TecnoLógicas

Unsupervised Analysis of Clinical Variables and Speech-Related Features in Patients with Functional Oropharyngeal Dysphagia

Análisis no supervisado de variables clínicas y características de habla en pacientes con disfagia orofaríngea funcional

Sebastián Roldán-Vasco[1]; Juan Salvador Mercado-Villegas[1]; Juan Pablo Restrepo-Uribe[1]

[1]Institución Universitaria ITM, Medellín-Colombia.

Correspondence: sebastianroldan@itm.edu.co

Received: 01 October 2025 **Accepted:** 22 September 2026 **Available:** 25 September 2026

How to cite / Cómo citar

S. Roldán-Vasco, J. S. Mercado-Villegas, and J. P. Restrepo-Uribe, "Unsupervised Analysis of Clinical Variables and Speech-Related Features in Patients with Functional Oropharyngeal Dysphagia," *TecnoLógicas*, vol. 29, no. 67, e3679, 2026. https://doi.org/10.22430/22565337.3679

CC BY NC SA

Abstract

Dysphagia is a syndrome related to swallowing impairment, which is a complex process that involves precise neuromuscular coordination and shares anatomical structures with the speech production process. Subsequently, the presence of dysphagia has been associated with speech alterations. The purpose of this work was to explore speech changes in patients with functional oropharyngeal dysphagia, before and after swallowing tasks. The recordings of 77 patients were analyzed and hand-crafted features were extracted from dimensions of phonation, articulation, and prosody. Feature selection was performed using AUC-ROC and hypothesis tests. k-Means, agglomerative clustering, and Gaussian-Mixture Models were used for clustering in the reduced feature space, using ten binarized clinical variables as comparison scenarios. The quality of the clusters was evaluated with the silhouette score, the Davies-Bouldin Index (BDI), and the Adjusted-Rank Index (ARI) for each scenario. The results showed good intra-cluster quality (silhouette ≥ 0.7 and BDI ≤ 0.2) for cough after swallowing, alteration in carrying the lips to the sides, and altered tongue strength, but only in speech recordings post-swallowing. Otherwise, a poor inter-quality cluster was found for both pre and post-swallowing tasks in all clinical variables, and patients with the less frequent condition tended to be grouped in the larger cluster. Additionally, the spatial distribution of the pre and post-swallowing features changed considerably. These findings suggest that swallowing tasks influence speech production in patients with dysphagia and that unsupervised clustering-based analysis of speech features could provide complementary insights for clinical assessment. More studies with larger and more balanced cohorts are needed to validate these results and explore their potential diagnostic applications.

Keywords

Clustering, dysphagia, speech processing, swallowing, unsupervised learning.

Resumen

La disfagia es un síndrome asociado a la alteración de la deglución, un proceso complejo que requiere una coordinación neuromuscular precisa y que comparte estructuras anatómicas con la producción del habla. De esta manera, la presencia de disfagia se ha relacionado con alteraciones del habla. El objetivo de este trabajo fue explorar los cambios en el habla de pacientes con disfagia orofaríngea funcional, antes y después de realizar tareas de deglución. Se analizaron grabaciones de 77 pacientes, y se extrajeron características en las dimensiones de fonación, articulación y prosodia. La selección de características se realizó mediante AUC-ROC y pruebas de hipótesis. Para el agrupamiento en el espacio de características reducido se emplearon k-Means, clustering aglomerativo y Modelos de Mezclas Gaussianas, utilizando diez variables clínicas binarizadas como escenarios de comparación. La calidad de los grupos se evaluó mediante el índice de silueta, el Índice de Davies-Bouldin (DBI por sus siglas en inglés) y el Índice de Rand Ajustado (ARI por sus siglas en inglés) para cada escenario. Los resultados mostraron una buena calidad intra-grupo (silueta ≥ 0.7 y DBI ≤ 0.2) para tos después de la deglución, alteración en llevar los labios hacia los lados y disminución de la fuerza lingual, pero únicamente en las grabaciones posteriores a la deglución. En contraste, se observó una baja calidad inter-grupo tanto en las tareas pre como post-deglución para todas las variables clínicas, y los pacientes con la condición menos frecuente tendieron a agruparse en el grupo más grande. Por otro lado, la distribución espacial de las características pre y post-deglución cambió de manera considerable. Estos hallazgos sugieren que las tareas de deglución influyen en la producción del habla en pacientes con disfagia, y que el análisis clúster no supervisado de características del habla podría aportar información complementaria para la evaluación clínica. Se requieren estudios adicionales con cohortes más grandes y balanceadas para validar estos resultados y explorar sus posibles aplicaciones diagnósticas.

Palabras clave

Clustering, disfagia, procesamiento de voz, deglución, aprendizaje no supervisado.

1. INTRODUCTION

1.1 Swallowing and dysphagia

Swallowing is a complex process that involves the passage of food and liquids from the mouth to the stomach. This process mainly involves six cranial nerves and the muscles of the head and neck [1]. Oral, pharyngeal, and esophageal are the three stages into which this procedure is typically separated. During the oral stage, food is brought to the mouth, chewed, and formed into a bolus before being forced into the pharynx by the tongue. The bolus travels through the throat during the pharyngeal stage, and the airway is closed as a preventative measure against aspiration. The food bolus finally travels into the esophagus during the esophageal phase. When peristalsis takes place, the esophageal sphincter opens, allowing the bolus to pass into the stomach [2].

Alterations in the normal swallowing process can lead to the development of a disorder known as dysphagia [1]. Depending on the location of the difficulty in performing the normal swallowing process, dysphagia can be classified as oral, pharyngeal (commonly called oropharyngeal dysphagia), or esophageal. Oropharyngeal dysphagia is a symptom that can occur as a result of various medical conditions, such as stroke, Parkinson's disease, and Alzheimer's disease, among others [3], [4]. It can cause serious health problems for patients, mainly associated with malnutrition, dehydration, bronchial aspiration, and even death [5].

One of the main drawbacks of dysphagia is related to its diagnosis. Among the techniques used is the EAT-10, which consists of a 10-question self-perception questionnaire, as well as the bedside test, where a swallowing specialist analyzes and issues a diagnosis that depends largely on the evaluator's experience. In addition, there are instrumental tests such as the Fiberoptic Endoscopic Evaluation of Swallowing (FEES), where an endoscope is introduced to evaluate the anatomical structures and the swallowing process. This method is highly invasive, causing discomfort to patients, and sometimes requires the use of anesthesia. Videofluoroscopic Swallow Study (VFSS) is also used as an instrumental diagnostic method, which is considered the gold standard for the diagnosis of dysphagia, but this test subjects the patient to X-rays, making it an invasive method and limited in the evolution of patients over time.

1.2 Speech for swallowing impairment assessment

Speech and voice recordings address the challenges associated with invasiveness, subjectivity, accessibility, and reliability of the swallowing evaluation techniques. Speech allows the evaluation of the acoustic dimension of swallowing [6]. Nevertheless, speech-language pathologists have traditionally relied on perceptual assessment of speech and voice quality as part of clinical swallowing evaluations, which is problematic when the signs are subtle and demand high professional expertise [7].

Although extensive research has been conducted on dysphagia and speech impairments as separate conditions, few studies have examined the degree of correlation between swallow dysfunction and voice changes [8]. An exception is the work of Festic et al. [9], which demonstrated a significant association between dysphagia and diadochokinetic performance in individuals with confirmed aspiration. Similarly, Park et al. [10] reported that speech features, when analyzed using conventional machine learning algorithms, can be used to predict both the severity of dysphagia and the likelihood of respiratory complications in post-stroke patients. In the same year, Zhao et al. [11] applied machine learning methods to identify dysphagia by utilizing traditional voice-related parameters collected from a throat-mounted vibrator, rather than microphones. Furthermore, Hwang et al. [12] reported a framework to detect dysphagia using self-supervised speech representation learning models with a pipeline that involves mel spectrograms, Wav2Vec 2.0, HuBERT, or WavLM models, achieving AUC scores up to 0.86 for aspiration detection using pre- and post-swallowing tasks. Kim et al. [13], applied aspiration detection models using Multi-Layer Perceptrons with the deep learning architecture MobileNet V3. They achieved an AUC of 0.8361 using a combination of pre-trained models of males and females. Roldan-Vasco et al. [14], [15] also reported that speech recordings are suitable for dysphagia detection, achieving classification scores up to 90% and 95% with good balance between sensitivity and specificity using classical machine learning methods and Gated-multimodal units, respectively. However, screening between individuals with and without dysphagia using speech recordings represents a markedly different challenge than distinguishing between dysphagia states in affected patients. While the former involves a binary classification task (healthy individuals vs. dysphagic patients), the latter requires sensitivity to subtle, within-speaker variations that reflect fluctuations in swallowing physiology or severity. This distinction is critical, as it highlights the need for analytical approaches capable not only of identifying broad pathological markers but also of capturing acoustic biomarkers that may shift dynamically across states. In this regard, and to the best of our knowledge, there is only one work that has addressed the spatial differences in acoustic biomarkers. Sarraf Shirazi et al. [16] evaluated the average power of three spectral sub-bands of breath sound signals acquired with a throat microphone in 21 dysphagic individuals. Using fuzzy k-means clustering algorithms, they applied a pipeline to detect aspiration vs. no-aspiration recordings with good results (86.5% accuracy). Nevertheless, breath sounds contain less information than speech, and it is limited to detect aspirations instead of subtle changes according to the dysphagia progression.

1.3 Contributions of this work

The goal of this work was to explore speech changes in patients with functional oropharyngeal dysphagia, before and after swallowing tasks, and this was addressed by asking the following open questions in the field of dysphagia research: 1) Are there voice changes associated with swallowing dysfunction? and 2) can speech recordings be used for screening of swallowing dysfunction? These questions motivated the following two scenarios of analysis, one oriented to determine changes in features extracted from speech recordings before and after swallowing tasks, and another focused on finding differences between features from two subsets according to dysphagia severity and other clinical variables.

2. METHODOLOGY

2.1 Patients characterization

A part of the MedSwallowDB database was used, as introduced in [15]. Information of 77 patients was analyzed (46 male and 31 female), all of them with diagnosed functional oropharyngeal dysphagia produced by neurological causes, i.e., neurogenic or neuromuscular. The etiologies included amyotrophic lateral sclerosis (n=11), multiple sclerosis (n=9), Parkinson's disease (n=17), stroke (n=7), multiple system atrophy (n=1), progressive supranuclear palsy (n=3), cerebral palsy (n=5), right recurrent laryngeal nerve neuropathy (n=1), ischemic stroke (n=3), autosomal recessive hereditary ataxia (n=3), secondary hydrocephalus (n=1), traumatic brain injury (n=4), muscular dystrophy (n=5), dermatomyositis (n=2), cognitive disorder (n=1), frontotemporal dementia (n=1), and motor neuron disease (n=3). All of them were Colombian Spanish native speakers with an average age of 56.8 ± 16.7 years old. Demographic and clinical information are provided in Table 1. Participants were excluded if they presented with esophageal or mechanical dysphagia, radiation-induced skin changes affecting the facial or cervical regions, orofacial or cervical edema or hematoma, previous surgical dissection involving the cervical skin, severe hypoxemia (SpO_2 < 80%), or a Deep Brain Stimulation implant. Written informed consent was obtained from all participants in accordance with the principles of the Declaration of Helsinki. The study protocol was reviewed and approved by the Institutional Ethics Committee of the Universidad Pontificia Bolivariana, Medellín, Colombia (Approval No. 07.01.2017).

Table 1. Data distribution for each clinical variable. Age is given as mean ± standard deviation M: male. F: female. Source: own elaboration.

Variable	Labels	Sex [M/F]	Age
EAT-10	EAT10≤16	25/16	57.0 ± 14.3
	EAT10>16	21/15	58.5 ± 17.3
Cough with yogurt	Yes	5/8	56.8 +15.8
	No	41/23	61.8+15.2
Cough before swallowing	Yes	7/5	61.2+14.2
	No	39/26	57.0+16.0
Intubation for more than one week	Yes	7/2	53.6+20.0
	No	39/29	58.2+15.1
Respiratory comorbidity	Yes	7/8	59.7+18.6
	No	39/23	57.2+15.0
Lingual fasciculations	Yes	13/7	61.9+14.2
	No	33/24	56.2+16.0
Multiple cookie swallows	Yes	28/22	58.2+15.2
	No	18/9	56.7+16.9
Mandibular nerve (V3) - sensory portion	Normal	37/28	58.2+15.6
	Impaired	9/3	55.0+16.7
Alteration in carrying lips to the sides	Yes	16/5	55.5+18.4
	No	30/26	58.5+14.6
Altered tongue strength	Normal	38/26	56.7+15.8
	Impaired	8/5	62.4+14.9

2.2 Speech and voice tasks

Speech and voice recordings from the MedSwallowDB were collected before (PRE) and after (POST) a series of swallowing tasks, which involved the intake of thin and thick liquids, a dry swallow, and a solid consistency. The MedSwallowDB database is available upon request, subject to certain academic restrictions. Further details of this protocol are provided in [15]. The following speech tasks were chosen for analysis:

- Sustained vowels \a\, \e\, \i\, \o\, and \u\, uttered during at least 3 seconds
- Diadochokinetic task: rapid repetition of the syllables pa-ta-ka
- Read text: a phonetically balanced text consisting of all Spanish sounds spoken in Colombia was read by the subjects. The text is as follows [17]: *Ayer fui al médico. ¿Qué le pasa? Me preguntó. Yo le dije: ay doctor, donde pongo el dedo me duele. - ¿Tiene la uña rota? - Sí. - Pues ya sabemos qué es. Deje su cheque a la salida* [Translation: Yesterday I went to the doctor. What happened to you? He asked me. I told him: ah doctor! Where I put my finger, it pains. - Do you have a broken nail? - Yes. - Then we now know what is happening. Leave your check at the exit.

2.3 Speech features

Features in four different speech dimensions were assessed: phonation, articulation, diadochokinesia and prosody [18]. An adapted version of the Neurospeech [18] and Praat [19] were used for feature extraction. Both tools are open-source software for linguistic analysis. The sliding window method was implemented with frames of 40 ms and 50% overlapping [18]. The extracted features are summarized below (see [14] for details):

- Phonation: fundamental frequency (F_0), Jitter, Shimmer, Amplitude Perturbation Quotient (APQ), Pitch Perturbation Quotient (PPQ), and the energy of the signal. For APQ and PPQ, the frame size was 150 ms.
- Articulation in sustained vowels: formants (F_1 and F_2), their first and second derivatives (ΔF_1, ΔF_2, $\Delta^2 F_1$, and $\Delta^2 F_2$), the Teager-Kaiser Energy Operator, 12 Mel-frequency cepstral coefficients (MFCC1~MFCC12) and their first and second derivatives, the vowel space area (VSA), vowel articulation index (VAI), and the formant centralization ratio (FCR).
- Diadochokinesia: F_0, logarithmic energy, syllable rate (syllable vocalizations per second - DDK), and pauses rate.
- Prosody: F_0, logarithmic energy, and duration of voiced and unvoiced segments.

Features such as F_0 and the logarithmic energy vary of dimension depending on the speech task performed.

2.4 Dimensionality reduction

The feature space includes 1126 features extracted from pre- and post-swallowing speech signals. Due to this high dimensionality when compared with the number of patients, a feature reduction method was employed based on the Mann-Whitney statistical test and the Area Under the Receiver Operating Characteristic -ROC- curve (AUC). Features were selected based on two criteria: a p-value smaller than 0.05 and an AUC value greater than 0.65. Regarding the AUC, the discriminatory capability of each feature was evaluated against the labels obtained from the binarization of the EAT-10 score and the other clinical variables described in Section 2.1. The Mann-Whitney test is a non-parametric statistical test used to evaluate the medians of unpaired data [20], [21]. On the other hand, the AUC is based on the ROC curve, which allows us to represent the relationship between the True Positive Rate (TPR) and the False Positive Rate (FPR) at different decision thresholds. The AUC can be interpreted as the probability that

the model assigns a higher score to a random positive observation than to a random negative observation [22], [23], [24].

The t-SNE algorithm was used for dimensionality reduction for visualization purposes. It seeks to preserve local proximity relationships between samples. To achieve this, similarities are modeled in a high-dimensional space by defining the conditional probability $p_{i|j}$ for each pair of points x_i and x_j. This allows to identify how similar is x_j to x_i according to a Gaussian distribution centered on x_i. This is achieved through the symmetric joint probability, according to ((1).

$$p_{ij} = \frac{p_{j|i} + p_{i|j}}{2n} \tag{1}$$

Subsequently, low-dimensional spaces are created by placing points y_i, and a probability distribution is defined using a Student t-test with one degree of freedom, as described in ((2).

$$q_{ij} = \frac{(1 + \|y_i - y_j\|^2)^{-1}}{\sum_{k \neq l}(1 + \|y_k - y_l\|^2)^{-1}} \tag{2}$$

Once the probability functions $P = p_{ij}$ and $Q = q_{ij}$ are obtained, the Kullback-Leibler divergence (KL) between both probabilities is minimized, as defined in (3). To achieve this, a gradient descent is implemented that adjusts the positions of y_i in a low-dimensional space.

$$\min_{\mathrm{Q}} (KL(P||Q)) = \min_{q_{ij}} \left(\sum_{i \neq j} p_{ij} log \frac{p_{ij}}{q_{ij}} \right) \tag{3}$$

2.5 Cluster analysis

2.5.1 Lateral deflection of composite piles

Clustering is the process of grouping data objects using a similarity measure [25]. Three types of unsupervised clustering algorithms were implemented in this work: k-Means, Gaussian Mixture Models, and Agglomerative Clustering. Analyses were performed in Python 3.11 using the scikit-learn library with a fixed random seed to ensure reproducibility. Missing values were imputed by the median, and features were scaled using RobustScaler, which performs a statistics-based scaling according to the first and third quartiles, making it more robust to a small number of large outliers. Clustering models included K-Means (k=2, k-means++ initialization), Gaussian Mixture Model (two components, initialized using k-means), and Agglomerative Clustering (k=2, Ward linkage with Euclidean distance). For visualization, a dimensionality reduction was conducted using t-SNE (two-dimensional embedding, perplexity=30, 1000 optimization iterations) and PCA (two components).

k-Means

k-Means groups data points according to how far apart they are from the cluster centroids. k-Means uses the squared Euclidean distance for cluster membership [25]. The number of clusters k is obtained by minimizing the sum of the squared error (SSE), defined in ((4), of each point x_i to its nearest centroid:

$$SSE = \sum_{i=1}^{n} \sum_{j=1}^{k} w_{ij} \|x_i c_j\|^2 \tag{4}$$

where c_j is the centroid of the j-th cluster, and w_{ij} is the binary assignment variable defined in ((5).

$$w_{ij} = \begin{cases} 0, & \text{if } x_i \text{ in cluster } j \\ 1, & \text{otherwise} \end{cases} \tag{5}$$

k-Means has some limitations: it cannot handle uncertainty when points are near multiple centroids, struggles with complex or non-linear cluster boundaries, and is sensitive to the initial guess of centroids [25].

Gaussian Mixture Models (GMM)

GMM uses a probabilistic method to allocate data points to clusters, with each cluster being represented by a unique Gaussian distribution [25]. It assumes data as a mixture of K Gaussian distributions. Each component k is parameterized by a mean vector μ_k, covariance matrix Σ_k, and a mixing coefficient π_k, where $\sum_{k=1}^{K} \pi_k = 1$ and $\pi_k \geq 0$.

The probability density function of the multivariate Gaussian distribution for a data point $x \in R^d$ is defined in ((6).

$$N(x|\mu_k, \Sigma_k) = \frac{1}{(2\pi)^{\frac{d}{2}}|\Sigma_k|^{\frac{1}{2}}} exp\left(-\frac{1}{2}(x-\mu_k)^T \sum_{k}^{-1}(x-\mu_k)\right) \tag{6}$$

The overall probability of a data point under the mixture model is computed using (7).

$$p(x) = \sum_{k=1}^{K} \pi_k N(x|\mu_k, \Sigma_k) \tag{7}$$

GMM employs the Expectation-Maximization algorithm, which alternates between estimating the posterior probabilities of data points belonging to each component (E-step) and updating the parameters of the Gaussian components—means, covariances, and mixing coefficients (M-step). This iterative process continues until convergence, typically defined by minimal change in log-likelihood. Unlike k-Means, GMM can model clusters with elliptical shapes and overlapping boundaries, making it more suitable for complex data distributions [26].

Agglomerative clustering

Agglomerative clustering is a type of hierarchical clustering, in which each pattern starts in its own separate cluster (singleton) and merges the clusters together one by one until a stopping condition is met [27]. In this work, the Ward linkage criterion was used, minimizing the variance of the clusters being merged (within-cluster variance). The Ward linkage method also uses the squared Euclidean distance [28].

2.5.2 Validation of clusters

Internal evaluation

Two clusters ($k = 2$) were chosen according to the labels generated by the clinical variables analyzed in this work (see Section 2.6). The intra-cluster quality was evaluated with the silhouette score and the Davies-Bouldin Index (DBI).

The silhouette score assesses how well data points fit within their assigned clusters compared to those in other clusters. The optimal number of clusters is the one that maximizes this score [29]. It ranges from -1 to 1, where values close to 1 indicate well-clustered

points, around 0 suggest overlapping or ambiguous clusters, and negative values imply possible misclassification [30].

On the other hand, the DBI evaluates the intra-cluster compactness and inter-cluster separation, favoring clusters that are both tight and well-separated. Notwithstanding, this index assumes spherical and equally sized clusters [29].

External evaluation

The Adjusted-Rand Index (ARI) was also used to assess the external quality of the clusters. The ARI has been widely recommended as the main criterion for external clustering assessment [31]. The ARI is an adjusted version of the original Rand Index, which measures the similarity between two clusters based on the proportion of object pairs that are consistently grouped together or apart in both clusters. Rand Index does not consider possible agreement by chance; in this regard, the ARI introduces a chance-corrected version. The ARI may take negative values for extremely discordant clusterings [31].

2.6 Voice changes associated with clinical screening-based variables

It has been demonstrated that certain clinical variables from the Bedside Clinical Evaluation of Swallowing are correlated with healthy and. dysphagic states for screening [32], [33], [34]. The following clinical variables were chosen for analysis accordingly: cough with yogurt, cough before swallowing, intubation greater than one-week, respiratory comorbidity, lingual fasciculations, piecemeal swallowing of cracker, sensory involvement of the cranial nerve V3, alteration in carrying lips to the sides, altered tongue strength, and the EAT-10 score. All of them were binary labeled as YES or NO, except for the EAT-10 score. Thus, the original dataset was split into two groups of patients, and the corresponding feature spaces were labeled as F^{yes} and F^{no}, respectively. Additionally, feature spaces extracted from recordings collected before and after swallowing tasks were also analyzed, onwards F_{pre} and F_{pos}. In this regard, the following comparisons were performed:

- Differences between patients with and without the clinical characteristic before swallowing, i.e. F_{pre}^{yes} and F_{pre}^{no}.
- Differences between patients with and without the clinical characteristic after swallowing, i.e. F_{pos}^{yes} and F_{pos}^{no}.

Otherwise, the EAT-10 tool validated for Colombian Spanish [35], was implemented to evaluate the dysphagia severity. Since this scale does not have well-defined thresholds for different severity levels, the median value of the computed scores was established. Thus, a similar analysis was performed for this characteristic, but divided into F^{mild} and F^{severe}.

These analyses were performed by hypothesis tests and by the Area Under the Receiver Operating Characteristic Curve (AUC). The hypothesis tests were carried out with the Mann-Whitney U test for unpaired scenarios (F^{yes} and F^{no}, or F^{mild} and F^{severe}), and with the Wilcoxon rank sum test for paired scenarios (F_{pre} vs. F_{pos}).

3 RESULTS AND DISCUSSION

A total of 1126 acoustic features were analyzed to explore potential changes in voice and speech after swallowing. Mann–Whitney U tests were applied to compare each feature between pre- and post-swallowing conditions, considering ten binarized clinical variables as reference labels. Features were retained only if they fulfilled the dual criteria of statistical significance $p < 0.05$ and $\mathrm{AUC} > 0.65$. In the pre-swallowing set, i.e. F_{pre}, 111 features met these conditions, while 100 features were identified in the post-swallowing set. Among these, the mean of $\Delta^2\mathrm{MFCC3}$ from the vowel /i/

in both pre- and post-swallowing tasks showed the highest discriminative power ($\mathrm{AUC} > 0.78$) with the strongest statistical significance ($p < 0.002$). Figure 1 shows the kernel density estimate (KDE) plot of this feature according to the clinical variable "alteration in carrying lips to the sides". The plot reveals clear distributional differences between participants with and without alterations in lateral lip movement. In both pre- and post-swallowing conditions, the distributions for participants with this alteration (Yes) were shifted toward higher values compared to those without it (No), showing only partial overlap between the two groups. Furthermore, it is also noteworthy that ranges of values differed considerably. The features selected using statistical tests and AUC are summarized in the Appendix.

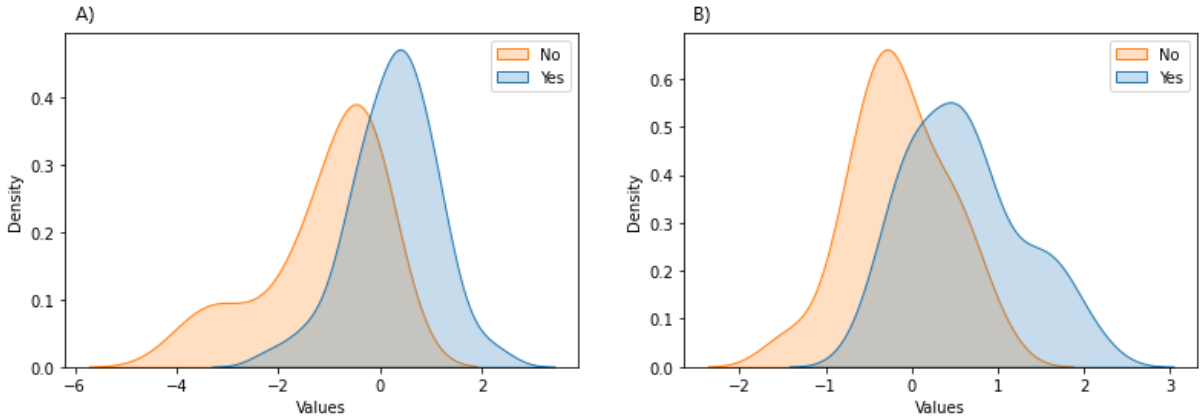


Figure 1. Kernel density estimates of Δ^2 $MFCC3$ from the vowel /i/ in patients with and without alteration in carrying lips to the sides in A) pre- and B) post-swallowing speech tasks
Source: own elaboration.

Otherwise, once features were selected according to Section 3.1, three unsupervised learning algorithms were applied, i.e. k-Means, agglomerative clustering, and GMM. In this regard, the internal and external quality of clusters were assessed, as shown in Table 2. Only one clinical variable retrieved Silhouette score ≥ 0.70 for pre-swallowing taks ($\mathcal{F}_{pre}$), i.e. the behavior of the sensory portion of the mandibular nerve. In post-swallowing tasks ($\mathcal{F}_{post}$), cough before swallowing, alteration in carrying lips to the sides, and alteration in the tongue strength retrieved Silhoutte score greater or equal than 0.70. This result is also confirmed by the Davies-Bouldin metric, which was the smallest in all the mentioned cases. Note that agglomerative clustering was the most consistent algorithm because all the mentioned cases involved such methods in contrast to GMM and k-Means. Since this is an expected behavior when data clusters are irregular or non-convex [36], a relevant inference from the analyzed data is the complex distribution of the feature space, which can make the detection of patterns difficult.

On the other hand, clusters that achieved high Silhouette scores and low Davies-Bouldin indices did not correspond to high ARI values (close to 1), indicating that strong internal quality did not necessarily translate into external validity. In fact, the ARI was negative for the aforementioned cases. The highest values were achieved for lingual fasciculations (k-Means) and multiple swallows of cookie (GMM) in $\mathcal{F}_{post}$, and for altered tongue strength in $\mathcal{F}_{pre}$ (agglomerative and GMM). Notwithstanding, unlike with internal quality metrics, strong conclusions cannot be drawn regarding algorithms. Additionally, no cluster achieved good external quality (the highest ARI was 0.28). In this regard, results indicate that clusters are moderately compact in the feature space, but they are not well separated. Bearing in mind that speech features have been shown to be suitable for dysphagia screening [6], it is clear that there is a well-defined differential pattern between healthy and dysphagic speech recordings; thus, discrepancies between internal and external clustering suggests that, even though speech features may capture an intrinsic functional pattern of dysphagia, they are not fully represented by the conventional clinical variables as the assessed in this research. These

findings also suggest the potential of speech-based clustering as a tool for refining patient stratification and guiding future studies aimed at developing clinically meaningful digital biomarkers.

Table 2. Results of clustering applied to clinically relevant variables. Best results are highlighted in bold
Source: own elaboration.

Clinical variable	Cluster Method	Pre-Swallowing			Post-Swallowing		
		Silhouette	Davies-Bouldin	ARI	Silhouette	Davies-Bouldin	ARI
Cough with yogurt	k-Means	0.25	1.89	-0.08	0.49	1.05	0.06
	Agglomerative	0.34	1.11	-0.08	0.40	1.07	-0.03
	GMM	0.28	1.53	0.05	0.38	1.13	0.01
Cough before swallowing	k-Means	0.36	1.14	-0.06	0.84	0.10	-0.02
	Agglomerative	0.36	1.02	-0.03	0.84	0.10	-0.02
	GMM	0.26	1.72	0.00	0.25	1.93	-0.03
Intubation greater than one week	k-Means	0.62	0.62	0.03	0.29	1.55	0.02
	Agglomerative	0.60	0.69	0.00	0.28	1.69	-0.06
	GMM	0.55	0.78	-0.02	0.31	1.52	-0.02
Respiratory comorbidity	k-Means	0.39	1.30	-0.08	0.66	0.70	-0.10
	Agglomerative	0.36	1.37	-0.05	0.66	0.70	-0.10
	GMM	0.34	1.39	-0.04	0.77	0.35	-0.04
Lingual fasciculations	k-Means	0.39	1.30	-0.07	0.21	1.43	0.24
	Agglomerative	0.43	1.13	-0.09	0.20	1.67	0.08
	GMM	0.32	1.49	-0.04	0.21	1.49	0.17
EAT-10 score	k-Means	0.33	1.00	0.00	0.34	1.09	0.02
	Agglomerative	0.37	0.99	-0.00	0.37	0.96	0.01
	GMM	0.16	2.05	0.06	0.20	1.93	0.27
Multiple cookie swallows	k-Means	0.19	2.11	0.13	0.27	1.67	0.15
	Agglomerative	0.20	2.31	0.07	0.26	1.50	0.11
	GMM	0.18	2.19	0.18	0.15	2.27	0.20
v3 sensory portion	k-Means	0.55	1.23	-0.08	0.34	1.52	0.13
	Agglomerative	0.71	0.20	-0.02	0.30	1.20	0.18
	GMM	0.55	1.23	-0.08	0.08	3.08	0.00
Alteration in carrying lips to the sides	k-Means	0.24	1.60	-0.05	0.78	0.15	-0.02
	Agglomerative	0.20	1.74	-0.06	0.78	0.15	-0.02
	GMM	0.23	1.66	-0.06	0.78	0.15	-0.02
Altered tongue strength	k-Means	0.14	2.24	0.19	0.76	0.16	-0.02
	Agglomerative	0.10	2.53	0.24	0.76	0.16	-0.02
	GMM	0.15	2.01	0.28	0.37	1.85	-0.11

In the case of EAT-10, two population groups called F^{mild} and F^{severe} were evaluated, defined from the binarization of EAT-10. Patients with EAT-10 < 16 were labeled as F^{mild}, while those with EAT-10 > 16 were labeled as F^{severe}. Table 2 shows that the k-Means algorithm exhibited unclear separation between clusters, as indicated by the Silhouette score and DBI, which means that loosely compact clusters in both pre- and post-swallowing conditions are generated. Furthermore, the ARI values (0.0 for pre-swallowing and 0.02 for post-swallowing) suggest that cluster assignment does not reproduce a clear structure for binarized data in mild and severe dysphagia. The behavior of the agglomerative algorithm was similar, with a slight improvement compared to k-Means. Furthermore, the ARI values show that the clusters did not group together within a clear structure for the binarized EAT-10 data. Finally, the GMM exhibited low compactness and limited cluster separation, but a slight increase in the post-

swallowing ARI was achieved (0.26), showing a greater but limited ability to reproduce a clear structure using the membership assignment.

It is noteworthy that the changes observed between pre- and post-swallow were minimal for the k-Means and Agglomerative models, suggesting that the variation between the pre- and post-swallow data was likely small. However, the GMM showed a significant improvement in ARI, increasing from 0.06 to 0.26. This may be related to the fact that probabilistic models such as the GMM are better able to capture the redistribution of clinical patterns, which is not reflected in distance-based methods. Furthermore, feature selection considered statistical difference as a key factor. All three clustering methods showed a slight improvement in post-swallowing results compared to pre-swallowing results. This improvement is especially evident in GMM, which shows a change in the clinical patterns detected through probabilistic approaches.

To observe such behavior, dimensionality reduction was performed with t-SNE with two components. This allows to figure out why the ARI was low for all the evaluated clusters. An example of the t-SNE visualization of patients with and without cough before swallowing is shown in Figures 2A and 2C ($\mathcal{F}_{pre}$ and $\mathcal{F}_{post}$, respectively). Observe that data distribution is quite different between $\mathcal{F}_{pre}$ and $\mathcal{F}_{post}$; this behavior was observed in all clinical variables, for both labels. It could be because the acoustic structure of speech is modulated by the swallowing act [37]. Divergence between point distribution in $\mathcal{F}_{pre}$ and $\mathcal{F}_{post}$ confirms that speech features suffer alterations when swallowing, in this case, after a series of swallowing tasks of thin liquids, thick liquids, and food (a cracker).

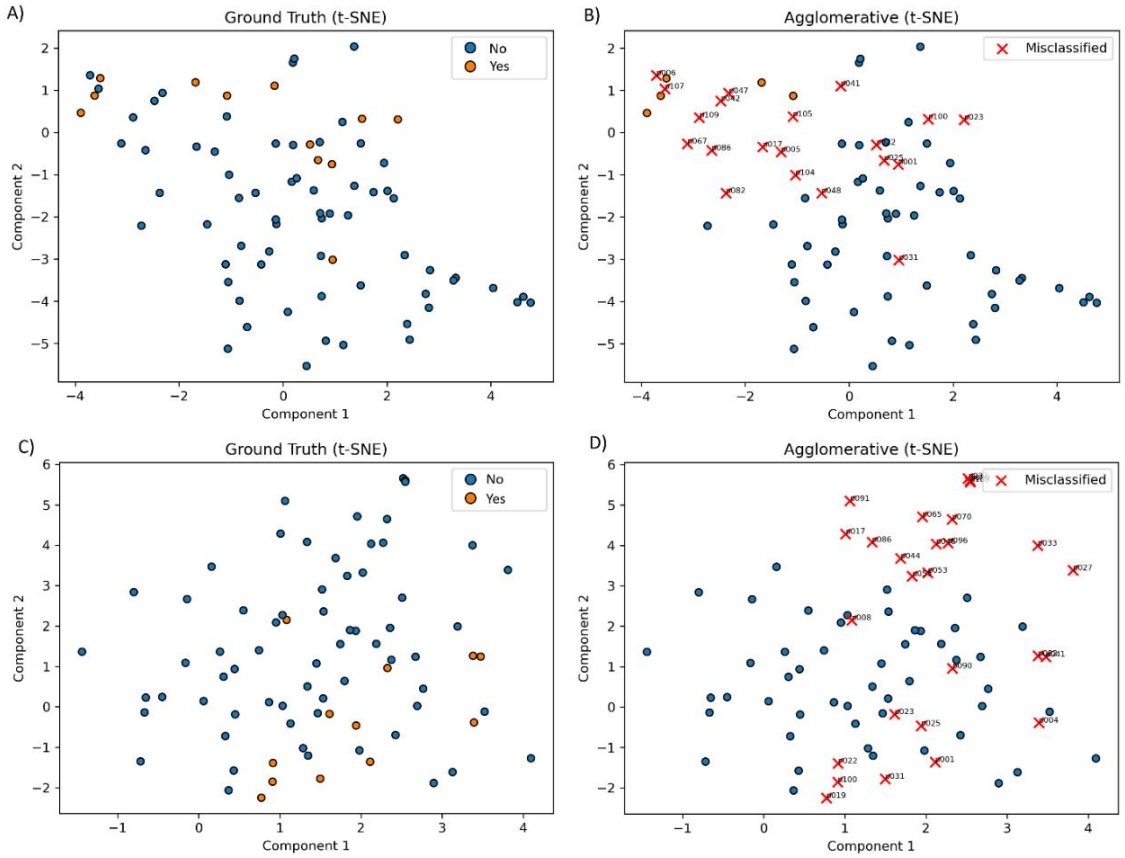


Figure 2. Clustering of patients using a 2D representation with t-SNE, for the clinical variable "cough before swallowing". A) and C) show label assignment for the YES and NO groups, in pre- and post-swallowing speech features, respectively. B) and D) show patients clustered incorrectly in pre- and post-swallowing speech features, respectively. Source: own elaboration.

Another key point is the cluster assignment, which can be seen in Figures 2B and 2D for $\mathcal{F}_{pre}$ and $\mathcal{F}_{post}$, respectively. Observe that almost all patients in the small cluster, i.e., with the condition $\mathcal{F}^{yes}$, were assigned with membership to the large cluster, mostly corresponding to $\mathcal{F}^{no}$. In other words, the distribution of clusters does not correspond to the presence or absence of the clinical condition. It could be explained by overlapping or insufficiently distinctive features that blur cluster boundaries.

Even though well-differentiated speech features have been reported for dysphagia discrimination, i.e. healthy vs. dysphagia [6], [14], [15], or aspiration vs. no-aspiration [12], [13], [16], results of the current work indicate that such differences are not necessarily useful for discrimination between different clinical manifestations in individuals with dysphagia. That is, classification of dysphagia states from speech recordings is inherently more complex than the binary discrimination of dysphagic vs. non-dysphagic speakers. Inter-state differences often manifest as subtle and transient acoustic cues, in contrast to the more robust inter-group contrasts typically observed between healthy and dysphagic populations. This suggests that models trained for diagnostic screening may not generalize well to state-level monitoring, emphasizing the importance of individualized approaches and temporally sensitive feature extraction. Consequently, further work should consider separate methodological frameworks for screening vs. follow-up applications.

On the other hand, the lack of a defined pattern could also be due to the clinical heterogeneity, in which the condition does not form a sharply distinct group. It is not surprising, because of the well-known complexity of the swallowing process impairment and its effect on speech characteristics [38]. Several clinical aspects play a role in speech changes in these individuals. For instance, it has long been believed that noticeable alterations in phonation characteristics following a swallow indicate that food or drink material has entered the laryngeal vestibule and is vibrating on the surface of the vocal folds [39]. However, it cannot be clinically validated, and it has not been correlated with other clinical manifestations. Additionally, because of the variety of vocal evaluation techniques, it has not been possible to reach a consensus on the recommendation of using vocal evaluation as an accurate method for identifying swallowing alterations [8].

3.1 Limitations and future work

The main limitations of this work were the sample size, the high dimensionality of the feature space, and the imbalance across clinical subgroups. Future work should therefore focus on increasing the sample size, addressing label imbalance, and the integration of modern feature selection models to reduce dimensionality and identify features that show stronger distributional differences across conditions (F_{pre} vs. F_{pos}, F^{mild} vs. F^{severe}, or cluster membership assignment). Further research will also explore the implementation of probabilistic clustering models, as these methods yielded the best results, and the development of unsupervised approaches capable of integrating additional clinical or instrumental variables to enrich the latent structure.

4 CONCLUSIONS

The internal structure of the clusters containing clinical variables like coughing before swallowing, lingual fasciculations, multiple cookie swallows, V3 sensory portion, change in carrying the lips to the sides, and altered tongue strength was compact and coherent, as evidenced by silhouette values above 0.70 and Davies–Bouldin indices near 0.0. When assessed using the ARI, external validity was not supported; results were close to zero and occasionally negative, indicating that the clusters were not able to consistently distinguish the pre-defined

clinical categories. While k-Means and GMM had increased variability, agglomerative clustering exhibited a superior ability to handle irregular or non-convex distributions; in particular, GMM significantly improved ARI in the post-swallowing case. The lack of concordance between internal and external quality highlights the complexity of the feature space and the challenge of achieving clear separability. Overall, while fully consistent clusters could not be identified with the current dataset, the results highlight promising directions for refining feature selection and modeling strategies, and they suggest that clustering on speech-related features remains a promising approach for advancing the clinical understanding of dysphagia.

5 ACKNOWLEDGMENTS

This work has been supported by MINCIENCIAS - República de Colombia, research project No. 121077758144, and partially supported by the Instituto Tecnológico Metropolitano (PCI 23207).

6 REFERENCES

[1] K. Shu *et al.*, "Recent advancements and future directions in automatic swallowing analysis via videofluoroscopy: A review," *Comput. Methods Programs Biomed.*, vol. 259, p. 108505, Feb. 2025. https://doi.org/10.1016/j.cmpb.2024.108505

[2] K. C. Wei *et al.*, "The Cortical and Subcortical Neural Control of Swallowing: A Narrative Review," *Dysphagia*, vol. 39, no. 2, pp. 177-197, Apr. 2024. https://doi.org/10.1007/s00455-023-10613-x

[3] B. Zhang *et al.*, "Effects of Non-Pharmacological Interventions on the Swallowing Function of Patients With Post-Stroke Dysphagia: A Systematic Review and Network Meta-Analysis," *J. Oral Rehabilit.*, vol. 52, no. 1. pp. 109-120, Jan. 2025. https://doi.org/10.1111/joor.13901

[4] L. Jia *et al.*, "Early application value of flexible laryngoscope swallowing function assessment in patients after partial laryngectomy," *Sci. Rep.*, vol. 15, no. 1, p. 6040, Feb. 2025. https://doi.org/10.1038/s41598-025-86799-x

[5] X. Liu *et al.*, "Comprehensive review of dysphagia and technological advances in dysphagia food," *Food Res. Int.*, vol. 199, p. 115354, Jan. 2025. https://doi.org/10.1016/j.foodres.2024.115354

[6] S. Roldán Vasco, "Assessment of oropharyngeal dysphagia using multimodal biosignals," Tesis de doctorado, Universidad de Antioquia, Medellín, Colombia, 2023. https://hdl.handle.net/10495/36438

[7] S. K. Daniels *et al.*, "Speech Pathology Reliability for Stroke Swallowing Screening Items," *Dysphagia*, vol. 30, no. 5, pp. 565-570, Oct. 2015. https://doi.org/10.1007/s00455-015-9638-x

[8] K. W. dos Santos *et al.*, "Using Voice Change as an Indicator of Dysphagia: A Systematic Review," *Dysphagia*, vol. 37, no. 4, pp. 736-748, Aug. 2022. https://doi.org/10.1007/s00455-021-10319-y

[9] E. Festic *et al.*, "Novel Bedside Phonetic Evaluation to Identify Dysphagia and Aspiration Risk," *Chest*, vol. 149, no. 3, pp. 649-659, Mar. 2016. https://doi.org/10.1378/chest.15-0789

[10] H. Y. Park *et al.*, "Author Correction: Post-stroke respiratory complications using machine learning with voice features from mobile devices," *Sci. Rep.*, vol. 12, no. 1, p. 21764, Dec. 2022. https://doi.org/10.1038/s41598-022-26224-9

[11] H. Zhao *et al.*, "Dysphagia diagnosis system with integrated speech analysis from throat vibration," *Expert Syst. Appl.*, vol. 204, p. 117496, Oct. 2022. https://doi.org/10.1016/j.eswa.2022.117496

[12] I. Hwang *et al.*, "Voice-Based Dysphagia Detection : Leveraging Self-Supervised Speech Representation," *Proc. Interspeech*, pp. 5683-5687, Aug. 2025. https://doi.org/10.21437/Interspeech.2025-761

[13] J. M. Kim *et al.*, "Prediction of dysphagia aspiration through machine learning-based analysis of patients' postprandial voices," *J. Neuroeng. Rehabil.*, vol. 21, no. 1, p. 43, Mar. 2024. https://doi.org/10.1186/s12984-024-01329-6

[14] S. Roldan-Vasco *et al.*, "Machine learning based analysis of speech dimensions in functional oropharyngeal dysphagia," *Comput. Methods Programs Biomed.*, vol. 208, p. 106248, Sep. 2021. https://doi.org/10.1016/j.cmpb.2021.106248

[15] S. Roldan-Vasco *et al.*, "Dysphagia screening with sEMG, accelerometry and speech: Multimodal machine and deep learning approaches," *Biomed. Signal Process. Control*, vol. 100, no. part B, p.

107030, Feb. 2025. https://doi.org/10.1016/j.bspc.2024.107030
[16] S. Sarraf Shirazi *et al.*, "Detection of swallows with silent aspiration using swallowing and breath sound analysis," *Med. Biol. Eng. Comput.*, vol. 50, pp. 1261-1268, Dec. 2012. https://doi.org/10.1007/s11517-012-0958-9
[17] J. R. Orozco-Arroyave *et al.*, "New Spanish speech corpus database for the analysis of people suffering from Parkinson's disease," in *Proc. 9th Int. Conf. Lang. Resour. Eval. (LREC'14)*, Reykjavik, Iceland, European Language Resources Association (ELRA), 2014, pp. 342-347. https://aclanthology.org/L14-1549/
[18] J. R. Orozco-arroyave *et al.*, "NeuroSpeech : An open-source software for Parkinson's speech analysis," *Digit. Signal Process.*, vol. 77, pp. 207-221, Jun. 2018. https://doi.org/10.1016/j.dsp.2017.07.004
[19] P. Boersma, and V. van Heuven, "Speak and unSpeak with Praat," *Glot Int.*, vol. 5, no. 9-10, pp. 341-347, Nov.-Dec. 2001. https://www.fon.hum.uva.nl/paul/papers/speakUnspeakPraat_glot2001.pdf
[20] D. S. Kerby, "The Simple Difference Formula: An Approach to Teaching Nonparametric Correlation," *Compr. Psychol.*, vol. 3, p. 11.IT.3.1, Jan. 2014. https://journals.sagepub.com/doi/10.2466/11.IT.3.1
[21] P. K. Sen, "On the Estimation of Relative Potency in Dilution (-Direct) Assays by Distribution-Free Methods," *Biometrics*, vol. 19, no. 4, p. 532, Dec. 1963. https://doi.org/10.2307/2527532
[22] J. A. Hanley, and B. J. McNeil, "The meaning and use of the area under a receiver operating characteristic (ROC) curve," *Radiology*, vol. 143, no. 1. pp. 29-36, Apr. 1982. https://doi.org/10.1148/radiology.143.1.7063747
[23] S. D. Walter, "The partial area under the summary ROC curve," *Stat. Med.*, vol. 24, no. 13, pp. 2025-2040, Jul. 2005. https://doi.org/10.1002/sim.2103
[24] A. P. Bradley, "The use of the area under the ROC curve in the evaluation of machine learning algorithms," *Pattern Recognit.*, vol. 30, no. 7, pp. 1145-1159, Jul. 1997. https://doi.org/10.1016/S0031-3203(96)00142-2
[25] E. Patel, and D. Singh Kushwaha, "Clustering Cloud Workloads: K-Means vs Gaussian Mixture Model," *Procedia Comput. Sci.*, vol. 171, no. 2019, pp. 158-167, 2020. https://doi.org/10.1016/j.procs.2020.04.017
[26] D. Reynolds, "Gaussian mixture models," in *Encyclopedia of biometrics*, S. Z. Li, and A. K. Jain, Eds., Boston, MA, USA: Springer, 2015, pp. 827-832. https://doi.org/10.1007/978-1-4899-7488-4_196
[27] A. K. Jain *et al.*, "Data clustering: a review," *ACM Comput. Surv.*, vol. 31, no. 3, pp. 264-323, Sep. 1999. https://doi.org/10.1145/331499.331504
[28] S. Miyamoto, *Theory of Agglomerative Hierarchical Clustering*. Singapore: Springer, 2022. https://doi.org/10.1007/978-981-19-0420-2
[29] H. Yin *et al.*, "A Rapid Review of Clustering Algorithms," *arXiv*, pp. 1-25, Jan. 2024. https://doi.org/10.48550/arXiv.2401.07389
[30] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *J. Comput. Appl. Math.*, vol. 20, pp. 53-65, Nov. 1987. https://doi.org/10.1016/0377-0427(87)90125-7
[31] J. E. Chacón, and A. I. Rastrojo, "Minimum adjusted Rand index for two clusterings of a given size," *Adv. Data Anal. Classif.*, vol. 17, no. 1, pp. 125-133, Mar. 2023. https://doi.org/10.1007/s11634-022-00491-w
[32] J. C. Suárez Escudero *et al.*, "Model based on clinical characteristics to identify patients with neurogenic oropharyngeal dysphagia," *Rev. Cuid.*, vol. 15, no. 3, p. e386, Sep. 2024. https://doi.org/10.15649/cuidarte.3861
[33] J. Suárez-Escudero *et al.*, "Modelo de integración entre variables clínicas y bioseñales de deglución computacional en disfagia orofaríngea neurogénica," *Rev. Chilena Fonoaud.*, vol. 24. pp. 1-14, Apr. 2025. https://doi.org/10.5354/0719-4692.2025.73503
[34] J. C. Suárez-Escudero *et al.*, "Findings in orofacial praxis in clinical swallow examination in patients with neurogenic oropharyngeal dysphagia," *Neurol. Perspect.*, vol. 5, no. 3, p. 100191, Jul. 2025. https://doi.org/10.1016/j.neurop.2025.100191
[35] L. F. Giraldo-Cadavid *et al.*, "Validation of the Spanish Version of the Eating Assessment Tool-10 (EAT-10spa) in Colombia. A Blinded Prospective Cohort Study," *Dysphagia*, vol. 31, no. 3, pp. 398-406, Jun. 2016. https://doi.org/10.1007/s00455-016-9690-1
[36] J. Choo *et al.*, "MOSAIC : A Proximity Graph Approach for Agglomerative Clustering," In *Data Warehousing and Knowledge Discovery. DaWaK 2007. Lecture Notes in Computer Science, vol 4654*. I. Y. Song *et al.*, Eds. Springer, Berlin, Heidelberg, 2007, pp. 231-240. https://doi.org/10.1007/978-3-540-74553-2_21
[37] Y. A. Kang *et al.*, "Detection of voice changes due to aspiration via acoustic voice analysis," *Auris*

Nasus Larynx, vol. 45, no. 4, pp. 801-806, Aug. 2018. https://doi.org/10.1016/j.anl.2017.10.007
[38] A. Miles, "Evaluation and Treatment of Swallowing Problems," in *Brain Injury Medicine, 3rd Edition. Principles and Practice*. A. Zasler *et al.*, Eds., New York, USA: Springer Publishing Company, 2021, ch. 68, pp. 1040-1047. https://doi.org/10.1891/9780826143051.0068
[39] S. Murugappan *et al.*, "Acoustic characteristics of phonation in 'wet voice' conditions," *J. Acoust. Soc. Am.*, vol. 127, no. 4, pp. 2578-2589, Apr. 2010. https://doi.org/10.1121/1.3308478

CONFLICT OF INTEREST

None declared.

AUTHORSHIP CONTRIBUTION

Sebastián Roldán-Vasco: Conceptualization, Design, Development of the research, as well as the Writing and Revision of the manuscript.

Juan Salvador Mercado-Villegas: Results and Writing.

Juan Pablo Restrepo-Uribe: Analysis of results, Manuscript writing.

IA USAGE STATEMENT

During the conduct of this research, the authors used Grammarly for language editing, including improvements in grammar, wording, and clarity. After using this tool/service, the authors carefully reviewed and edited the content as necessary and assume full responsibility for the content of the publication.