# Parallel Representation Learning for the Classification of Pathological Speech: Studies on Parkinson's Disease and Cleft Lip and Palate

J. C. Vasquez-Correa[a,b], T. Arias-Vergara[a,b,c], M. Schuster[c], J. R. Orozco-Arroyave[a,b], E. Nöth[a]

[a] *Pattern Recognition Lab, Friedrich-Alexander-University Erlangen-Nuremberg, Germany.*
[b] *Faculty of Engineering, University of Antioquia UdeA, Medellín, Colombia.*
[c] *Department of Otorhinolaryngology, Ludwig-Maximilians-University, Munich, Germany.*

**Abstract**

Speech signals may contain different paralinguistic aspects such as the presence of pathologies that affect the proper communication capabilities of a speaker. Those speech disorders have different origin depending on the type of the disease. For instance, diseases with morphological origin such as cleft lip and palate that causes hypernasality, or with neurodegenerative origin such as Parkinson's disease that generates hypokinetic dysarthria on the patients. Automatic assessment of pathological speech allows to support the diagnosis and/or the evaluation of the disease severity. Conventional methods are based on the manually applied assessment of single features such as jitter, shimmer, or formant frequencies that may not completely model all of the phenomena that appear due to the disease. This paper introduces a novel strategy based on unsupervised representation learning for automatic detection of pathological speech. The proposed approach is based on the use of recurrent and convolutional autoencoders trained to extract informative features to characterize the presence of pathologies in speech. A novel feature set based on the reconstruction error of the autoencoders is also proposed. The performance of the introduced models is evaluated classifying pathological speech signals recorded from people suffering

*Corresponding author
*Email address:* juan.vasquez@fau.de (J. C. Vasquez-Correa)

from Parkinson's disease, and children with cleft lip and palate. All participants from this study were Spanish native speakers. The proposed models are accurate to classify the speech signals of both kinds of diseases, with an accuracy of up to 97% for cleft lip and palate, and up to 84% for the case of Parkinson's disease. We also show that the reconstruction error from the autoencoders in different frequency regions contain information related to specific speech symptoms of both diseases.
*Keywords:* Pathological speech, unsupervised representation learning, convolutional autoencoders, recurrent autoencoders Parkinson's disease, Cleft lip and palate

## 1. Introduction

The speech signals contain paralinguistic information with specific cues about the speaker, including their identity, mood, age, gender, and the presence of diseases that may alter their communication capabilities. The automatic classification of paralinguistic aspects has many potential applications, and has received a lot of attention by the research community (Schuller and Batliner, 2013; Schuller et al., 2019; Cummins et al., 2018). Potential applications include the assessment of pathological speech, which allows the development of computer aided tools to support the diagnosis and the prediction of the disease severity (Orozco-Arroyave et al., 2015). Particularly, the assessment of pathological speech has focused on the analysis of diseases with different origin such as morphological, or neurological.

One of the speech pathologies caused by morphological changes is hypernasality, which appears in about 90% of patients with cleft lip and palate (CLP), even after surgical correction of the palate (Vijayalakshmi et al., 2007). CLP patients may experience feeding and swallowing difficulties, hearing loss, and different speech disorders such as soft voice and omission or substitution of sounds. CLP also causes excess of nasalization, which is characterized by the presence of additional resonances in the nasal cavity during the speech production (Wy-

att et al., 1996). The velopharyngeal dysfunction may also cause nasal emission of the air stream, resulting in weak consonant production, short utterance length, and the development of compensatory articulation movements (Wyatt et al., 1996). In addition to the diseases with morphological origin, patients with neurological disorders such as Parkinson's disease (PD) develop hypokinetic dysarthria, which appears in about 90% of the patients (Ho et al., 1999). Speech symptoms caused by PD include rigidity of the vocal folds, bradykinesia, and reduced control of muscles and limbs involved in the speech production. The effects of the dysarthria in the speech of PD patients also include increased acoustic noise (Hornykiewicz, 1998), reduced intensity (Baker et al., 1998), harsh and breathy voice quality, increased nasality (Spencer and Rogers, 2005), monopitch, monoludness, speech rate disturbances (Skodda et al., 2011), imprecise articulation of consonants (Tykalova et al., 2017), and involuntary introduction of pauses (Moretti et al., 2003).

CLP and PD patients share common speech symptoms due to the presence of the disease. For instance, PD patients can exhibit hypernasality because their reduced control of the nasal cavity (Saxon et al., 2019). In addition, the speech of CLP patients is affected by different articulation disorders, similar to those observed in PD patients, such as problems in the pronunciation of fricatives or weakened plosives (Maier et al., 2009). Both CLP and PD patients have compensatory articulation disorders. Those movements in CLP patients include the substitution of glottal stops by voiced stops, or nasal fricatives by oral fricatives (Prathanee et al., 2014). For the case of PD patients, they usually have incomplete vocal closure by maintaining a continuous level of vocal fold activity to avoid the difficulty of initiating the phonation (Blanchet and Snyder, 2009). This behavior causes that voiceless stops such as /p/, /t/, and /k/ are replaced by /b/, /d/, and /g/.

Clinical observations in the speech of patients can be objectively and automatically measured by using computer aided methods supported in signal processing and pattern recognition methods with the aim to address two main aspects: (1) to support the diagnosis of the disease by classifying healthy control

(HC) subjects and patients, and (2) to predict the level of degradation of the speech of the patients according to a specific clinical scale.

Different approaches have been proposed to automatically model pathological speech with different origins. One of the first studies to automatically assess the speech of CLP patients was performed by Schuster et al. (2006), who evaluated the intelligibility of CLP patients using the word accuracy obtained from a speech recognition system. The authors showed that there is significant difference between the word accuracy obtained between children with isolated cleft lip, isolated cleft palate, unilateral cleft lip and palate, and bilateral cleft lip and palate. The results also indicated that there is a significant negative correlation between the word accuracy and a perceptual evaluation of the intelligibility of the children (Pearson correlation $r = -0.90$). Vijayalakshmi et al. (2007) improved the resolution of the speech spectrum using the modified group delay functions to find a peak located in 250 Hz, which exhibited a higher intensity in hypernasal voices than in healthy ones. Later, Maier et al. (2009) aimed to detect the presence of distinct articulation disorders in CLP patients such as hypernasality in vowels, nasalized consonants, pharyngealization, glottal articulation, among others. The authors computed features such as the word accuracy, several prosody features from the pitch and energy contours, the Teager energy profile, Mel-frequency cepstral coefficients (MFCCs), and the goodness of pronunciation. The proposed method achieved moderate to good agreement (kappa index $\kappa \approx 0.6$) for the detection of all articulation disorders, using several classification methods. Orozco-Arroyave et al. (2013) computed several features based on non-linear dynamics analysis to discriminate between children with CLP and HC, when they pronounced sustained vowels and isolated words. The authors reported an accuracy of up to 92%, using a classifier based on support vector machines (SVM). Golabbakhsh et al. (2017) detected hypernasality in CLP children using acoustic features such as jitter, shimmer, and MFCCs, combined with features extracted from wavelet decompositions. The participants were asked to read six sentences in Persian language. The authors reported accuracies of up to 85% when MFCCs were combined with wavelet-based features,

using also an SVM classifier. Vikram et al. (2018) computed MFCCs and their derivatives only in segments with glottal activity to classify children with CLP vs. HC. The features were classified with a deep neural network (DNN) and Gaussian mixture model–based classifiers with an accuracy of 93.3%. Recently, Dubey et al. (2019) introduced the use of constant Q-Cepstral coefficients to classify normal, moderate, and severe hypernasality levels of CLP patients and HC children, when they pronounce sustained vowels. The authors reported an accuracy of up to 83.3% using an SVM classifier.

Regarding the assessment of the speech of PD patients, several studies have described the associated speech symptoms considering features based on phonation, articulation, prosody, and intelligibility (Orozco-Arroyave et al., 2018; Orozco-Arroyave, 2016; Bocklet et al., 2013; Rusz et al., 2017; Moro-Velázquez et al., 2018). Phonation features model the stability and periodicity of the vocal fold vibration, and include perturbation features such as jitter, shimmer, amplitude perturbation quotient, pitch perturbation quotient, and non-linear dynamics measures (Orozco-Arroyave et al., 2015; Sakar et al., 2013; Naranjo et al., 2016). Phonation features also include noise measures such as harmonics to noise ratio, glottal to noise excitation ratio, and voice turbulent index, among others (Tanaka et al., 2011). Articulation symptoms are related to the modification of the position, constriction, and shape of several limbs and muscles to produce speech. These symptoms have been modeled with features such as vowel space area, vowel articulation index, formant centralization ratio, voiced onset time, and onset energy (Orozco-Arroyave, 2016; Rusz et al., 2013; Novotnỳ et al., 2014; Montaña et al., 2018). Additional articulation features included a non-linear dynamics analysis on the amplitude envelope of diadochokinetic (DDK) exercises (Godino-Llorente et al., 2017), models of the bio-mechanical systems of the jaw-tongue movement (Gómez-Vilda et al., 2017), or the evaluation of the pronunciation of specific phonetic units using posterior probabilities from Gaussian mixture models (Moro-Velazquez et al., 2019). Prosody deficits in PD are manifested as monotonous speech, monoloudness, reduced stress, and changes in speech rate and pauses (Skodda et al., 2011). In addition, the bradykinesia

and freezing of movement sometimes cause difficulty in the initiation of voluntary speech and inappropriate long silences. Prosody features are based on the contour of the fundamental frequency, energy, duration, and voiced rate (Bocklet et al., 2013; Galaz et al., 2016). Finally, intelligibility is a measure of how comprehensible is the speech of a person. Intelligibility assessment is commonly performed using automatic speech recognition systems, and the word error rate has been used to discriminate between PD and HC speakers (Orozco-Arroyave et al., 2016a; Barnish et al., 2016; Dimauro et al., 2017).

In addition to the hand-crafted feature extraction models, there is a growing interest in the research community to consider deep learning models in the assessment of the speech of PD patients. The "2015 computational paralinguistic challenge (ComParE)" (Schuller et al., 2015) had one of the sub-challenges about the automatic estimation of the neurological state of PD patients. The ground-truth was given according to the part III of the movement disorder society - unified Parkinson's disease rating scale (MDS-UPDRS), which is focused on the evaluation of motor capabilities of the patients, including one specific item for speech production. The winners of the challenge (Grósz et al., 2015) reported a Spearman's correlation of 0.65 when grouping automatically the speech tasks per speaker and using Gaussian processes and DNNs to perform the prediction of the clinical score. A deep learning based articulation approach was proposed in (Vásquez-Correa et al., 2017) to model the difficulties of the patients to stop/start the vibration of the vocal folds. Onset and offset transitions were modeled with time-frequency representations to be used as input for a convolutional neural network (CNN). The authors considered speech recordings of PD patients and HC speakers in three languages: Spanish, German, and Czech, and reported accuracies ranging from 70% to 89%, depending on the language. Tu et al. (2017) proposed a deep learning model to predict the dysarthria severity adding an intermediate interpretable hidden layer with four perceptual dimensions: nasality, vocal quality, articulatory precision, and prosody. The authors obtained an interpretable output highly correlated (Spearman's correlation $\rho$=0.82) with a subjective evaluation of the dysarthria severity of the

patients provided by speech and language pathologists. Zhang (2017) combined perturbation and articulation features with a deep learning model based on autoencoders to classify PD patients and HC subjects. Different acoustic features were used as input for the autoencoder. The bottleneck features from the autoencoder were used to feed a K-nearest neighbor (KNN) classifier. The authors considered the data from Sakar et al. (2013), which include 20 PD and 20 HC subjects, all of them Turkish native speakers. The reported accuracy was 94%; however, the results were slightly optimistic because the hyper-parameters of the autoencoder were optimized on the test set. A different approach was proposed by Zhang et al. (2019), where the authors considered non-speech body sounds such as breathing, clearing throat, and swallowing to classify PD vs. HC subjects. The non-speech body sounds were modeled using a deep learning strategy based on ResNet architectures. The proposed method achieved an accuracy of up to 83.3% in a dataset formed with 321 PD patients and 569 HC subjects. The results were comparable to the ones obtained with normal speech sounds. However, the speaker independence was not guaranteed in the training process, which leads to biased and optimistic results.

Conventional hand-crafted features extracted in the literature may not adequately capture enough information to characterize the speech signals associated with different speech disorders. Methods based on feature representation learning have the potential to extract more abstract and robust features than those manually computed. These features could help to improve the accuracy of different models to classify pathological speech (Cummins et al., 2018). There are recent studies focused on extracting features based on deep learning strategies for assessment of pathological speech. However, there are still several strategies that can be addressed, especially in aspects related to unsupervised representation learning to extract suitable features for pathological speech classification.

This paper introduces a parallel representation learning strategy to model two kinds of pathological speech signals with different origin: dysarthric speech due to PD and hypernasal speech due to CLP. For both applications, this a first step in developing robust speech-based technology to evaluate the degree

of affection of the speech, i.e., the dysarthria level of PD patients, and the hypernasality level of CLP children after surgical intervention, which can be helpful to evaluate the progress of speech therapy. In addition, the classification methods are particularly useful for CLP patients because in some cases the cleft palate is hidden by the skin. Two types of autoencoders were implemented to compute low-dimensional feature representations of the speech frames: (1) a convolutional autoencoder (CAE) to learn a representation of the spatial distribution of the energy content in a spectrogram, and (2) a recurrent autoencoder (RAE) to model the temporal evolution of the spectral components of a speech frame. We considered the features from the hidden representation in the bottleneck space, and a proposed feature set based on the reconstruction error of the autoencoder in different spectral components of the speech signal. These features are used to classify PD and CLP patients vs. HC subjects, all of them Spanish native speakers. The aim of choosing these pathologies is to test whether the proposed methods are accurate to model speech disorders with different origin, i.e., morphological and neurodegenerative, and for speakers with different ages, i.e., children affected by CLP and elderly persons affected by PD. In addition, due to the fact that CLP and PD patients exhibit common speech problems such as the presence of hypernasality or compensatory articulatory movements (Saxon et al., 2019; Maier et al., 2009; Blanchet and Snyder, 2009), we think that the methods proposed here are applicable and robust to model the speech disorders from both diseases.

The rest of the paper is organized as follows. Section 2 describes the data used for training the autoencoders and to classify PD patients and CLP children. Section 3 describes the proposed convolutional and recurrent autoencoders to characterize pathological speech, and the classification and validation strategies. Section 4 shows the results obtained in this paper to classify both PD and CLP patients. Finally, Section 5 shows the conclusions obtained from the study and further experiments to be performed.

## 2. Data

### *2.1. Training data*

The CIEMPIESS corpus (Hernández-Mena and Herrera-Camacho, 2014) was used to train the convolutional and recurrent autoencoders. The data consist of 17 hours of FM radio podcasts in Mexican Spanish. The data consider only "clean" utterances, i.e., those made by only one person, with no background noise, foreign accents, or music. The data are formed with 16717 audio files with a sampling frequency of 16 kHz and 16-bit resolution. The speech signals were uttered by 96 male and 45 female speakers. 700 utterances from the entire corpus (speaker independent) were subtracted to be used as the validation set for the training process of the autoencoders.

### *2.2. CLP data*

Data for CLP was provided by *Grupo de procesamiento y reconocimiento de señales (GPRS)* of the National University from Colombia. The data contain utterances from 135 children with repaired CLP and 58 HC. The age of the children ranges from 5 to 15 years old. All the patients were evaluated by speech therapists and they were diagnosed with hypernasal speech, mainly because changed nasality is often still present after surgical therapy of the cleft. The tasks performed by the participants include the pronunciation of isolated Spanish words such as /bola/, /chuzo/, /coco/, /gato/, /jugo/, /mano/, /papa/, and /susi/. These words contain different groups of phonemes to characterize properly the different place and manner of articulation of the patients (Orozco-Arroyave et al., 2016b).

### *2.3. PD Data*

Data for PD include recordings of the PC-GITA database (Orozco-Arroyave et al., 2014). The data contain speech utterances from 50 PD and 50 HC Colombian Spanish native speakers. Each speaker performed different speech exercises, including DDK tasks i.e., rapid repetition of syllables such as /PA-TA-KA/,

reading of ten isolated sentences, reading of a phonetic balanced text with 36 words, and a spontaneous monologue about their daily activities. Additional information from the participants is shown in Table 1. The results from the statistical tests show that the data are balanced in gender, age, and education level. The values of the MDS-UPDRS-III (Goetz et al., 2008) scores, which range from 0 to 132 indicate that most of the patients are in intermediate state of the disease.

Table 1: General information of the subjects. Time since diagnosis, age and education are given in years. **[F/M]**: Female/Male. Mean(Standard deviation). *$p$–value calculated through Chi–square test. **$p$–value calculated through t-test.

| | PD patients | Healthy controls | Patients vs. controls |
|---|---|---|---|
| Gender [F/M] | 25/25 | 25/25 | $^{*}p = 1.00$ |
| Age [F/M] | 60.7(7.3)/61.3(11.7) | 61.4(7.1)/60.5(11.6) | $^{**}p = 0.98$ |
| Education level [F/M] | 11.5(4.1)/10.9(4.5) | 11.5(5.2)/10.6(4.4) | $^{**}p = 0.88$ |
| Time since diagnosis [F/M] | 12.6(11.5)/8.7(5.8) | – | |
| MDS–UPDRS–III [F/M] | 37.6(14.0)/37.8(22.1) | – | |

## 3. Methods

The general methodology addressed in this study is shown in Figure 1. The training data are used to train convolutional and recurrent autoencoders, which are used to extract features from the utterances of the test data. The extracted features from both autoencoders are classified using SVM and DNN methods to discriminate between PD and HC subjects, and between CLP and HC children. Additional details about each stage of the methodology are explained in the following sections.

### *3.1. Convolutional autoencoder*

The CAE is considered to characterize the spatial information embedded in the time-frequency representation from the input. The architecture of the CAE is shown in Figure 2. The input is a spectrogram with 128 frequency bins distributed according to the Mel-scale and 126 time steps. The speech signal is segmented into "chunks" of 500 ms with a time-shift of 250 ms. The

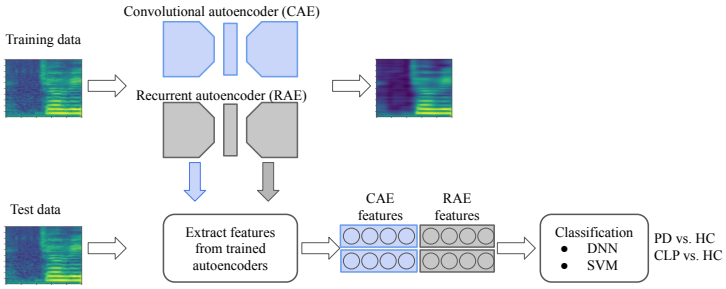


Figure 1: General methodology followed in this study.

short time Fourier transform (STFT) is computed for each "chunk" with a window length of 32 ms and a step-size of 4 ms, forming the 126 time-steps. The STFT is computed with 512 frequency points, which are transformed into the Mel-scale using 128 filters, forming the input spectrogram observed in the left part of Figure 2. In these spectrograms we did not lose information about the fundamental frequency contour because the high number of Mel filters and the large frame size. The input spectrogram is encoded with a four layer CNN with leaky RELU activation functions and a fully connected layer to form the bottleneck representation $\mathbf{h}$. Each layer of the CNN consists of a $3 \times 3$ kernel to map the fine structures of the time-frequency representation into different feature maps. The number of feature maps on each convolutional layer is twice the previous one in order to get more detailed representations of the input space in the deeper layers. In addition, max-pooling operation is also performed after each convolutional layer. The decoder is formed with a set of four transposed convolutional layers to map the bottleneck representation into the reconstructed version of the input spectrogram.

### *3.2. Recurrent autoencoder*

Besides the CAE described above, we also considered an RAE to characterize the temporal structures of the input spectrogram. For this case the input spectrogram is the same as the one in the CAE. Each column of the $n = 126$ time steps of the spectrogram serves as input for a sequence to one recurrent

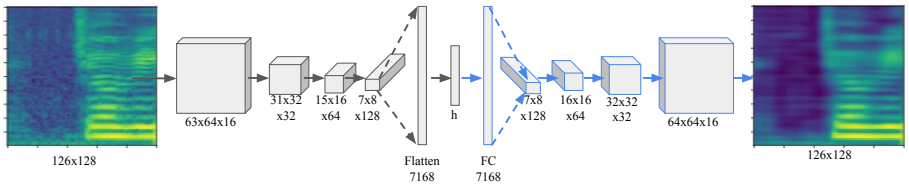


Figure 2: Scheme of the CAE considered in this study. **FC**: fully connected layer, **h**: bottleneck representation.

neural network in the encoder, which is formed with a bidirectional LSTM (BLSTM) with 128 cells to model information from the past (backward) and future (forward) states of the sequence, simultaneously. The output sequence of the BLSTM layer at the last time step $\mathbf{x}_n$ is stacked with the hidden state of the layer at the last time step $\mathbf{s}_n$ because they have observed and carry information about the whole input sequence. This stacked vector then passes trough a fully connected layer to get the bottleneck representation **h**. The decoder is formed with a sequence of 2 LSTM layers to retrieve the original spectrogram from the bottleneck representation. The bottleneck features were replicated 126 times for the decoder. This part was necessary since every LSTM cell in the decoder requires an input vector. The complete architecture of the RAE is shown in Figure 3. The Pytorch (Paszke et al., 2017) implementation of the trained models are available online[1] for the research community. The repository also contains scripts to train the autoencoders with different datasets, and methods to use the trained autoencoders to extract the proposed features.

### *3.3. Feature Extraction with Autoencoders*

Two different feature sets are extracted from the trained autoencoders, according to Figure 4. The first set consists of the bottleneck features **h** obtained from the CAE and the RAE, computed from the "chunks" with 500 ms length. The second feature set is based on the mean square error (MSE) between the input and the decoded spectrograms, computed for each frequency

[1] `https://github.com/jcvasquezc/AEspeech`

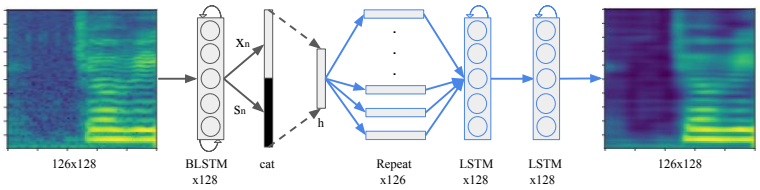


Figure 3: Scheme of the RAE considered in this study. $\mathbf{h}$: bottleneck representation. $\mathbf{x}_n$ output of the BLSTM layer at the last time step, $\mathbf{s}_n$ hidden state of the BLSTM at the last time step.

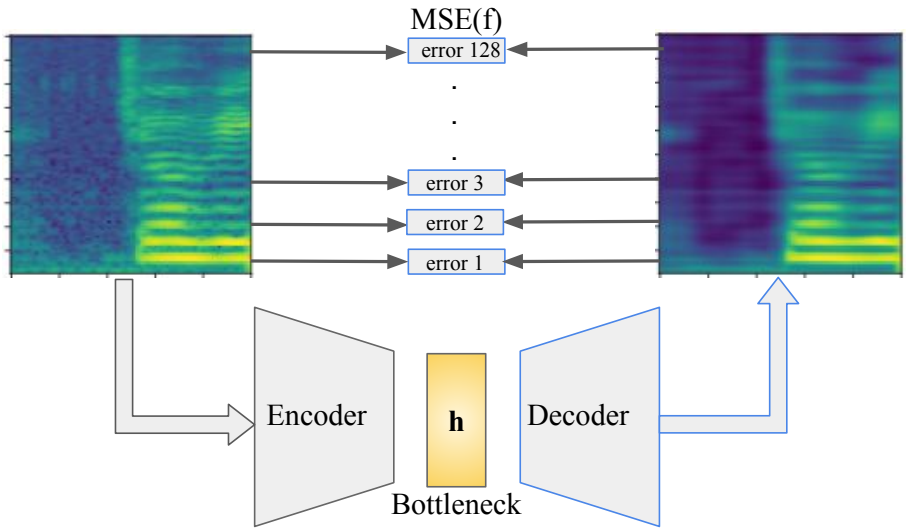


Figure 4: Features extracted from the autoencoders in this study.

band (MSE($f$)). The main hypothesis for the second feature set is that not all frequency regions of the spectrogram can be reconstructed with the same error, and such a reconstruction error is related to the presence of paralinguistic aspects such as different speech disorders.

### *3.4. Classification and validation*

Two classification strategies were considered to evaluate the proposed approach. The first classifier is based on a fully connected DNN to process the extracted $M$-dimensional feature vectors from the autoencoders. The network

is formed with two fully connected layers with $M/2$ neurons, followed by the classification layer with a Softmax activation function to make the final decision. A dropout layer with a rate of 0.5 is considered between the fully connected layers to avoid over-fitting. The classifier was trained with an ADAM optimizer (Kingma and Ba, 2014). All features from the test subjects were classified individually, to get local decisions for each feature vector. The final decision for each speaker was made based on a majority vote strategy. The second classification strategy is based on an SVM with a Gaussian kernel. The features extracted from a single utterance $X \in \mathbb{R}^{M \times N}$ (being $N$ the number of 500 ms length spectrograms extracted from the utterance) are concatenated in the second dimension, and four statistical functions are extracted (mean, standard deviation, skewness, and kurtosis), forming a vector $x_u \in \mathbb{R}^{4 \times M}$ to represent the complete utterance. These vectors are classified with an SVM with a Gaussian kernel to get a global decision for the complete utterance. Since each speaker pronounces several utterances, the decision for each speaker was also made based on a majority vote strategy. The SVM and DNN are considered because they are currently state-of-the-art methods in different applications related to recognition of paralinguistic aspects from speech, including pathological speech classification (Orozco-Arroyave et al., 2015; Berus et al., 2019; Viswanathan et al., 2020; Novotný et al., 2020). In addition, SVMs are robust enough to model high dimensional feature spaces and small datasets (Scholkopf and Smola, 2001)

The validation process for the classifiers follows a nested speaker independent 10-fold cross-validation strategy. We consider 80% of the data for training (155 subjects for the case of CLP data and 80 speakers from the PD corpus), 10% for development (19 subjects from the CLP and 10 for the case of PD corpus), and the remaining 10% for test.

For the SVM classifier, the complexity hyper-parameter $C$ and the bandwidth of the kernel $\gamma$ were optimized in a randomized search strategy, as follows: the values of $C$ and $\gamma$ are modeled with an exponential probability density function, which generates values for each hyper-parameter to be evaluated according

to the performance in the development set. After several iterations with differently generated values from the probability functions, the hyper-parameters that produced the highest accuracy are stored. The optimal hyper-parameters are found based on the median of the values of the hyper-parameters obtained for each fold. Finally, the 10-fold cross-validation is repeated in order to guarantee that all test samples are evaluated with the optimal hyper-parameters, which leads to more realistic and stable results.

## 4. Experiments and Results

### *4.1. Analysis of the reconstruction error*

Figure 5 shows an example of the spectrograms decoded by the convolutional and recurrent autoencoders. The input spectrogram from Figure 5(a) corresponds to a sample of the validation set of the training data, which is encoded and decoded with the convolutional and recurrent autoencoders, producing the spectrograms depicted in Figures 5(b) and 5(c), respectively. Note how the harmonic components related to the fundamental frequency are well reconstructed. Additionally, note that the autoencoder is removing most of the background noise of the speech utterance.

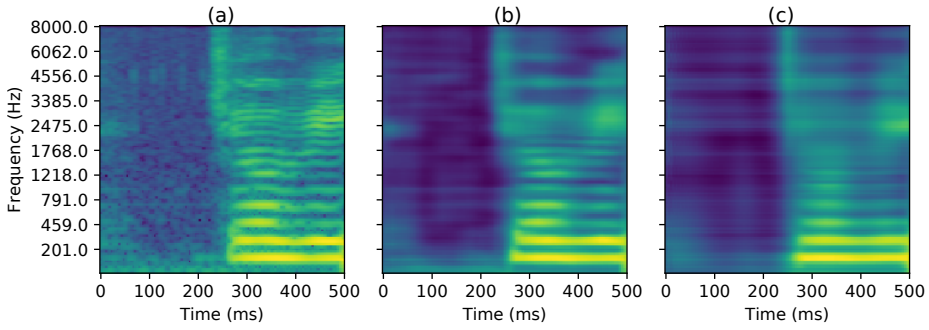


Figure 5: (a) Input spectrogram. (b) Decoded spectrogram with the CAE. (c) Decoded spectrogram with the RAE.

The difference in the reconstruction error of the CAE between the pathological and healthy speakers is observed in Figure 6. The average and the standard

deviation of the reconstruction error in Figure 6(b) is higher for CLP than for HC children. Conversely, for the case of PD patients and their corresponding HC subjects in Figure 6(a), the reconstruction error is higher for the HC subjects than for the PD patients, especially in the area inside the red square that corresponds to frequencies below 3000 Hz. The Wilcoxon signed-rank test was applied to evaluate whether the differences in the reconstruction error between healthy speakers and patients are significant. The results from the tests indicate that the differences are significant in both cases: PD vs. HC ($p \ll 0.005$, W=549.0) and CLP vs. HC ($p \ll 0.005$, W=182.0). Note also that for frequencies lower than 2 kHz the reconstruction error increases linearly with the frequency in both cases, PD (Pearson's correlation $r$=0.554), and CLP ($r$=0.765). This behavior is valid for the case of PD patients not only for the area highlighted with the read square but for the complete frequency range ($r$=0.885). The difference in the reconstruction error for different frequency regions is expected since most of the information from the speech signals is in the low part of the spectrum, and the bottleneck features assign higher weights to reconstruct that part of the spectrogram. Due to the same reason, the standard deviation of the reconstructed error is higher in the upper frequencies.

The reconstruction error for the RAE is observed in Figure 7 for the PD and CLP databases. For the case of PD in Figure 7(a), note that the error is higher for HC than for PD speakers, as it was observed for the CAE, especially in frequencies below 5 kHz. This effect could be explained because monotonicity and monoloudness, which make patients to produce speech with less variability than healthy people. These aspects make PD speech easier to be reconstructed than healthy speech when using both autoencoders. A contrary effect is observed for the CLP patients in Figure 7(b), where the error is higher for CLP patients than for HC subjects, as it was observed as well with the CAE. This behavior may be explained because CLP patients have problems to produce phonemes with more energy content in higher frequencies, like sibilants (Peterson-Falzone et al., 2016). Note also that the relation between the error and the frequency is linear for frequencies below 3 kHz, approximately. The Wilcoxon signed-rank

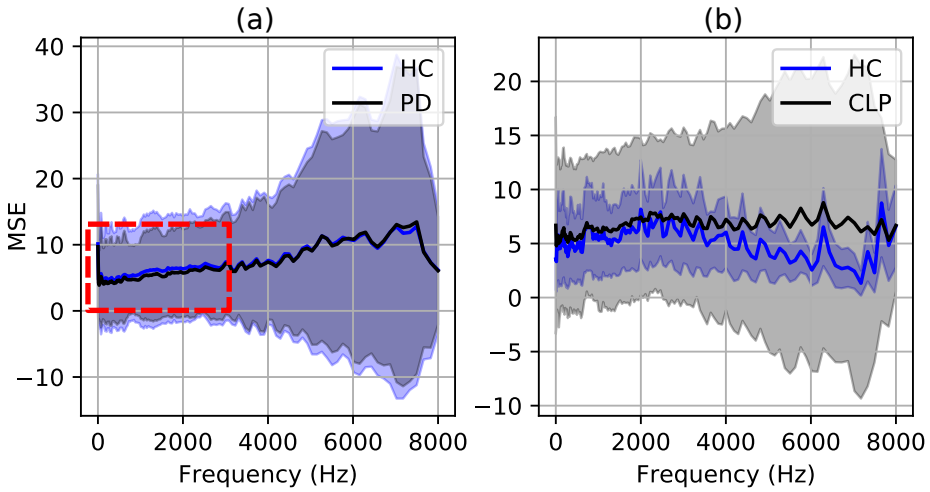


Figure 6: Average reconstruction error per frequency band of the convolutional autoencoder for the speakers of the **(a)** PD and **(b)** CLP databases.

test was also applied to evaluate whether the differences in the reconstruction error are significant. The results from the tests indicate that the differences are significant in both cases: PD vs. HC ($p \ll 0.005$, W=2.0) and CLP vs. HC ($p \ll 0.005$, W=767.0).

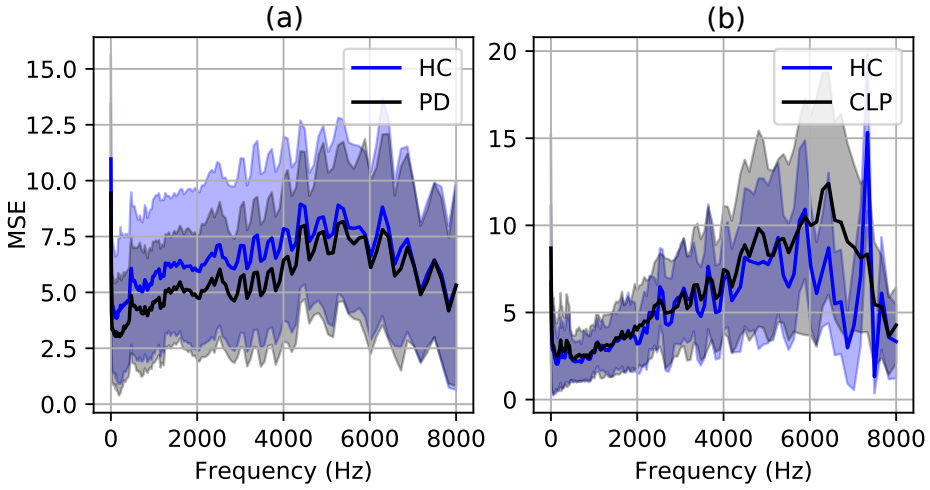


Figure 7: Average reconstruction error per frequency band of the recurrent autoencoder for the speakers of the **(a)** PD and **(b)** CLP databases.

### *4.2. Classification CLP vs HC*

Table 2: Area under the ROC curve obtained to classify CLP vs. HC children using the features extracted from the convolutional and recurrent autoencoders with different number of hidden nodes in the bottleneck layer using a DNN and an SVM classifiers.

| Feature set | | DNN | | | | | SVM | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 64 | 128 | 256 | 512 | 1024 | 64 | 128 | 256 | 512 | 1024 |
| CAE | bottleneck | 0.923 | 0.927 | 0.948 | 0.960 | 0.939 | 0.959 | 0.971 | 0.956 | 0.958 | 0.940 |
| CAE | error | 0.919 | 0.939 | 0.925 | 0.938 | 0.936 | 0.966 | 0.962 | 0.951 | 0.954 | 0.973 |
| CAE | bottleneck & error | 0.951 | 0.956 | 0.960 | 0.954 | 0.955 | 0.964 | 0.971 | 0.977 | 0.972 | 0.950 |
| RAE | bottleneck | 0.862 | 0.929 | 0.934 | 0.931 | 0.915 | 0.655 | 0.773 | 0.730 | 0.744 | 0.811 |
| RAE | error | 0.947 | 0.949 | 0.962 | 0.933 | 0.956 | 0.927 | 0.946 | 0.938 | 0.936 | 0.943 |
| RAE | bottleneck & error | 0.955 | 0.950 | 0.956 | 0.958 | 0.953 | 0.935 | 0.938 | 0.935 | 0.921 | 0.927 |
| CAE & RAE | bottleneck & error | 0.962 | 0.964 | 0.965 | 0.956 | 0.976 | 0.944 | 0.952 | 0.964 | 0.959 | 0.947 |
| | Avg. AUC | 0.931 | 0.945 | **0.950** | 0.947 | 0.947 | 0.907 | 0.930 | 0.921 | 0.921 | 0.927 |
| | Avg. diff. with max. of each row | 0.170 | 0.075 | **0.039** | 0.059 | 0.059 | 0.230 | 0.067 | 0.129 | 0.136 | 0.089 |

The results to discriminate between children with CLP and HC subjects using the convolutional and the recurrent autoencoders are shown in Table 2. The area under the ROC curves (AUC) is included. The features were combined using an early fusion strategy in all experiments. In general the results using SVM and DNN are comparable, except for the bottleneck RAE features, where the DNN is better. The highest AUC is 0.977, obtained with the combination of bottleneck and error features from the CAE, and with the SVM classifier. These results suggest that bottleneck and the proposed error-based features are complementary for the automatic classification of the disease. In addition, the combination of features from the convolutional and recurrent autoencoders improve the results with the DNN but not with the SVM classifier, where the best results were obtained only with features extracted from the CAE. This behavior can be explained because the DNN is more robust to classify the high dimensional feature vector obtained with the combination of the bottleneck and error features from both autoecoders, which is not case of the SVM. These results also suggest that features extracted from both autoencoders are complementary. Differences in the classification with different numbers of nodes in the bottleneck layer are not significant; however, the results with $h = 256$ are better because on average they provide the highest AUC with the DNN classifier (highlighted

in bold). In addition, the last row of the table includes the average difference of each result per column with the best AUC of each row. A lower average difference indicates a more stable result across all feature sets. The results with $h = 256$ also have the lowest difference.

Details of the results obtained with $h = 256$ nodes in the bottleneck layer are observed in Table 3, which includes values of accuracy, sensitivity, and specificity. The number of features and the associated AUC are also included. The feature set that produces the higest accuracy with the DNN and SVM classifiers are highlighted in bold. The highest accuracy (93.6%) is obtained with the combination of bottleneck and error features from the CAE, which confirms that those features are complementary for the addressed problem. In addition, note that for the DNN classifier, the highest accuracy is observed for the combination of all features from both autoencoders, while the combination of the features from the CAE achieved the highest accuracy with the SVM.

Table 3: Accuracy (ACC %), Sensitivity (SENS %), specificity (SPEC %), and area under the ROC curve for the results obtained with $h = 256$ nodes in the bottleneck layer of the autoencoders, classifying CLP vs. HC subjects.

| Feature set | | Num. features | ACC (%) | SENS (%) | SPEC (%) | AUC |
|---|---|---|---|---|---|---|
| | | | **DNN** | | | |
| CAE | bottleneck | 256 | 84.1 | 86.5 | 81.1 | 0.948 |
| CAE | error | 128 | 85.5 | 86.5 | 81.1 | 0.925 |
| CAE | bottleneck & error | 384 | 88.5 | 86.5 | 92.5 | 0.960 |
| RAE | bottleneck | 256 | 84.8 | 86.5 | 79.2 | 0.934 |
| RAE | error | 128 | 86.2 | 84.6 | 86.8 | 0.962 |
| RAE | bottleneck & error | 384 | 87.3 | 86.5 | 86.7 | 0.956 |
| **CAE & RAE** | **bottleneck & error** | **768** | **90.9** | 87.5 | 96.2 | 0.965 |
| | | | **SVM** | | | |
| CAE | bottleneck | 4x256 | 89.2 | 87.5 | 89.9 | 0.956 |
| CAE | error | 4x128 | 89.2 | 83.3 | 92.3 | 0.951 |
| **CAE** | **bottleneck & error** | **4x384** | **93.6** | 89.1 | 96.1 | 0.977 |
| RAE | bottleneck | 4x256 | 70.7 | 70.6 | 70.7 | 0.730 |
| RAE | error | 4x128 | 90.5 | 85.2 | 93.2 | 0.938 |
| RAE | bottleneck & error | 4x384 | 89.2 | 86.0 | 90.7 | 0.935 |
| CAE & RAE | bottleneck & error | 4x768 | 92.0 | 93.0 | 89.0 | 0.964 |

The result obtained with the best model, i.e., CAE with bottleneck and error features with the SVM classifier, is observed with more detail in Figure 8. The reduced false positive rate in the ROC curve of Figure 8(a) shows the capability

of the model to detect the target class i.e., CLP children. On the other hand, the histograms and the fitted probability density functions in Figure 8(b) show the scores assigned by the classifier to predict the corresponding class for each sample in the test set. Note that the equal error rate is slightly deviated to the left in the decision threshold.

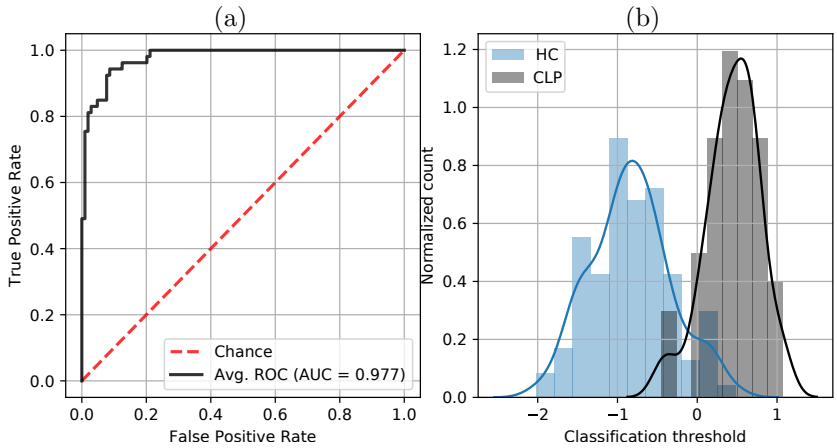


Figure 8: (a) ROC curve for the best result obtained classifying CLP vs. HC speakers. (b) Histograms and their corresponding density estimation for the scores obtained from the classifier.

An additional experiment was performed to evaluate which are the most accurate speech tasks for the addressed problem. We performed a separate classification of CLP vs. HC using each word included in the corpus with all the proposed features and the SVM classifier. The results are available in Table 4. The best result with each feature set is highlighted in bold to detail which are the most discriminant words in the classification problem. The results indicated that the most accurate words were those with sibilant fricative phonemes such as /CHUZO/ and /SUSI/. This result is in line with observations reported in the literature, where the excess in the nasal air emission of CLP patients, i.e., hypernasality, is associated with the abnormal production of fricative sounds, especially the sibilants (Kalita et al., 2019).

The results from Figure 9 show the Top-10 feature sets and speech exercises that produced the highest accuracies for the addressed problem. The top ac-

Table 4: AUCs obtained with the different feature sets from the autoencoders classifying CLP speakers with the different speech tasks available in the corpus. **Avg.** average

| Speech exercise | Feature set | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | CAE | | | RAE | | | Fusion | |
| | bottleneck | error | bottleneck & error | bottleneck | error | bottleneck & error | bottleneck & error | Avg. |
| **BOLA** | 0.938 | 0.970 | 0.960 | 0.533 | **0.963** | 0.939 | 0.947 | 0.893 |
| **CHUZO** | **0.974** | 0.976 | **0.971** | **0.681** | **0.963** | **0.956** | **0.971** | **0.927** |
| **COCO** | 0.936 | 0.935 | 0.945 | 0.569 | 0.877 | 0.856 | 0.926 | 0.863 |
| **GATO** | 0.813 | 0.892 | 0.875 | 0.641 | 0.881 | 0.877 | 0.848 | 0.832 |
| **JUGO** | 0.891 | **0.978** | 0.889 | 0.506 | 0.826 | 0.856 | 0.911 | 0.837 |
| **MANO** | 0.883 | 0.869 | 0.909 | 0.583 | 0.871 | 0.849 | 0.913 | 0.840 |
| **PAPA** | 0.850 | 0.903 | 0.900 | 0.625 | 0.891 | 0.890 | 0.927 | 0.855 |
| **SUSI** | 0.970 | 0.963 | **0.971** | 0.592 | 0.977 | 0.952 | 0.957 | 0.912 |

curacy is 97.0%, obtained with the fusion of error and bottleneck features from the CAE. Figure 9 confirmed that the words with the highest accuracies are those with sibilant fricatives such as /SUSI/ and /CHUZO/.

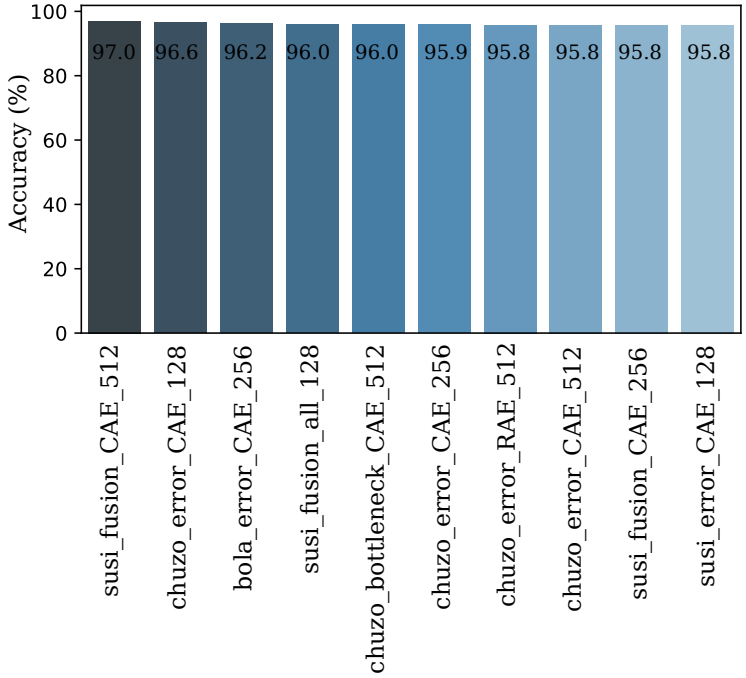


Figure 9: Top-10 accuracies obtained with the different feature sets and speech exercises classifying CLP vs HC subjects.

The results obtained in this paper are comparable with others reported in

previous studies where the same data were used, and which consider traditional hand-crafted features based on periodicity, noise, and spectral content (Orozco-Arroyave et al., 2015). However, it should be noted that the experiments addressed in (Orozco-Arroyave et al., 2015) are biased, since the hyper-parameters of the classifier were optimized according to the accuracy on the test set, which is not reliable and makes their results optimistic. The results from this paper are more realistic for the considered data. The results from this paper are also higher than the ones reported in (Carvajal-Castaño and Orozco-Arroyave, 2019), which consider also the same data, and used hand-crafted articulation features. The authors reported an accuracy up to 93.6% compared to the 97.0% obtained in this paper.

### *4.3. Classification PD vs. HC*

Table 5: Area under the ROC curve obtained to classify PD vs. HC subjects using the features extracted from the convolutional and recurrent autoencoders with different number of hidden nodes in the bottleneck layer using a DNN and an SVM classifiers.

| | | **DNN** | | | | | **SVM** | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 64 | 128 | 256 | 512 | 1024 | 64 | 128 | 256 | 512 | 1024 |
| CAE | bottleneck | 0.709 | 0.756 | 0.758 | 0.728 | 0.742 | 0.846 | 0.862 | 0.852 | 0.844 | 0.837 |
| CAE | error | 0.807 | 0.796 | 0.762 | 0.790 | 0.814 | 0.781 | 0.836 | 0.780 | 0.763 | 0.841 |
| CAE | Bottleneck & error | 0.801 | 0.795 | 0.808 | 0.760 | 0.815 | 0.868 | 0.870 | 0.857 | 0.870 | 0.839 |
| RAE | bottleneck | 0.717 | 0.749 | 0.753 | 0.752 | 0.708 | 0.796 | 0.745 | 0.691 | 0.806 | 0.767 |
| RAE | error | 0.790 | 0.753 | 0.760 | 0.784 | 0.832 | 0.877 | 0.866 | 0.852 | 0.871 | 0.874 |
| RAE | Bottleneck & error | 0.811 | 0.801 | 0.822 | 0.841 | 0.816 | 0.893 | 0.908 | 0.764 | 0.854 | 0.783 |
| CAE & RAE | Bottleneck & error | 0.833 | 0.795 | 0.793 | 0.822 | 0.791 | 0.878 | 0.892 | 0.850 | 0.870 | 0.828 |
| | Avg. AUC | 0.781 | 0.778 | 0.779 | 0.782 | 0.788 | 0.848 | **0.854** | 0.807 | 0.840 | 0.824 |
| Avg. diff. with max. of each row | | 0.178 | 0.201 | 0.190 | 0.169 | 0.128 | 0.117 | **0.077** | 0.410 | 0.178 | 0.287 |

The results classifying PD vs. HC subjects are shown in Table 5 using the DNN and SVM classifiers and the different feature sets extracted with the autoencoders. For the DNN classifier, note that the highest average AUC is obtained with 512 and 1024 nodes, and the best result is observed with the fusion of bottleneck and error features from the RAE with 512 nodes (AUC=0.841) and with the RAE error features with 1024 nodes (AUC=0.832). Conversely, for the SVM, the highest average AUC is obtained with a lower number of hidden nodes (128). The best results are also observed with the fusion of bottleneck

and error features from the RAE (AUC=0.908) and the fusion of all feature sets from both autoencoders (AUC=0.892). In general, the SVM is more accurate than the DNN, for this application.

More detailed results for the classification of PD and HC subjects are shown in Table 6. The feature set that produces the higest accuracy with the DNN and SVM classifiers are highlighted in bold. The best results are obtained with features derived from the RAE, for both classifiers. Particularly, the proposed reconstruction error features from the RAE are the most accurate when we consider the SVM classifier (accuracy=84%).

Table 6: Accuracy (ACC %), Sensitivity (SENS %), specificity (SPEC %), and area under the ROC curve for the results obtained with $h = 512$ nodes in the bottleneck layer of the autoencoders for the DNN classifier and with $h = 128$ nodes for the SVM classifier, , classifying PD vs. HC subjects.

| Feature set | | Num. features | ACC (%) | SENS (%) | SPEC (%) | AUC |
|---|---|---|---|---|---|---|
| | | | **DNN** | | | |
| CAE | bottleneck | 512 | 63 | 54 | 72 | 0.728 |
| CAE | error | 128 | 72 | 64 | 80 | 0.790 |
| CAE | bottleneck & error | 640 | 69 | 68 | 70 | 0.760 |
| RAE | bottleneck | 512 | 67 | 76 | 58 | 0.752 |
| RAE | error | 128 | 74 | 58 | 90 | 0.784 |
| **RAE** | **bottleneck & error** | **640** | **76** | 70 | 82 | 0.841 |
| CAE & RAE | bottleneck & error | 1280 | 76 | 72 | 80 | 0.822 |
| | | | **SVM** | | | |
| CAE | bottleneck | 128×4 | 79 | 80 | 78 | 0.862 |
| CAE | error | 128×4 | 72 | 71 | 73 | 0.836 |
| CAE | bottleneck & error | 256×4 | 78 | 80 | 76 | 0.870 |
| RAE | bottleneck | 128×4 | 64 | 68 | 62 | 0.745 |
| **RAE** | **error** | **128×4** | **84** | 85 | 83 | 0.866 |
| RAE | bottleneck & error | 256×4 | 81 | 81 | 81 | 0.908 |
| CAE & RAE | bottleneck & error | 512×4 | 77 | 80 | 75 | 0.892 |

Details of the most accurate model, i.e., RAE error with the SVM classifier, are observed in Figure 10. The ROC curve in Figure 10(a) shows the capability of the model to detect the target class i.e., PD speakers. On the other hand, the histograms and the fitted probability density functions in Figure 10(b) show the scores assigned by the classifier to predict each sample of the database. Although there are other feature sets with higher AUC than the one displayed in Figure 10, e.g., the combination of bottleneck and error features from the RAE

(AUC=0.908), we selected only the RAE error features because their stability in the prediction of the class per speaker. The ROC curve and histograms of the predictions for the fusion of features from the RAE are shown in Figure 11. Note that in this case, there are more errors spread over the decision space in Figure 11(b), e.g, the black bar at the left part of the figure.

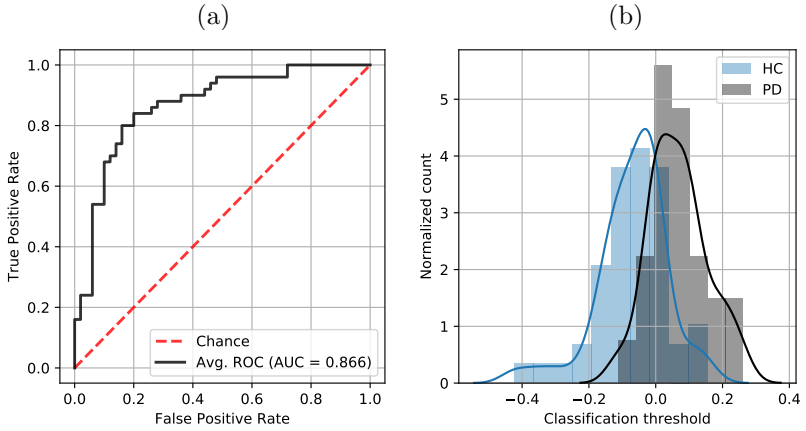


Figure 10: (a) ROC curve for the error-based features obtained from the RAE classifying PD vs. HC speakers. (b) Histograms and their corresponding density estimation for the scores obtained from the classifier.

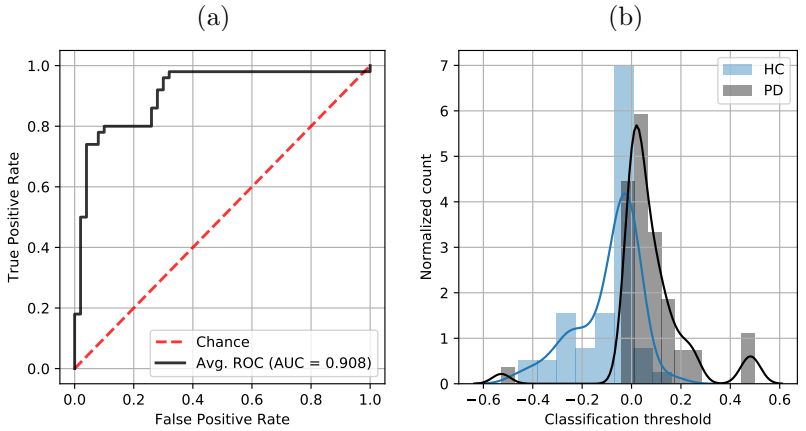


Figure 11: (a) ROC curve for the fusion of bottleneck and error features from the RAE classifying PD vs. HC speakers. (b) Histograms and their corresponding density estimation for the scores obtained from the classifier.

We performed also classification experiments with the separate speech tasks pronounced by the patients. The results are available in Table 7. The best result with each feature set is highlighted in bold to detail which are the most discriminant speech tasks in this classification problem. We observed that the most accurate exercises correspond to the read text and the monologue, which is expected because the data used to train the autoencoders correspond also to continuous speech utterances. On the other hand, the most discriminant DDK exercises are the rapid repetition of PA-TA-KA (average AUC=0.699) and PA-KA-TA (average AUC=0.694), which agrees with previous studies about the sequence of bilabial, alveolar, and velar stops has more discriminating power than other DDK sequences (Rueda et al., 2019).

Table 7: AUCs obtained with the different feature sets from the autoencoders classifying PD speakers with the different speech tasks available in the corpus. **Avg.** average

| **Speech exercise** | **Feature set** | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | **CAE** | | | **RAE** | | | **Fusion** | |
| | **bottleneck** | **error** | **bottleneck & error** | **bottleneck** | **error** | **bottleneck & error** | **bottleneck & error** | **Avg.** |
| **PA-TA-KA** | 0.719 | 0.744 | 0.754 | **0.601** | 0.588 | 0.764 | 0.726 | 0.699 |
| **PA-KA-TA** | 0.749 | 0.701 | 0.775 | 0.536 | 0.597 | 0.758 | 0.741 | 0.694 |
| **PE-TA-KA** | 0.591 | 0.657 | 0.706 | 0.599 | 0.520 | **0.789** | 0.760 | 0.657 |
| **PA** | 0.762 | 0.620 | 0.729 | 0.502 | 0.596 | 0.656 | 0.742 | 0.658 |
| **TA** | 0.668 | 0.619 | 0.708 | 0.518 | **0.668** | 0.761 | 0.773 | 0.674 |
| **KA** | 0.841 | 0.627 | 0.792 | 0.599 | 0.505 | 0.534 | 0.748 | 0.664 |
| **read sentences** | 0.746 | 0.671 | 0.749 | 0.553 | 0.612 | 0.646 | 0.715 | 0.670 |
| **read text** | 0.859 | **0.755** | **0.836** | 0.588 | 0.638 | 0.720 | **0.842** | **0.748** |
| **monologue** | **0.874** | 0.639 | **0.836** | 0.569 | 0.599 | 0.755 | 0.818 | 0.727 |

The results from Figure 12 show the Top-10 feature sets and speech tasks that produced the highest accuracies for the classification of PD vs. HC. The top accuracy is 84.0%, obtained with the fusion of all speech tasks and features from the autoencoders, and the error based features from the RAE. Note that in this case, the best result is almost always obtained with the combination of all speech tasks rather than with individual tasks.

The results obtained with our proposed method are better than others reported in the literature when the same data were used, and which considered different sets of hand-crafted features. For instance, features based on periodicity, noise, and spectral content (Orozco-Arroyave et al., 2015), features based

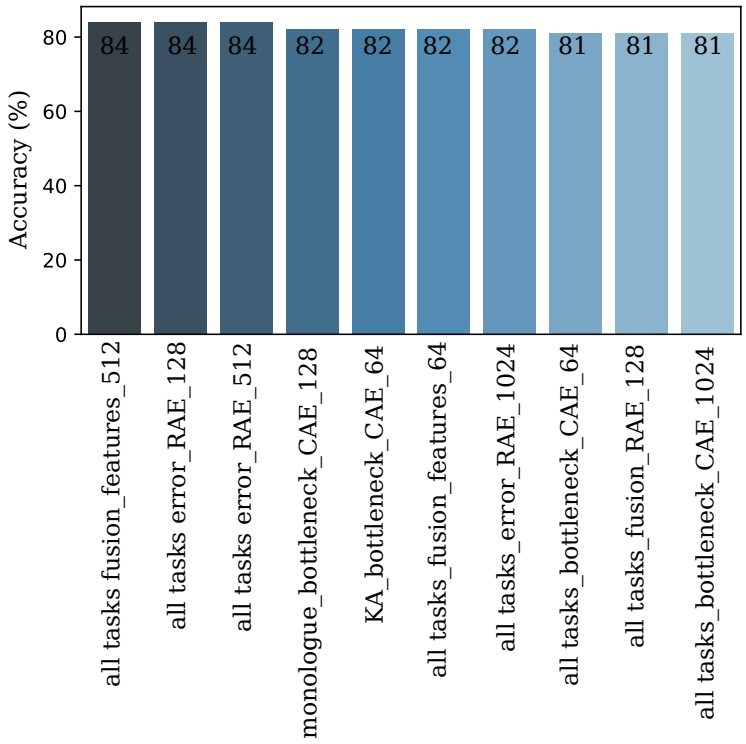


Figure 12: Top-10 accuracies obtained with different feature sets and speech tasks classifying PD vs HC subjects.

on articulation (Vásquez-Correa et al., 2017), features based on Gaussian mixture models representations (Moro-Velazquez et al., 2019), or features based on non-linear dynamics (Godino-Llorente et al., 2017), empirical mode decomposition (Rueda et al., 2019), among others.

## 5. Conclusion

This paper introduced a parallel representation learning scheme to model pathological speech signals with morphological and neurological origins. Two types of autoencoders were implemented to compute low-dimensional feature representations of the speech frames from each disease: (1) a CAE to learn a representation of the spatial distribution of the energy content in a spectrogram, and (2) a RAE to model the temporal evolution of the spectral components

of a speech frame. Both autoencoders were trained with a corpus of healthy speech in order to have a robust model that could serve as a reference for a posterior extraction of features related with the presence of speech disorders. We computed two different feature sets extracted from the trained autoencoders: (1) features from the hidden representation in the bottleneck space, and (2) a proposed feature set based on the reconstruction error of the autoencoder in different spectral components of the speech signal. The extracted features were used to classify speech utterances from PD and CLP patients vs. their corresponding HC subjects using two classification strategies: DNN and SVM.

The reconstruction error from the autoencoders contains information about paralinguistic aspects of the speakers such as the presence of speech disorders. Higher reconstruction errors were observed for CLP patients than for healthy subjects, especially in frequencies above 2 kHz, where usually sibilant fricatives appear in the speech. This is an important aspect because those sounds are typically associated with the impairments suffered by CLP patients due to the excess of air coming out through the nasal cavity. Conversely, the reconstruction error for the PD patients is lower for patients than for healthy speakers, in particular for frequencies below 2.5 kHz, approximately. This aspect can be likely explained due to the monotonicity in the speech of PD patients, which makes them to produce slower speech with less variability.

The proposed models were accurate to model speech signals from patients of both diseases: PD and CLP. Accuracies of up to 97.0% were obtained in the classification of CLP vs. HC speakers. The accuracies observed classifying PD and HC speakers were up to 84.0%. In addition, the information obtained from the error-based features was complementary to the information extracted from the bottleneck features to classify pathological speech signals. The features extracted from the RAE were more accurate than those obtained from the CAE in both scenarios: PD vs. HC and CLP vs. HC. In general, classifying PD vs. HC speakers is more challenging than the problem of discriminating between CLP vs. HC children. The main reason is because elderly healthy speakers in general, also exhibited articulation and phonation problems due to the normal

aging process (Arias-Vergara et al., 2017). . Additionally, elderly people could also be or have been smokers or drinkers.

Finally, the results obtained in this study are comparable or better than those observed in the literature when the same data are considered, both for classification of CLP (Orozco-Arroyave et al., 2015, 2013, 2016b; Carvajal-Castaño and Orozco-Arroyave, 2019), and for the classification of PD (Orozco-Arroyave, 2016; Vásquez-Correa et al., 2017; Godino-Llorente et al., 2017; Moro-Velazquez et al., 2019; Rueda et al., 2019).These comparisons show that our method based on unsupervised feature learning produces state-of-the-art results for the addressed datasets.

Further experiments with the proposed approach will include the prediction of the neurological state and the dysarthria level of the PD patients, and the level of nasality in CLP children. The models will also be considered to evaluate speech disorders with different origin than the ones addressed here. For instance speech disorders with laryngeal origin like those developed by patients with larynx cancer, or speech impairments with perceptual origin like the exhibited by cochlear implants users. The aim is to test whether the proposed methods are general to model pathological speech. In addition, the methods are going to be evaluated with speech corpora in different languages to test their reliability and generalization, in a similar way to the experiments addressed in (Vásquez-Correa et al., 2019).

**Acknowledgment**

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie Grant Agreement No. 766287. T. Arias-Vergara is also under grants of Convocatoria Doctorado Nacional-785 financed by COLCIENCIAS. The work reported here was also financed by CODI from University of Antioquia by grant Number 2017–15530.

## References


Arias-Vergara, T., Vásquez-Correa, J.C., Orozco-Arroyave, J.R., 2017. Parkinson's disease and aging: analysis of their effect in phonation and articulation of speech. Cognitive Computation 9, 731–748.

Baker, K.K., Ramig, L.O., Luschei, E.S., Smith, M.E., 1998. Thyroarytenoid muscle activity associated with hypophonia in parkinson disease and aging. Neurology 51, 1592–1598.

Barnish, M.S., Whibley, D., et al., 2016. Roles of cognitive status and intelligibility in everyday communication in people with Parkinson's disease: A systematic review. Journal of Parkinson's disease 6, 453–462.

Berus, L., Klancnik, S., Brezocnik, M., Ficko, M., 2019. Classifying parkinson's disease based on acoustic measures using artificial neural networks. Sensors 19, 16.

Blanchet, P.G., Snyder, G.J., 2009. Speech rate deficits in individuals with parkinson's disease: a review of the literature. Journal of Medical Speech-Language Pathology 17, 1–7.

Bocklet, T., Steidl, S., Nöth, E., Skodda, S., 2013. Automatic Evaluation of Parkinson's Speech – Acoustic, Prosodic and Voice Related Cues, in: Proceedings of INTERSPEECH, pp. 1149–1153.

Carvajal-Castaño, H.A., Orozco-Arroyave, J.R., 2019. Articulation analysis in the speech of children with cleft lip and palate, in: Iberoamerican Congress on Pattern Recognition, Springer. pp. 575–585.

Cummins, N., Baird, A., Schuller, B., 2018. Speech analysis for health: Current state-of-the-art and the increasing impact of deep learning. Methods .

Dimauro, G., Di-Nicola, V., et al., 2017. Assessment of speech intelligibility in parkinson's disease using a speech-to-text system. IEEE Access 5, 22199–22208.

Dubey, A.K., Mahadeva-Prasanna, S.R., Dandapat, S., 2019. Hypernasality Severity Detection Using Constant Q Cepstral Coefficients, in: Proceedings of INTERSPEECH, pp. 4554–4558.

Galaz, Z., Mekyska, J., et al., 2016. Prosodic analysis of neutral, stress-modified and rhymed speech in patients with parkinson's disease. Computer Methods and Programs in Biomedicine 127, 301–317.

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

Goetz, C., et al., 2008. Movement Disorder Society-sponsored revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS): Scale presentation and clinimetric testing results. Movement Disorders 23, 2129–2170.

Golabbakhsh, M., Abnavi, F., Kadkhodaei-Elyaderani, M., et al., 2017. Automatic identification of hypernasality in normal and cleft lip and palate patients with acoustic analysis of speech. The Journal of the Acoustical Society of America 141, 929–935.

Gómez-Vilda, P., Mekyska, J., et al., 2017. Parkinson disease detection from speech articulation neuromechanics. Frontiers in neuroinformatics 11, 56.

Grósz, T., Busa-Fekete, R., et al., 2015. Assessing the degree of nativeness and Parkinson's condition using Gaussian processes and deep rectifier neural networks, in: Proceedings of INTERSPEECH, pp. 1339–1343.

Hernández-Mena, C.D., Herrera-Camacho, J., 2014. CIEMPIESS: A new open-sourced mexican spanish radio corpus, in: Proceedings of the ninth international conference on language resources and evaluation (LREC'14), European Language Resources Association (ELRA) Reykjavik, Iceland. pp. 371–375.

Ho, A.K., Iansek, R., Marigliani, C., Bradshaw, J.L., Gates, S., 1999. Speech impairment in a large sample of patients with parkinson's disease. Behavioural neurology 11, 131–137.

Hornykiewicz, O., 1998. Biochemical aspects of Parkinson's disease. Neurology 51, S2–S9.

Kalita, S., Sudro, P.N., Prasanna, S.M., Dandapat, S., 2019. Nasal Air Emission in Sibilant Fricatives of Cleft Lip and Palate Speech, in: Proceedings of INTERSPEECH, pp. 4544–4548.

Kingma, D.P., Ba, J., 2014. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 .

Maier, A., Hönig, F., Bocklet, T., Nöth, E., Stelzle, F., Nkenke, E., Schuster, M., 2009. Automatic detection of articulation disorders in children with cleft lip and palate. The Journal of the Acoustical Society of America 126, 2589–2602.

Montaña, D., Campos-Roca, Y., Pérez, C.J., 2018. A diadochokinesis-based expert system considering articulatory features of plosive consonants for early detection of parkinson's disease. Computer Methods and Programs in Biomedicine 154, 89–97.

Moretti, R., et al., 2003. Speech initiation hesitation'following subthalamic nucleus stimulation in a patient with parkinson's disease. European Neurology 49, 251–253.

Moro-Velázquez, L., et al., 2018. Analysis of speaker recognition methodologies and the influence of kinetic changes to automatically detect parkinson's disease. Applied Soft Computing 62, 649–666.

Moro-Velazquez, L., et al., 2019. A forced gaussians based methodology for the differential evaluation of parkinson's disease by means of speech processing. Biomedical Signal Processing and Control 48, 205–220.

Naranjo, L., Pérez, C.J., Campos-Roca, Y., Martín, J., 2016. Addressing voice recording replications for parkinson's disease detection. Expert Systems with Applications 46, 286–292.

Novotný, M., Dusek, P., Daly, I., Ruzicka, E., Rusz, J., 2020. Glottal source analysis of voice deficits in newly diagnosed drug-naïve patients with parkinson's disease: Correlation between acoustic speech characteristics and non-speech motor performance. Biomedical Signal Processing and Control 57, 101818.

Novotnỳ, M., Rusz, J., et al., 2014. Automatic evaluation of articulatory disorders in Parkinson's disease. IEEE/ACM Trans. on Audio, Speech and Language Processing 22, 1366–1378.

Orozco-Arroyave, J.R., 2016. Analysis of speech of people with Parkinson's disease. 1st ed., Logos-Verlag, Berlin, Germany.

Orozco-Arroyave, J.R., Vargas-Bonilla, J.F., Arias-Londoño, J.D., Murillo-Rendón, S., Castellanos-Domínguez, G., Garcés, J.F., 2013. Nonlinear dynamics for hypernasality detection in spanish vowels and words. Cognitive Computation 5, 448–457.

Orozco-Arroyave, J.R., Vásquez-Correa, J.C., et al., 2016a. Towards an automatic monitoring of the neurological state of the Parkinson's patients from speech, in: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6490–6494.

Orozco-Arroyave, J.R., Vásquez-Correa, J.C., et al., 2018. Neurospeech: An open-source software for Parkinson's speech analysis. Digital Signal Processing 77, 207–221.

Orozco-Arroyave, J.R., et al., 2014. New Spanish speech corpus database for the analysis of people suffering from Parkinson's disease, in: Language Resources and Evaluation Conference, (LREC), pp. 342–347.

Orozco-Arroyave, J.R., et al., 2015. Characterization methods for the detection of multiple voice disorders: Neurological, functional, and laryngeal diseases. IEEE Journal of Biomedical and Health Informatics 19, 1820–1828.

Orozco-Arroyave, J.R., et al., 2016b. 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 Ingeniería Universidad de Antioquia 80, 109–123.

Paszke, A., Gross, S., Chintala, S., et al., 2017. Automatic differentiation in pytorch, in: Conference on Neural Information Processing Systems (NIPS), pp. 1–4.

Peterson-Falzone, S.J., Trost-Cardamone, J., Karnell, M.P., Hardin-Jones, M.A., 2016. The clinician's guide to treating cleft palate speech. Elsevier Health Sciences.

Prathanee, B., Seepuaham, C., Pumnum, T., 2014. Articulation disorders and patterns in children with a cleft. Asian Biomedicine 8, 699–706.

Rueda, A., Vásquez-Correa, J.C., Rios-Urrego, C.D., Orozco-Arroyave, J.R., Krishnan, S., Nöth, E., 2019. Feature representation of pathophysiology of parkinsonian dysarthria, in: Proceedings of INTERSPEECH, pp. 3048–3052.

Rusz, J., Cmejla, R., et al., 2013. Imprecise vowel articulation as a potential early marker of parkinson's disease: Effect of speaking task. The Journal of the Acoustical Society of America 134, 2171–2181.

Rusz, J., et al., 2017. Comparative analysis of speech impairment and upper limb motor dysfunction in parkinson's disease. Journal of Neural Transmission 124, 463–470.

Sakar, B.E., et al., 2013. Collection and analysis of a parkinson speech dataset with multiple types of sound recordings. IEEE Journal of Biomedical and Health Informatics 17, 828–834.

Saxon, M., Tripathi, A., Jiao, Y., Liss, J., Berisha, V., 2019. Robust estimation of hypernasality in dysarthria. arXiv preprint arXiv:1911.11360 .

Scholkopf, B., Smola, A.J., 2001. Learning with kernels: support vector machines, regularization, optimization, and beyond. MIT press.

Schuller, B., Batliner, A., 2013. Computational paralinguistics: emotion, affect and personality in speech and language processing. John Wiley & Sons.

Schuller, B., Steidl, S., Batliner, A., Hantke, S., Hönig, F., Orozco-Arroyave, J.R., Nöth, E., Zhang, Y., Weninger, F., 2015. The INTERSPEECH 2015 computational paralinguistics challenge: Nativeness, Parkinson's & eating condition, in: Proceedings of INTERSPEECH, pp. 478–482.

Schuller, B., et al., 2019. Affective and behavioural computing: Lessons learnt from the first computational paralinguistics challenge. Computer Speech & Language 53, 156–180.

Schuster, M., Maier, A., Haderlein, T., Nkenke, E., Wohlleben, U., Rosanowski, F., Eysholdt, U., Nöth, E., 2006. Evaluation of speech intelligibility for children with cleft lip and palate by means of automatic speech recognition. International Journal of Pediatric Otorhinolaryngology 70, 1741–1747.

Skodda, S., Grönheit, W., Schlegel, U., 2011. Gender-related patterns of dysprosody in Parkinson disease and correlation between speech variables and motor symptoms. Journal of Voice 25, 76–82.

Spencer, K.A., Rogers, M.A., 2005. Speech motor programming in hypokinetic and ataxic dysarthria. Brain and Language 94, 347–366.

Tanaka, Y., Nishio, M., Niimi, S., 2011. Vocal acoustic characteristics of patients with parkinson's disease. Folia Phoniatrica et logopaedica 63, 223–230.

Tu, M., Berisha, V., Liss, J., 2017. Interpretable objective assessment of dysarthric speech based on deep neural networks, in: Proceedings of INTERSPEECH, pp. 1849–1853.

Tykalova, T., Rusz, J., Klempir, J., Cmejla, R., Ruzicka, E., 2017. Distinct patterns of imprecise consonant articulation among parkinson's disease, progressive supranuclear palsy and multiple system atrophy. Brain and language 165, 1–9.

Vásquez-Correa, J.C., Arias-Vergara, T., Rios-Urrego, C.D., Schuster, M., Rusz, J., Orozco-Arroyave, J.R., Nöth, E., 2019. Convolutional neural networks and a transfer learning strategy to classify parkinson's disease from speech in three different languages, in: Iberoamerican Congress on Pattern Recognition, pp. 697–706.

Vásquez-Correa, J.C., Orozco-Arroyave, J.R., Nöth, E., 2017. Convolutional neural network to model articulation impairments in patients with Parkinson's disease, in: Proceedings of INTERSPEECH, pp. 314–318.

Vásquez-Correa, J.C., Serra, J., Orozco-Arroyave, J.R., Vargas-Bonilla, J.F., Nöth, E., 2017. Effect of acoustic conditions on algorithms to detect parkinson's disease from speech, in: Proceedings of ICASSP, IEEE. pp. 5065–5069.

Vijayalakshmi, P., Reddy, M.R., O'Shaughnessy, D., 2007. Acoustic analysis and detection of hypernasality using a group delay function. IEEE Transactions on Biomedical engineering 54, 621–629.

Vikram, C.M., Tripathi, A., Kalita, S., Mahadeva-Prasanna, S.R., 2018. Estimation of hypernasality scores from cleft lip and palate speech. Proceedings of INTERSPEECH , 1701–1705.

Viswanathan, R., Arjunan, S.P., et al., 2020. Complexity measures of voice recordings as a discriminative tool for parkinson's disease. Biosensors 10, 1.

Wyatt, R., Sell, D., Russell, J., Harding, A., Harland, K., Albery, L., 1996. Cleft palate speech dissected: a review of current knowledge and analysis. British Journal of Plastic Surgery 49, 143–149.

Zhang, H., Song, C., Wang, A., Xu, C., Li, D., Xu, W., 2019. Pdvocal: Towards privacy-preserving parkinson's disease detection using non-speech body sounds, in: Proceedings of Mobicom, pp. 1,17.

Zhang, Y.N., 2017. Can a smartphone diagnose parkinson disease? a deep neural network method and telediagnosis system implementation. Parkinson's Disease 2017, 1–11.