A preliminary study of Clinical Plausability for Primary Progressive Aphasia Spanish Speakers through Shallow Learning and Acoustic Features eGeMAPS

Enrique A. de la Cal[1], Jimuel Jr. Celeste[2], Mashrura Tasnim[2], Amable J. Valdés[1], Eleni Stroulia[2], and Elena Herrera[3]

[1] University of Oviedo, Computer Science Department,
Faculty of Geology, Oviedo, Spain
UO232486@uniovi.es, delacal@uniovi.es
[2] Computing Dpt.,
Faculty of Science,
University of Alberta,
Edmonton, Canada {jimueljr,mashrura,stroulia}@ualberta.ca
[3] University of Oviedo, Psychology Department,
Faculty of Psychology, Oviedo, Spain
herreraelena@uniovi.es

Abstract. Primary Progressive Aphasia (PPA) is a neurodegenerative disorder characterized by a progressive decline in language abilities. It is clinically categorized into three variants: non-fluent/agrammatic (nfvPPA), semantic (svPPA), and logopenic (lvPPA). While functional MRI remains the diagnostic gold standard, its cost and limited accessibility hinder widespread early screening, motivating the development of alternative, speech-based approaches using acoustic and linguistic features.
This study presents a variant-specific classification pipeline designed for Spanish-speaking patients, a population largely underrepresented in prior research. 12 participants from a clinical trial dataset were selected, each performing an structured cognitve test. The pipeline consists of three binary classifiers—one per variant—each trained to distinguish the corresponding PPA subtype from healthy controls. We evaluated three representative shallow learning methods with different hyperparameter configurations, including support vector machines (SVM), Random Forests (RF), and feedforward neural networks (FNN), all trained using eGeMAPS acoustic features. A leave-one-speaker-out (LOSO) validation strategy ensured speaker-independent evaluation.
Results show that classifiers for nfvPPA and lvPPA achieved F1-scores between 94% and 87%, outperforming the svPPA classifier (70%). And finally, and at global level a clinical plausibility analysis was conducted, correborating that the model performance of the three healthy-variant followed the expected acoustic severity gradient (nfvPPA $>$ lvPPA $>$ svPPA) with a robust statistically significant difference of the performance (Kruscal-Wallis+Dunn's post-hoc tests). These findings support the feasibility of an acoustic-based pipeline for early, variant-sensitive

PPA screening in Spanish, contributing to the development of scalable, clinically informed diagnostic tools.

Keywords: Intelligent Speech Analysis, Neurodegenerative disorders, Primary Progressive Aphasia, eGeMAPS, Leave One Speaker Out

1 Introduction

Primary Progressive Aphasia (PPA) is a neurodegenerative disorder characterized by the progressive deterioration of language abilities, while other cognitive domains remain relatively preserved in the early stages [8]. Clinically, PPA is classified into three main variants: non-fluent/agrammatic (nfvPPA), semantic (svPPA), and logopenic (lvPPA), each associated with distinct linguistic and anatomical profiles [8, 9].

Although functional magnetic resonance imaging (fMRI) remains a widely accepted gold standard for diagnosis due to its ability to reveal atrophy patterns and functional connectivity disruptions [2], its high cost and limited accessibility. Consequently, there is a growing demand for faster, cost-effective, and more accessible diagnostic alternatives.

Several studies have explored early automated PPA diagnosis using acoustic and linguistic features, ML, and DL techniques [15, 17, 21]. Most existing studies have focused on English-speaking populations with limited datasets, often excluding widely spoken languages such as Spanish [5]. There remains a notable gap in Intelligent Data Analysis research evaluating the clinical plausibility[4] of PPA variant classification, both from an acoustic perspective and the expected holistic impairments across the different variants.

Two clinical trial conducted between 2022–2025 [11, 22] implemented a PPA-specific cognitive protocol including 19 tasks from validated tests (ACE-III [13], MLSE [16], and BETA [7]). This produced a dataset of speech recordings from 20 participants (11 PPA, 9 controls), each contributing roughly one hour of audio. For this study, we selected a balanced subset of 12 participants (2 per PPA variant, 6 controls), ensuring comparable segment distributions.

This study addresses two main objectives: i) to propose a variant-specific classification pipeline, consisting of a set of binary machine learning models trained on the selected 12 participants data; and ii) to conduct a global clinical plausibility analysis, evaluating whether the best-performing models for each PPA variant align with prior expert judgments regarding expected speech impairments.

The structure of the paper is as follows: Section 2 describes the dataset, feature extraction process, and machine learning methods used in the pipeline. Section 3 presents and discusses the model performance results. Section 4 focuses on the clinical plausibility analysis of the models obtained in Section 3, considering the severity impairments. Finally, Section 5 outlines the main conclusions and future research directions.

[4] Clinical plausibility refers to the extent to which a model's outputs align with known clinical patterns and expert expectations.

2 Materials and methods

This section present the subdataset selected from a global PPA datataset collected through the clinical trials CEImPA 2022.459/2023.227 [11, 22], as well as the methods that compose the pipeline proposed to process the data and train the clasification models.

2.1 Dataset Description

Table 1 summarizes demographic and clinical data for 20 participants recruited from the Universitary Asturias Hospital (Spain) as part of a recent PPA-focused clinical trials [11, 22]. The dataset includes 9 healthy controls and 21 patients diagnosed with one of the three PPA variants: nfvPPA, svPPA, or lvPPA, according to the specific MLSE [16] cognitve test diagnostic criteria.

Controls were all male and exhibited high MLSE-GLOBAL (MLSEG) scores (mean = 97.9), while patient groups showed more variability in age, gender, education, and cognitive reserve. In particular, patients tended to be older and included more female participants. As expected, MLSE-GLOBAL scores were lower among patients (mean = 83.1), aligning with clinical severity.

ID	Hospital	Age	Gender	HandD	ScholarS	CognitiveR	MLSEG	MLSED
control1-control9	HUCA	66.6	9×Male	Right	12.4	10.9	97.9	Healthy
pac1-pac11	HUCA	74.7	10×Male, 1×Female	Right	11.2	11.6	83.1	3×lvPPA, 3×nfvPPA, 5×svPPA

Table 1: Dataset of controls and PPA patients with average demographic and cognitive variables.

The voice recordings of the 20 participants were collected as part of a cognitive assessment protocol conducted under the approved clinical trial CEImPA 2022.459 [11]. Each recording session lasted between 30 to 45 minutes and focused exclusively on 19 structured cognitive speech tasks derived from a curated subset of ACE-III, MLSE, BETA, and FREE test batteries. These tasks specifically involved overt voice production and targeted key cognitive-linguistic domains such as naming, fluency, reading, spontaneous speech, and repetition.

Although the combined cognitive battery includes a broader range of tasks, only those requiring spoken responses were selected for this study. Tasks without voice production (e.g., visuospatial or written components) were excluded to ensure uniformity in audio-based cognitive assessment.

For the sake of simplicity, the present study treats all recordings as belonging to a single, generic cognitive task. Task-specific analyses will be addressed in future work.

All recordings were conducted using a Yotto YDM-20 USB microphone connected to a MacBook Pro, captured in mono at a sampling rate of 44.1 kHz.

2.2 Methods

A typical pipeline of intelligent audio aNalysis has been used, composed of the following steps:

Data Preprocessing: denoising and audio segmentation Each raw audio file—containing the full interview with both interviewer and participant speech—was manually timestamped and labeled to extract only the participant's responses for each cognitive task.

The resulting audio segments were preprocessed using all combinations of two denoising methods and two normalization settings. For denoising, we considered two well-established techniques: Spectral Gating and LogMMSE. In both cases, default parameters were selected to ensure conservative noise reduction without distorting speech. For normalization, we applied amplitude normalization targeting an RMS level of –20 dBFS to improve signal comparability while avoiding aggressive peak limiting. The "None" options preserved the raw audio signal.

Several segmentation strategies exist in the literature, including speaker-based, silence-based, punctuation-based, fixed-window, and model-based methods [12, 19, 20, 23]. As a first approach, we adopted pause-based segmentation using known inter-utterance pause durations in Spanish (700 ms) [4, 14].

Each of the six preprocessed versions of the recordings was segmented using pydub.split_on_silence. To optimize RMS thresholds for silence detection, we performed a grid search across seven SNR values (–15 to –50 dBFS), considering the –20 dBFS normalization target. As a fitness function, we evaluated the number and total duration of resulting utterances. In total, 48 configurations were tested from the cross-combination of six preprocessing pipelines and eight threshold levels.

Feature extraction To train the models, we selected the well-established Extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) [6], consisting of 88 acoustic functional features computed using the openSMILE toolkit. In this context, a functional acoustic feature set refers to the extraction of 88 summary features per utterance, where each utterance constitutes one sample in the dataset.

Participants selection and dataset partitioning Six healthy participants (HP) and six PPA individuals, comprising two cases each of the three variants, lvPPA, nfvPPA anbd svPPA. A Leave-One-Speaker-Out (LOSO) cross-validation strategy was applied, forming folds of four participants—three for training and one for testing. Based on the six HP and two participants per PPA variant, all possible 3+1 combinations were generated, yielding 60 folds per PPA variant. The participants have been selected attending to a similar number of samples. So, each fold constitutes a two-class problem.

Model selection, hyperparameter optimization and feature selection Three classical machine learning techniques have been selected belonging to three of the most representative ML approaches, RandomForest (RF), Support Vector Classifier (SVC), and FeedForward Neural Networks (FFNN) from the Sklearn tool-kit. And three configurations have been selected for the execution of the three models: i) baseline, ii) hyperparameters tuning and ii) hyperparameters tuning with anova selection (See table 2).

Model	Setup	Hyperparameters
SVC	Baseline	MinMax scaling; default: $C = 1.0$, kernel=`rbf`, γ = `scale`, max_iter=-1
	Tuning	$C \in \{0.1, 1, 10, 100\}$, $\gamma \in \{1, 0.1, 0.01, 0.001, \texttt{scale}\}$, max_iter $\in \{50, 100, 150\}$
	Tuning + ANOVA	Same as above with ANOVA feature selection (`selector_k` $\in \{80\%, 85\%, 90\%, 95\%\}$ of features)
Random Forest	Baseline	No scaling; default: $n\_estimators = 100$, max_depth=None
	Tuning	$n\_estimators \in \{50, 100, 150\}$, max_depth $\in \{10, 20, 30, \texttt{None}\}$, min_samples_split $\in \{2, 5, 10\}$, min_samples_leaf $\in \{1, 2, 4\}$
	Tuning + ANOVA	Same as above with ANOVA feature selection (`selector_k` $\in \{80\%, 85\%, 90\%, 95\%\}$ of features)
FFNN	Baseline	MinMax scaling; default: 1 hidden layer, 64 units, dropout=0.2, learning_rate=1e-3, epochs=100, batch_size=32
	Tuning	dense_units $\in \{64, 128, 176, 256, 512\}$, dropout_rate $\in \{0.1, 0.2, 0.3\}$, learning_rate $\in \{10^{-2}, 10^{-3}, 10^{-4}\}$
	Tuning + ANOVA	Same as above with ANOVA feature selection (`selector_k` $\in \{80\%, 85\%, 90\%, 95\%\}$ of features)

Table 2: Hyperparameter configurations for SVC, Random Forest, and FFNN across three experimental setups.

Evaluation Metrics All the models have been trained using accuracy as a training metric, while the performance of each model has been shown f1-score, accuracy, precision and recall.

3 Results and Performance Analysis

The results section is arranged in three parts: i) the results and decision about the selection of the preprocessing and utterances split silence configuration, ii) the selection of participants, and iii) the comparison of the performance of the trained models.

3.1 Selection of utterances split configuration

Figure 1 contains the left) number of utterances and right) average duration of utterances by participant obtained after the denoising, normalization and the

utterance split processes with the different configurations. It can be stated that the two configurations spectral_gating and logmmse with amplitude normalization get the higher number of utterances (blue square), but the logmmse gets a closer average duration to the real average duration in spanish of a utterance, -30dBFS (blue bar = uttersplit_thres_30). So from now on the dataset configuration used is Logmmse denoising, Amplitude normalization and Silence threshold of -30dBFS.

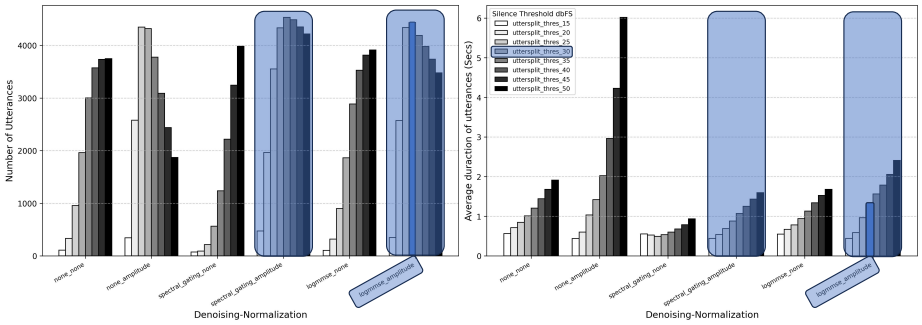

Fig. 1: left) Avg. Duration of utterances by participant, right) total number of utterances)

3.2 Selection of participants

In order to select a homogenius group of participants, we have depicted the average number of utterances per participant (see Figure 2). So, healthy participants pac2 to pac8 have been selected (yellow), and the three pairs of ppa, pac11/pac2 (orange), pac6/pac7 (blue), and pac8/pac9 (green).

3.3 Comparison of the performace of the models

Figure 3 (left) shows the F1-score performance of the different model-configuration combinations across 60 folds for each PPA variant. Overall, the Random Forest (RF) approach outperforms the other models in the majority of configurations and across all PPA variants. In addition, it is evident that the best-performing configuration for the svPPA variant achieves clearly lower F1-scores compared to the best results obtained for lvPPA and nfvPPA.

To statistically assess the differences between model configurations within each PPA variant, we applied a paired t-test or Wilcoxon signed-rank test, depending on the normality of the F1-score differences (verified using the Shapiro-Wilk test). A Bonferroni correction was applied to adjust for multiple comparisons. Each configuration was compared against the best-performing configuration for its respective variant (highlighted in bold in Figure 3 (right)).

The significance of each comparison is indicated in the figure using the following symbols: * for $p < 0.05$, ** for $p < 0.01$, and *** for $p < 0.001$. The

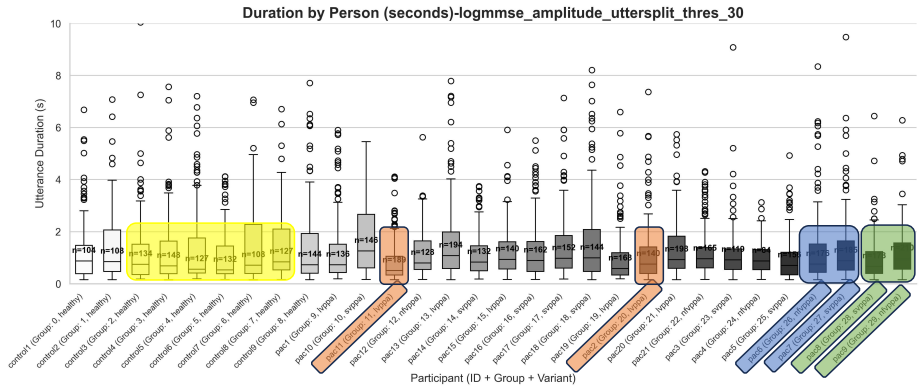

Fig. 2: Number of utterances by participant and boxplot of the durations in second by participant

corrected p-values led to rejection of the null hypothesis in all configurations except those based on Random Forest, across all variants. This indicates that all FNN and SVM configurations (except for SVM with hyperparameter tuning) are statistically significantly worse than the corresponding RF baseline.

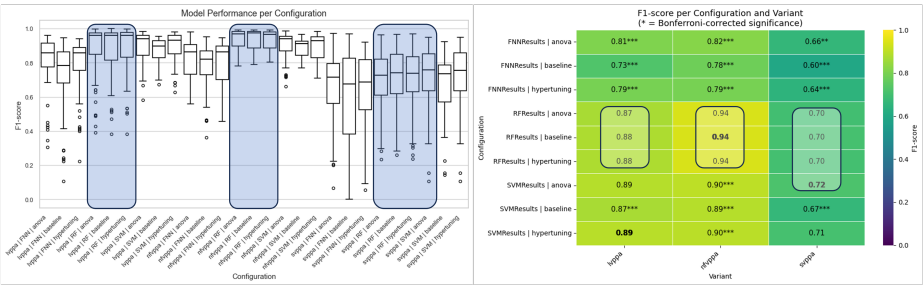

Fig. 3: left) Boxplot with the F1 of all the configurations, right) significance study with ttest/wilcoxon

Based on both predictive performance and statistical significance, we conclude that the Random Forest configurations consistently represent the best-performing models across the three PPA variants. Since no statistically significant differences were found among the RF configurations themselves (i.e., absence of *, **, or *** in the post-hoc comparison), any of these configurations may be selected as the final model without compromising performance (see table 3).

Table 3: Performance metrics (accuracy, F1-score, precision, recall) of the best Random Forest configurations for each PPA variant. CI 95%

Variant	Accuracy	F1-score	Precision	Recall
lvPPA	0.8122 ± 0.0558	0.8777 ± 0.0415	1.0000 ± 0.0000	0.8122 ± 0.0558
nfvPPA	0.8923 ± 0.0277	0.9395 ± 0.0535	1.0000 ± 0.0000	0.8923 ± 0.0277
svPPA	0.5725 ± 0.0535	0.7048 ± 0.0472	1.0000 ± 0.0000	0.5725 ± 0.0535

4 Holistic clinical plausibility analysis

We define holistic clinical plausibility as the extent to which the performance of the trained classifiers aligns with the degree of acoustic impairment associated with each PPA variant, considering a comprehensive assessment of speech beyond specific cognitive tasks. This concept integrates model accuracy with the global severity of acoustic deficits in speech production for each variant —the higher the severity, the higher the classifier's performance—, thus offering a clinically grounded validation framework.

4.1 Acoustic differences between PPA and Healthy Speech

Considering only acoustic speech characteristics derived from the eGeMAPS feature set, which includes prosodic, spectral, temporal, and voice quality transforms, the degree of similarity between the three PPA variants and the healthy speech varies drastically (see table 4).

Patients with semantic variant PPA (svPPA) typically exhibit fluent speech with preserved articulation, prosody, and voice quality, despite deep semantic impairments. This results in an acoustic profile that closely resembles that of healthy speakers, particularly in prosodic and spectral domains [1].

In logopenic variant PPA (lvPPA), speech is marked by increased pausing and hesitation due to word-finding difficulties, which introduces greater variability in temporal and prosodic features, though articulation and voice quality remain relatively unaffected [10, 18].

In contrast, non-fluent/agrammatic variant PPA (nfvPPA) is characterized by effortful, dysfluent speech, often accompanied by apraxia of speech and phonetic distortions, leading to degraded pitch contours, increased jitter and shimmer, and altered speech timing—all of which are directly reflected in eGeMAPS features [3].

Based on this acoustic profile, the variants can be ordered from most to least acoustically similar to healthy controls as: svPPA $>$ lvPPA $>$ nfvPPA (see table 4). This gradient supports the clinical plausibility of our model's ability to distinguish PPA variants using only acoustic features.

Accordingly, we expect the binary classification performance to follow the clinical severity of acoustic deviations: models targeting nfvPPA should yield the highest accuracy, followed by lvPPA, while svPPA—being acoustically more similar to healthy speech—may pose greater challenges for accurate classification.

Table 4: Qualitative comparison of speech similarity between PPA and healthy speech

Variant	Acoustic Characteristics	Severity	ML Perf.
svPPA	Fluent, stable prosody, normal articulation and voice quality	Low	Low
lvPPA	Some pauses, increased hesitation, but prosody/articulation mostly intact	Moderate	Medium
nfvPPA	Reduced speech rate, distorted prosody, voice harshness, apraxia effects	High	High

4.2 Numerical results

In order to analize the order relation between the performance of the best models obtained for each PPA variant, we need to guarantee that the results are comparable. So we have followed the next steps:

- Assess the normality of F1 score distributions using the Shapiro-Wilk test.
- Based on the results, select either a one-way ANOVA for normally distributed data or the Kruskal-Wallis test for non-parametric analysis.
- These tests evaluate whether statistically significant differences existed among group means or medians. If p-value is lower to 0.05, the conclusion is that the at least one model differs from the others in terms of the central tendency (median)
- Finally, following a significant global result (global p-value < 0.05), we applied Dunn's post-hoc test with Bonferroni correction to identify specific pairwise differences while controlling for multiple comparisons (Dunn's p-value < 0.05).

And the final results, shows that after the normality shapiro test, the Kruskall-Wallis was calculated sporting a p-value of 0.000 (see Table 4), that correborates the at least one model differs from the other two respect to the median. And finally, to check that all three models, have statistically significant difference respect to the others, the Dunn's post-hoc test with Bonferroni got the results of table 5. It can be stated that the three algorithm performance can be compared. So, these findings are consistent with, and further support, the severity gradient among PPA variants reported in the literature. Table 6 presents the models ordered by F1-score. The results indicate that the performance of the variant-specific classifiers correlates with the severity of acoustic impairment, following a descending gradient from high to low severity.

5 Conclusion and Future Work

This work presents a ppa variant-specific classification pipeline for the detection of Primary Progressive Aphasia (PPA) variants in Spanish-speaking patients using shallow learning techniques applied to acoustic features derived from structured cognitive tasks. RF, FFNN and SVC approaches were trained using

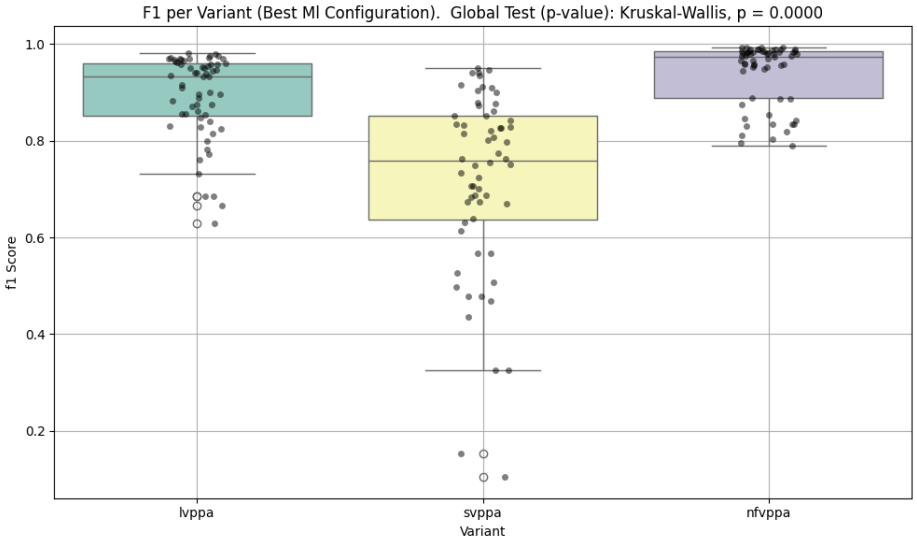

Fig. 4: Global clinical plausability computed with the Kruskal-Wallis test and the Dunn's post-hoc test with Bonferroni correction

Table 5: Global test (Kruskal–Wallis) and pairwise post-hoc comparison (Dunn + Bonferroni)

	lvppa	**nfvppa**	**svppa**
lvppa	1.000000	0.001849	0.0000002
nfvppa	0.001849	1.000000	4.25×10^{-18}
svppa	0.0000002	4.25×10^{-18}	1.000000

Global test (Kruskal–Wallis): statistic = 78.6353, p-value = 0.0000

eGeMAPS features and evaluated through a rigorous LOSO cross-validation protocol, showing that Random Forest consistently outperforms other models across all variants.

The clinical plausibility of the best RF model was validated an acoustic perspective, and the model performance order, aligned well with severity gradient among the variants—nfvPPA > lvPPA > svPPA—, reinforcing the relevance of eGeMAPS features in capturing clinically meaningful signals.

Despite promising results, this study also highlights key limitations. First, the small sample size restricts generalizability and statistical power. Second, the variability in utterance counts and task durations across participants may introduce bias into performance estimates.

Future work will address these limitations by expanding the dataset to include more participants and exploring automatic speaker diarization and transcription pipelines for more scalable annotation. Additionally, task-aware multiclass classifiers will be introduced to directly model the PPA subtype in a uni-

Table 6: Ordered list of models by F1-score for each of the PPA variant. CI 95%

Order	Variant	Configuration	F1-score	Clinical Severity impairment
1	nfvPPA	RF | baseline	0.9395 ± 0.0535	High
2	lvPPA	RF | baseline	0.8777 ± 0.0415	Medium
3	svPPA	RF | baseline	0.7048 ± 0.0472	Low

fied architecture. We also aim to incorporate additional linguistic and syntactic features and evaluate the potential of transformer-based acoustic encoders. Ultimately, this line of research contributes to the development of interpretable, language-specific AI tools to support earlier, non-invasive diagnosis of neurodegenerative language disorders.

6 Acknowledgement

This research was supported by Enrique's Santander Bank Foundation Excellence International Fellowship Grants 2023 (PPA SSRg-GMM proposal), by Spanish Missions Science and Innovation call under project MIG-20211008 (INMERBOT consortium), by the Spanish Ministry of Economic Affairs and Industry grant MCINN-24-PID2023-146257OB-I00, and the Foundation for the Promotion of Applied Scientific Research and Technology in Asturias (FICYT) under the GRUPIN program with grant SEK-25-GRU-GIC-24-055, and by Stroulia's NSERC Discovery grant RGPIN-2020-05033 ("Smart indoor Spaces (SiS): Towards an Integrated Framework for the Internet of Things in the Built Environment"), Tasnim's Alberta Graduate Excellence Scholarship and GRA Rice Graduate Scholarship in Communications.

References

1. Ash, S., McMillan, C., Gunawardena, D., Avants, B., Morgan, B., Khan, A., Grossman, M.: Speech errors in progressive non-fluent aphasia. Brain and Language **109**(1), 22–29 (2009)
2. Bonner, M.F., Ash, S., Grossman, M.: The new neuroscience of semantic memory. Annual Review of Psychology **67**, 381–407 (2016). https://doi.org/10.1146/annurev-psych-122414-033715
3. Cordella, C., Bertinetto, P.M., Riccadonna, S., Vignando, M., Ferrari, C., Catricalà, E., Cappa, S.F.: Automated speech analysis for the assessment of patients with predementia and Alzheimer's disease. Alzheimer's & Dementia **15**(7), 1063–1073 (2019)
4. Cuartero, S., Zeller, M., Shea, C.: Silent Pauses in the Speech of Monolingual and Heritage Speakers of Spanish. Languages **8**(3), 173 (2023). https://doi.org/10.3390/languages8030173
5. Cuetos, F., Arango-Lasprilla, J.C., Valencia, C.: Linguistic and cognitive assessment in Spanish-speaking patients with primary progressive aphasia. Journal of Neurolinguistics **59**, 100997 (2021). https://doi.org/10.1016/j.jneuroling.2021.100997

6. Eyben, F., Scherer, K.R., Schuller, B.W., Sundberg, J., André, E., Busso, C., Devillers, L., Epps, J., Laukka, P., Narayanan, S.S., Truong, K.P.: The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. IEEE Transactions on Affective Computing **7**(2), 190–202 (2016)
7. F., Vega, C., González Nosti, M.: Beta, batería para la evaluación de los trastornos afásicos, vol. 29. Instituto de Orientación Psicológica EOS., Madrid, instituto edn. (2009)
8. Gorno-Tempini, M.L., Hillis, A.E., Weintraub, S., Kertesz, A., Mendez, M., Cappa, S.F., Ogar, J.M., Rohrer, J.D., Black, S., Boeve, B.F., Manes, F., Dronkers, N.F., Vandenberghe, R., Rascovsky, K., Patterson, K., Miller, B.L., Knopman, D.S., Hodges, J.R., Mesulam, M.M., Grossman, M.: Classification of primary progressive aphasia and its variants. Neurology **76**(11), 1006–1014 (2011). https://doi.org/10.1212/WNL.0b013e31821103e6
9. Gorno-Tempini, M.L., Dronkers, N.F., Rankin, K.P., Ogar, J.M., Phengrasamy, L., Rosen, H.J., Johnson, J.K., Weiner, M.W., Miller, B.L.: Cognition and anatomy in three variants of primary progressive aphasia. Annals of Neurology **55**(3), 335–346 (2004). https://doi.org/10.1002/ANA.10825
10. Gorno-Tempini, M.L., Hillis, A.E., Weintraub, S., Kertesz, A., Mendez, M., Cappa, S.F., Ogar, J.M., Rohrer, J.D., Black, S., Boeve, B.F., others: Classification of primary progressive aphasia and its variants. Neurology **71**(16), 1220–1226 (2008)
11. Herrera, E.: Evolución neuropsicológica de las tres variantes de la Afasia Progresiva Primaria (CEImPA 2022.459). Tech. rep. (2022)
12. Hirschberg, J., Litman, D.: Empirical studies on the disambiguation of cue phrases. Computational Linguistics **19**(3), 501–530 (1993)
13. Hsieh, S., Schubert, S., Hoon, C., Mioshi, E., Hodges, J.R.: Validation of the Addenbrooke's Cognitive Examination III in frontotemporal dementia and Alzheimer's disease. Dementia and geriatric cognitive disorders **36**(3-4), 242–250 (2013). https://doi.org/10.1159/000351671
14. Llisterri, J.: Methodological Issues in the Study of Suprasegmentals. https://joaquimllisterri.cat/phonetics/LBASS_21/ (2021)
15. Merhbene, G., Lecron, F., Fortemps, P., Dickerson, B.C., Kurpicz-briki, M., Rezaii, N.: Detecting Primary Progressive Aphasia (PPA) from Text : A Benchmarking Study (Ml) (2025)
16. Patel, N., Peterson, K.A., Ingram, R.U., Storey, I., Cappa, S.F., Catricala, E., Halai, A., Patterson, K.E., Lambon Ralph, M.A., Rowe, J.B., Garrard, P.: A 'Mini Linguistic State Examination' to classify primary progressive aphasia. Brain Communications **4**(2) (2022). https://doi.org/10.1093/braincomms/fcab299
17. Rezaii, N., Hochberg, D., Quimby, M., Wong, B., Brickhouse, M., Touroutoglou, A., Dickerson, B.C., Wolff, P.: Artificial intelligence classifies primary progressive aphasia from connected speech. Brain **147**(9), 3070–3082 (2024). https://doi.org/10.1093/BRAIN/AWAE196
18. Rohrer, J.D., Ridgway, G.R., Crutch, S.J., Hailstone, J., Goll, J.C., Clarkson, M.J., Mead, S., Beck, J., Cipolotti, L., Houlden, H., others: Progression of regional grey matter atrophy in non-fluent aphasia and semantic dementia. NeuroImage **49**(1), 669–676 (2010)
19. Tavafi, M.A., Fukuda, T.: A speech segmentation method using silence detection and Gaussian mixture models. In: 2013 IEEE International Conference on Mechatronics and Automation. pp. 540–545. IEEE (2013)
20. Terpstra, C., Khebour, I., Bradford, M., Wisniewski, B., Krishnaswamy, N., Blanchard, N.: How Good is Automatic Segmentation as a Multimodal Discourse Anno-

tation Aid? In: Proceedings of the 19th Joint ACL-ISO Workshop on Interoperable Semantics (ISA-19). pp. 75–81 (2023)
21. Tetzloff, K.A., Wiepert, D., Botha, H., Duffy, J.R., Clark, H.M., Whitwell, J.L., Josephs, K.A., Utianski, R.L.: Automatic Speech Recognition in Primary Progressive Apraxia of Speech. Journal of Speech, Language, and Hearing Research **67**(9), 2964–2976 (2024)
22. Valdés, A.J., de la Cal, E., Herrera, E.: AI Techniques Based on Oral Production for Automatic Support in the Differential Diagnosis of Primary Progressive Aphasia. Approved Clinical Trial (6 2023)
23. Yang, S., Zhang, Z., Jiang, Y., Qin, C., Liu, S.: A Unified Supervised and Unsupervised Dialogue Topic Segmentation Framework Based on Utterance Pair Modeling. In: Proceedings of the 2025 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 4898–4908 (2025)