Nasality Detection in Parkinson's Disease

Matilde Peixoto
Instituto Superior Técnico
Portugal
matilde.peixoto@tecnico.ulisboa.pt

***Abstract*—Neurodegenerative diseases, such as Parkinson's disease, are disabling conditions that constrain the daily lives of those who suffer from them. With the ageing of the general population they are expected to become more predominant. The advances made in speech analysis and machine learning, have revealed the potential to automatically detect these disorders through speech.**

The objective of this Master Thesis is to assess the viability of automatically detecting hypernasality in Parkinson's disease for European Portuguese. Existing machine learning models for hypernasality detection for other languages are mostly trained on disordered speech databases. Given that European Portuguese is a low-resourced language, the approach used in this thesis was to use a healthy speech database to model the characteristics of nasalisation. We created a deep neural network classifier that divides the sounds into oral or nasal consonants and vowels. Our best results for this classifier (accuracy of 84.42%) were achieved with a time-delay neural network. Using the output of the classifier as acoustic correlates of nasality, we created a nasality score for the prediction of hypernasality in Parkinson's disease patients. The analysis made to the results shows that there are statistically significant differences between the control group and the Parkinson's group.

***Index Terms*—Parkinson's disease, Speech, Nasalisation, Deep Learning**

I. Introduction

Dealing with any type of long-term health condition can create stress and burden in the lives of the patients and of those around them. Neurodegenerative diseases can greatly affect every level of the patient's life and can also result in a total ineptitude to perform everyday tasks. These patients may have: breathing difficulties, motor problems, cognitive problems or gradual memory loss (with the possibility of affecting long-term memory) [1].

According to the Institute for Health Metrics and Evaluation (IHME), the third most prevalent cause of disability and premature death in the EU are neurological disorders. A burden that, with the continuing ageing of the European population, is expected to increase. In 2017, the total number of people in the EU that had a neurological disorder was approximately 21 million [2].

Speech has been identified as a potential biomarker for diseases that affect the organs involved in its planning and production. These diseases include neurodegenerative disorders such as Parkinson's Disease (PD) and Huntington's Disease (HD).

Neurodegenerative diseases don't have a cure yet, but with early detection and prevention the symptoms can be treated and delay disease development. Monitoring symptoms might also help understand disease progression. Here, speech matters. Speech is a non-invasive, cost-effective way to diagnose and monitor several disorders. Speech, combined with expert evaluation, can help those who deal with these conditions.

Digital biomarkers, such as speech, may give a complementary assessment approach to traditional clinical examinations since they are non-invasive, objective, and ecologically valid. These assessments may be done remotely, which increases accessibility and lowers the inherent risks of going to healthcare centres. The use of these biomarkers eases long-term follow-up with periodic checks, which may result in more comprehensive data that may help physicians make better informed decisions. [3].

One of the many voice qualities that can be perceived through speech is nasalisation. Nasal resonance can be defectively produced in speech by those who have velopharyngeal mechanism problems (hypernasal speech), resulting in a part of the air exiting through the nose in non-nasal sounds. As hypernasality can be a sign of problems with the peripheral nervous system, the central nervous system, or both, it is important for clinicians to identify it [4].

The main goal of this project is to develop a system that automatically detects hypernasality in Parkinson's Disease (PD) for European Portuguese (EP). This and other diseases that affect the respiratory organs may cause this feature of a person's voice to be more prominent than in healthy individuals. To accomplish the main goal, we created a model that classifies phonemes as nasal or non-nasal consonants and vowels. Using the output of the classifier as acoustic correlates of nasality, we computed a nasality score for each utterance of our corpus. The validity of this method as a means of detecting and monitoring hypernasality for PD patients was assessed by comparing the results of the nasality score in the control and non-control groups to determine whether there is a statistically significant difference between the scores of the two groups.

II. Background

A. Nasalisation

Nasalisation is a voice quality that translates in the production of a sound where part of the air exits through the nose instead of the mouth. This phenomenon happens due to the lowering of the velum which opens an airway (velopharyngeal port) to the nasal cavity. In figure 1 it is possible to see how the position of the velum differs when producing a nasal or an oral sound.

The great majority of languages have nasal sounds. In Portuguese there are three nasal consonants which are /m/, /n/ and /η/. Regarding vowel nasalisation, Portuguese presents a phonetic nasalisation, which means that there are distinctively nasalised vowels that are not in the context of a nasal consonant [5].

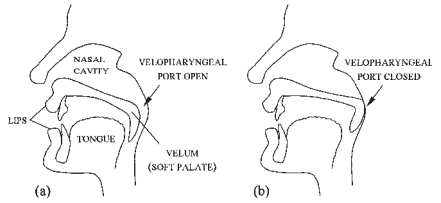

Fig. 1. (a) nasal sound (b) oral sound.

When there is a perceived excessiveness of nasal resonance in speech production, it is considered hypernasality. This can occur due to an inability to control the changes in airflow between the oral and nasal cavities. As velar movement requires dexterity, hypernasality is a typical sign of motor-speech disorders that include PD, HD, amyotrophic lateral sclerosis(ALS), and cerebellar ataxia. It is also the distinctive perceptual characteristic of speech with a cleft palate [6].

The modelling of the vocal tract, which begins in the glottis and continues to the lips, can be an indicator of nasality. It is comparable to a linear, time-invariant filter and its shape and dimensions are essential for the study of articulatory processes in speech production. In hypernasal speech, the shape of the vocal tract is altered as a consequence of the coupling of the cavities (nasal and oral), the disparity between the two passageways of the nasal cavity, and the sinuses branching from the nasal cavity wall [7].

The detection of hypernasality is a complex task that requires the computation of the ratio of resonances across the mouth cavities. When there is a disproportionately high amount of nasal resonance it is considered hypernasal. This can be a difficult estimation to make since it is dependent on a number of different variables such as word choice and the size and shape of each individual's cavities. It results in a highly nonlinear and complex mapping between the perceived and the actual acoustic nasal resonance [8].

B. Acoustic Correlates for Parkinson's Disease

Neurodegenerative disorders affect primarily the brain cells. They are disabling conditions that progressively degenerate nerve cells that when dead cannot be regenerated. There are several symptoms experienced when suffering from one of these diseases, being one of them difficulty in communication, either due to memory loss and forgetfulness or for motor reasons such as dysarthria. Dysarthria is frequently present in PD patients. It is defined as damage, weakness or paralysis in the muscles used to produce speech and can affect respiration, phonation, articulation and prosody [9].

Even though patients suffering from PD can present impairments in all speech dimensions, several studies show that the speech dimension that is the most correlated to PD patients is phonation, followed by articulation [10] [11]. However, Rusz et al. [12], determined that for early untreated PD, the most predominantly affected speech dimension appears to be prosody.

To examine phonation, the most usual measures are extracted from sustained vowels. Due to elevated laryngeal tension and low laryngeal stability, PD patients have shown to have higher mean and variation of the fundamental frequency, higher jitter and shimmer and a smaller NHR when compared to healthy subjects [13] [14] [15].

Articulation features are most commonly extracted from diadochokinetic (DDK) tasks which consist in rapid syllable repetitions such as /pa/, /ta/, /ka/. Studies have shown that when examining articulation the stop consonants of PD patients were imprecise and similar to fricatives [12]. These articulatory deficits may result from small range of motion of lips and tongue and also from articulatory weakness [13].

Prosodic features are usually determined from continuous speech, either from reading tasks or spontaneous speech. Patients with PD tend to present a decreased intensity and F0 variability and decreased rhythm when compared to control groups [12] [13] [14].

For PD patients that develop dysarthria, which approximately 90% of patients do as the disease progresses [16], mild to moderate hypernasality is a symptom that has been shown to appear due to reduced control of the nasal cavity [17].

C. Deep Learning

Deep learning (DL) is a branch of machine learning (ML) that has this designation due to the greater depth of layers in its neural networks (NN). Feedforward Neural Networks, are the basis of DL models. They are inspired by neuroscience, since the perceptrons, the basic units of these models, are conceptually modelled from biological neurons [18].

A neural network is composed of several nodes (or perceptrons) that, in turn, are organised in layers of one or more nodes. They usually consist of an input layer, an output layer and a number of hidden layers. In feedforward NN the information moves only forward, from the input nodes, to the hidden nodes and finally to the output nodes.

Each node receives a vector of features as input. This vector is then multiplied by a vector of weights and a bias value is added. The result of this computation is then passed through an activation function returning one single value. This process can be described by equation 1 where y is the output, ϕ is the activation function, w is the weights, x the input vector and b the bias.

$$y = \phi(w^T x + b) \tag{1}$$

D. Time Delay Neural Network

Time delay neural networks (TDNN) [19] are multilayer neural networks. The two main properties that make the TDNN a good fit for speech-related tasks are its ability to learn the temporal structure of acoustic events and their relationships,

and the fact that it is shift-invariant, which means that the features learned by the network are independent of their precise location on the input data. Since the TDNN detects phonemes and their underlying acoustic properties regardless of temporal location, it outperforms static classification.

In the TDNN architecture each layer has a different temporal resolution, that increases as we go through the network's layers. The initial transforms of the TDNN are learnt on narrower contexts whilst the deeper layers process the hidden activations from a wider temporal context [20].

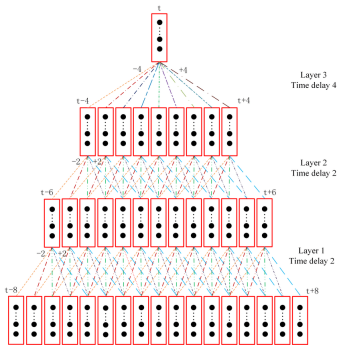

Fig. 2. Time Delay Neural Network structure.

Figure 2 shows that each layer's nodes depends on adjacent nodes from the previous layer. The number of nodes is determined by a delay offset $[d1, d2]$ called the context. For each t, $d1$ and $d2$ represent the number of context frames before and after t, respectively. At each time step t, the TDNN performs a convolution operation, which consists on an element-wise multiplication of the kernel weights and input beneath it, followed by the sum of each element multiplication.

III. Related Work

Clinically, hypernasality is commonly assessed with perceptual evaluations done by speech-language pathologists. Alternatively, there are a few instrumental methods that can be used, such as magnetic resonance imaging (MRI), which can be used to observe the velum directly, and the nasometer [8].

The nasometer is the most used instrument in clinical settings and evaluates the extent of nasal escape in cleft palate. It consists on an headset with two microphones separated by a plate that measure the nasal and oral sound pressure levels of speech. These measurements are then used to obtain the acoustic energies and calculate the nasalance which corresponds to the ratio between the nasal and the total (sum of nasal and oral) acoustic energies [21].

These methods all rely on the availability of trained clinicians to manoeuvre the devices and to read and interpret the results. In the last decades, several studies that use speech signal analysis to detect nasality have emerged.

In Kataoka et al. [22], the authors study hypernasality in children by assessing different spectral features in the isolated vowel [i]. Using perceptual ratings as ground truth, they concluded that there is a high correlation between the amplitudes of 1/3-octave bands and the ratings. When compared to a control group, the hypernasal group presented increased amplitudes between the first and second formant (F1 and F2) and decreased amplitudes near F2.

Pruthi et al. [23] proposed a set of acoustic parameters (APs) for the detection of vowel nasalisation. There are nine APs that include A1-P0 and A1-P1, where A1 corresponds to the amplitude of the first formant and P0 and P1 correspond to the amplitude of an extra peak below and above F1, respectively. It also includes nPeaks40dB which counts the number of peaks within 40dB of the maximum dB amplitude in a frame and F1BW which is the bandwidth of F1. The authors achieved accuracies ranging between 69% and 96% for different datasets using a support vector machine as classifier.

A few studies ([24], [21]) used the voice low tone to high tone ratio (VLHR) as a rating for nasality. VLHR corresponds to the ratio between low-frequency power and high-frequency power of the sound power spectrum. This division between low and high is done by a specific cut-off frequency which can differ depending on the task. In Tsai et al. [21], the authors used this method in read speech and obtained significantly greater VLHR scores in the nasal sentences when compared to the non-nasal sentences.

There has also been a number of works relying on machine learning to determine hypernasality in children with cleft lip and palate. In Orozco-Arroyave et al. [25], the authors present a method that detects if the speech is hypernasal or healthy by using Nonlinear Dynamics (NLD) features and entropy measurements as input features to a support vector machine. They tested their system with a German and a Spanish databases that contained recordings of healthy and hypernasal children and found that the combination of NLD features and entropy measurements provided the best results.

In Wang et al. [7], the authors describe a feature-independent algorithm that uses a convolutional neural network (CNN) to detect hypernasality. The input of the CNN is a speech spectrogram and the dataset used is comprised of Chinese children and adults with cleft palate. The authors achieve accuracies of 93% and infer that /i/ is the most sensitive vowel to hypernasality.

IV. System Overview

The main objective of the system implemented is to predict hypernasality in pathological speech for EP. The method used to achieve this is based on the proposed method by Mathad et al. in [26], where the authors conducted a similar study for the English language. An overview of the system developed can be seen in figure 3.

Scarcity of data for pathologies is a barrier for clinical speech processing applications. The main advantage of this technique over other machine learning works is that the training data originates from healthy people. This makes the solution disease-independent and makes it easier to find datasets with a more reasonable size.

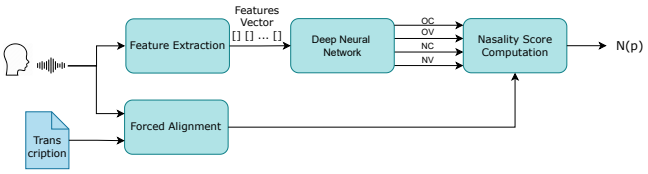

Fig. 3. Overview of the proposed system.

The inputs of the system are an audio file of continuous speech and the text transcription of that file. The final output consists of a nasality score which is a value comprised between 0 and 1. The system as a whole consists of four main blocks that will be explained below.

A. Feature Extraction

The features extracted from the speech file are mel-frequency cepstral coefficients (MFCC) and their first and second order derivatives (Δ and $\Delta\Delta$MFCC). There are a total of 39 features for each frame: 13 MFCC, from coefficients 0^{th} to 12^{th}, and the same number of Δ and $\Delta\Delta$.

Speech is a nonstationary random signal that can only be presumed stationary in short-time intervals of 10 to 30 ms. Hamming windows are used to split the signal into 20ms frames with 10ms overlap.

The reason for using MFCC as features is that they have the ability to model both the vocal chords and the vocal tract. When presented with disordered speech, FFT-based MFCC have the intrinsic capacity to represent either an irregular movement of the vocal chords or a lack of closure caused by compensatory movements in the vocal tract as a result of velopharyngeal incompetence [27].

The delta (differential) and delta-delta (acceleration) coefficients are used in order to better understand the speech signal. They provide an insight into the dynamics of the power spectrum, i.e., the trajectories of MFCC over time.

B. Forced Alignment

Forced alignment is a special mode of automatic speech recognition (ASR). In ASR, the sole input is an audio file, and the algorithm predicts the words based on acoustic features and a language model. In forced alignment ASR, a written transcript of the audio is also needed. In this mode, the algorithm just evaluates word and phoneme boundaries.

This block was performed using TRIBUS [28] which is an ASR system for European Portuguese developed at INESC-ID. This system returns the alignment of the audio with the transcript at a phoneme-level.

Forced alignment was performed on the test data so that when the system is calculating the nasality score, it knows which phoneme corresponds to each point in time.

C. Deep Neural Network

After feature extraction, the feature vectors go into a neural network. The NN was developed to identify phonemes as oral or nasal. For this purpose, it was trained to classify each frame as one of the five following classes: oral consonants (OC), oral vowels (OV), nasal consonants (NC), nasal vowels (NV) and silence.

The NN was trained on an healthy speech database called BD-PUBLICO that is described below. The idea explored here is that, when compared to healthy speech, pathological speech will present a higher rate of nasality, $P(NC)$ or $P(NV)$, in oral sounds.

Three different neural network architectures were developed in order to see which one could achieve the best performance.

D. Nasality Score

The output of the NN is used to calculate the nasality score. Since the last layer of the NN is a softmax, its output can be considered as the probability of each class. Hence, for each input frame x, the output of the NN will be $P(OC|x)$, $P(OV|x)$, $P(NC|x)$, $P(NV|x)$, $P(sil|x)$, which corresponds to the probability of the input being an oral consonant, oral vowel, nasal consonant, nasal vowel and silence, respectively. The nasality score, $N(p)$, is computed based on these probabilities and the phonemes provided by the transcript. For each oral phoneme, p, a score is computed by averaging the corresponding nasal probability, i.e., $P(NC)$ for oral consonants and $P(NV)$ for oral vowels, of all the frames that belong to phoneme p. A formal equation of how $N(p)$ is calculated can be seen in 2 where x_i is the i^{th} frame in phoneme p_j and $|X^{p_j}|$ the total number of frames of the phoneme.

$$N(p_j) = \begin{cases} \sum_i P(NC|x_i)/|X^{p_j}|, if p_j \in OC \\ \sum_i P(NV|x_i)/|X^{p_j}|, if p_j \in OV \end{cases}. \quad (2)$$

Finally, in order to have one single score per audio file, the nasality scores of the upper quartile, i.e., the 25% highest nasality scores per phoneme are averaged.

V. Corpus

Two different corpora were used in the execution of this work: BD-PUBLICO and FraLusoPark. BD-PUBLICO was used to train the neural networks and test its performance, whereas FraLusoPark was used to evaluate the overall performance of the whole system for the task in hand.

A. BD-PUBLICO

BD-PUBLICO [29] is a healthy speech corpus which was owned by INESC-ID for automatic speech recognition purposes. This database consists of read speech from the Portuguese newspaper Público. It was created in 1997 and contains recordings from 120 different speakers, with ages ranging from 19 to 28 years old. The corpus is already divided into training, validation and test sets. The size of each subset is presented in table I. All utterances are sampled at 16kHz.

Since this corpus was not created for the specific purpose of studying nasalisation, the read speech does not have particular emphasis on sentences with large amounts of nasal sounds. Thus, the dataset is naturally unbalanced with oral phonemes

TABLE I
BD-PUBLICO SPEECH CORPUS.

	# utters	# speakers	# hours
train set	8388	100	23
validation set	584	10	2
test set	592	10	2

being the great majority. Table II shows the distribution of frames per class in the training set.

TABLE II
TRAINING SET FRAMES PER CLASS.

oral cons	oral vowels	nasal cons	nasal vowels
2972190	2512535	313440	607554

B. FraLusoPark Corpus

The FraLusoPark corpus [30] is a database of Portuguese and French Parkinson's speech recordings. The EP subset includes 120 speakers, 60 PD patients and 60 age- and gender-matched controls, aged 35 to 85. The PD patients are at various stages of the disease. They can be divided into three subsets: *early stage*, for patients with 0 to 3 years of PD and no motor fluctuations; *medium stage*, for individuals with 4 to 9 years of disease or 0 to 3 years and experiencing motor fluctuations and *advanced stage* for patients with more than 10 years of disease duration. Each subset has a total of 20 individuals.

All participants were recorded doing several tasks, including steady vowel /a/ phonation, maximum phonation time, oral diadochokinesia and several continuous speech tasks. For the purpose of this work, the task that is relevant is a reading task of a short text called "The North Wind and the Sun" (European Portuguese adaptation). The PD patients undertook this assessment in two distinct occasions: (1) at least 12 hours after withdrawal of all anti-Parkinsonian drugs and (2) following at least 1 hour after the administration of the usual medication.

This dataset is not divided in training, validation and test sets and the total duration of the EP subset is around 4.5 hours. All utterances are sampled at 16kHz. The size of the set can be seen in table III.

TABLE III
FRALUSOPARK SPEECH CORPUS.

	# utters	# speakers	# minutes
control	110	60	80
G1	91	20	68
G2	85	20	65
G3	69	20	55
total PD	245	60	188
total	355	120	268

VI. ARCHITECTURE

As previously stated, three distinct neural network architectures were implemented: a feedforward neural network, a hierarchical neural network, and a time-delay neural network.

All of the models underwent a similar training and testing process, shown in figure 4. The NN was trained using supervised learning, which requires labelled inputs so the model can learn input-output relationships. For that end, forced alignment was performed on the training dataset.

During the training phase, features were extracted from the data and afterwards are fed into the model along with the labels. After training, the best performing models were saved. In the test phase, the features were inputted into the stored model, which, depending on which mode it is, can provide either the accuracy and loss of the model for that dataset (evaluation mode) or the probabilities of each class being the correct label (prediction mode).

The data used for the training phase were the training and validation sets of BD-Público. For the test phase we used the test set of BD-Público as well as the controls of the FraLusoPark corpus to see how well the model adapts to recordings produced in different conditions from those of the training data.

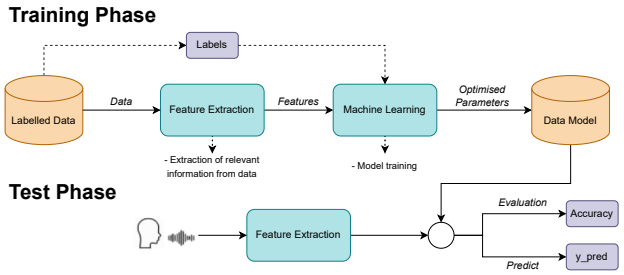

Fig. 4. Training and testing procedures scheme.

All the models were trained using the Adam optimiser and the He method as a weight initialiser. The loss function chosen was the cross entropy loss function which measures the performance of both binary and multi-class models that have as output a probability between 0 and 1. For each of the models created we did an ablation study of the following parameters:

- **Learning rate** - The learning rate specifies how much the model is adjusted in response to the predicted error when weights are updated. Too low a learning rate might lead to a lengthy training process, while too high can result in a sub-optimal weight set or an unstable training process.
- **Batch size** - Batch size is the amount of samples used to train a model before updating biases and weights. Very large batch sizes may result in poor generalisation, meaning the neural network performs poorly beyond the training set. Smaller batch sizes converge faster to good solutions, although they're not necessarily optimal. Most used batch sizes are powers of 2, which maximise CPU/GPU processing.
- **Epochs** - An epoch specifies the amount of times the learning algorithm loops over the entire dataset. Each training sample can adjust model parameters once each epoch. When training a model, it is important to find

the right number of epochs because with the rise in this number, the models weights are adjusted more times, and the curve progresses from underfitting to optimal to overfitting.

- **Dropout** - Dropout is a technique used during the training of DNN that have a large number of parameters with the goal of avoiding overfitting. The idea is to drop a portion of the units and their connections at each step of the training, preventing units from co-adapting excessively.

A. FeedForward Neural Network

The first model trained was a feedforward neural network, whose topology is depicted in figure 5. It consists of 5 layers with dropout being applied between each layer. The hidden layers have a decreasing number of nodes, with the first having 1024 nodes, the second 512 nodes and the third 256 nodes. Before the output layer batch normalisation is applied. The different layers of this architecture are described below.

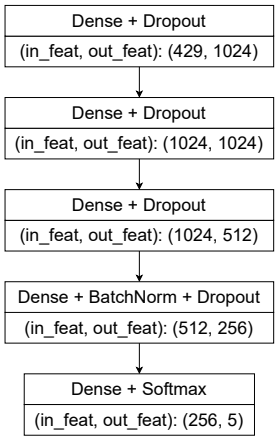

Fig. 5. FeedForward Neural Network architecture.

- **Input Layer** - For each input frame there is a context applied at the entrance of the neural network. This means that we use the adjacent frames in time to enhance the number of features. In this particular case the context is 11 frames, five preceding frames of the current input and five subsequent ($[t-5, t+5]$). Given that each frame has 39 features, the first layer has an input vector with a total of 39 x 11, or 429 features. Also, in order to avoid border effect, the first and last 5 frames are truncated. This results in an input shape of $(W-10)$ x 429 where W is the total number of frames.
- **Dense** - All the layers in this architecture are dense layers which consist in regular fully-connected layers that apply the computation shown in 1. The hidden layers all have a ReLU activation function.
- **Batch Normalisation** - This method was introduced in [31] and resorts to the concept of normalising the input data to zero-mean and constant standard deviation, which has been known for decades to be beneficial to neural network training, and applies it to the intermediate layers [32].

In order to minimise the cost of using batch normalisation there are two simplifications that are made. The first is that each scalar feature is normalised independently and the second is that the normalisation is made per mini-batch and not with the whole training set. The normalisation step is presented in 3 where $E[x]$ represents the mean, $Var[x]$ the variance and ϵ a constant that is added for numerical stability.

$$\hat{x} = \frac{x - E[x]}{\sqrt{Var[x] + \epsilon}} \tag{3}$$

BatchNorm improves training speed, allows greater learning rates, and enhances generalisation accuracy. By normalising activations throughout the network, it keeps the little changes to the layer parameters from escalating as the data is propagated through the DNN.

- **Output Layer** - This layer is also a dense layer, with the distinction that its activation function is softmax rather than ReLU. The shape of the output is $(W-10)$ x 5.

B. Hierarchical Neural Network

The second model trained was an hierarchical neural network. This architecture consists in a sequence of neural networks that form an acyclic graph. In this particular case, a binary tree.

Our model consists of three neural networks that have the same structure as the feedforward neural network. The main difference is that instead of having a single multi-class classifier, it divides the task into three smaller problems and uses one binary classifier for each of the sub-tasks.

As can be seen in figure 6 the first DNN divides each feature vector into consonants or vowels, while the remaining two split the consonants and vowels, respectively, into nasal or oral, resulting in the final labels: oral consonant, nasal consonant, oral vowel and nasal vowel.

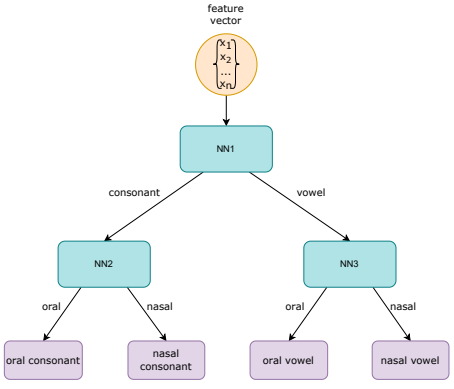

Fig. 6. Hierarchical Neural Network scheme.

C. Time-Delay Neural Network

Figure 7 depicts the architecture of the final model trained, a time-delay neural network. All of the layers in this network are TDNN layers, which means that context is applied to each

layer. We used 3 as the context size for all layers, with the exception of the output layer where this value corresponds to 1. Consequently, and given that each frame has 39 features, the first layer's input vector contains 39 x 3, or 117 features. As shown in the figure, the number of input features for each hidden layer is equal to three times the number of output features of the preceding layer. The first two hidden layers contain 1024 output nodes, whereas the latter two have 512 output nodes each. The output layer is also a TDNN layer with the difference that it uses softmax as the activation function. Given that the dilation is 1 for all layers, for each time step t the layers are passing frames $[t-1, t+1]$.

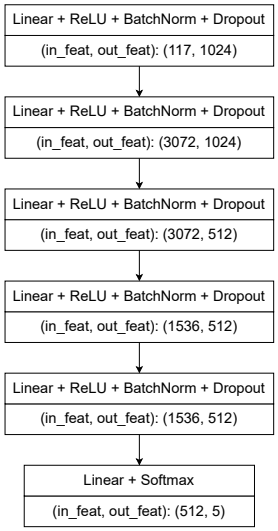

Fig. 7. Time Delay Neural Network architecture.

A standard TDNN layer architecture consists of a one dimension convolution, followed by a rectified linear unit (ReLU) activation function and batch normalisation. In our model dropout is also applied after the activation function. The input layer and the hidden layers all follow this layout. As stated above, the output layer does have a different activation function and does not have batch normalisation or dropout.

VII. Results

A. FeedForward Neural Network

Since the feedforward neural network was the first to be trained, each hyperparameter was fine-tuned independently, i.e., only one parameter was adjusted for each training run. This gave for a better understanding of each parameter's importance within the network. The most significant results obtained can be seen in the tables that follow.

TABLE IV
BATCH SIZE FINE-TUNING RESULTS.

	Train		Test	
Batch size	**Accuracy (%)**	**Loss**	**Accuracy (%)**	**Loss**
32	81.53	0.4828	82.04	0.4499
64	**81.52**	**0.4833**	**82.72**	**0.4424**
128	81.57	0.4817	82.52	0.4624

TABLE V
EPOCHS FINE-TUNING RESULTS.

	Train		Test	
Epochs	**Accuracy (%)**	**Loss**	**Accuracy (%)**	**Loss**
20	**82.48**	**0.4517**	**83.15**	**0.4291**
30	82.76	0.4437	83.11	0.8544
50	83.09	0.4352	83.10	0.4312

TABLE VI
LEARNING RATE FINE-TUNING RESULTS.

	Train		Test	
Learning Rate	**Accuracy (%)**	**Loss**	**Accuracy (%)**	**Loss**
0.01	81.52	0.4833	82.72	0.4424
0.001	82.48	0.4517	83.15	0.4291
0.0001	**82.9**	**0.4395**	**83.2**	**0.4286**

TABLE VII
DROPOUT FINE-TUNING RESULTS.

	Train		Test	
Dropout	**Accuracy (%)**	**Loss**	**Accuracy (%)**	**Loss**
0.2	**82.48**	**0.4517**	**83.15**	**0.4291**
0.3	81.62	0.4764	82.94	0.4377
0.5	80.00	0.5239	82.32	0.4622

The results obtained for the accuracies and losses on the train and test sets show that the best combination of parameters is a batch of 64 utterances, a learning rate of 0.0001, a dropout of 0.2 (or 20%) and 20 epochs. With this combination we achieve an accuracy of 83.2% and the confusion matrix presented in figure 8 for the BD-Publico test set. Even though, for some of the parameters (batch size and number of epochs) the training accuracies were higher for the bigger values, the test accuracies were lower which might indicate that the model was starting to overfit to the training dataset.

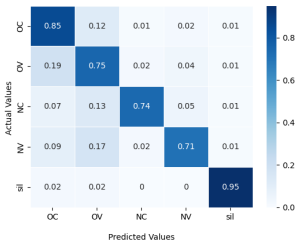

Fig. 8. Feedforward NN confusion matrix: OC - oral consonant, OV - oral vowel, NC - nasal consonant, NV - nasal vowel.

B. Hierarchical Neural Network

For the hierarchical neural network, we used the same parameters that produced the best results for the feedforward neural network, and experimented with varying the number of hidden nodes per layer. Four distinct models were developed, and the number of nodes for each dense layer of each model is shown in figure 9.

Since the hierarchical neural network consists of three different NN, different combinations of the three models

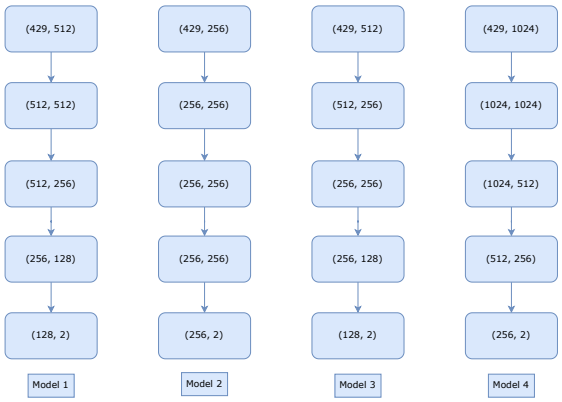

Fig. 9. Number of input and output nodes per layer for each model.

presented in figure 9 were tried. The results can be seen in table VIII, where the first column indicates which models were used for each of the NN, i.e., how many nodes each dense layer had. The numbering of the NN in this table is the same as in figure 6.

TABLE VIII
HIERARCHICAL NEURAL NETWORK RESULTS.

	Train Accuracy (%)			Test
Layers NN1->NN2->NN3	**NN1**	**NN2**	**NN3**	**Accuracy (%)**
1 -> 1 -> 1	82.24	97.51	92.29	77.04
2 -> 2 -> 2	80.73	95.76	91.08	75.78
1 -> 2 -> 2	82.84	98.22	93.76	77.99
3 -> 2 -> 2	82.67	97.95	93.03	77.72
4 -> 4 -> 4	**83.84**	**98.31**	**94.38**	**78.49**

Even though the results for the two networks that classify phonemes as oral or nasal are both high (above 90%), there is a 4-5% discrepancy between the consonant classifier (NN2) and the vowel classifier (NN3) in each iteration. This indicates that the network has a harder time differentiating nasal vowels from oral vowels than nasal consonants from oral consonants.

The architecture that achieved the best results is the one that uses the fourth model for all the networks. It obtained an accuracy of 78.49% in the test set and the confusion matrix presented in figure 10.

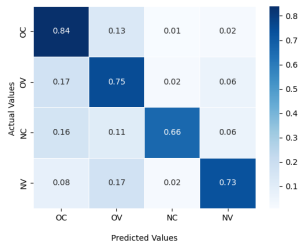

Fig. 10. Hierarchical NN confusion matrix: OC - oral consonant, OV - oral vowel, NC - nasal consonant, NV - nasal vowel.

C. Time Delay Neural Network

For the TDNN, since it is a much more memory costly model, the batch size had to be small in order to guarantee that the model could run without surpassing the available memory. For this reason, the batch size used was 16. The remaining parameters were fine-tuned and some of the results can be found in table IX. These results show that the configurations that achieved the best results had learning rate (LR) and dropout values of 0.001 and 0.3, respectively.

TABLE IX
TDNN FINE-TUNING RESULTS.

Parameters			Train		Test	
Epoch	**LR**	**Dropout**	**Acc.(%)**	**Loss**	**Acc.(%)**	**Loss**
15	0.01	0.3	83.57	9.17e-3	84.00	2.31e-5
25	0.01	0.3	84.06	9.12e-3	84.05	2.30e-5
25	0.01	0.2	83.57	9.17e-3	84.00	2.31e-5
25	0.001	0.2	85.73	8.98e-3	84.15	2.30e-5
35	0.001	0.2	86.53	8.90e-3	84.16	2.30e-5
35	**0.001**	**0.3**	**85.61**	**8.99e-3**	**84.42**	**2.30e-5**
35	0.0001	0.3	85.06	9.03e-3	84.13	2.30e-5

Figure 11 shows the effect that the number of epochs has on the training and validation sets of BD-Publico. The training accuracy continues growing for the whole 100 epochs. However the validation accuracy stabilises after a few dozen epochs which indicates that training for a very high number of epochs would only get the model to fit better with the training data and would not bring any benefits to other data (it would not generalise better). This is also visible in table IX: with the increase in number of epochs the train accuracy goes up but the test accuracy does not have a significant change ($\pm 0.05\%$).

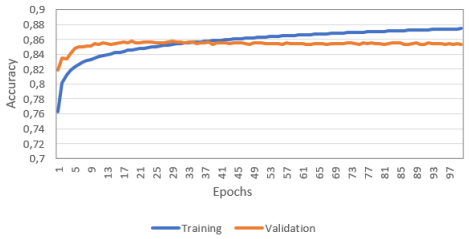

Fig. 11. TDNN training and evaluation accuracies over 100 epochs.

Overall, the best iteration of the TDNN model had the following parameters: batch size of 16, 35 epochs, a learning rate of 0.001 and a dropout of 30%. This model achieved a test accuracy of 84.42% for the BD-Publico test set and a confusion matrix that can be seen in figure 12.

D. Networks Comparison

The hierarchical neural network only has four output classes, whereas the other two networks each have five (the same four plus the silence class). This is due to the fact that during the first round of training experiments, we eliminated the frames corresponding to silences. With the development of

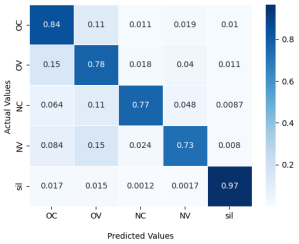

Fig. 12. TDNN confusion matrix: OC - oral consonant, OV - oral vowel, NC - nasal consonant, NV - nasal vowel, sil - silence.

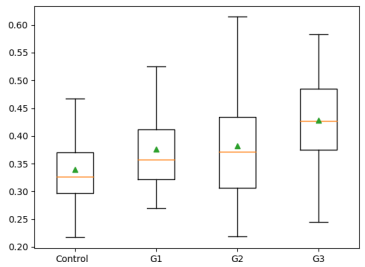

Fig. 13. Boxplot of nasality score for Parkinson: G1 - early stage, G2 - medium stage, G3 - advanced stage.

this work we determined that it would be beneficial to include the silences, since they are always present in the audio files.

For this reason, the results of the hierarchical neural network cannot be directly compared with the results of the other two NN. However, the silence class was not included in the feedforward neural network's first training iterations. These outcomes are not discussed in this report, although they were identical to those achieved by the hierarchical neural network. Due to the fact that the latter did not provide any advances in comparison to the former, along with limited time and resources, it was decided to concentrate on improving the feedforward neural network and developing the TDNN to determine which of these two achieved better results. Since the TDNN obtained the best results, this model was selected to perform the remaining tests.

E. Parkinson's Results

In order to understand how the neural network adapted to the FraLusoPark corpus recordings we ran an evaluation using only the controls of this corpus. The accuracy for this subset was 77.48%. This loss in accuracy, when compared to the BD-Publico dataset, might be explained by the difference in recording conditions.

The next analysis consisted in running the whole system as explained in IV. The results for each of the groups can be seen in figure 13. The figure shows a boxplot that represents the distribution in the nasality score per quartile for each group. The yellow lines represent the medians and the green triangle the mean values. Both of these ratios increase visibly from group to group ($Controls < G1 < G2 < G3$) with groups $G1$ and $G2$ having the lowest discrepancy.

To validate these results we conducted an analysis of variance (ANOVA) followed by a *post-hoc* analysis to determine which groups differ from each other. The Tamhane *post-hoc* test was chosen since it is insensitive to unequal variances between groups [33].

The ANOVA results showed that there are differences between the groups, $F(3, 349) = 13.58, p < 0.001$. The *post-hoc* analysis (table X) shows that, if we consider $p < 0.05$, there are significant differences between the control group and all the Parkinson's groups and there are also significant differences between the group in early stages of Parkinson's and the group in advanced stages.

It is also important to note that groups 2 and 3 have virtually no differences between them ($p = 0.978$), which suggests that nasalisation is as prevalent in Parkinson's patients in medium stages as it is in advanced stages.

This is in accordance with the literature. In [34], the authors do an assessment of speech and voice quality in PD patients and mention two surveys, each done on 200 PD patients, that are based on perceptual ratings. These studies found that voice abnormalities are often noticed even in the early stages of Parkinson's disease and are present in almost all patients towards the late stages. In the later stages of PD, articulatory impairments, including hypernasality, become more apparent.

TABLE X
RESULTS OF TAMHANE *post-hoc* TESTS.

Group (I)	Group (J)	Significance	Cohen's *d*
Control	G1	.003	0.391
	G2	$< .001$	0.723
	G3	$< .001$	0.866
G1	G2	.298	0.332
	G3	.014	0.475
G2	G3	.978	0.143

In Mathad et al. [26], the authors had 14 speech-language pathologists rank hypernasality present in the audios on a scale from 1 to 7. These perceptual rankings were then compared with the results from their system. In our work, due to time and resources constraints, we could not do perceptual rating tests with trained clinicians. However, two (non-clinical) INESC-ID researchers performed an informal test. This revealed that while evaluating the audios, other affected characteristics of the speech were more prominent than nasalisation and, hence, more discernible in the audio.

Nevertheless, the effect sizes presented in table X, showed small (differences between control and $G1$, $G1$ and $G2$) to large effect sizes (difference between control group and $G3$). Moreover, aggregating the three groups with PD into one group and comparing it with the control group, resulted in an effect size in the moderate to large range, $t(351) = 5.51, p < .001, d = 0.63$.

VIII. Conclusions

This thesis aimed to construct a deep-learning system that could differentiate nasal from non-nasal sounds in EP. For this end, three neural network architectures were developed.

Creating deep learning models has some challenges including the need for large amounts of data to train the models and the need for good GPU with a considerable memory size. Despite these limitations we were able to create models that achieved good results.

The TDNN was the neural network that achieved the highest level of accuracy in the BD-Publico test set, with a score of 84.42%. The success of this design in the field of speech processing may be attributed, in large part, to its shift-invariance nature and the fact that it adds context to every layer.

The second goal of this thesis was to assess if this model would be able to detect hypernasality in PD patients. The results showed that there are significant differences between the control group and the three groups of Parkinson's patients. The effect sizes between the control group and the aggregation of the three PD groups was of moderate size. This indicates that, even though nasality is not the most prominent characteristic of PD speech, it is more prevalent in PD patients than it is in healthy individuals.

The results also demonstrated that the effect sizes between the control and the three stages of Parkinson's were increasingly higher. This indicates that this characteristic could also be used to measure the progression of the disease.

References

[1] P. Batista and A. Pereira, "Quality of life in patients with neurodegenerative diseases," Dimensions, vol. 1, p. 3, 2016.

[2] G. Deuschl, et al., "The burden of neurological diseases in Europe: an analysis for the global burden of disease study 2017," The Lancet Public Health, vol. 5, no. 10, pp. e551–e567, 2020.

[3] J. Robin, et al., "Evaluation of speech-based digital biomarkers: Review and recommendations," *Digital Biomarkers*, vol. 4, no. 3, pp. 99–108, 2020.

[4] D. A. Cairns, et al., "A noninvasive technique for detecting hypernasal speech using a nonlinear operator," *IEEE transactions on biomedical engineering*, vol. 43, no. 1, p. 35, 1996.

[5] T. Pruthi, *Analysis, vocal-tract modeling and automatic detection of vowel nasalization.* University of Maryland, College Park, 2007.

[6] M. Saxon, et al., "Robust estimation of hypernasality in dysarthria." *CoRR,* vol. abs/1911.11360, 2019.

[7] X. Wang, et al., "Automatic hypernasality detection in cleft palate speech using CNN," *Circuits, Systems, and Signal Processing*, vol. 38, no. 8, pp. 3521–3547, 2019.

[8] M. Saxon, et al., "Objective measures of plosive nasalization in hypernasal speech," in ICASSP 2019-2019 *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6520–6524.

[9] A. Pompili, et al., "Automatic detection of Parkinson's disease: an experimental analysis of common speech production tasks used for diagnosis," in *International Conference on Text, Speech, and Dialogue*. Springer, 2017, pp. 411–419

[10] C. L. Ludlow, et al., "Speech timing in Parkinson's and Huntington's disease," *Brain and language*, vol. 32, no. 2, pp. 195–214, 1987.

[11] P. Zwirner and G. J. Barnes, "Vocal tract steadiness: a measure of phonatory and upper airway motor control during phonation in dysarthria," *Journal of Speech, Language, and Hearing Research*, vol. 35, no. 4, pp. 761–768, 1992.

[12] J. Rusz, et al., "Quantitative acoustic measurements for characterization of speech and voice disorders in early untreated Parkinson's disease," *The journal of the Acoustical Society of America*, vol. 129, no. 1, pp. 350–367, 2011.

[13] A. M. Goberman and C. Coelho, "Acoustic analysis of parkinsonian speech i: Speech characteristics and l-dopa therapy," *NeuroRehabilitation*, vol. 17, no. 3, pp. 237–246, 2002.

[14] R. J. Holmes, et al., "Voice characteristics in the progression of Parkinson's disease," *International Journal of Language & Communication Disorders*, vol. 35, no. 3, pp. 407–418, 2000.

[15] S. Skodda and U. Schlegel, "Speech rate and rhythm in Parkinson's disease," *Movement disorders: official journal of the Movement Disorder Society*, vol. 23, no. 7, pp. 985–992, 2008.

[16] K. Tjaden, "Speech and swallowing in Parkinson's disease," *Topics in geriatric rehabilitation*, vol. 24, no. 2, p. 115, 2008.

[17] J. C. Vasquez-Correa, et al., "Parallel representation learning for the classification of pathological speech: studies on Parkinson's disease and cleft lip and palate," *Speech Communication*, vol. 122, pp. 56–67, 2020.

[18] I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning. MIT Press, 2016.

[19] A. Waibel, et al., "Phoneme recognition using time-delay neural networks," *IEEE transactions on acoustics, speech, and signal processing*, vol. 37, no. 3, pp. 328–339, 1989.

[20] V. Peddinti, et al., "A time delay neural network architecture for efficient modeling of long temporal contexts," in *Sixteenth annual conference of the international speech communication association*, 2015.

[21] Y.-J. Tsai, et al., "Voice low tone to high tone ratio, nasalance, and nasality ratings in connected speech of native mandarin speakers: a pilot study," *The Cleft palate-craniofacial journal*, vol. 49, no. 4, pp. 437–446, 2012.

[22] R. Kataoka, et al., "The relationship between spectral characteristics and perceived hypernasality in children," *The Journal of the Acoustical Society of America*, vol. 109, no. 5, pp. 2181–2189, 2001.

[23] T. Pruthi and C. Y. Espy-Wilson, "Acoustic parameters for the automatic detection of vowel nasalization," in *Eighth Annual Conference of the International Speech Communication Association*, 2007.

[24] G.-S. Lee, et al., "Evaluation of hypernasality in vowels using voice low tone to high tone ratio," *The Cleft Palate-Craniofacial Journal*, vol. 46, no. 1, pp. 47–52, 2009.

[25] J. R. Orozco-Arroyave, et al., "Automatic detection of hypernasal speech signals using nonlinear and entropy measurements," in *Thirteenth Annual Conference of the International Speech Communication Association*, 2012.

[26] V. C. Mathad, et al., "Deep learning based prediction of hypernasality for clinical applications," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6554–6558

[27] J. R. Orozco-Arroyave, et al., "Automatic detection of hypernasal speech of children with cleft lip and palate from spanish vowels and words using classical measures and nonlinear analysis," *Revista Facultad de Ingenier 1a Universidad de Antioquia*, no. 80, pp. 109–123, 2016.

[28] C. Carvalho and A. Abad, "Tribus: An end-to-end automatic speech recognition system for european portuguese," *IberSpeech*, 2021.

[29] J. P. Neto, et al., "The design of a large vocabulary speech corpus for portuguese," in *Fifth European Conference on Speech Communication and Technology*, 1997.

[30] S. Pinto, et al., "Dysarthria in individuals with Parkinson's disease: a protocol for a binational, cross-sectional, case-controlled study in french and european portuguese (fralusopark)," *BMJ open*, vol. 6, no. 11, p. e012885, 2016.

[31] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International conference on machine learning*. PMLR, 2015, pp. 448–456.

[32] N. Bjorck, et al., "Understanding batch normalization," *Advances in neural information processing systems*, vol. 31, 2018.

[33] "One-way anova post hoc tests," https://www.ibm.com/docs/en/spss-statistics/saas?topic=anova-one-way-post-hoc-tests, accessed: 2022-10-18.

[34] S. Skodda, "Analysis of voice and speech performance in Parkinson's disease: a promising tool for the monitoring of disease progression and differential diagnosis," *Neurodegenerative Disease Management*, vol. 2, no. 5, pp. 535–545, 2012.