Alejandro Guerrero-López[1], Julián D. Arias-Londoño[1], Stefanie Shattuck-Hufnagel[2], and Juan I. Godino-Llorente[1]

[1]Universidad Politécnica de Madrid
[2]Massachusetts Institute of Technology

March 19, 2024



**Abstract**

Parkinson's disease significantly impacts speech, particularly affecting phonemic groups like stop-plosives, fricatives, and affricates. However, its objective impact on the different phonemic groups has been briefly addressed in the past.

This study introduces a new model, called MARTA, built upon a Gaussian Mixture Variational AutoEncoder with metric learning to measure the disease's impact on the phonemic grouping automatically and objectively. MARTA was trained on normophonic speech before adapting it to parkinsonian speech. The model effectively clusters phonemic groups unsupervised and demonstrates enhanced discriminative power when supervised using forced-aligned labels. Our findings reveal that beyond the traditionally affected phonemes, Parkinson's disease not only affects stop-plosives, voiced-plosives, and nasals, but also significantly impacts liquids, vowels, and fricatives, with the model achieving a benchmarking $91\% \pm 9$ discrimination capability. An in-depth evaluation of the impact of the disease on the different phonemic groups represents an advance in the current knowledge of its effects on the speech, and has clear implications in the speech therapy of people with Parkinson's disease.

Moreover, regardless of the specific application domain presented, the model introduced has potential downstream utility in assessing the manner of articulation, whether influenced by other medical conditions or certain dialectal variations.

# MARTA: a model for the automatic phonemic grouping of the parkinsonian speech

Alejandro Guerrero-López , Julián D. Arias-Londoño , *Senior Member, IEEE*, Stefanie Shattuck-Hufnagel , and Juan I. Godino-Llorente , *Senior Member, IEEE*

**Abstract**—Parkinson's disease significantly impacts speech, particularly affecting phonemic groups like stop-plosives, fricatives, and affricates. However, its objective impact on the different phonemic groups has been briefly addressed in the past.
This study introduces a new model, called MARTA, built upon a Gaussian Mixture Variational AutoEncoder with metric learning to measure the disease's impact on the phonemic grouping automatically and objectively. MARTA was trained on normophonic speech before adapting it to parkinsonian speech. The model effectively clusters phonemic groups unsupervised and demonstrates enhanced discriminative power when supervised using forced-aligned labels. Our findings reveal that beyond the traditionally affected phonemes, Parkinson's disease not only affects stop-plosives, voiced-plosives, and nasals, but also significantly impacts liquids, vowels, and fricatives, with the model achieving a benchmarking $91\% \pm 9$ discrimination capability. An in-depth evaluation of the impact of the disease on the different phonemic groups represents an advance in the current knowledge of its effects on the speech, and has clear implications in the speech therapy of people with Parkinson's disease.
Moreover, regardless of the specific application domain presented, the model introduced has potential downstream utility in assessing the manner of articulation, whether influenced by other medical conditions or certain dialectal variations.

**Index Terms**—Speech, Gaussian Mixture Variational Autoencoder, Parkinson's Disease, Phonemic Grouping, Manner of articulation, Manner classes, Unsupervised clustering, Supervised clustering, Downstream, Parkinson's discrimination

## 1 Introduction

Parkinson's disease is a chronic condition resulting from the gradual death of brain cells, particularly those in the substantia nigra responsible for dopamine production [1]. The decrease in dopamine levels in Parkinson's Disease (PD) patients leads to noticeable motor symptoms such as lack of coordination, muscle rigidity, and slowed movements [2].

Diagnostic criteria for PD are mainly based on the observation of motor signs and non-motor indicators, with neuropathological diagnosis during autopsy considered the gold standard [3]. Studies suggest that standard clinical criteria can produce 90% precision in diagnosis in an average of 2.9 years [4]. Yet, recent evaluations have underscored the complexities associated with PD diagnosis, highlighting the need for innovative approaches, including genetic analysis, machine learning techniques, and biomarker identification, to improve diagnostic precision and accelerate the process [5]. Consequently, the development of novel diagnostic tools is essential to improve early detection and intervention strategies [6].

In this context, the automatic analysis of the speech emerges as an efficient and timely alternative for the diagnosis and evaluation of PD [7]. Since the speech production requires precise and complex movements, it is sensitive to the early effects of neurodegenerative processes associated with PD, resulting in dysphonia, dysarthria, and disprosody [8]. Empirical evidence supports these findings, reporting articulatory deficits in PD patients, manifested as lower precision, amplitude, velocity, and variability in the opening of the tongue, jaw, and lower lip during articulation, leading to imprecise consonant or vowel production [9]–[11].

Acoustic analyses have identified several key indicators of PD in speech patterns. The variation in voice onset time for voiced and voiceless stops is one of such markers, suggesting the potential presence of PD [12]. Spirantisation is also a characteristic effect present in the speech of patients with PD [13], [14]. Moreover, certain studies suggest differences in the slopes and variability of the formant frequencies between patients and controls [15], and also in the vowel space area of both groups [16]. There is also evidence of significant alterations in speech rate [17]. A comprehensive overview of these and other articulatory deficits in PD is provided in [18].

Previous research has examined the perceptual impact of various phonemes in different languages on the detection of PD. Distinct phonemic groups are affected differently by PD, with notable impacts on stop-plosives, fricatives, and affricates. This claim aligns with previous studies that conducted perceptual analyses of Parkinsonian speech, highlighting these specific phonemic irregularities [19]–[21].

*This work was supported by the Ministry of Economy and Competitiveness of Spain under Grants PID2021-128469OB-I00 and TED2021-131688B-I00, and by Comunidad de Madrid, Spain. This is also supported by a grant of the MISTI MIT Global Experiences. Universidad Politécnica de Madrid also supports Julián D. Arias-Londoño through a María Zambrano 2021 grant. Finally, the authors would like to thank the Madrid ELLIS unit (European Laboratory for Learning & Intelligent Systems) for its indirect support.*
*Alejandro Guerrero-López, Julián D. Arias-Londoño, and Juan I. Godino-Llorente are with Escuela Técnica Superior de Ingenieros (ETSI) de Telecomunicación, Universidad Politécnica de Madrid, 28040, Madrid, Spain (e-mail: alejandro.guerrero@upm.es; julian.arias@upm.es, ignacio.godino@upm.es).*
*Stefanie Shattuck-Hufnagel is with the Speech Communication Group, Research Laboratory of Electronics. Massachusetts Institute of Technology, Cambridge, MA 02139, USA (e-mail: sshuf@mit.edu).*
*Manuscript submitted February xx, 2024*

Recent research has further advanced our understanding of PD detection through speech analysis. The study in [22] highlights the potential of fricative sounds —parameterised by duration, intensity, and/or spectral moments— to assess co-articulation capabilities and to evaluate the patient's ability to perform complex movements that are impaired in PD. This finding is also supported by [23], suggesting that fricatives are more discriminative than vowels for the detection of PD. In addition, the study in [24] explores the use of occlusive consonants for the detection of PD, achieving a classification accuracy of 94.4% with the plosive /k/, being particularly discriminative in a dataset of Spanish speakers. On the same line, [14] reports that voiced consonants and consonants in the word medial position are prone to distortion in PD, also concluding that these distortions contribute strongly to the perceived intelligibility of the PD group. The importance of individual fricatives and stop-plosives is further confirmed in additional cross-language experiments [25], using Spanish and Czech speakers, leading to accuracies of 79% and 94%, respectively. Subsequent work [26] introduces features extracted from relevant articulation moments, highlighting the importance of studying transitions between specific phonemes to evaluate the speech of patients with PD.

Automatic systems are also proposed for the detection of PD using articulatory and phonatory features extracted with signal processing techniques, reporting accuracies —in the most robust cases, methodologically speaking— below 90% [27], [28]. These systems aim for a binary categorisation of PD vs. controls, with no detailed analysis of the discrimination capabilities of the distinct Manner Classes (MC) of phonemes (i.e., categories with the same manner of articulation), a crucial aspect for future system developments, to adequately select the speech tasks to be employed, and to design appropriate speech rehabilitation techniques.

To our knowledge, only the preliminary study in [29] has addressed an automatic analysis of different phonemic groups in the context of PD speech. The results suggest that the plosive segments tend to provide a better accuracy for the detection of PD, followed by vowels and fricative segments. Nevertheless, despite its findings, the methodology outlined in [29] does not follow current methodological trends in the design of AI systems, so it has limitations in codifying the phonemic information in a way that can be used for downstream tasks associated with treatment assessment or integrated into current proposals of multimodal systems for PD detection and evaluation [30].

In a broader context, not dedicated to the analysis of the speech of PD patients, a recent work [31] proposed a CLIP-like [32] model architecture called SCRAPS. This model codifies phonetic and acoustic information into a unified latent space. SCRAPS employs distinct encoders for phonetics and acoustics, minimising the distance between these modalities to establish a cohesive phonetic-audio representation. This approach provides flexibility, allowing potential downstream applications to be built upon it and to develop other models/solutions across a spectrum of speech and audio tasks, such as speech recognition and mispronunciation detection. However, SCRAPS focusses mainly on a generic phonetic and acoustic analysis of normophonic speakers without ensuring the discriminative power of each phonemic group or MC, hindering the applicability to biomedical contexts where characterising the effects of a certain pathology on each phonemic group is crucial for the detection and assessment of the disease. SCRAPS allows the mapping of a particular utterance to a latent space aligned with phonetic characteristics, but the information about to what extent a particular manner class deviates from its theoretical (or canonical) phonemic structure is lost. This limitation hinders its direct applicability to PD speech research, where the analysis of MC could provide critical insights.

In this context, this study introduces Manner of ARTiculation Analysis (MARTA), an innovative tool powered by a Gaussian Mixture Variational AutoEncoder (GMVAE) [33], which seeks to bridge these gaps by focussing on a class-driven approach to map the latent acoustic space, without the need for a separate encoder for each MC. Consequently, the study aims to pinpoint the speech segments holding greater relevance for the automatic detection of PD. The underlying hypothesis posits that the relevance of acoustic segments varies, as each one stems from distinct vocal tract narrowing, configuration, and articulation. By focussing on MC, MARTA aims not only to improve discrimination between normophonic and parkinosnian speech, but also to create a latent space that could support a variety of speech-related applications. This approach extends beyond the scope of SCRAPS, offering specificity for PD research through a MC analysis, and going beyond the limitations of [29] by providing a versatile framework for further exploration to develop downstream applications in speech processing, multimodal models, speech therapy, and dialectal evaluation of the speech.

MARTA is initially trained using the Albayzin dataset, which is a phonetically balanced dataset of Castillian Spanish normophonic (i.e., Healthy Controls (HC)) speakers. MARTA effectively clustered various MC, such as vowels, fricatives, stop-plosives, liquids, and nasals. Extending the research, the model was adapted to the NeuroVoz dataset. This new domain contains different acoustic materials from parkinsonian patients and controls, and was used to evaluate the quality and discrimination capabilities of the MC clustering in presence of PD.

The experimental approach was twofold. Initially, MARTA was trained in a non-supervised way with normophonic speakers from both datasets, with the aim of validating its clustering capability for different phonemic groups. The effectiveness of this approach was quantified using Jensen-Shannon Distance (JSD) between the clusters formed by the different phonemic groups in the latent space. The results are visualised in two-dimensional (2D), three-dimensional (3D), and 32-dimensional (32D) spaces.

In the subsequent phase, MARTA was tailored to discriminate between PD and HC speech. The methodology encompasses two analysis angles: first, an unsupervised examination of the JSD within parkinsonian MC; and second, a supervised classification technique, named Manner of ARTiculation Analysis with Supervision (MARTA-S), for detecting PD from the speech. Despite the application field, the techniques and the model presented in this paper could be used to evaluate the manners of articulation due to the presence of other diseases or due to dialectical differences.

The rest of the paper is organised as follows. Section 2 is dedicated to introducing the materials and methods used to develop the system; Section 3 presents the results; and Section 4 is dedicated to a discussion and drawing the main conclusions.

## 2 Materials and methods

This section describes the corpora used for the experiments and presents in detail the architecture of MARTA and methods used for processing the audio recordings.

### 2.1 Materials: Speech corpora

In this work, we have used two different speech corpora, namely: Albayzin and NeuroVoz. The first is a widely used corpus of normophonic speakers that is used to warm-up the model, and the second is a corpus of parkinsonian speech which is used to adjust it to the specific field of application. Both corpora are presented next.

#### *2.1.1 Albayzin corpus*

The Albayzin corpus [34] is a phonetically balanced dataset consisting of recordings from 200 Text-Dependent Utterances (TDU) sampled at 16 kHz and quantised with 16 bits. The corpus contains audio recordings from four individuals who uttered all 200 TDU, in addition to recordings from 160 speakers, each articulating a subset of 25 sentences of the total. The complete dataset contains 4.800 recordings, equivalent to approximately 4.1 hours of speech, along with their corresponding transcriptions.

The sentences recorded were chosen to follow the typical distribution of the different phonemes in Castilian Spanish. To ensure a comprehensive representation of less common sounds, a minimum requirement of 40 occurrences was set, resulting in a minimum of 960 appearances across the subcorpus. In terms of contextual relevance, the sentences were chosen to ensure that every phoneme is accompanied by its most relevant contexts at least four times. Generally, a context is considered relevant if it occurs in at least 10% of the appearances. Additionally, contexts that may lead to significant phonetic modifications were also deemed relevant. Regarding syllabic proportions, the distribution of each sound in stressed and unstressed syllables was intentionally crafted to reflect the natural proportions found in Castilian Spanish.

All speakers in this corpus are considered normophonic and, therefore, are associated with the HC group.

#### *2.1.2 NeuroVoz corpus*

The NeuroVoz corpus [35], [36] comprises speech samples from 57 individuals with PD and 44 HC subjects, all native Castilian Spanish speakers. Speech signals were recorded using an AKG C420 headset microphone connected to a phantom power preamplifier and recorded with the Medivoz software [37] with a total of 3.77 hours of audio data. The sampling rate was 44.1 kHz, and quantisation was performed with 16 bits.

The corpus contains different speech materials, including sustained vowel phonations, twelve TDU, and a monologue. Our focus was narrowed to the TDU of the dataset, allowing a controlled analysis of the speech patterns. The utterances were produced in a natural, comfortable tone and intensity of speech. Detailed transcriptions of these utterances, along with their corresponding representations in the International Phonetic Alphabet (IPA) [38], are provided in Table 1.

All speakers with PD were under pharmacological treatment and took the medication between 2 and 5 hours before the recordings. The research protocol, including the speech recording process, was approved by the Ethics Committee of Hospital General Universitario Gregorio Marañón. The protocol was in accordance with the Declaration of Helsinki as formulated by the World Medical Association, as well as the relevant European Directives. Informed consent was duly obtained from every participant.

### 2.2 Methods

This section presents the pre-processing methods applied to audio recordings, the methods used to label speech frames into their phonemic class (MC), and a detailed view of the architecture used for processing the audio signals.

#### *2.2.1 Pre-processing*

The pre-processing consists on three different steps, namely: phonetic alignment, audio pre-processing, and phonemic grouping. The specificities of each of these steps are presented next.

2.2.1.1 Phonetic alignment across datasets: The audio data were pre-processed to estimate the time instants corresponding to the beginning of each acoustic segment following a forced alignment phonetic procedure. This was carried out automatically using 'faseAlign' [39], a Python® tool that automates the forced alignment of Spanish speech from their text transcriptions, producing a phonetic label for each timestamp. This tool is integrated into the Hidden Markov Model Speech Recognition Toolkit (HTK) [40]. 'faseAlign' inherently includes acoustic models and dictionaries suited for a wide range of Latin American Spanish dialects, ensuring precise phonetic mappings and timing. Although the Albayzin corpus contains a manual alignment of the audio recordings, this tool was used in the search for a completely automated procedure and to ensure that the same processing is applied to both corpora used. This step was crucial in introducing a uniform alignment error across both datasets, thereby standardising the pre-processing pipeline.

2.2.1.2 Audio pre-processing: The pre-processing of the audio files followed a uniform pipeline applicable to both Albayzin and NeuroVoz corpora. Initially, audio files were processed using the *librosa* Python library [41], where they were downsampled to a standard rate of 16 kHz. Subsequently, each audio file underwent a normalisation procedure that divided the amplitude by its maximum absolute value. Following normalisation, the audio files were segmented into 400 ms frames with 50% of overlap. Audio files shorter than 400 ms were excluded to maintain consistency in frame length. Correspondingly, HTK *TextGrids* were also segmented following the same procedure, ensuring that they matched their respective audio segments.

TABLE 1: Transcriptions and translations of selected sentences

| Sent. # | Spanish transcription | IPA transcription | English translation |
|---|---|---|---|
| 1 | *La patata no está bien ablandada* | [la pa'ta ta no es 'ta βjen a βlan 'da ðа] | "The potato is not soft enough" |
| 2 | *Mañana vamos de acampada* | [ma 'ɲa na 'βa moz ðe a kam 'pa ða] | "Tomorrow we are going camping" |
| 3 | *Cuando las barbas de tu vecino veas pelar pon las tuyas a remojar* | ['kwan do laz 'βar βaz ðe tu βe 'θi no 'βe as pe 'lar pon las 'tu yas a re mo 'xar] | "When your neighbor's beard you see peeling, put yours to soak" |
| 4 | *Burro grande ande o no ande* | ['bu řo 'ɣran de 'an de o no 'an de] | "Big donkey walk or not walk" |
| 5 | *De la calle vendrá quien de tu casa te echará* | [de la 'ka ʎe βen 'dra kjen de tu 'ca sa te e tʃa 'ra] | "From outside will come that who will kick you out from your house" |
| 6 | *Carmen baila el mambo* | ['kar men 'baj la el 'mam bo] | "Carmen dances the mambo" |
| 7 | *Cuando el diablo no sabe qué hacer con el rabo mata moscas* | ['kwan do el 'dja βlo no 'sa βe 'ke a 'θer kon el 'řa βo ma ta 'mos kas] | "When the devil does not know what to do, it kills flies with its tail" |
| 8 | *Esto es una ganga* | ['es to es 'u na 'ɣaŋ ga | "This is a bargain" |
| 9 | *Juan tira de la manga* | [xwan 'ti ra ðe la 'maŋ ga] | "Juan pulls the sleeve" |
| 10 | *Dame pan y llámame perro* | ['da me pan i ʎa ma me 'pe řo] | "Give me bread and call me dog" |
| 11 | *No pidas a quien pidió, ni sirvas a quien sirvió* | [no 'pi ðas a kjem pi 'ðjo ni 'sir βas a kjen sir 'βjo] | "Do not beg the one who begged, nor serve the person who served" |
| 12 | *Tomás tira de la manta* | [to 'mas 'ti ra ðe la 'man ta] | "Tomás pulls the blanket" |

Further pre-processing was conducted on a per-frame basis. For each frame, a spectrogram was generated using the following parameters: a window length of 30 ms with 50% overlapping, Fast Fourier Transform (FFT) of 512 points, and a resolution of 65 mel frequency bands. The amplitude of the spectrogram was then converted to decibels (dB) and normalised relative to itself. This process resulted in a dataset comprising 400 ms spectrogram segments, each featuring 65 mel bands. Moreover, each 30 ms window within these spectrograms was associated with its corresponding phoneme.

2.2.1.3 Phonemic grouping: Building upon the pre-processing steps outlined earlier, the study also involved a phonemic grouping phase. In this process, each phoneme identified was labelled as one of eight distinct MC [29], i.e., a one-hot vector representation $\mathbf{m}_c \in \{0,1\}^8$ was assigned to each 30 ms window in the spectrogram according to its corresponding phoneme, as defined in Table 2. To do so, the label assigned to each window corresponds to the most represented phoneme (previously proposed by the forced alignment procedure).

This process resulted in a dual-layered labelling system: one label identifying the speaker's condition (HC or PD), consistent across all spectrograms for a given patient; and the other label denoting the MC for each acoustic feature for every window of the spectrogram. This dual labelling approach was crucial to link the phonetic details with the corresponding health condition of the patient.

For the purpose of this study, affricates and silences/short pauses were excluded from the experimental analysis and results. The reason for excluding affricates is their under-representation in the TDU from the NeuroVoz dataset. This under-representation is mainly attributed to the natural composition of the Spanish language, where affricates constitute the least common phonemic group, accounting for less than 3% of all phonemes [42]. By focussing on the six remaining MC, the analysis provides a more robust understanding of the speech characteristics prevalent in the studied datasets, particularly those relevant to the speech patterns of individuals with PD.

#### 2.2.2 Manner of ARTiculation Analysis (MARTA)

This study introduces MARTA, a tailored Gaussian Mixture (GM) Variational AutoEncoder (VAE) combined with metric learning, specifically designed to cluster spectrogram features into distinct MC for analysing misarticulations. This approach, drawing inspiration from [43], [44] and integrating metric learning techniques from [45], allows the effective clustering of MC. The functionality and architecture of MARTA are illustrated in Figure 1. This integration aims to improve the identification and analysis of speech patterns, focussing particularly on the nuances of pronunciation corresponding to patients with PD.

In MARTA's framework, the initial step involves a characterisation of the speech by means of spectrograms, denoted as $\mathbf{X} \in \mathbb{R}^{M,W_s}$ (see details of the mathematical notation in the Appendix). Here, $M$ corresponds to the number of mel-frequency bands, and $W_s$ to the number of temporal windows, each spanning 30 ms with 50% of overlap. For this study, these parameters are set as follows: $M$=65 mel bands, and $W_s$=25 windows.

A convolutional block acts as a bottleneck, transforming input spectrograms, $\mathbf{X}$, into a compact encoded representation, denoted as $\mathbf{X}_e \in \mathbb{R}^{W_s,H}$ where the temporal window dimension is kept to ensure the correspondence between each window and their associated MC. Here, $H$ represents the flattened dimension of the output of the convolutional block (channels $\times$ rows).

The core of the GMVAE architecture is the processing of encoded data ($\mathbf{X}_e$), with the objective of reconstructing it ($\tilde{\mathbf{X}}_e$) following the principles of an autoencoder. This generative mechanism incorporates two distinct latent categorical discrete random variables $\mathrm{Y} \in \{0,\ldots,G\}$, where $G$ represents the maximum number of Gaussian components in the mixture. The probability distribution $p(\mathrm{Y})$ is defined by a categorical distribution:

$$p(\mathrm{Y}) = Cat(\mathrm{Y}|\boldsymbol{\pi}), \tag{1}$$

where the parameter $\boldsymbol{\pi}$ is a vector specifying the probability associated with each potential category within Y.

TABLE 2: Categorisation of phonemes into MC (IPA)

| $\mathbf{m}_c$ codification | Manner Class | Phonemes (IPA) |
|---|---|---|
| $[0,0,0,0,0,0,0,1]$ | Plosives | /p/, /t/, /k/ |
| $[0,0,0,0,0,0,1,0]$ | Voiced Plosives | /b/, /β/, /d/, /ð/, /g/, /ɣ/ |
| $[0,0,0,0,0,1,0,0]$ | Nasals | /n/, /ŋ/, /m/, /ɲ/, /ɟ/ |
| $[0,0,0,0,1,0,0,0]$ | Fricatives | /f/, /s/, /z/, /ʃ/, /h/, /θ/ |
| $[0,0,0,1,0,0,0,0]$ | Liquids | /ɾ/, /r/, /ɻ/, /l/, /j/, /ʒ/, /ʎ/ |
| $[0,0,1,0,0,0,0,0]$ | Vowels | /a/, /e/, /i/, /o/, /u/, /w/, /eː/ |
| $[0,1,0,0,0,0,0,0]$ | Affricates | /tʃ/, /ts/ |
| $[1,0,0,0,0,0,0,0]$ | Silences and Short Pauses | [silence], [short pause] |

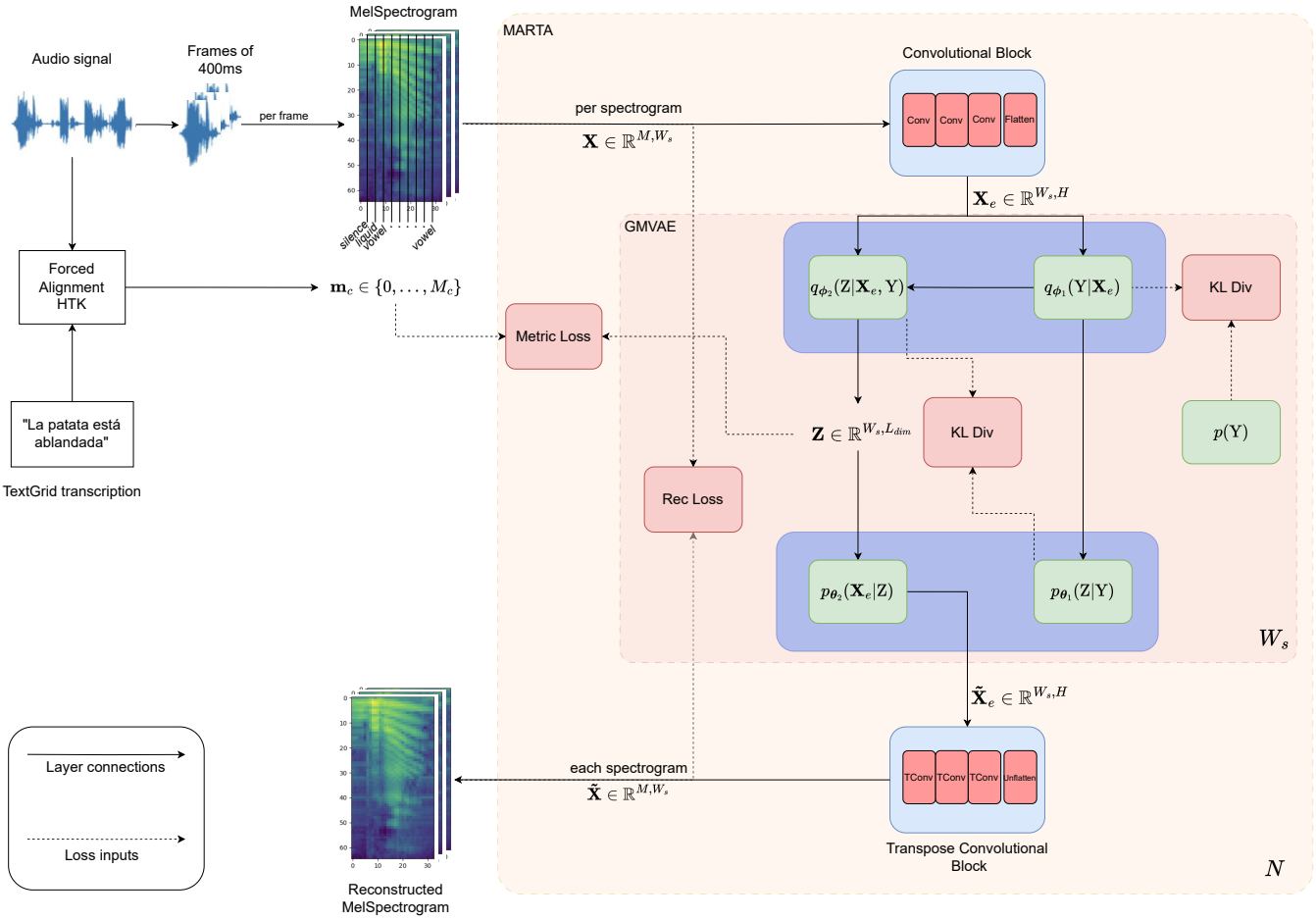



Fig. 1: Overview of MARTA. Green blocks are Neural Network (NN)s used to characterise the probabilistic processes. Red blocks denote terms of the loss function $\mathcal{L}$.

The continuous random variable $\mathrm{Z} \in \mathbb{R}^{W_s,L_{dim}}$ encapsulates the information of each 30 ms window ($\mathbf{w}_s$) of the encoded spectrogram ($\mathbf{X}_e$) in an embedded latent space ($\mathbf{z}_i$) characterised by the dimension $L_{dim}$. Where each $\mathbf{z}_i$, corresponding to a specific MC ($\mathbf{m}_c$, that is, a one-hot vector of size 8), is designed to retain essential audio reconstruction data while simultaneously embodying information pertinent to the MC.

The joint probability distribution of the data and latent variables is given by $p_{\mathbf{\Theta}}(\mathbf{X}_e,\mathrm{Z},\mathrm{Y})$, which is decomposed into the likelihood of the data given the latent continuous state, i.e., $p_{\boldsymbol{\theta}_2}(\mathbf{X}_e|\mathrm{Z})$, the distribution of the latent continuous variable given the latent discrete state, i.e., $p_{\boldsymbol{\theta}_1}(\mathrm{Z}|\mathrm{Y})$, and the prior distribution of the latent discrete variable, i.e., $p(\mathrm{Y})$. The generative model is formally defined as:

$$p_{\mathbf{\Theta}}(\mathbf{X}_e,\mathrm{Z},\mathrm{Y}) = p_{\boldsymbol{\theta}_2}(\mathbf{X}_e|\mathrm{Z})p_{\boldsymbol{\theta}_1}(\mathrm{Z}|\mathrm{Y})p(\mathrm{Y}) \tag{2}$$

where each term is expressed as:

$$\begin{aligned} p_{\boldsymbol{\theta}_1}(\mathrm{Z}|\mathrm{Y}) &= \mathcal{N}(\mathrm{Z}|\mu_{\boldsymbol{\theta}_1}(\mathbf{y}),\sigma^2_{\boldsymbol{\theta}_1}(\mathbf{y})), \\ p_{\boldsymbol{\theta}_2}(\mathbf{X}_e|\mathrm{Z}) &= \mathcal{N}(\mathrm{Z}|\mu_{\boldsymbol{\theta}_2}(\mathbf{z}),\sigma^2_{\boldsymbol{\theta}_2}(\mathbf{z})). \end{aligned} \tag{3}$$

Following the GMVAE [33], the distributions $p_{\boldsymbol{\theta}_1}(\mathrm{Z}|\mathrm{Y})$ and $p_{\boldsymbol{\theta}_2}(\mathbf{X}_e|\mathrm{Z})$ are represented by dense layers whose size is set according to the dimension of the latent space. Specifically, in the generative phase, the decoder network $p_{\boldsymbol{\theta}_2}(\mathbf{X}_e|\mathrm{Z})$ is tasked with generating $\mathbf{X}_e$ samples from their latent continuous representation Z. The last step of the

MARTA architecture is the transpose convolutional block, which mirrors the bottleneck in its structure and learns to reconstruct the original spectrograms, $\tilde{\mathbf{X}} \in \mathbb{R}^{M,W_s}$, from their encoded reconstructed states, $\tilde{\mathbf{X}}_e \in \mathbb{R}^{W_s,H}$. In summary, MARTA serves both to reconstruct input spectrograms and to infer the corresponding MC for each window, all within an unsupervised learning context with respect to the health condition.

In the inference stage, the true posterior distribution $p_\eta(\mathrm{Z},\mathrm{Y}|\mathbf{X}_e)$ is intractable [43]. Consequently, an approximation with a tractable variational posterior $q_{\mathbf{\Phi}}(\mathrm{Z},\mathrm{Y}|\mathbf{X}_e)$ is proposed. This is achieved through variational inference [46], where the inference model is defined as:

$$q_{\mathbf{\Phi}}(\mathrm{Z},\mathrm{Y}|\mathbf{X}_e) = q_{\boldsymbol{\phi}_2}(\mathrm{Z}|\mathbf{X}_e,\mathrm{Y})q_{\boldsymbol{\phi}_1}(\mathrm{Y}|\mathbf{X}_e) \tag{4}$$

with $q_{\boldsymbol{\phi}_1}(\mathrm{Y}|\mathbf{X}_e) = Cat(\mathrm{Y}|\pi_{\boldsymbol{\phi}_1}(\mathbf{X}_e))$ being a categorical distribution represented by a Gumbel-Softmax [47] and $q_{\boldsymbol{\phi}_2}(\mathrm{Z}|\mathbf{X}_e,\mathrm{Y}) = \mathcal{N}(\mathrm{Z}|\mu_{\boldsymbol{\phi}_2}(\mathbf{x}_e+\mathbf{y}_e),\sigma^2_{\boldsymbol{\phi}_2}(\mathbf{x}_e+\mathbf{y}_e))$ a Gaussian distribution. The term $\mathbf{y}_e = h(\mathbf{y})$ is an expanded version of $\mathbf{y}$ where $h(\cdot)$ is a dense layer that transforms $\mathbf{y}$ to align it with the dimensionality of $\mathbf{x}_e$ for proper summation.

The optimisation of the model is driven by the maximisation of the Evidence Lower Bound (ELBO):

$$\begin{aligned}\log p_{\mathbf{\Theta}}(\mathbf{X}_e) \geq \mathbb{E}_{q_{\boldsymbol{\phi}_1}(\mathrm{Y}|\mathbf{X}_e)}\Big[\mathbb{E}_{q_{\boldsymbol{\phi}_2}(\mathrm{Z}|\mathbf{X}_e,\mathrm{Y})}\left[\log p_{\boldsymbol{\theta}_2}(\mathbf{X}_e|\mathrm{Z})\right] - \\ KL\left(q_{\boldsymbol{\phi}_2}(\mathrm{Z}|\mathbf{X}_e,\mathrm{Y})||p_{\boldsymbol{\theta}_1}(\mathrm{Z}|\mathrm{Y})\right)\Big] - \\ KL\left(q_{\boldsymbol{\phi}_1}(\mathrm{Y}|\mathbf{X}_e)||p(\mathrm{Y})\right)\end{aligned} \tag{5}$$

which comprises a reconstruction loss for $\mathbf{X}_e$, a Gaussian mixture loss, and a categorical loss term.

To enhance the clustering of the MC, we introduce an auxiliary metric learning loss inspired by [44], specifically the Lifted Structured loss as detailed in [45]. This loss function is defined as:

$$\tilde{J} = \frac{1}{2|\mathbb{P}|}\sum_{(i,j)\in\mathbb{P}} \max(0,\tilde{J}_{\mathbf{z}_i,\mathbf{z}_j})^2, \tag{6}$$

where $\tilde{J}_{\mathbf{z}_i,\mathbf{z}_j}$ is computed using the log-sum-exp trick over distances between points in the latent space, enforcing the model to learn a discriminative latent space where distances reflect the distinction among MC:

$$\begin{aligned}\tilde{J}_{\mathbf{z}_i,\mathbf{z}_j} = \log\left(\sum_{\forall k\in\mathbb{N}}\exp\{\alpha - D_{\mathbf{z}_i,\mathbf{z}_k}\} + \sum_{\forall l\in\mathbb{N}}\exp\{\alpha - D_{\mathbf{z}_j,\mathbf{z}_l}\}\right) \\ + D_{\mathbf{z}_i,\mathbf{z}_j}\end{aligned} \tag{7}$$

$D_{\mathbf{z}_i,\mathbf{z}_j}$ is the Euclidean distances between latent representations, with $\mathbb{P}$ and $\mathbb{N}$ representing sets of positive (they share same $\mathbf{m}_c$ label) and negative pairs (they do not share the $\mathbf{m}_c$ label) respectively, and the margin $\alpha$ set to 1, to separate the distributions in the latent space as per the referenced article.

Consequently, the loss to optimise within MARTA framework comprises four distinct components. Each of these components serves a specific role in the learning process: the reconstruction loss $\mathcal{L}_{Rec}$ ensures the model's output fidelity to the original input; the Gaussian mixture loss $\mathcal{L}_{GM}$ facilitates the learning of a continuous latent space structured by the underlying Gaussian mixture model; the categorical loss $\mathcal{L}_{Cat}$ enables the model to utilise the discrete latent variable effectively; and the metric learning loss $\mathcal{L}_{Metric}$ imposes a structured similarity measure in the latent space, helping the clustering of the MC. Collectively, these components form the comprehensive loss function $\mathcal{L}$, which is formally expressed as:

$$\mathcal{L} = \mathcal{L}_{Rec} + \mathcal{L}_{GM} + \mathcal{L}_{Cat} + \mathcal{L}_{Metric} \tag{8}$$

This loss function defines the optimisation strategy for our model, guiding the learning process toward a representation that is both generative and discriminative, aligning with our goal of accurate and meaningful clustering of phonemes inside the spectrograms.

#### 2.2.3 MARTA with Supervision

Building upon our initial unsupervised approach, we propose a supervised variant that we termed Manner of ARTiculation Analysis with Supervision (MARTA-S). The supervised version (MARTA-S) requires two different steps: clustering, and classification. These steps are presented next.

2.2.3.1 MARTA-S Clustering: Manner of ARTiculation Analysis with Supervision (MARTA-S) clustering retains the original structure of MARTA but increased the MC from eight to sixteen ($M_c = 16$), splitting each of the eight MC into two, corresponding to the PD and HC clusters. The primary objective of this clustering was to push the cluster boundaries further, thereby increasing the model's ability to detect misarticulations associated with the health condition.

2.2.3.2 MARTA-S Classifier: Taking advantage of the latent space learnt by MARTA-S, a classifier was trained to categorise the speech into two classes: PD or HC. To do so, the parameters of MARTA-S clustering were frozen, and only those layers corresponding to the classifier were modified in the training stage. This progression aimed to evaluate the categorisation accuracy of the clusters defined by MARTA-S' latent space to separate PD and HC speech traits. The design and functionality of this classifier are shown in Figure 2, which highlights its integral role.

The inputs to the MARTA-S classifier are samples from the latent space representation, $\mathbf{Z} \in \mathbb{R}^{W_s,L_{dim}}$, augmented with the label corresponding to the MC. We explored three NN architectures for classification, namely: a Convolutional Neural Network (CNN), a Multilayer Perceptron (MLP), and a time-distributed CNN combined with a Time-CNN Long Short-Term Memory (LSTM) network, following [48]. In the CNN-based approaches, the kernel size was tailored to $\{L_{dim},3\}$ to capture the entire latent space considering adjacent windows. On the other hand, the MLP used a simple flattening operation, applying a direct multiplication of $L_{dim} \times W_s$. For the time-distributed approach, all consecutive frames of the mel spectrograms are used as input to the network, including a zero padding to ensure a fixed sequence length.

Different strategies were tested to combine the latent space representation $\mathbf{Z}$ from the static MARTA-S with the MC label: (i) utilising the latent encoding directly $\mathbf{Z}$; (ii) using a static torch embedding layer to transform the one-hot encoding ($\mathbf{m}_c \in \{0,1\}^8$) into a matrix $\mathbf{M}_c = d(\mathbf{m}_c)$ representation where $\mathbf{M}_c \in \mathbb{R}^{W_s,L_{dim}}$; and, (iii) allowing the embedding layer to learn an optimal transformation

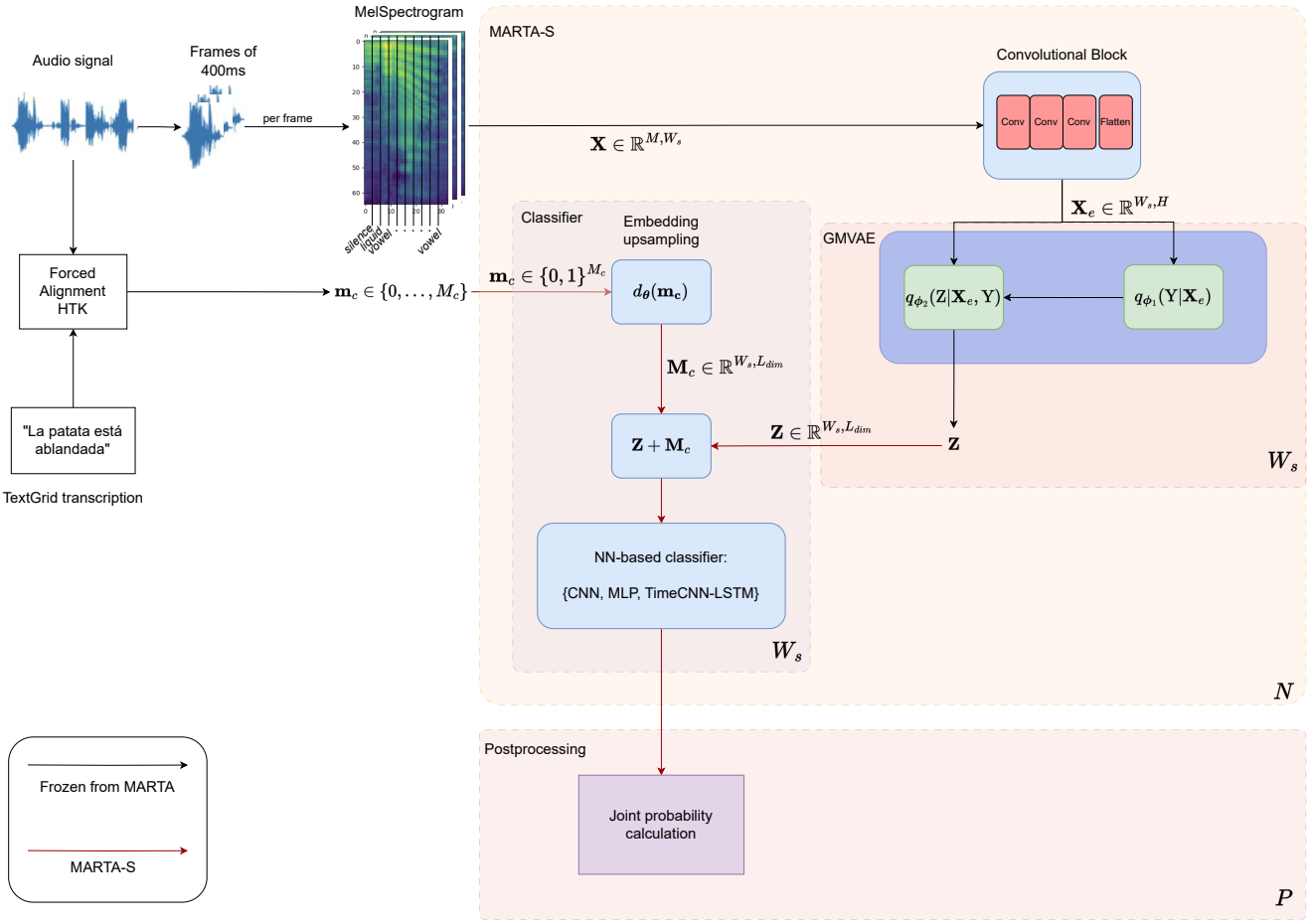



Fig. 2: MARTA-S classifier. Only red-line connections are performed.

$\mathbf{M}_c = d_{\boldsymbol{\theta}}(\mathbf{m}_c)$. Consequently, the input of the classifier was formed as one of the following: (i) $\mathbf{Z} = \mathbf{Z}$, (ii) $\mathbf{Z} = \mathbf{Z} + d(\mathbf{m}_c)$, or (iii) $\mathbf{Z} = \mathbf{Z} + d_{\boldsymbol{\theta}}(\mathbf{m}_c)$, depending on the embedding method chosen. The classifier was trained using a weighted cross-entropy loss to balance the representation of PD and HC samples into the training partition. In the inference phase, the model assigns a label to each 400 ms audio segment, categorising them as either PD or HC. To determine an overall diagnosis for each subject, we implemented a post-processing stage. This stage calculates the joint probability across all spectrograms for an individual patient and assigns a definitive class based on the highest aggregated probability. This approach ensures a coherent and probabilistically justified classification for each subject.

## 2.3 Experimental setup

An experimental methodology was designed to evaluate the capabilities of the MARTA model in processing and classifying speech, particularly focussing on the distinction between PD and HC speech patterns.

### *2.3.1 Unsupervised experiments*

Initially, we performed an unsupervised MARTA warm-up using the Albayzin corpus. The precise articulation of the normophonic speakers recorded in this data set make it perfect for initialising a model designed to detect misarticulations.

Subsequently, a domain adaptation MARTA was performed using utterances from NeuroVoz HC patients. For this purpose, ten folds of the HC subjects were randomly selected for training while one fold was kept for testing. This step was crucial in adapting the model to a more realistic and varied dataset, which, as described in Section 2.1.2, differs in the quality of the recordings and the equipment used. For validation and to implement early stopping, we used a combined set comprising Albayzin and NeuroVoz HC data.

The efficacy of the unsupervised model was evaluated on unseen portions of Albayzin and NeuroVoz datasets, encompassing both HC and PD speakers. We used two primary evaluation metrics: the reconstruction Mean Squared Error (MSE) of the spectrograms and the JSD [49], [50] for various comparative analyses. These analyses included a comparison of the healthy speech from Albayzin with the HC and PD speech from NeuroVoz; as well as a comparison of the speech belonging to the HC and PD groups within NeuroVoz itself. For these comparisons, we implemented a non-parametric Kernel Density Estimation (KDE) for each MC cluster, differentiating by health condition. This resulted in eight KDEs for the Albayzin dataset (one for each MC)

and sixteen KDEs for NeuroVoz (one for each MC and health condition). Subsequently, JSD calculations were performed between these KDEs. These analyses are showcased for three different values of $L_{dim}$. The initial pair seeks to illustrate the alignment between the model's internal mapping and the MC clusters, while the final one aims to leverage the model's ability to enhance the separation among these clusters. For clarity, this experiment will henceforth be referred to as `Exp_Unsup_`$L_{dim}$`D`, where $L_{dim}$ denotes the specific dimension evaluated.

This approach was designed to evaluate the model's ability to group healthy speech and its ability to detect the eight MC.

#### 2.3.2 Supervised experiments

Afterwards, a supervised study was performed to explore the potential of MARTA to differentiate between PD and HC speech. This phase required a new training with an expanded number of MC: from eight to sixteen. The expansion was due to a subdivision of each MC into two subgroups: PD and HC. This retraining aims to increase the distance (measured with the JSD) between both classes.

Subsequent evaluations mirrored those of the unsupervised phase. We specifically focused on how the increase in the number of groups affected the JSD, and the overall distinction between the healthy and pathological categories. Similar to the former case, this experiment will henceforth be referred to as Exp_Sup_$L_{dim}$D, where $L_{dim}$ again denotes the specific dimension evaluated.

The final stage aimed to assess the efficacy of these clusters for the screening of PD. For this purpose, NeuroVoz data were used exclusively. For a more robust approach, data augmentation techniques were used, such as frequency masking on the spectrograms, and a balanced dataset is ensured through stratification. With this dataset, various classifiers were trained to determine the effectiveness of MARTA-S for the screening of PD. This stage was crucial to validate whether the supervised clustering had practical applications for the screening of PD. This experiment will henceforth be referred to as `Exp_Classif`.

## 3 Results

### 3.1 Unsupervised analysis of the manner of articulation

#### 3.1.1 Constrained interpretable latent spaces: 2D and 3D spaces

Figure 3 displays the latent space distribution of the different phonemic groups from test patients, with distinct colours representing the different MC. This plot evidences MARTA's ability to cluster several classes —namely vowels (in cyan), stop-plosives (in blue), and fricatives (in purple)— according to the setup defined for `Exp_Unsup_2D`. However, the overlap of nasals (in green), liquids (in orange), and voiced-plosives (in red) with the vowels cluster indicates a potential limitation.

This observation suggests that while the model demonstrates competence in distinguishing three out of the six MC in the constrained 2D space, an increase in the dimensionality of the latent space might be necessary to differentiate the remaining classes effectively.

Figure 4 provides a rendering of the latent space obtained by applying the `Exp_Unsup_3D` configuration, which shows the dispersion of the different MC. Expansion to a 3D domain significantly improves the separation of vowels clusters. Notably, distinct clusters corresponding to liquids (in orange) and voiced-plosives (in red) are discovered from the expanded latent space. This enhanced visualisation indicates that the capacity of the model to differentiate the different MC has improved, with five out of six classes now clearly distinguishable, leaving only the nasals (in green) less clustered.

To complement this visual interpretation, Figure 5 presents a quantitative measure of class separation, depicting the JSD calculated among the KDEs estimated for each MC. Since, in this case, the MC clusters are defined according to the `Exp_Unsup_3D` setup, several comparisons among clusters from Albayzing and HC/PD NeuroVoz samples were performed independently. This metric provides an additional layer of analysis, quantifying the distinctness of clusters within the three-dimensional latent space.

Figure 5a illustrates the JSD matrix for the MC between Albayzin and the NeuroVoz HC cluster. The lowest JSD values are found along the matrix diagonal, suggesting a high degree of similarity between corresponding MC in the training and test datasets for HC patients. Despite this, a significant overlap is observed between voiced-plosives, liquids, and nasals, as they exhibit comparable JSD values, which is consistent with the clustering patterns observed in Figure 4. Furthermore, the matrix indicates that the fricatives in the NeuroVoz data exhibit the greatest dissimilarity relative to the Albayzin training samples, with a JSD value of 0.31, which means that the discrepancy between the datasets is predominantly manifested within the cluster of fricatives.

Figure 5b displays the JSD matrix for MC comparing the Albayzin and NeuroVoz PD datasets. The matrix reveals the lowest JSD values along its diagonal, indicating a stronger correspondence between each MC in the test set with its counterpart in the training set. The overlap among nasals, voiced-plosives, and liquids remains, consistent with previous observations in the latent space (Figure 4).

Additionally, Figure 5c compares the JSD values between the HC and PD clusters in the NeuroVoz dataset. This comparison shows a negligible distinction between both, suggesting that unsupervised analysis alone is not enough to differentiate parkinsonian from normophonic subjects.

#### 3.1.2 Optimal model configuration

To ascertain the optimal clusterisation, we undertook a systematic cross-validation, focussing on the latent dimension ($L_{dim}$), the number of Gaussian components $G$), and the configuration of the hidden layers within the encoder ($f(\mathbf{X})$) and decoder ($g(\tilde{\mathbf{X}}_e)$) networks. The top ten configurations, as delineated in Table 3, were selected based on their reconstruction and metric loss performance.

A Pearson correlation analysis was performed to evaluate the relationship between the number of Gaussians, the latent dimensions, and the observed loss metrics. The latent dimension exhibited a slight negative correlation with the reconstruction and metric losses, with coefficients of $-0.15$ and $-0.13$, respectively. This suggests a modest inverse

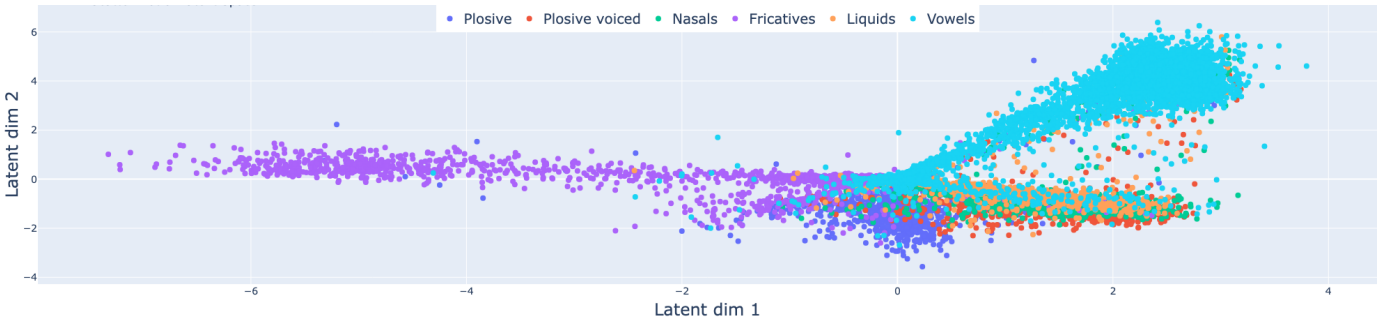
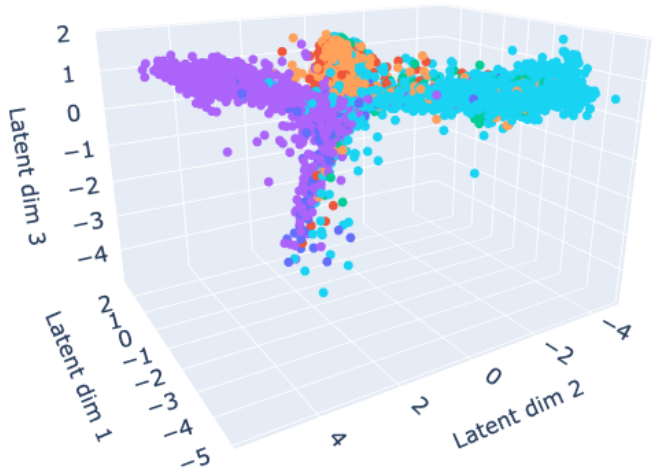
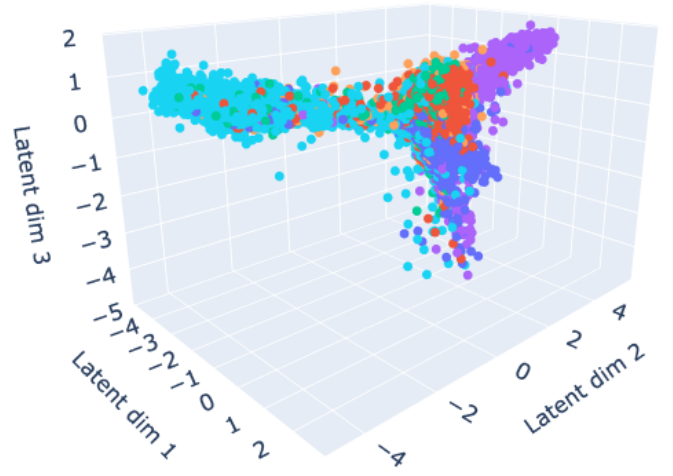



Fig. 3: Unsupervised MARTA. 2D representation of the different MC in the latent space obtained according to the `Exp_Unsup_2D` experiment.




(a) 3D latent space. First point-of-view




(b) 3D latent space. Second point-of-view

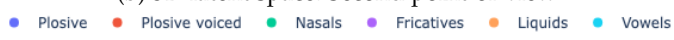



Fig. 4: Unsupervised MARTA. Visualisations of the different MC in the latent space of `Exp_Unsup_3D`, from different points-of-view.

relationship, indicating that larger latent dimensions could potentially contribute to a reduction of the loss. On the contrary, the number of Gaussian components ($G$) showed negligible correlations, with coefficients of $-0.04$ for reconstruction loss and $-0.03$ for metric loss, suggesting a minimal impact on performance.

TABLE 3: The ten best models measured by reconstruction and metric losses in the validation set.

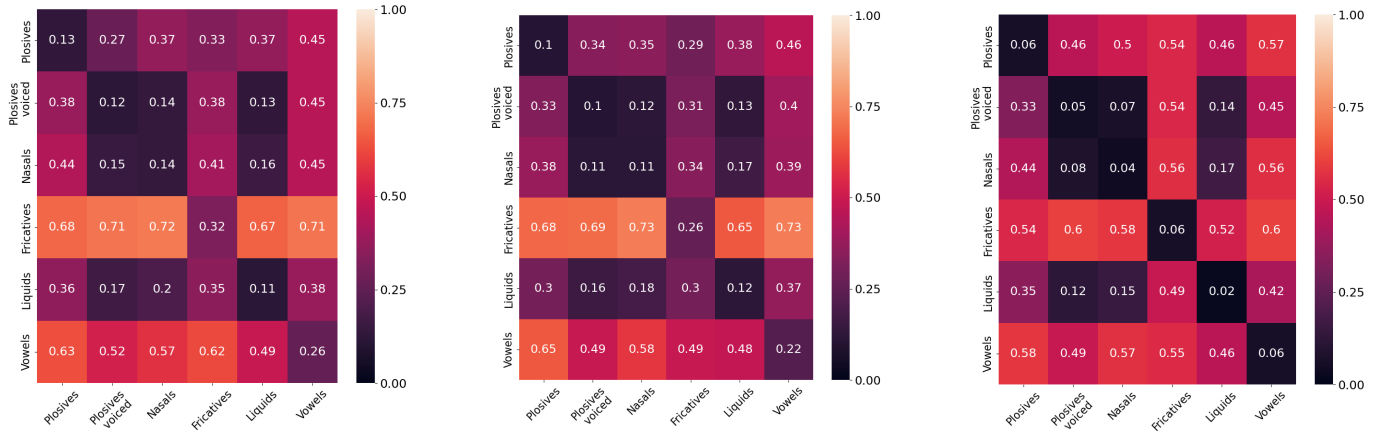
| $G$ | $L_{dim}$ | Encoder/Decoder | Reconstruction loss | Metric loss |
|---|---|---|---|---|
| 128 | 32 | [64, 1024, 64] | 219.21 | 1315.42 |
| 128 | 32 | [64, 1024, 64] | 218.04 | 1315.79 |
| 1024 | 32 | [64, 256, 64] | 210.95 | 1315.31 |
| 512 | 32 | [64, 256, 64] | 209.05 | 1315.80 |
| 32 | 64 | [64, 1024, 64] | 226.16 | 1316.28 |
| 64 | 32 | [64, 1024, 64] | 225.91 | 1316.34 |
| 512 | 64 | [64, 1024, 64] | 222.84 | 1316.37 |
| 512 | 32 | [64, 256, 64] | 215.37 | 1316.38 |
| 512 | 64 | [64, 256, 64] | 216.25 | 1316.48 |
| 32 | 128 | [64, 512, 64] | 225.07 | 1316.48 |
| 16 | 32 | [64, 128, 64] | 228.38 | 1317.40 |

Given these insights, we opted for a configuration of $G = 16$ Gaussian components, which is equivalent to one Gaussian per MC and condition. This maintains a simple model while ensuring an adequate representational capacity. In addition, a $L_{dim} = 32$ was chosen to balance the trade-off between achieving satisfactory metric scores and keeping the simplicity of the model. Such a configuration defines the `Exp_Unsup_32D` setup. Finally, the encoder/decoder architecture was set to [64, 1024, 64] in order to limit the complexity of the model and to place the focus on feature extraction. Following the previous results, the subsequent analysis aims to quantitatively compare the performance of the model chosen ($L_{dim} = 32$) against the variant previously evaluated ($L_{dim} = 3$), employing distance metrics to study the degree of improvement in unsupervised learning tasks and their difference in absolute and Mean Absolute Percentage Error (MAPE) value.

TABLE 4: Comparison of 3D and 32D models in terms of reconstruction power measured in MSE.

| Metric | 3D Model | 32D Model |
|---|---|---|
| MSE | $0.14 \pm 0.02$ | $0.11 \pm 0.01$ |
| MSE decrease MAPE | – | 19.35 |

The comparison between the 32D and 3D models reveals notable differences. As indicated in Table 4, the 32D model shows a marked improvement in spectrogram reconstruction, evidenced by a reduction of 19.35 points in MAPE for MSE. Distance analysis, as illustrated in Figure 6, highlights the improved ability of the 32D model to separate the voiced-plosives cluster from the nasals cluster. While in the 3D model, these clusters were closely associated, with a JSD




(a) JSD between the canonical normophonic clusters (columns) and the test clusters of HC from NeuroVoz (rows).

(b) JSD between the canonical normophonic clusters (columns) and the test clusters of PD from NeuroVoz (rows).

(c) JSD between the HC test clusters of NeuroVoz (columns) and PD test clusters of NeuroVoz (rows).

Fig. 5: JSD of all clusters (HC and PD) in the 3D latent space following the training scheme in `Exp_Unsup_3D`.

of 0.14 (as seen in Figure 5a), the 32D model distinguishes them with a JSD of 0.61, which represents a substantial increase in MAPE pf 335.71 points. Despite these improvements, the distinction between voiced-plosives and liquids persists as a challenge, with a JSD of 0.1, still indicating close proximity.

Regarding PD speech, the 32D model reveals a slight increase in the separation of voiced-plosives, nasals, and liquids, as shown in Figure 6b. Furthermore, the contrast between HC and PD speech within the NeuroVoz dataset reinforces the previous observations seen in the 3D analysis: the `Exp_Unsup_32D` setup alone may not be sufficient for the direct identification of parkinsonian manners of articulation.

### 3.2 Supervised analysis of the parkinsonian manner of articulation

In this section, the model used before, characterised by a 32D latent space ($L_{dim} = 32$), an encoder/decoder configuration of [64, 1024, 64] and 16 Gaussian components ($G = 16$), was adapted following the `Exp_Sup_32D` training scheme. This adaptation aims to optimise the model to distinguish between 16 MC, 8 MC per health condition (i.e., HC and PD). The efficacy of the model was further assessed through the analysis of distances between HC and PD clusters within the NeuroVoz dataset.

Subsequently, the discriminative power of the model was assessed using a classification task (`Exp_Classif`). For this purpose, a 10-fold cross-validation method was implemented. An ablation study was conducted to ascertain the optimal input for the classifier, ultimately guiding the determination of the most effective training approach for a classifier tasked with differentiating HC and PD pronunciations of MC.

#### *3.2.1 Effects in the distance between clusters*

This analysis focuses on the spatial separation of speech clusters belonging to PD and HC. This distance is quantitatively evaluated using JSD metrics. The analysis also includes an evaluation of the absolute differences and MAPE values, providing a detailed evaluation of the dispersion of the groups within the latent space.

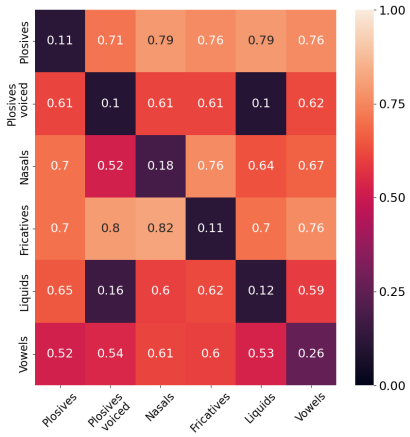
Figure 7a presents a JSD distance matrix which reports the lowest values along its diagonal, which means that each healthy MC cluster of the NeuroVoz dataset is proximal to its canonical counterpart in Albayzin. This clustering suggests a successful grouping of the different MC when following the `Exp_Sup_32D` setup. Compared to the `Exp_Unsup_32D` (Figure 6a), there is an increase in the diagonal values from $0.15 \pm 0.06$ to $0.21 \pm 0.06$, which is equivalent to an MAPE of 58.07%. This change suggests more difficulties in differentiating HC clusters in the supervised framework.

In contrast, the `Exp_Sup_32D` setup shows a notable increase in the distances of the main diagonal for the PD clusters, as shown in Figure 7b. This increase suggests a clearer distinction between the parkinsonian MCs of NeuroVoz and those normophonic of the Albayzin corpus. Specifically, the diagonal values have increased from $0.17 \pm 0.05$ (in the unsupervised analysis) to $0.40 \pm 0.06$ (using the supervised model), which represents an MAPE of 159.75%. This significant increase underscores the feasibility of identifying differences in a supervised setting between the MC corresponding to the PD and HC clusters.

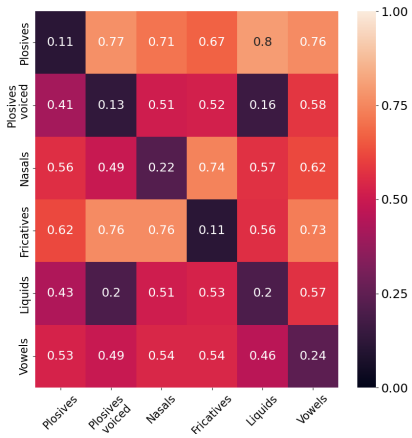
The potential for domain shift, given that the analyses involve different datasets, is addressed in Figure 7c, which delineates the separation of PD and HC clusters within the NeuroVoz data set. This figure reveals an increase in the values of the main diagonal from $0.07 \pm 0.02$ to $0.36 \pm 0.07$ (for unsupervised and supervised models, respectively), which represents an MAPE of 450.69 points. This increase (particularly in the context of stop-plosives (0.45), voiced-plosives (0.40), and nasals (0.40)), confirms that parkinsonian clusters are distinctly separable from healthy ones.

#### *3.2.2 Discrimination of parkinsonian and healthy speech using the NeuroVoz dataset*

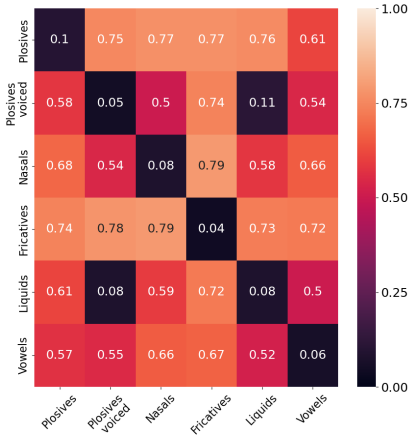
Once the clusterisation efficacy has been established, the subsequent analysis assesses the capability of the model for the automatic screening of PD. Results are calculated




(a) JSD between the canonical healthy clusters (columns) and the HC test clusters of NeuroVoz (rows).

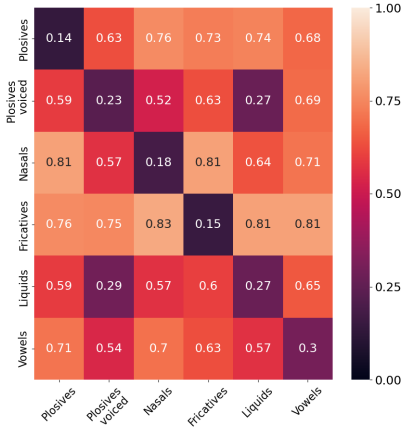



(b) JSD between the canonical healthy clusters (columns) and the PD test clusters of NeuroVoz (rows).

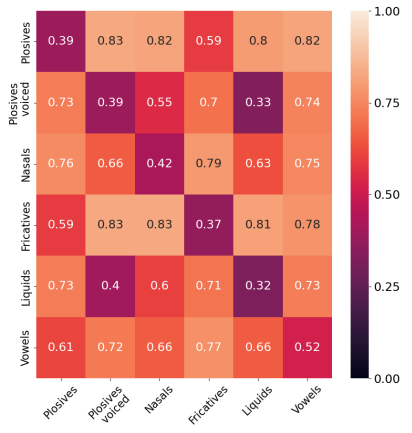



(c) JSD between the HC test clusters of NeuroVoz (columns) and PD test clusters of NeuroVoz (rows).

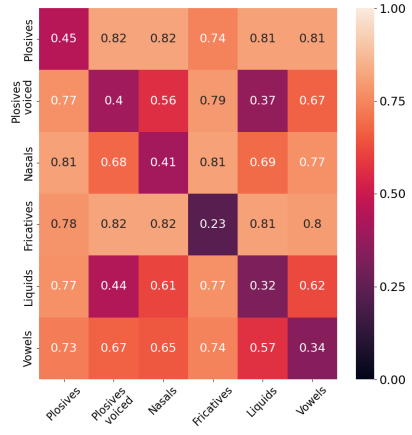
Fig. 6: JSD between clusters in the 32D latent space following the `Exp_Unsup_32D` training scheme.




(a) JSD between the canonical healthy clusters (columns) and the HC test clusters of NeuroVoz (rows.




(b) JSD between the canonical healthy clusters (columns) and the PD test clusters of NeuroVoz (rows).




(c) JSD between the HC test clusters of NeuroVoz (columns) and PD test clusters of NeuroVoz (rows).

Fig. 7: JSD between clusters in the 32D latent space following the `Exp_Sup_32D` training scheme.

at a patient-level, i.e., the joint probability across all manner class latent representation for an individual patient is calculated assigning then a definitive class based on the highest aggregated probability. This approach ensures a coherent and probabilistically justified classification for each subject.

Table 5 displays the Balanced Accuracy (BACC) and Area under the ROC curve (AUC) scores derived from a 10-fold cross-validation test used to identify patients as PD or HC. During the `Exp_Classif` setup, various architectures, as referenced in Section 2.2.3, were explored. The CNN architecture that incorporated the latent space ($\mathbf{Z}$) alongside a learnable embedding of the MC ($d_{\boldsymbol{\theta}}(\mathbf{m}_c)$) and Data Augmentation (DA) achieved a BACC of $0.91 \pm 0.09$.

### 3.2.3 *Accuracy analysis masking the different MC*

On the other hand, Table 6 illustrates the impact on classification performance when each MC is selectively omitted from the inference phase. Notably, the exclusion of vowels reports the most significant decrease in MARTA-S's classification performance, dropping from $0.91 \pm 0.09$ to $0.70 \pm 0.12$.

TABLE 5: Discrimination capabilities of the model to separate PD and HC. Test metrics are given for different configurations in terms of mean and standard deviation obtained following a 10-fold cross-validation.

| Classifier | Input | Metric | Mean $\pm$ Std |
|---|---|---|---|
| CNN | $\mathbf{Z} + d_{\boldsymbol{\theta}}(\mathbf{m}_c)$ + DA | BACC | **0.91 $\pm$ 0.09** |
| | | AUC | 0.86 $\pm$ 0.13 |
| TimeCNN-LSTM | $\mathbf{Z} + d_{\boldsymbol{\theta}}(\mathbf{m}_c)$ + DA | BACC | 0.83 $\pm$ 0.11 |
| | | AUC | **0.88 $\pm$ 0.12** |
| MLP | $\mathbf{Z} + d_{\boldsymbol{\theta}}(\mathbf{m}_c)$ + DA | BACC | 0.88 $\pm$ 0.11 |
| | | AUC | 0.85 $\pm$ 0.15 |
| CNN | $\mathbf{Z} + d_{\boldsymbol{\theta}}(\mathbf{m}_c)$ | BACC | 0.86 $\pm$ 0.10 |
| | | AUC | 0.85 $\pm$ 0.15 |
| CNN | $\mathbf{Z} + d(\mathbf{m}_c)$ | BACC | 0.84$\pm$ 0.12 |
| | | AUC | 0.83 $\pm$ 0.14 |
| CNN | $\mathbf{Z}$ | BACC | 0.81 $\pm$ 0.01 |
| | | AUC | 0.84 $\pm$ 0.11 |

Similarly, the removal of stop-plosives, voiced-plosives, liquids and fricatives results in a uniform reduction in model performance to an average 0.79 in across these cases. The

TABLE 6: Discrimination downgrading of the model by masking one manner class. Note: % indicates the percentage of each manner class in the test set. Affricates (1.2%) and silences (22.4%) were omitted.

| Classifier | Masked | % | Metric | **Mean ± Std** |
|---|---|---|---|---|
| CNN | Plosives | 14.0 | BACC | $0.79 \pm 0.15$ |
| | | | AUC | $0.86 \pm 0.14$ |
| CNN | Plosives voiced | 6.0 | BACC | $0.79 \pm 0.14$ |
| | | | AUC | $0.86 \pm 0.14$ |
| CNN | Nasals | 11.0 | BACC | $0.80 \pm 0.12$ |
| | | | AUC | $0.85 \pm 0.16$ |
| CNN | Fricatives | 6.5 | BACC | $0.79 \pm 0.14$ |
| | | | AUC | $0.86 \pm 0.14$ |
| CNN | Liquids | 5.9 | BACC | $0.79 \pm 0.14$ |
| | | | AUC | $0.86 \pm 0.14$ |
| CNN | Vowels | 33.0 | BACC | $\mathbf{0.70 \pm 0.12}$ |
| | | | AUC | $0.85 \pm 0.15$ |

impact of omitting liquids, voiced-plosives, and fricatives is particularly notable, given their minor representation in the test spectrograms, comprising only 6% of the sounds.

## 4 Discussion and Conclusions

The paper introduces MARTA, a novel tool powered by a GMVAE aimed at improving the automatic detection of PD through speech analysis. MARTA creates a latent acoustic space, which allows the discrimination between the different MC, but also between normophonic and parkinsonian speech segments without the need for different encoders. MARTA also opens the door to evaluate the effect of the disease on different MC, which is considered crucial for further detection of parkinsonian speech and for determining appropriate rehabilitation methods. The analysis presented underscores the contribution of each MC for the screening of PD, with vowels, stop-plosives, and nasals emerging as particularly influential in distinguishing HC and PD manners of articulation.

The model automatically processes the audio signal to clusterise 30 ms windows based on their MC, providing embedding vectors that codify phonetic and acoustic information able to characterise differences between PD patients and HC.

MARTA was initially trained using the Albayzin dataset, which is composed of 4800 utterances belonging to Castilian Spanish normophonic speakers. This canonical model can effectively cluster the different normophonic MC in an unsupervised way, such as vowels, fricatives, stop-plosives, liquids, and nasals. This capability was further validated through its application to the normophonic recordings of the NeuroVoz dataset, and is evidenced by a small average JSD along the diagonal between both healthy clusters.

Extending the research, the model was then adapted to the domain represented by the NeuroVoz dataset, encompassing both PD and HC speakers. The goal was to assess its discriminative power across the different MC for the detection of PD. MARTA showed an interesting ability to cluster the PD and HC manners of articulation in a non-supervised way, achieving an average Jensen-Shannon distance of $0.15 \pm 0.06$ between the same phonemic groups (but with different health conditions).

The experimental approach involved a second phase of supervised classification. The results in this second phase (using MARTA-S) suggest that the supervised scheme is capable of detecting the speakers with PD with high accuracy (91%±9). In terms of JSD the supervised model reports a significant increase of $0.36 \pm 0.07$ for the same phonemic groups but different health conditions. This represents an improvement of 450.69 points in terms of MAPE with respect to the unsupervised scenario.

The supervised and unsupervised results also suggest that PD not only affects the manner of articulation of stop-plosives, voiced-plosives, and nasals, but also significantly affects liquids, vowels and fricatives.

In terms of comparison, the overall discrimination results obtained are aligned with those reported in [29] using the same corpus of speakers, and using different models for each MC. Notably, MARTA-S' approach registers a modest improvement in BACC, escalating from 89% to 91%, with the additional advantage of using a single integrated model (instead of one per MC). However, a direct comparison of the unsupervised performance of MARTA with that of SCRAPS [31] was not feasible, due to the inaccessibility of SCRAPS as an open source model. A recent work published in [51], evaluated some speech-based self-supervised embedding methods, such as Wav2Vec [52] or HuBERT [53], in the context of parkinsonian speech. However, those approaches are intended for more general speech recognition tasks and are not driven by a phonemic perspective, and therefore, they do not provide an interpretable latent space as MARTA does.

The high accuracy obtained opens the door to the development of new downstream applications in speech processing. As an example, it could improve the accuracy of speech recognition in patients with Parkinson's disease and help develop novel comprehensive multimodal PD assessment tools. Furthermore, it opens the door to better and more sophisticated methods for speech therapy of patients with PD. In this sense MARTA is expected to mark a significant breakthrough by offering a sophisticated, data-driven approach to identify and understand the effect of PD in the speech and the degree of affection, thus providing speech therapists with precise information to effectively address these challenges.

Regardless of the specific application domain presented in this paper, the methods and models outlined have potential utility in generically discerning the manner of articulation, whether influenced by other medical conditions or dialectal variations. The application to other domains of application remains as a future work.

## Code availability

The source code supporting the findings of this study is openly available. The interested reader can access the models and scripts used in this research in the following repository: https://github.com/BYO-UPM/MARTA.

## References

[1] D. Aarsland, B. Creese, M. Politis, K. R. Chaudhuri, D. H. Ffytche, D. Weintraub, and C. Ballard, "Cognitive decline in parkinson disease," *Nature Reviews Neurology*, vol. 13, no. 4, pp. 217–231, 2017.

[2] S. Sveinbjornsdottir, "The clinical symptoms of parkinson's disease," *Journal of Neurochemistry*, vol. 139, pp. 318–324, 2016.

[3] S. Koga, N. Aoki, R. J. Uitti, J. A. Van Gerpen, W. P. Cheshire, K. A. Josephs, Z. K. Wszolek, J. W. Langston, and D. W. Dickson, "When dlb, pd, and psp masquerade as msa: an autopsy study of 134 patients," *Neurology*, vol. 85, no. 5, pp. 404–412, 2015.

[4] A. J. Hughes, S. E. Daniel, Y. Ben-Shlomo, and A. J. Lees, "The accuracy of diagnosis of parkinsonian syndromes in a specialist movement disorder service." *Brain: a Journal of Neurology*, vol. 125 Pt 4, pp. 861–70, 2002.

[5] E. Tolosa, A. Garrido, S. W. Scholz, and W. Poewe, "Challenges in the diagnosis of parkinson's disease," *The Lancet Neurology*, vol. 20, no. 5, pp. 385–397, 2021.

[6] J. Pujols, S. Peña-Díaz, D. F. Lázaro, F. Peccati, F. Pinheiro, D. González, A. Carija, S. Navarro, M. Conde-Giménez, J. García, S. Guardiola, E. Giralt, X. Salvatella, J. Sancho, M. Sodupe, T. F. Outeiro, E. Dalfo, and S. Ventura, "Small molecule inhibits $\alpha$-synuclein aggregation, disrupts amyloid fibrils, and prevents degeneration of dopaminergic neurons," *Proceedings of the National Academy of Sciences*, vol. 115, pp. 10 481 – 10 486, 2018.

[7] Q. C. Ngo, M. A. Motin, N. D. Pah, P. Drotár, P. Kempster, and D. Kumar, "Computerized analysis of speech and voice for parkinson's disease: A systematic review," *Computer Methods and Programs in Biomedicine*, p. 107133, 2022.

[8] J. R. Duffy, *Motor Speech Disorders: Substrates, Differential Diagnosis, and Management*, third ed. ed. St. Louis Mo: Elsevier Mosby, 2013.

[9] B. Walsh and A. Smith, "Basic parameters of articulatory movements and acoustics in individuals with parkinson's disease," *Movement Disorders*, vol. 27, 2012.

[10] K. Forrest and G. Weismer, "Dynamic aspects of lower lip movement in parkinsonian and neurologically normal geriatric speakers' production of stress." *Journal of Speech and Hearing Research*, vol. 38 2, pp. 260–72, 1995.

[11] P. Svensson, C. Henningson, and S. Karlsson, "Speech motor control in parkinson's disease: a comparison between a clinical assessment protocol and a quantitative analysis of mandibular movements." *Folia Phoniatrica*, vol. 45 4, pp. 157–64, 1993.

[12] K. V. Chenausky, J. MacAuslan, and R. S. Goldhor, "Acoustic analysis of pd speech," *Parkinson's Disease*, vol. 2011, 2011.

[13] J. I. Godino-Llorente, S. Shattuck-Hufnagel, J.-Y. Choi, L. Moro-Velázquez, and J. A. Gómez-García, "Towards the identification of idiopathic parkinson's disease from the speech. new articulatory kinetic biomarkers," *PLoS ONE*, vol. 12, 2017.

[14] T. K. Antolík and C. Fougeron, "Consonant distortions in dysarthria due to parkinson's disease, amyotrophic lateral sclerosis and cerebellar ataxia," in *Interspeech*, 2013.

[15] G. Weismer and J. Wildermuth, "Formant trajectory characteristics in persons with parkinson, cerebellar, and upper motor neuron disease," *Journal of the Acoustical Society of America*, vol. 103, pp. 2892–2892, 1998.

[16] J. Rusz, R. Cmejla, T. Tykalová, H. Ruzickova, J. Klempír, V. Majerova, J. Picmausova, J. Roth, and E. Ružička, "Imprecise vowel articulation as a potential early marker of parkinson's disease: effect of speaking task." *The Journal of the Acoustical Society of America*, vol. 134 3, pp. 2171–81, 2013.

[17] P.-A. McRae, K. Tjaden, and B. Schoonings, "Acoustic and perceptual consequences of articulatory rate change in parkinson disease." *Journal of Speech, Language, and Hearing Research*, vol. 45 1, pp. 35–50, 2002.

[18] L. Moro-Velázquez, J. A. G. García, J. D. Arias-Londoño, N. Dehak, and J. I. Godino-Llorente, "Advances in parkinson's disease detection and assessment using voice and speech: A review of the articulatory and phonatory aspects," *Biomedical Signal Processing and Control*, vol. 66, p. 102418, 2021.

[19] J. Logemann and H. B. Fisher, "Vocal tract control in parkinson's disease," *Journal of Speech and Hearing Disorders*, vol. 46, pp. 348–352, 1981.

[20] J. Robbins, J. Logemann, and H. S. Kirshner, "Swallowing and speech production in parkinson's disease," *Annals of Neurology*, vol. 19, 1986.

[21] E. Q. Wang, L. V. Metman, R. Bakay, J. Arzbaecher, B. A. Bernard, and D. M. Corcos, "Hemisphere-specific effects of subthalamic nucleus deep brain stimulation on speaking rate and articulatory accuracy of syllable repetitions in parkinson's disease." *Journal of Medical Speech-Language Pathology*, vol. 14 4, pp. 323–334, 2006.

[22] Y. Kim, "Acoustic characteristics of fricatives/s/and//produced by speakers with parkinson's disease," *Clinical archives of communication disorders*, vol. 2, no. 1, p. 7, 2017.

[23] T. Bhattacharjee, Y. Belur, A. Nalini, R. Yadav, and P. K. Ghosh, "Exploring the role of fricatives in classifying healthy subjects and patients with amyotrophic lateral sclerosis and parkinson's disease," *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, 2023.

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

[25] L. Moro-Velázquez, J. A. G. García, J. I. Godino-Llorente, J. Villalba, J. Rusz, S. Shattuck-Hufnagel, and N. Dehak, "A forced gaussians based methodology for the differential evaluation of parkinson's disease by means of speech processing," *Biomed. Signal Process. Control.*, vol. 48, pp. 205–220, 2019.

[26] A.-M. Tăuțan, B. Ionescu, and E. Santarnecchi, "Artificial intelligence in neurodegenerative diseases: A review of available tools with a focus on machine learning techniques," *Artificial Intelligence in Medicine*, vol. 117, p. 102081, 2021.

[27] J. Hlavnička, R. Cmejla, T. Tykalová, K. Šonka, E. Růžička, and J. Rusz, "Automated analysis of connected speech reveals early biomarkers of parkinson's disease in patients with rapid eye movement sleep behaviour disorder," *Scientific Reports*, vol. 7, 2017.

[28] L. Moro-Velázquez, J. A. G. García, J. I. Godino-Llorente, J. Villalba, J. R. Orozco-Arroyave, and N. Dehak, "Analysis of speaker recognition methodologies and the influence of kinetic changes to automatically detect parkinson's disease," *Applied Soft Computing*, vol. 62, pp. 649–666, 2018.

[29] L. Moro-Velázquez, J. A. Gómez-García, J. I. Godino-Llorente, F. Grandas-Perez, S. I. Shattuck-Hufnagel, V. Yagüe-Jiménez, and N. Dehak, "Phonetic relevance and phonemic grouping of speech in the automatic detection of parkinson's disease," *Scientific Reports*, vol. 9, 2019.

[30] V. Skaramagkas, A. Pentari, Z. Kefalopoulou, and M. Tsiknakis, "Multi-modal deep learning diagnosis of parkinson's disease-a systematic review," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 31, 2023.

[31] I. Vallés-Pérez, G. Beringer, P. Bilinski, G. Cook, and R. Barra-Chicote, "Scraps: Speech contrastive representations of acoustic and phonetic spaces," *arXiv preprint arXiv:2307.12445*, 2023.

[32] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.

[33] N. Dilokthanakul, P. A. M. Mediano, M. Garnelo, M. C. Lee, H. Salimbeni, K. Arulkumaran, and M. Shanahan, "Deep unsupervised clustering with gaussian mixture variational autoencoders," 2017.

[34] M. A. Moreno Bilbao, D. Poig, A. Bonafonte Cávez, E. Lleida, J. Llisterri, J. B. Mariño Acebal, and C. Nadeu Camprubí, "Albayzin speech database: Design of the phonetic corpus," in *EUROSPEECH 1993: 3rd European Conference on Speech Communication and Technology: Berlin, Germany: September 22-25, 1993*. . EUROSPEECH, 1993, pp. 175–178.

[35] J. Mendes-Laureano, J. A. Gómez-García, A. Guerrero-López, E. Luque-Buzo, J. D. Arias-Londoño, F. J. Grandas-Pérez, and J. I. Godino-Llorente, "Neurovoz: a castillian spanish corpus of parkinsonian speech," 2024.

[36] ——, "Neurovoz: a castillian spanish corpus of parkinsonian speech dataset," 2024. [Online]. Available: https://doi.org/10.5281/zenodo.10777657

[37] J. I. Godino-Llorente, N. Saenz-Lechon, V. Osma-Ruiz, S. Aguilera-Navarro, and P. Gómez-Vilda, "An integrated tool for the diagnosis of voice disorders," *Medical Engineering & Physics*, vol. 28, no. 3, pp. 276–289, 2006.

[38] E. J. Vajda, "Handbook of the international phonetic association: A guide to the use of the international phonetic alphabet," *Language*, vol. 76, p. 928, 2000.

[39] E. Wilbanks, "faseAlign (version 1.1.14)," https://github.com/EricWilbanks/faseAlign, 2022, [Computer software]. [Online]. Available: https://github.com/EricWilbanks/faseAlign

[40] S. Young, G. Evermann, M. Gales, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, "The htk book version 3.4 manual," *Cambridge University Engineering Department, Cambridge, UK*, 2006.

[41] B. McFee, C. Raffel, D. Liang, D. P. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "Librosa: Audio and music signal analysis

in python," in *Proceedings of the 14th Python in Science Conference*, vol. 8, 2015, pp. 18–25.
[42] H. E. Pérez, "Frecuencia de fonemas," *Revista Electrónica de la Red Temática en Tecnologías del Habla*, vol. 1, 2003.
[43] N. Dilokthanakul, P. A. Mediano, M. Garnelo, M. C. Lee, H. Salimbeni, K. Arulkumaran, and M. Shanahan, "Deep unsupervised clustering with gaussian mixture variational autoencoders," *arXiv preprint arXiv:1611.02648*, 2016.
[44] J. A. Figueroa, "Semi-supervised learning using deep generative models and auxiliary tasks," in *NeurIPS Workshop on Bayesian Deep Learning*, 2019.
[45] H. Oh Song, Y. Xiang, S. Jegelka, and S. Savarese, "Deep metric learning via lifted structured feature embedding," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4004–4012.
[46] D. M. Blei, A. Kucukelbir, and J. D. McAuliffe, "Variational inference: A review for statisticians," *Journal of the American Statistical Association*, vol. 112, no. 518, pp. 859–877, 2017.
[47] E. Jang, S. Gu, and B. Poole, "Categorical reparameterization with gumbel-softmax," *arXiv preprint arXiv:1611.01144*, 2016.
[48] E. J. Ibarra, J. D. Arias-Londoño, M. Zañartu, and J. I. Godino-Llorente, "Towards a corpus (and language)-independent screening of parkinson's disease from voice and speech through domain adaptation," *Bioengineering*, vol. 10, no. 11, p. 1316, 2023.
[49] J. Lin, "Divergence measures based on the shannon entropy," *IEEE Transactions on Information Theory*, vol. 37, no. 1, pp. 145–151, 1991.
[50] F. Pérez-Cruz, "Kullback-leibler divergence estimation of continuous distributions," in *2008 IEEE International Symposium on Information Theory*. IEEE, 2008, pp. 1666–1670.
[51] A. Favaro, Y.-T. Tsai, A. Butala, T. Thebaud, J. Villalba, N. Dehak, and L. Moro-Velázquez, "Interpretable speech features vs. dnn embeddings: What to use in the automatic assessment of parkinson's disease in multi-lingual scenarios," *Computers in Biology and Medicine*, vol. 166, p. 107559, 2023.
[52] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, vol. 33, pp. 12 449–12 460, 2020.
[53] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "HuBERT: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.

**Alejandro Guerrero-López** received the B.Sc. degree in Telematics Engineering from Universitat de les Illes Balears (UIB), Mallorca, Spain, in 2019. He received the M.Sc. in Information Health Engineering from Universidad Carlos III de Madrid (UC3M) , Spain, in 2020 and the dual Ph.D degree in Probabilistic Machine Learning from UC3M and Universidad Rey Juan Carlos (URJC), Madrid, Spain, in 2023. Since 2023, he is Postdoctoral Researcher at Universidad Politécnica de Madrid (UPM), Madrid, Spain, with the Department of Signal, Systems and RadioCommunication, ETSIT-UPM. His research interests include probabilistic machine learning, generative models and their application to biomedical problems.

**Julián D. Arias-Londoño** (Senior Member, IEEE) received the B.S. and the M.Eng. degrees from Universidad Nacional de Colombia (UNAL), Manizales, Colombia, in 2005 and 2007, respectively, and the dual Ph.D. degree in Computer Science and Automatics from Universidad Politécnica de Madrid (UPM), Spain, and UNAL, in 2010. Since 2012, he has been with the Department of Systems Engineering and Computer Science, Universidad de Antioquia (UdeA), Medellín, Colombia, where he has been appointed as a Full Professor since 2020. He is currently a Visiting Researcher with the Signals, Systems and Radiocommunications Dept., ETSIT-UPM, funded by a María Zambrano Grant. He is part of the Intelligence Information Systems Laboratory (UdeA) and of the Applications of Signal Processing Group (UPM). His research interests include the areas of computational intelligence, machine learning, and signal processing applied to biomedical and biological data analysis. Since 2022, he has been a member of the European Laboratory for Learning and Intelligent Systems (ELLIS) Society and of its Madrid unit.

**Stefanie Shattuck-Hufnagel** is a Principal Investigator in the Research Laboratory of Electronics (RLE) at the Massachusetts Institute of Technology (MIT). She received her B.A. from Wellesley College in 1965 and her Ph.D. in psychology from MIT in 1975. She was Assistant Professor of Psychology at Cornell University from 1974 to 1979 before joining RLE in 1980. Dr. Shattuck-Hufnagel investigates the cognitive structures and processes involved in speech production planning, particularly at the level of speech sound sequencing. Her work with speech error patterns and with the acoustic analyses of prosody has implications for cognitive models of speech production and for phonological theory, as well as applications in speech recognition and synthesis. She leads the Speech Communication Group of the MIT.

**Juan Ignacio Godino-Llorente** (Senior Member, IEEE) was born in Madrid, Spain. He received the B.Sc. and M.Sc. degrees in Telecommunications Engineering and the Ph.D. degree in Computer Science from the Universidad Politécnica de Madrid (UPM), Spain, in 1992, 1996, and 2002, respectively. Since 2011, he has been a Full Professor at the Signal Systems and Radiocommunications Department, UPM. He has also been the Spanish Coordinator of the 2103 COST Action funded by the European Science Foundation, and the General Chairman of the Third Advanced Voice Function Assessment Workshop. During his career, he has led more than 20 research projects funded by national or international public bodies and industry. During the academic term 2003–2004, he was a Visiting Professor with Salford University, Manchester, U.K. In 2016, he was a Visiting Researcher with the Massachusetts Institute of Technology, Cambridge, MA, USA, funded by a Fulbright grant. He has served as an Editor for the Speech Communication Journal, the EURASIP Journal of Advances in Signal Processing, the IEEE Journal of Selected Topics in Signal Processing (JSTSP), and the IEEE Transactions on Audio, Speech, and Language Processing (TASLP). Since 2021, he has been a member of the European Laboratory for Learning and Intelligent Systems (ELLIS) Society and of its Madrid unit.