rooden, A. M. Stiggelbout, H. A. M. Middelkoop, van J. J. Hilten, Cognitive impairment in Parkinson's disease, Journal of Neurology, Neurosurgery & Psychiatry 78 (11) (2007) 1182–1187.

[45] A. A. Kehagia, R. A. Barker, T. W. Robbins, Neuropsychological and clinical heterogeneity of cognitive impairment and dementia in patients with Parkinson's disease, The Lancet Neurology 9 (12) (2010) 1200–1213. `doi:10.1016/S1474-4422(10)70212-X`.

[46] D. Aarsland, K. Bronnick, C. Williams-Gray, D. Weintraub, K. Marder, J. Kulisevsky, D. Burn, P. Barone, J. Pagonabarraga, L. Allcock, et al., Mild cognitive impairment in parkinson disease: a multicenter pooled analysis, Neurology 75 (12) (2010) 1062–1069.

[47] A. M. Goberman, M. Blomgren, E. Metzger, Characteristics of speech disfluency in Parkinson disease, Journal of Neurolinguistics 23 (5) (2010) 470–478. `doi:10.1016/j.jneuroling.2008.11.001`.

[48] C. Szegedy, S. Ioffe, V. Vanhoucke, A. A. Alemi, Inception-v4, inception-ResNet and the impact of residual connections on learning, in: 31st AAAI Conference on Artificial Intelligence, AAAI 2017, 2017, pp. 4278–4284. `arXiv:1602.07261`.

[49] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Vol. 2016-Decem, 2016, pp. 770–778. `arXiv:1512.03385`, `doi:10.1109/CVPR.2016.90`.

[50] L. Sifre, PhD thesis Rigid-Motion Scattering For Image Classification, Ph. D. thesis.

[51] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural computation 9 (8) (1997) 1735–1780.

[52] A. Howard, M. Sandler, G. Chu, L. C. Chen, B. Chen, M. Tan, W. Wang, Y. Zhu, R. Pang, V. Vasudevan, Q. V. Le, H. Adam, Searching for MobileNetV3, in: arXiv, 2019, pp. 1314–1324.

[53] S. Ioffe, C. Szegedy, Batch normalization: Accelerating deep network training by reducing internal covariate shift, 32nd International Conference on Machine Learning, ICML 2015 1 (2015) 448–456. `arXiv:1502.03167`.

Appendix A. Model architecture

The convolutional feature extraction part was constructed with two major building blocks inspired by the Inception architectures presented for image classification [48]. The core idea was to apply multiple convolution kernels of varying sizes in parallel such that the network itself would learn which kernel worked best for what task. Figure A.18a depicts a residual inception block. The initial 1x1 convolutions perform a channel reduction to make the following convolution operations less parameter-intensive. The 3x3 and 5x5 convolutions which operated in parallel were realized with separated kernels to further reduce the required number of parameters. A final 1x1 convolution projected the concatenated results from the three branches back to the original number of channels. Before the output was then added to the input to realize a residual connection [49], we applied activation scaling ($s = 0.3$) as proposed in [48] to stabilize training. The second important building block of our network was the reduction inception block shown in Figure A.18b. The major purpose of that component was to reduce the remaining number of frequency bins while leaving the time dimension unchanged. Reduction was performed with parallel max-pooling and strided convolution layers and their outputs were concatenated. The architectures of the CTC and the framewise model are described in Table A.9. In the first step, we used several 2D-convolutions to reduce the number of frequency bands and encode the corresponding information in the channel dimension. This part of a convolutional neural network (CNN) is commonly referred to as the stem. Note that the very first layer halved the number of time steps. Our dual-channel spectrograms were computed with a temporal resolution of 5 ms, which helped to reliably detect very short PHONES. For the final prediction however, a resolution of 10 ms was sufficient. After the stem, we used the residual and reduction inception blocks to learn feature maps and gradually reduce the number of remaining frequency bands. A depthwise separable convolution [50] was used to project the four remaining frequency bands and their 580 channels after the last inception block down to 300 values per time step.

(a) Residual inception block used in the PHONE recognition network.

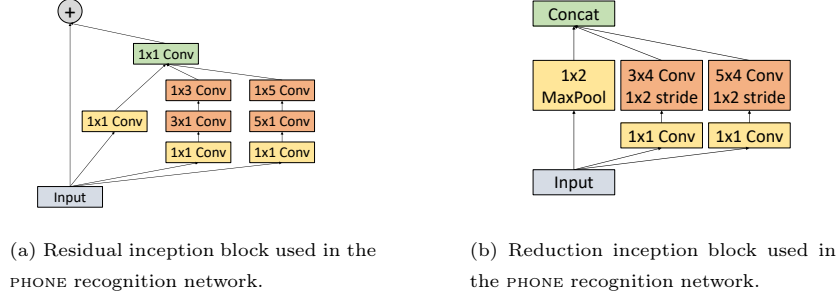(b) Reduction inception block used in the PHONE recognition network.

Figure A.18: Two types of inception blocks used in the PHONE recognition network

The initial stem, the inception blocks and the depthwise separable convolution were considered the feature extraction part of the architecture. To perform a sequence analysis, we then applied a stack of two bidirectional Long Short-Term Memories (BiLSTM) [51]. Each BiLSTM used a hidden state vector with 240 units in the case of CTC, 200 units during the pre-training. In the final step, the result for each time step was linearly projected to the number of targets, 35 for CTC, 44 for aligned pre-training. All hyperparameters, for example kernel sizes or channel configurations, were tuned for the best result in PHONE sequence prediction. Thus, the presented architecture was never optimized for PD speech analysis.

Hard swish [52] was used as activation function. The outputs of the convolutional stem-layers, the residual inception blocks and the depthwise separable convolution were normalized through batch normalization [53]. After each layer normalization, we applied a dropout of 10% to prevent overfitting. The model for aligned pre-training comprised around 6.6 million parameters, the final CTC model was a bit larger with roughly 7.2 million parameters.

Table A.9: Outline of the CTC PHONE recognition model. Output size depended on the length of the sample (T). #c indicates number of channels. #x# denotes kernel size in temporal (first) and frequency (second) domain. [#, #] denotes the stride in the respective domain. For the recurrent neural network (RNN) layers and linear projection layers, numbers in brackets denote the configuration of the pretrained model with alignment information.

Output size	Layer
2Tx64, 60	60c 1x4 Conv [1, 2]
Tx64, 120	120c 5x1 Conv [2, 1]
Tx32, 160	160c 1x4 Conv [1, 2]
Tx16, 200	200c 1x4 Conv [1, 2]
Tx16, 200	2 x Residual Inception Block ch. reduced: 70
Tx8, 340	Reduction Inception Block ch. reduced: 70
Tx8, 340	2 x Residual Inception Block ch. reduced: 120
Tx4, 580	Reduction Inception Block ch. reduced: 120
Tx4, 580	2 x Residual Inception Block ch. reduced: 200
Tx300	Depthwise separable convolution
Tx480 (400)	2 x BiLSTM 240 (200) hidden units
Tx35 (44)	Linear projection to 35 (44) target PHONES