

UNIVERSIDAD TÉCNICA FEDERICO SANTA MARÍA
DEPARTAMENTO DE ELECTRÓNICA
VALPARAÍSO - CHILE



**“REDUCCIÓN DE DIMENSIÓN Y CLUSTERING
NO SUPERVISADO PARA EL ANÁLISIS DE
SEÑALES DE VOZ AMBULATORIAS CAPTADAS
POR ACELERÓMETRO”**

GABRIEL DÍAZ

**MEMORIA DE TITULACIÓN PARA OPTAR AL TÍTULO DE INGENIERO
CIVIL ELECTRÓNICO**

PROFESOR GUÍA:

JUAN PABLO CORTÉS

PROFESOR CORREFERENTE:

MATÍAS ZAÑARTU



CONSTANCIA DE VALIDACIÓN Y CONFIDENCIALIDAD DE MONOGRAFÍA A REPOSITORIO ACADÉMICO

1.- IDENTIFICACIÓN DEL TRABAJO ACADÉMICO

Tipo de monografía (marcar una opción): ☒ Memoria o trabajo de título ☐ Tesis de Postgrado

Título del trabajo: Reducción de dimensión y clustering no supervisado para el análisis de señales de voz ambulatorias captadas por acelerómetro

Nombre del candidato(a): Gabriel Andrés Díaz Lizana

Carrera / Grado: Ingeniería Civil Electrónica

Campus: Casa Central

Departamento: Electrónica

2.- VALIDACIÓN DEL PROFESOR GUÍA/DIRECTOR DE TESIS

Yo, Matías Zañartu, en mi calidad de profesor(a) guía/director(a) del trabajo académico mencionado anteriormente

DEJO CONSTANCIA que:

- He revisado esta versión del documento y corresponde a la versión final aprobada del trabajo.
- El trabajo cumple con los requisitos académicos y de formato establecidos por la institución.

3.- EVALUACIÓN DE CONFIDENCIALIDAD POR PROPIEDAD INDUSTRIAL (marcar una opción)

☒ El trabajo **NO contiene** información que amerite confidencialidad y puede ser publicado de inmediato en repositorio con acceso abierto.

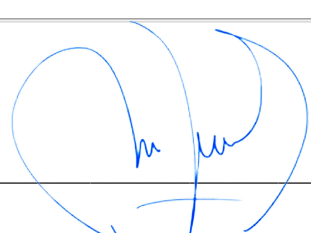
☐ El trabajo **CONTIENE** información con potenciales implicancias de propiedad industrial o intelectual y requiere un periodo de confidencialidad (**embargo**) por (**marcar una opción**):

☐ 6 meses ☐ 12 meses ☐ 2 años ☐ 3 años ☐ 5 años ☐ 10 años

Fundamentación de la necesidad de confidencialidad (obligatorio si se solicita embargo):

4.- FIRMAS

Profesor(a) guía o director(a) de memoria o tesis:

Fecha: 20/10/2025 Firma: 

Estudiante o Candidato(a):

Fecha: 14/10/2025

Firma: 

Este formulario debe ser insertado como página 2 de la memoria o tesis, completado y firmado por estudiante y profesor(a) antes de la entrega en portal PRISMA de Biblioteca USM.

Resumen

Los trastornos de la voz asociados a hiperfunción vocal representan un desafío diagnóstico y terapéutico debido a la ausencia de marcadores objetivos estables y a la variabilidad del comportamiento vocal diario. En este contexto, el monitoreo ambulatorio con acelerometría laríngea permite registrar señales de fonación durante la vida cotidiana, generando grandes volúmenes de datos que requieren nuevas estrategias de análisis.

El presente trabajo propone un pipeline no supervisado para caracterizar el comportamiento vocal ambulatorio en mujeres con patología fonotraumática y sus controles, combinando reducción de dimensión mediante UMAP y clusterización con HDBSCAN. Se utilizaron registros de 13 pares de sujetos (pacientes y controles) obtenidos en el Massachusetts General Hospital, segmentados en ventanas de 50 ms y subsecuencias de 1 s. La calidad del embedding y de los agrupamientos se evaluó con métricas internas (DBCV, Silhouette, Davies-Bouldin) y de preservación espacial-temporal.

Los resultados revelan una estructura latente estable que distingue tres patrones principales: (i) un clúster basal asociado a fonación modal cotidiana, (ii) un clúster vinculado a voz cantada y (iii) un clúster minoritario con flujos glóticos anómalamente altos, posiblemente atribuibles a artefactos de modelado. La comparación entre condiciones mostró redistribuciones moderadas tras la terapia vocal, así como la emergencia de patrones adicionales al aumentar la escala temporal de análisis. Asimismo, los estudios de caso individuales evidenciaron la capacidad del método para detectar eventos vocales transitorios de interés clínico.

Estos hallazgos demuestran que la combinación UMAP + HDBSCAN constituye una herramienta robusta y escalable para explorar millones de ventanas de datos de voz ambulatoria, aportando bases objetivas para la caracterización de patrones vocales, el monitoreo terapéutico y la detección temprana de comportamientos hiperfuncionales en la vida diaria.

Abstract

Voice disorders associated with vocal hyperfunction represent a diagnostic and therapeutic challenge due to the lack of stable objective markers and the high variability of daily vocal behavior. In this context, ambulatory monitoring using a neck-surface accelerometer enables long-term phonation recording in everyday life, generating large volumes of data that demand advanced analysis strategies.

This work proposes an unsupervised pipeline to characterize ambulatory vocal behavior in women with phonotraumatic pathology and matched controls, combining dimensionality reduction through UMAP with clustering via HDBSCAN. Data from 13 subject pairs (patients and controls) collected at Massachusetts General Hospital were analyzed, segmented into 50-ms windows and 1-s subsequences. The quality of embeddings and clustering was evaluated using internal metrics (DBCV, Silhouette, Davies-Bouldin) as well as spatial-temporal preservation indices.

Results revealed a stable latent structure distinguishing three main patterns: (i) a basal cluster associated with everyday modal phonation, (ii) a cluster strongly linked to singing voice, and (iii) a minority cluster with abnormally high glottal airflow values, likely attributable to modeling artifacts. Comparisons across clinical conditions showed moderate redistributions following voice therapy, while longer temporal windows allowed the emergence of additional phonatory patterns. Case studies further demonstrated the method’s ability to detect transient vocal events of clinical interest.

These findings demonstrate that the combination of UMAP and HDBSCAN provides a robust and scalable framework for exploring millions of windows of ambulatory voice data, offering objective foundations for the characterization of vocal patterns, therapy monitoring, and early detection of hyperfunctional behaviors in daily life.

Glosario

ACFL Amplitud del Flujo Glótico (AC Flow)

AMV Monitoreo Ambulatorio de la Voz (Ambulatory Voice Monitoring)

CPP Prominencia del Pico Cepstral (Cepstral Peak Prominence)

DBCV Validación de Clustering Basado en Densidad (Density-Based Clustering Validation)

DTW Algoritmo de Distancia Temporal Dinámica (Dynamic Time Warping)

f_0 Frecuencia Fundamental (Fundamental Frequency)

H1-H2 Diferencia entre el Primer y Segundo Armónico (First and Second Harmonic Difference)

HDBSCAN Clustering Espacial de Aplicaciones con Ruido Basado en Densidad Jerárquica (Hierarchical Density-Based Spatial Clustering of Applications with Noise)

IBIF Filtrado Inverso Basado en Impedancia Subglotal (Impedance-Based Inverse Filtering)

LH Ratio Relación Espacial Bajas-Altas Frecuencias (Low-High Spectral Ratio)

MFDR Tasa Máxima de Declinación del Flujo (Maximum Flow Declination Rate)

MGH Hospital General de Massachusetts (Massachusetts General Hospital)

NAQ Cociente de Amplitud Normalizado (Normalized Amplitude Quotient)

NPVH Hiperfunción Vocal No Fonotraumática (Non-Phonotraumatic Vocal Hyperfunction)

OQ Cociente de Abertura Glótica (Open Quotient)

PCA Análisis de Componentes Principales (Principal Component Analysis)

Ps Presión Subglótica (Subglottal Pressure)

PVH Hiperfunción Vocal Fonotraumática (Phonotraumatic Vocal Hyperfunction)

RMS Valor Cuadrático Medio (Root Mean Square)

SPL Nivel de Presión Sonora (Sound Pressure Level)

t-SNE Embedding de Vecinos Estocásticos Distribuidos en t (t-Distributed Stochastic Neighbor Embedding)

UMAP Aproximación y Proyección Uniforme de Variedades (Uniform Manifold Approximation and Projection)

VH Hiperfunción Vocal (Vocal Hyperfunction)

ZCR Tasa de Cruces por Cero (Zero Crossing Rate)

Índice general

1..	Introducción	3
1.1.	Objetivos	4
1.1.1.	Objetivo general	4
1.1.2.	Objetivos específicos	4
2..	Marco Teórico	6
2.1.	Bases clínicas y adquisición de datos	6
2.1.1.	Hiperfunción Vocal	6
2.1.2.	Monitoreo Ambulatorio	7
2.1.3.	Parámetros acústicos	9
2.1.4.	Parámetros aerodinámicos	10
2.2.	Técnicas de análisis de datos	12
2.2.1.	Reducción de Dimensión	12
2.2.2.	UMAP	14
2.2.3.	Clustering de Series temporales	15
2.2.4.	HDBSCAN	17
2.2.5.	Métricas	18
3..	Metodología	21
3.1.	Datos	22
3.2.	Pre-Procesamiento	23
3.2.1.	Selección de Ventanas de 50 ms	23
3.2.2.	Construcción de subsecuencias de 1 s	24

3.3. Clustering	25
3.4. Visualización	27
3.5. Cálculo de Métricas de Preservación Espacial y Temporal	28
4.. <i>Resultados y Análisis</i>	30
4.1. Resumen de la calidad del embedding y del clustering	31
4.2. Visualización del espacio latente	32
4.3. Perfil de features por clúster	33
4.4. Diferencias entre condiciones clínicas	35
4.5. Efecto del tamaño de ventana	37
4.6. Preservación Espacial y Temporal	40
4.7. Análisis de caso individual	42
5.. <i>Conclusiones y Trabajo futuro</i>	44

1. INTRODUCCIÓN

Los trastornos de la voz relacionados con la *hiperfunción vocal* representan un desafío clínico considerable debido a la ausencia frecuente de indicadores objetivos estables. Tradicionalmente, la evaluación de estos trastornos se ha basado en herramientas subjetivas como la autoevaluación del paciente o la percepción clínica, lo que puede limitar la precisión diagnóstica. Esta subjetividad, sumada a la variabilidad diaria del comportamiento vocal, dificulta la detección de patrones relevantes durante una evaluación médica breve.

Con el objetivo de superar estas limitaciones, en los últimos años se han desarrollado tecnologías de *monitoreo ambulatorio* utilizando acelerómetros de cuello, que permiten registrar señales de fonación durante largos periodos y en condiciones de vida diaria. Este enfoque puede permitir a los clínicos evaluar de una mejor manera el papel de los comportamientos vocales en el desarrollo de los trastornos de voz, ubicar con precisión temporal los comportamientos abusivos o inadecuados y evaluar sistemáticamente el grado de cumplimiento del paciente con los objetivos de la terapia vocal.

Estos sistemas de monitoreo generan grandes volúmenes de datos al registrar de manera continua la vibración de la piel del cuello, asociada a la actividad vocal, lo cual abre nuevas oportunidades para el análisis cuantitativo y objetivo del comportamiento vocal.

Debido a la alta dimensionalidad y densidad de los datos obtenidos, su análisis eficiente y comprensible presenta un gran desafío. En este contexto, las técnicas de aprendizaje automático no supervisado, particularmente la *reducción de dimensión* y el *clustering*, surgen como herramientas fundamentales para descubrir patrones la-

tentes en los datos sin necesidad de etiquetas predefinidas. Estas técnicas permiten explorar la estructura subyacente de las señales y podrían facilitar la identificación de dinámicas vocales asociadas a comportamientos patológicos.

1.1. *Objetivos*

1.1.1. *Objetivo general*

Caracterizar los patrones latentes del comportamiento vocal ambulatorio captado por acelerómetro adherido a la piel en la zona anterior del cuello para mujeres con y sin patología fonotraumatológica, mediante la combinación de reducción de dimensionalidad y clustering.

1.1.2. *Objetivos específicos*

- Diseñar un *pipeline* que integre UMAP (reducción de dimensión) y HDBSCAN (clustering), evaluando la calidad de este con Density Based Cluster Validation, Silhouette y Davies-Bouldin Score, así como métricas de preservación espacial y temporal.
- Comparar la estabilidad de los clústeres y la estructura del espacio latente entre dos escalas temporales ventanas atómicas de 50 ms, y subsecuencias de 1 s con *stride* de 0.5 s), determinando la sensibilidad del método frente al tamaño de ventana.
- Cuantificar las diferencias en la distribución de clústeres entre sujetos no patológicos y sujetos patológicos, tanto antes como después del tratamiento vocal, utilizando frecuencias relativas y tamaños de efecto de Cohen *d*.
- Visualizar los *embeddings* en dos dimensiones para facilitar la interpretación y validar, de forma exploratoria, la coherencia de las agrupaciones detectadas.

- Analizar casos individuales representativos para identificar eventos vocales transitorios de interés clínico.

2. MARCO TEÓRICO

2.1. *Bases clínicas y adquisición de datos*

2.1.1. *Hiperfunción Vocal*

La hiperfuncion vocal (VH, por sus siglas en inglés) se define como un patrón crónico de uso excesivo y/o descoordinado de las fuerzas musculares laríngeas que compromete la eficiencia fonatoria, generalmente provocada por comportamientos del uso de voz ineficientes o abusivos y que pueden desencadenar en alteraciones orgánicas o funcionales de la voz. La VH se considera un componente etiológico en los tipos de trastornos de voz que ocurren con mayor frecuencia [1].

Desde un punto de vista clínico, existen dos principales tipos de VH: La hiperfunción vocal fonotraumática (PVH) y no-fonotraumática (NPVH). La PVH está asociada a la presencia de lesiones benignas vocales, como pueden ser nódulos o pólipos. Estas lesiones pueden impedir el correcto cierre de las cuerdas vocales, lo que reduce la eficiencia de la producción del sonido y provoca que la persona compense, provocando así un daño adicional al tejido [2]. Por otro lado, la NPVH se asocia a la fatiga vocal y a la tensión excesiva en los músculos del cuello, sin la presencia de una lesión tisular visible.

El presente estudio se enfocará únicamente en pacientes PVH para así trabajar con un conjunto de personas con una condición similar.

Diversas teorías han intentado explicar el origen de la PVH a través de factores fisiológicos, conductuales y psicológicos. Desde el punto de vista fisiológico, se postula que el microtrauma repetitivo por colisión glótica en las cuerdas vocales genera inflamación y engrosamiento progresivo del tejido, produciendo la formación de nódulos

o pólipos [3, 4]. Este proceso se ve reforzado por un “círculo vicioso” donde la lesión incipiente reduce la eficiencia vibratoria obligando al paciente a aumentar la presión subglótica y el esfuerzo muscular, incrementando aún más las fuerzas de impacto sobre los pliegues [1].

En el ámbito conductual, estudios han demostrado que la PVH es más prevalente en profesionales de la voz como docentes o cantantes, donde el uso prolongado a altos volúmenes se combina con técnicas fonatorias inadecuadas, promoviendo el desarrollo de lesiones [5]. Por otro lado, factores psicológicos como el estrés crónico o rasgos de personalidad también han sido vinculados a un mayor riesgo de PVH [6].

Por lo tanto, estos enfoques sugieren que la PVH no surge de un único factor, sino de la interacción entre una predisposición individual y hábitos vocales.

2.1.2. *Monitoreo Ambulatorio*

El análisis del comportamiento vocal para la identificación de patrones hiperfuncionales requiere de una evaluación prolongada, ya que estos comportamientos pueden variar significativamente a lo largo del día y en diferentes contextos para un mismo individuo [7]. Las evaluaciones clínicas tradicionales, realizadas en una consulta, capturan solo una instantánea limitada del rendimiento vocal, pudiendo no ser representativas del uso vocal en la vida diaria [8]. Es por esto que surge el monitoreo ambulatorio de la voz (AMV, por sus siglas en inglés), una metodología que permite capturar la producción vocal de forma naturalista y extendida en el tiempo.

La evolución del AMV se puede rastrear desde dispositivos iniciales como el *Vox-Log* [9] y el *Voice Accumulator* [10], que se centraban principalmente en medir el tiempo de fonación y la intensidad vocal. Si bien estos sistemas sentaron las bases, su capacidad para caracterizar la fisiología laríngea era limitada.

Una línea de investigación pionera y fundamental en este campo ha sido la desarrollada por el grupo del Massachusetts General Hospital (MGH) y el MIT. Su trabajo ha sido crucial para trasladar el monitoreo ambulatorio de la mera dosificación vocal a la evaluación de la función vocal. Su innovación central fue la incorporación de

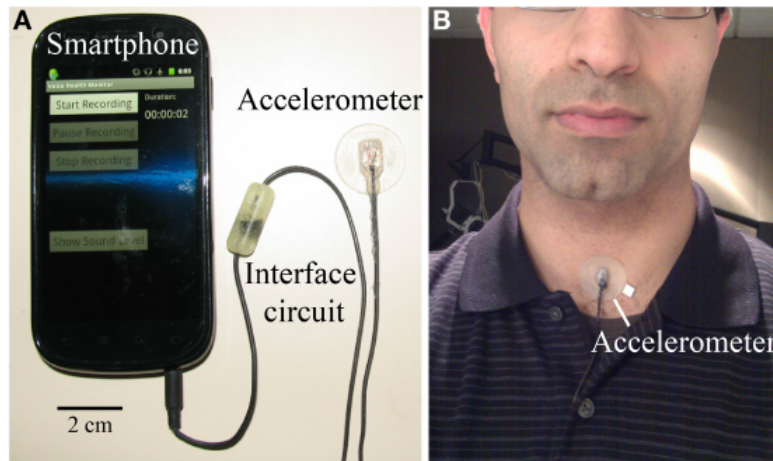


Fig. 2.1: Sistema ambulatorio de monitoreo de voz: (a) smartphone, acelerómetro y cable con circuito de interfaz encapsulado en resina epóxica (b) acelerómetro con cable montado sobre una almohadilla de silicona, fijado al cuello a la altura intermedia entre la manzana de Adán y la escotadura en forma de “V” de la clavícula. [2].

un acelerómetro de cuello de alta sensibilidad como indicador del comportamiento laríngeo, complementando la medición acústica [11]. Este enfoque permitió, por primera vez, estimar parámetros aerodinámicos glóticos de forma ambulatoria mediante técnicas de filtrado inverso basado en impedancias subglóticas (IBIF) [12].

La herramienta utilizada en el presente estudio se enmarca dentro de esta última y más avanzada línea de investigación. Corresponde a un dispositivo móvil que emplea un acelerómetro de alto ancho de banda y utiliza un smartphone como plataforma de adquisición de datos [11]. El smartphone contiene un codec que graba la señal del acelerómetro a una frecuencia de muestreo de 11025 Hz [2]. El acelerómetro se adhiere a la base del cuello del sujeto, permitiendo captar de manera no invasiva la aceleración subglotal de la superficie del cuello (ver Figura 2.1). Esta señal es relativamente inmune a los sonidos ambientales y proporciona una medida relacionada con la voz que no está filtrada por el tracto vocal, lo que mitiga las preocupaciones de confidencialidad [2].

De la señal obtenida mediante el acelerómetro se pueden extraer distintos parámetros basados en el flujo de aire glótico, tanto de tipo acústico como aerodinámico. [2] Entre los parámetros acústicos se encuentran por ejemplo el nivel de presión sonora

(SPL), la frecuencia fundamental (f_0), la diferencia armónica (H1-H2), todos útiles para evaluar la eficiencia fonatoria y el cierre glótico [2]. Por otra parte, mediante un modelado del sistema vocal, que luego pasa por un filtrado inverso subglotal basado en impedancias (IBIF) [12], es posible estimar parámetros aerodinámicos como el flujo de aire máximo (ACFL), la pendiente máxima de cierre glótico (MFDR), entre otros. Estos permiten caracterizar de forma más directa la carga biomecánica sobre los pliegues vocales.

A continuación, se presentan de forma más detallada los distintos parámetros utilizados para el análisis del comportamiento vocal. Todos estos calculados en base a la señal del acelerómetro.

2.1.3. Parámetros acústicos

Los parámetros acústicos permiten caracterizar la calidad de la voz y han mostrado diferencias significativas entre individuos con hiperfunción vocal y sujetos sanos, cuando se utilizan estadísticas derivadas de estos parámetros, como la asimetría (*skewness*) y la curtosis [13]. En general, las voces hiperfuncionales presentan menor periodicidad y alteraciones espectrales evidentes comparadas con voces normales [13]. Dentro de los parámetros acústicos extraídos por el dispositivo de monitoreo ambulatorio, se encuentran [2]:

- Nivel de presión sonora a 15 cm (SPL): Amplitud de aceleración mapeada a nivel de presión sonora acústica, medida en dB.
- Frecuencia fundamental (f_0): Inverso de la ubicación del primer pico distinto de cero en la función de autocorrelación normalizada, se mide en Hz.
- Amplitud del pico de autocorrelación: Amplitud relativa del primer pico distinto de cero en la función de autocorrelación normalizada.
- Pico subarmónico: Amplitud relativa de un pico secundario, si existe, ubicado aproximadamente a la mitad del camino hacia el pico de autocorrelación.

- **Inclinación espectral armónica:** Pendiente de regresión lineal sobre los primeros 8 armónicos espectrales, medida en dB/octava.
- **Relación espectral bajas-altas frecuencias (LH Ratio):** Diferencia entre la potencia espectral por debajo y por encima de los 2000Hz, medida en dB.
- **Prominencia del pico espectral (CPP):** Magnitud del pico más alto en el cepstrum de potencia, medida en dB.
- **Tasa de cruces por cero (ZCR):** Proporción del marco en que la señal cruza su valor medio.

2.1.4. Parámetros aerodinámicos

Estos parámetros describen el comportamiento del flujo de aire y la presión a través de la laringe durante la fonación. Dado que la producción de voz es fundamentalmente un fenómeno aerodinámico, las medidas de flujo y presión están directamente ligadas a la función glótica. Estudios recientes han demostrado que pacientes con PVH tienden a exhibir valores normalizados de parámetros aerodinámicos más bajos en comparación con sujetos sanos para lograr un mismo SPL, lo que sugiere un uso menos eficiente del sistema fonatorio [14]. En [15] usaron un dispositivo con acelerómetro de cuello en monitoreo ambulatorio y estimaron parámetros de flujo glotal (usando IBIF) cada 50 ms, de los cuales calcularon estadísticas para comparar grupos pacientes vs controles. Reportan que ocho estadísticas resumen de flujo glotal difieren significativamente entre pacientes y controles y que una regresión logística logra una AUC de 0,82 para separar PVH/control, destacando el valor clínico de estas medidas para evaluar la hiperfunción vocal.

Dentro de estos parámetros aerodinámicos, se encuentran [15]:

- **Amplitud del flujo glótico (AC Flow):** Amplitud pico a pico del flujo de aire a través de la glotis en cada vibración. Es indicativa de cuánto aire pasa cuando las cuerdas se separan.

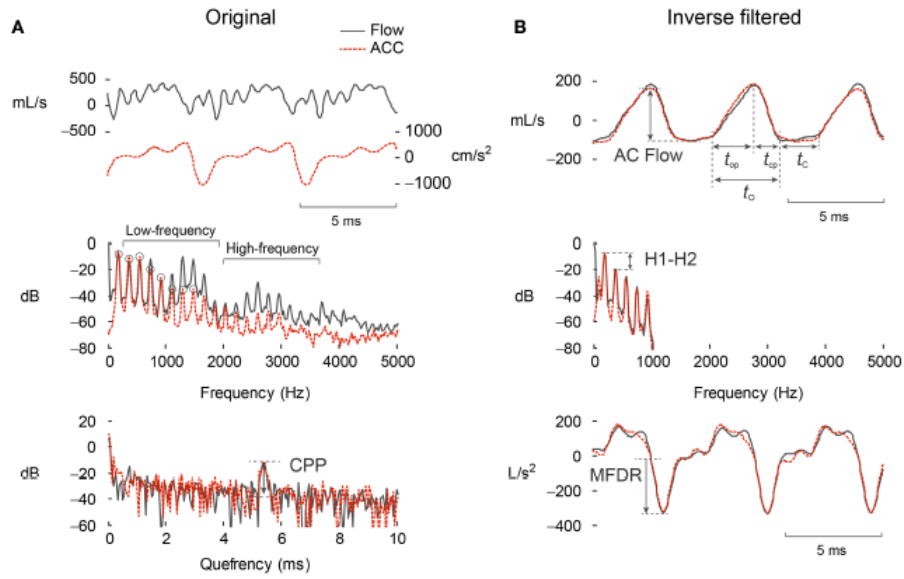


Fig. 2.2: Parametrización de las formas de onda (A) original y (B) filtrada inversamente a partir del flujo de aire oral (negro) y de la aceleración de superficie del cuello (ACC, rojo punteado), procesadas con filtrado inverso basado en impedancia subglótica. [2]

- Tasa máxima de declinación del flujo (MFDR): Pico negativo de la primera derivada de la forma de onda glotal. Asociada con la velocidad de cierre de las cuerdas.
- Cociente de apertura glótica (OQ): Fracción de ciclo vibratorio en que la glotis permanece abierta.
- H1-H2: Diferencia entre la magnitud de los dos primeros armónicos.
- Cociente de amplitud normalizado (NAQ): Ratio de ACFL a MFDR dividido por el periodo glotal.

En [16] se valida un método para estimar la presión subglótica (P_s) a partir de la vibración de superficie del cuello medida con acelerómetro, mediante modelos de regresión múltiples e IBIF, complementando las métricas mencionadas anteriormente. Los resultados muestran diferencias claras entre grupos, con valores promedio y máximos significativamente más altos en pacientes PVH que en controles.

Posteriormente, en [17] se aplicaron técnicas de machine learning con validación cruzada anidada para cuantificar qué tan consistentes son las diferencias entre pacientes y controles a través de 16 parámetros extraídos de acelerómetro de cuello y aerodinámicas estimadas vía IBIF, además de métricas de uso vocal como duración de fonación y de reposo. Los resultados mostraron que los patrones característicos (*signatures*) más consistentes de la PVH fueron el H1-H2 reducido, el SPL más alto, la inclinación espectral más baja y descansos vocales más cortos. Estas medidas lograron clasificaciones con precisiones entre 70-75 % y grandes tamaños de efecto. Además, se confirmó que la duración de los periodos de voz y descanso por sí sola también discrimina con alta consistencia, aportando evidencia directa de que los patrones de uso vocal (más allá de los promedios de intensidad o frecuencia) son marcadores fundamentales de PVH.

2.2. Técnicas de análisis de datos

2.2.1. Reducción de Dimensión

El volumen de datos al monitorear de manera constante el comportamiento vocal de una persona diariamente puede llegar a ser muy elevado. Es por esto que surge la necesidad de recurrir a herramientas como la reducción de dimensión para el análisis y comprensión de los datos obtenidos. Estas técnicas logran proyectar datos de alta dimensión en un espacio 2D o 3D manteniendo las relaciones estructurales, permitiendo así un análisis visual de datos [18] que en un inicio tienen demasiadas dimensiones para poder ser visualizados en un gráfico.

Entre las principales técnicas de reducción de dimensionalidad se encuentra PCA, que condensa la mayor varianza global con gran rapidez e interpretabilidad, pero sólo captura relaciones lineales [19]. También t-SNE, excelente para revelar clústers locales en 2-3 D, aunque costosos y poco fiables para distancias globales [20]; UMAP, que preserva estructura local y global con mayor velocidad y escalabilidad que t-SNE [21], siendo hoy la opción favorita para visualizar grandes volúmenes no lineales y,

finalmente, los autoencoders basados en redes profundas, que aprenden representaciones altamente no lineales incluso capturando dependencias temporales a costa de mayor tiempo de entrenamiento y necesidad de datos abundantes.

Tanto UMAP como los autoencoders son potentes alternativas para la reducción de los datos ambulatorios, pero debido al interés en la visualización se opta por utilizar UMAP ya que este algoritmo fue diseñado específicamente para proyectar datos de alta dimensión a espacios visuales. En cambio, los autoencoders generan un código latente de dimensionalidad arbitraria, cuya interpretación no es trivial; al forzarlo a representaciones de 2 o 3 dimensiones para fines de visualización, se corre el riesgo de distorsionar o degradar la estructura del espacio latente y, por ende, su capacidad representacional [22].

UMAP fue utilizado para reducir la dimensión con el fin de representar datos espacio-temporales multivariados provenientes de modelos hidrológicos en [23]. En este estudio se compararon PCA, GTM, t-SNE y UMAP, concluyéndose que UMAP logra una representación precisa de las estructuras globales y locales en los datos, con una eficiencia computacional superior a t-SNE y una menor sensibilidad a los valores atípicos en comparación con GTM. Además, se demostró que las visualizaciones obtenidas mediante UMAP permiten identificar patrones sutiles en el comportamiento espacial del sistema, así como trayectorias temporales bien definidas a lo largo de múltiples escalas. Este tipo de análisis visual facilitó la detección de regiones con comportamientos hidrológicos similares, sin la necesidad de aplicar previamente técnicas de clustering.

En [24], UMAP fue evaluado como técnica de reducción de dimensión para facilitar el análisis visual de series temporales, tanto univariadas como multivariadas. Los resultados mostraron que UMAP tuvo un desempeño competitivo, superando a t-SNE en aspectos como la preservación de relaciones entre grupos de datos (estructura global), la claridad de las transiciones temporales y la capacidad de escalamiento para grandes volúmenes de datos. En el mismo estudio, se analizan datos de respiración humana, donde UMAP permitió identificar patrones anómalos de manera más

clara que t-SNE, al representar los ciclos respiratorios como bucles bien definidos en el espacio reducido. Además, en la aplicación a series multivariadas captadas por acelerómetros (actividad de aves), UMAP facilitó la visualización de seis patrones de comportamiento y sus transiciones, mostrando trayectorias más armónicas que PCA y mejor organizadas que t-SNE.

La reducción de dimensión no se limita a proveer una visualización de datos de dimensión elevada. También puede ser utilizada como un paso previo a algoritmos de clustering, mejorando tanto la eficiencia computacional como la calidad de los resultados al eliminar redundancias y ruido en los datos. Un estudio reciente demostró que aplicar técnicas de reducción de dimensión como UMAP antes del clustering mejora significativamente el rendimiento de algoritmos como K-means, DBSCAN y clustering espectral, tanto en precisión como en velocidad, especialmente en conjuntos de datos de alta dimensión [25].

A continuación, se ahonda en las principales características del algoritmo UMAP.

2.2.2. UMAP

UMAP (Uniform Manifold Approximation and Projection) es una técnica de reducción de dimensión no lineal basada en el aprendizaje de variedades (manifold learning), que asume que los datos residen sobre una estructura de baja dimensión embebida en el espacio original. Posee distintos hiperparámetros que tienen un impacto significativo en la proyección de los datos, se presentan los cuatro más relevantes [26]:

- **n_neighbors:** Controla cómo el algoritmo equilibra la estructura local versus la estructura global de los datos. Valores pequeños realzan micro-estructuras, mientras que valores mayores favorecen la coherencia global .
- **min_dist:** Regula cuán compactos pueden quedar los puntos, evitando la formación de conglomerados artificialmente densos.
- **n_components:** Determina la dimensión a la que se reducirán los datos.

- **metric:** Define la métrica de distancia utilizada para calcular las relaciones entre puntos; UMAP permite usar métricas clásicas como la euclídea o métricas especializadas como la DTW para series temporales (con un costo computacional elevado).

Al emplear índices de vecinos aproximados y algoritmos de optimización estocásticos, UMAP puede reducir millones de muestras con complejidad sub-cuadrática, superando en tiempo de cómputo a otras técnicas como t-SNE sin restringir la dimensión de la proyección. Además, tiene la capacidad de transformar nuevos datos no vistos en una proyección previamente entrenada [27], permitiendo así ahorrar mucho tiempo de cómputo entrenando el algoritmo en una porción de los datos para luego generalizarlo con el resto de estos.

Debido a que se ha comprobado que UMAP es una herramienta eficaz para representar dinámicamente series temporales en un espacio 2D y mejora el rendimiento de algoritmos de clustering es que se opta como método de pre-procesamiento y visualización para los datos obtenidos mediante el sistema de monitoreo ambulatorio.

2.2.3. Clustering de Series temporales

El clustering es una técnica de aprendizaje no supervisado que agrupa datos en función de su similitud, con el objetivo de descubrir estructuras o patrones en conjuntos de datos no etiquetados [28]. Así, toda la información captada por el sensor de acelerómetro puede ser analizada de forma exploratoria en base a los clústers formados por el algoritmo utilizado, identificando agrupaciones con patrones similares que podrían corresponder por ejemplo a comportamientos vocales saludables o patológicos. Además, el clustering facilita la detección de estructuras latentes y variaciones no evidentes a simple vista, lo que puede ser clave para el descubrimiento de nuevos patrones de uso vocal en entornos ambulatorios.

El clustering en series temporales se divide en dos categorías principales [29]: clustering de series temporales completas y clustering de sub-secuencias en una serie

temporal. Donde la primera categoría consiste en agrupar distintas series de tiempo en clústers distintos, mientras que la segunda consiste en agrupar distintas sub-secuencias o puntos de una misma serie de tiempo en distintos clústers. El presente estudio cae en la segunda categoría, ya que se busca agrupar porciones de voz de menor duración dentro de una señal temporal más extensa correspondiente a los datos ambulatorios de uno o más sujetos.

Entre los principales enfoques de clustering para series temporales se encuentran los *métodos de partición* (como K-means) que se caracterizan por su rapidez y simplicidad, pero exigen conocer el número de clústers y sufren con secuencias de distinta longitud [30]. Los *esquemas fuzzy* permiten que cada subsecuencia pertenezca a más de un clúster, algo valioso cuando los cambios de comportamientos son graduales [30]. Estos métodos no son sencillos de aplicar cuando se trata de series temporales de distinta longitud debido a la ambigüedad al medir los centros de los clústers [31]. Los *métodos jerárquicos* utilizan dendogramas, que resultan útiles cuando no se conoce el número de clústers y se necesita examinar los datos a múltiples escalas; su desventaja es igualmente su elevado costo. Los *algoritmos basados en densidad* (como OPTICS o HDBSCAN) buscan regiones densas separadas por vacíos, detectan clústers de forma arbitraria y ruido sin fijar el número de clústers, tienen un costo de cómputo menor y requieren parámetros sensibles [31].

En la práctica, la elección del algoritmo de clustering depende tanto de los fines analíticos como de la naturaleza del conjunto de datos, ya que cada familia presenta fortalezas y limitaciones propias. Cuando se trabaja con volúmenes masivos y, además de agrupar, se desea una representación visual que facilite la interpretación, resulta beneficioso incorporar primero una reducción de dimensionalidad. En particular, la combinación de UMAP para proyectar la estructura no lineal a un espacio de baja dimensión y HDBSCAN para detectar agrupaciones de densidad variable aporta eficiencia computacional y claridad visual. Un estudio reciente de [32], que evaluó 85 bases de datos de series temporales univariadas, demostró que aplicar UMAP seguido de HDBSCAN elevó la calidad del agrupamiento (índice V-measure) en el 75 %

de los casos frente a ejecutar HDBSCAN directamente sobre los datos originales, superando también a métodos clásicos como K-means y el clustering jerárquico tanto en rendimiento medio como en número de casos con el mejor resultado absoluto.

Además, se ha demostrado que la combinación de UMAP + HDBSCAN incrementa la calidad de agrupamiento (medidas V-score y silhouette) entre un 15 % y un 25 % respecto a DBSCAN y k-means, manteniendo una escalabilidad lineal hasta millones de muestras [24].

Estos hallazgos respaldan la elección de UMAP + HDBSCAN en el presente trabajo, donde se dispone de un conjunto de datos extenso y se privilegia la interpretabilidad visual sin sacrificar la robustez del agrupamiento. A continuación, se ahonda en las principales características de HDBSCAN.

2.2.4. *HDBSCAN*

HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise) se extiende del algoritmo DBSCAN, construyendo primero una jerarquía de densidad y extrayendo de ella los grupos más estables. A diferencia de su predecesor DBSCAN, HDBSCAN selecciona automáticamente los clusters significativos según su estabilidad, sin necesidad de fijar a priori el número de grupos [33]. Esto presenta una gran ventaja debido a que permite detectar agrupaciones de distinta densidad y forma sin requerir intervención manual o supuestos fuertes sobre la estructura de los datos.

Los principales hiperparámetros que rigen el algoritmo son [34] :

- **min_cluster_size:** Especifica el tamaño mínimo de un cluster y controla la granularidad del resultado.
- **min_samples:** Actúa como parámetro de suavizado, afectando la robustez frente al ruido.
- **metric:** Define la métrica de distancia.

En registros ambulatorios de acelerómetro laríngeo, la densidad de puntos puede variar según la intensidad vocal o la duración de episodios fonatorios, por lo que la identificación automática del número de clusters por parte del algoritmo es de gran valor. Además, cuando se aplica HDBSCAN sobre un embedding obtenido con UMAP, el algoritmo opera en un espacio de menor dimensión con clusters más compactos, lo que reduce el tiempo de cómputo, mejora la diferenciación entre subgrupos y produce resultados fácilmente visualizables [26].

2.2.5. Métricas

Calidad de Agrupamiento

En este trabajo se busca caracterizar el comportamiento vocal diario de distintos sujetos mediante análisis de clustering, enfocado en subsecuencias o ventanas temporales de corta duración de la señal. En este contexto, es importante destacar que no se dispone de etiquetas que describan el tipo específico de comportamiento en cada instante o subsecuencia (por ejemplo, un grito, un susurro o una emisión patológica). Las únicas etiquetas disponibles corresponden al nivel global del sujeto, como la condición patológica o la realización de terapia vocal.

Esta limitación implica que no es posible evaluar directamente la precisión del método de clustering, ya que no se cuenta con una *ground truth* detallada. En consecuencia, la validación debe apoyarse en métricas internas que no dependen de etiquetas externas y que permiten estimar la calidad de las agrupaciones generadas por los algoritmos no supervisados.

En [25], evalúan el desempeño de distintos algoritmos de clustering sobre datos de alta dimensión, utilizando las métricas Silhouette Score y Davies-Bouldin Index para comparar la coherencia y separación de los grupos obtenidos.

El *Silhouette Score* mide qué tan similar es un punto a los elementos de su mismo clúster en comparación con los puntos de otros clústeres, tomando valores entre -1 y 1 , donde valores más altos indican mejor separación y cohesión [35]. Por otro lado,

el *Davies-Bouldin Index* evalúa la relación entre la dispersión dentro de los clústeres y la distancia entre ellos; en este caso, valores más bajos indican mejor calidad de agrupamiento [25].

Estas métricas son muy utilizadas para evaluar el desempeño del clustering generado cuando no se dispone de etiquetas, por lo que se optará por utilizarlas para este trabajo. Se debe destacar que el *Silhouette Score* está especialmente diseñado para evaluar clústeres convexos o esféricos. Por este motivo, tiende a subvalorar la calidad cuando los clústeres tienen formas irregulares, alargadas o de tamaño variable [36].

Dado que HDBSCAN tiende a formar clusters irregulares, se opta a complementar la métrica de Silhouette Score con Density-Based Clustering Validation (DBCV). Esta técnica está diseñada específicamente para validar agrupamientos generados por algoritmos basados en densidad. El DBCV evalúa la calidad del clustering considerando tanto la densidad interna de los clústeres como su separación relativa. Sus valores van entre -1 y 1, donde valores más altos indican una mejor definición estructural de los clústeres, es decir, regiones densas bien separadas entre sí [37].

Preservación Espacial y Temporal

Para evaluar el desempeño de un algoritmo de reducción de dimensión, es común emplear métricas que cuantifican cómo se preservan las relaciones de vecindad o distancias entre puntos al pasar del espacio original de alta dimensión al espacio proyectado. Estas métricas de calidad espacial incluyen indicadores clásicos como *Trustworthiness*, *Continuity* y *Normalized Stress*, los cuales miden, respectivamente, la fidelidad local, la ausencia de vecinos falsos y la conservación de distancias globales [38, 39].

Sin embargo, dado que en este contexto se trabaja con series temporales multivariadas, también resulta fundamental evaluar cómo se preserva la continuidad temporal en el espacio reducido. En este marco, se adoptan las métricas propuestas en [40], que evalúan la estabilidad temporal de las proyecciones. Estas incluyen el *Temporal Stress*, que mide cuán proporcional es el desplazamiento de un punto entre dos

tiempos consecutivos respecto a su cambio en el espacio original, y las correlaciones de Shepard temporales (*Temporal Pearson*, *Spearman* y *Kendall*), que cuantifican la monotonicidad entre el cambio real y el cambio proyectado a lo largo del tiempo.

El uso conjunto de métricas espaciales y temporales permite tener una noción más completa sobre cómo afecta la reducción de dimensión a la estructura y evolución a los datos de voz captados por acelerómetro.

3. METODOLOGÍA

La metodología propuesta se basa en los trabajos previos mencionados. Se utilizaron las señales ambulatorias obtenidas en [15], a partir de las cuales se extrajeron distintos parámetros acústicos y aerodinámicos para 13 sujetos patológicos y sus respectivos controles. Para cada paciente, se recopilaban registros diarios durante una semana previa al tratamiento y otra semana posterior al tratamiento, usualmente 6 meses después de la primera semana de grabación. En el caso de los controles sanos, se obtuvieron datos durante una única semana.

Con los datos procesados, se aplicó primero un algoritmo de reducción de dimensionalidad como etapa previa al clustering, siguiendo el enfoque descrito en [25]. A continuación, se utilizó un algoritmo de agrupamiento no supervisado para identificar patrones latentes en el espacio reducido.

De forma paralela, se proyectaron los datos a dos dimensiones con fines visuales, coloreando cada punto según el clúster asignado. Esta representación permite explorar de forma intuitiva la estructura de los grupos y evaluar la coherencia visual de la segmentación generada por el algoritmo.

Todo este procedimiento se aplicó tanto a los datos originales segmentados en ventanas de 50 ms como a un conjunto alternativo construido con ventanas de 1 s de duración, con el objetivo de evaluar la sensibilidad del método frente a cambios en la escala temporal de análisis.

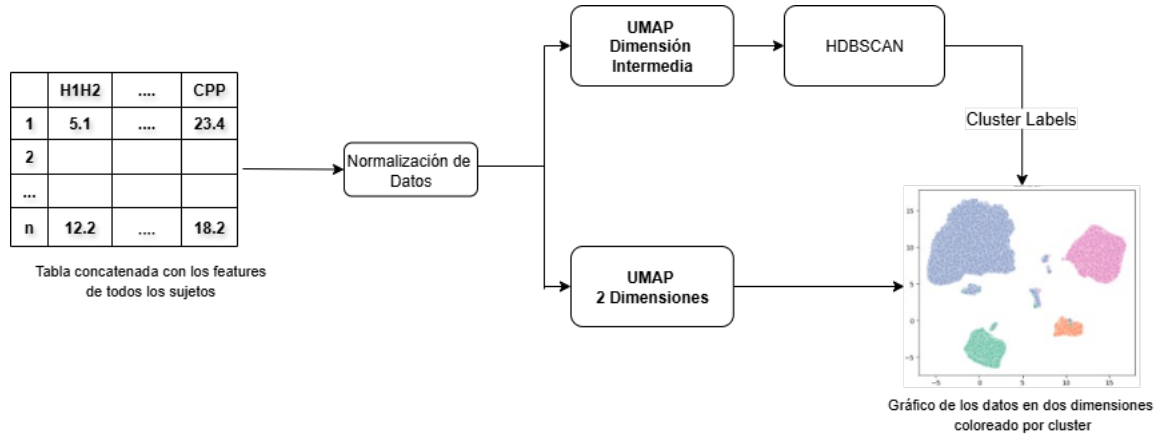


Fig. 3.1: Diagrama de la metodología propuesta

3.1. Datos

Los datos utilizados en este trabajo provienen de la base de datos del Hospital General de Massachusetts (MGH), correspondiente a la utilizada en [15]. Esta base se generó a partir de un estudio en el que participaron 96 participantes, conformados por 48 pares de mujeres adultas. Cada par se compone de un paciente con PVH y un control, emparejado según la edad y ocupación.

Dentro de esta base de datos existen carpetas para cada sujeto, con archivos MAT separados por día de registro. En el caso de los pacientes patológicos, se dispone de mediciones diarias correspondientes a una semana antes del tratamiento y otra semana posterior al tratamiento. Para el caso de los controles, existen mediciones diarias a lo largo de una semana. Esta estructura permite contrastar las diferencias entre tres estados clave: pacientes pre-tratamiento, pacientes post-tratamiento y controles sanos.

Cada archivo MAT almacena diversos parámetros extraídos de la señal del acelerómetro laríngeo, muestreada a 11025 Hz y segmentada en ventanas no superpuestas de 50 ms, donde se incluyen todos los parámetros acústicos y aerodinámicos mencionados anteriormente.

Adicionalmente, existen señales auxiliares como 'voiced'. Esta señal binaria indica si la ventana es considerada o no como un segmento de voz. Cada ventana es clasificada como vocalizada si los parámetros acústicos derivados de la misma se en-

cuentran dentro de ciertos rangos definidos como se especifica en [2]. De esta manera se logran eliminar varios tipos de actividad no fonatoria, como tocar el sensor o niveles extremadamente altos de ruido ambiental y también evitar consonantes u otro tipo de fonemas que no contengan vibración de los pliegues vocales.

Existe además la señal 'voicedRMS', subgrupo del vector 'voiced' descarta las ventanas donde la razón absoluta de los valores RMS de la primera mitad divididos por la segunda mitad del cuadro fuera mayor que un umbral (1.5); de este modo, no se consideran las ventanas que muestran inicios o finales (onsets u offsets), ya que típicamente resultan en estimaciones incorrectas del filtrado inverso [15] lo que afecta directamente a los parámetros aerodinámicos ya que son calculados con el método IBIF.

La señal 'breathGroup' surge con el objetivo de cuantificar aún más la carga vocal, suavizando sobre el vector binario 'voiced' de modo que los segmentos vocalizados contiguos sean conectados si se encuentran lo suficientemente próximos entre sí según un umbral de duración predefinido (típicamente menor a 0,5 segundos) [2]. De esta forma, el vector suavizado sirve para no tomar en cuenta ciertas ventanas vocalizadas aisladas, que pueden corresponder a sonidos no relacionados con el habla, aclaraciones de garganta, respiraciones ruidosas u otros eventos no lingüísticos.

Para este estudio se utilizan los datos de 13 pares de sujetos, rescatando de estos los valores de sus parámetros acústicos y aerodinámicos.

3.2. Pre-Procesamiento

3.2.1. Selección de Ventanas de 50 ms

Una vez recopilados los datos de los 13 pares de sujetos, es necesario realizar un pre-procesamiento antes de utilizarlos como entrada en los algoritmos de reducción de dimensionalidad o clustering. Para cada semana de cada sujeto, se crea una tabla con los datos pre-procesados, incluyendo información adicional como su estado (control, pre o post tratamiento), el ID del sujeto y la fecha.

timestamp	date	ID	status	CPP	ZCR	...	H1-H2
14:25:05.36	2012-01-01	NF021	Control	20.360238	0.059891	...	31.233570
14:25:05.41	2012-01-01	NF021	Control	21.666720	0.059891	...	30.776362
...	...	...	...	...	...	...	...
19:25:58.16	2012-01-13	NF021	Control	20.921265	0.0061706	...	31.236905

Tab. 3.1: Tabla de ejemplo con datos para un sujeto control.

Dado que resulta relevante analizar únicamente los segmentos correspondientes al habla, se emplea la información del vector 'breathGroup' para seleccionar únicamente aquellas ventanas que pertenecen a eventos fonatorios de mayor extensión. Específicamente, se conservan las ventanas que forman parte de un segmento de 'breathGroup' con una duración superior a 30 ventanas consecutivas (equivalente a 1,5 segundos), mientras que las ventanas que no cumplen este criterio son descartadas del análisis.

Luego, para cada ventana rescatada, se extraen los parámetros acústicos, aerodinámicos y metadata adicional como la marca de tiempo, la fecha, el ID del sujeto, entre otros. Finalmente, para cada sujeto se juntan todas las ventanas rescatadas en una tabla, como se muestra en 3.1.

Se crea una tabla para cada sujeto y, finalmente, todas las tablas se unen formando una única tabla con todos los datos. Esta, de aproximadamente 7 millones de filas, es la que se utilizará como entrada para los distintos algoritmos utilizados.

Dentro de los parámetros recopilados, existen algunos altamente correlacionados entre sí. Por ejemplo, está la variable `cppall` y `cppall_2048`. Ambas son medidas de CPP y su diferencia destaca en la ventana en la que fue calculada. En estos casos, se descarta una de estas (en este caso, `cppall_2048`) debido a su alta correlación ya que presentaría redundancia y no aporta información adicional.

3.2.2. Construcción de subsecuencias de 1 s

Además del análisis basado en ventanas de 50 ms, se implementó un segundo pipeline que agrupa dichas ventanas en *subsecuencias* de 1 s (equivalentes a 20 ventanas), con un *stride* de 10 ventanas (0.5 s). Esta variante tiene como objetivo explorar

cómo varía el comportamiento del modelo de reducción de dimensión y clustering al aumentar la escala temporal de análisis, permitiendo capturar patrones que podrían manifestarse únicamente en intervalos más largos.

Para asegurar continuidad temporal, se estableció que todas las ventanas dentro de cada subsecuencia debían ser consecutivas; no obstante, se toleró un único salto de hasta 50 ms (una ventana) por subsecuencia, con el fin de compensar posibles pérdidas puntuales de señal.

El procedimiento consistió en recorrer, para cada sujeto y cada día, el vector de ventanas válidas mediante un deslizador de tamaño 20 con un paso de 10 ventanas. Solo se conservaron aquellas subsecuencias que cumplieran con el criterio de contigüidad temporal mencionado.

Luego, sobre las 12 *features* originales de cada subsecuencia, se calcularon dos estadísticas descriptivas: la mediana y la desviación estándar. Esto dio lugar a un total de 24 columnas por subsecuencia (12 medianas + 12 desviaciones), permitiendo capturar no solo el valor central de cada parámetro, sino también su variabilidad local.

Como resultado, se obtuvo un nuevo conjunto de datos compuesto por aproximadamente 150 000 subsecuencias.

3.3. *Clustering*

El proceso de clustering consta de dos etapas: La etapa de reducción de dimensión utilizando UMAP como pre-procesamiento y la etapa de clusterización aplicando HDBSCAN.

De la tabla formada con todos los datos para todos los sujetos, se extraen únicamente las columnas correspondientes a las features (descartando meta-datos como fecha, ID del sujeto, etc.) para ser usadas como entrada al algoritmo de HDBSCAN y UMAP.

Dado que se recomienda normalizar los datos antes de aplicar algoritmos como UMAP o HDBSCAN, especialmente cuando las variables se encuentran en distintas

escalas [26, 33], como ocurre en el presente conjunto de datos, se procede a aplicar una estandarización mediante z-score. Esta técnica transforma cada variable para que tenga media cero y desviación estándar uno, sin alterar la forma de la distribución ni eliminar valores atípicos.

Para escoger los parámetros adecuados de cada algoritmo, se realiza un barrido sistemático sobre un conjunto de valores relevantes para los principales hiperparámetros de cada técnica. Luego, se mide el desempeño de cada agrupación realizada utilizando las métricas DBCV, Silhouette Score y Davies-Bouldin Index que permiten cuantificar la calidad de la agrupación considerando cohesión intra-clúster y separación inter-clúster.

La elección inicial de los valores a explorar se basó tanto en la literatura como en criterios prácticos derivados del tamaño y naturaleza del dataset

- **n_neighbors:** Se exploraron valores de 10, 20, 30, 50 y 100. Valores bajos favorecen la formación de clústeres más pequeños y definidos localmente, mientras que valores altos permiten la detección de agrupaciones más amplias. Esta variación permite evaluar si los sujetos se organizan en patrones finos o globales.
- **min_dist:** Se probaron valores de 0.001, 0.05 y 0.1. Un *min_dist* bajo favorece clústeres más compactos (mayor densidad), mientras que valores más altos permiten una representación más dispersa, lo que puede facilitar la separación de grupos cuando las fronteras son menos nítidas.
- **n_components:** Se utilizaron valores de 5, 9 y 12 para el dataset con ventanas de 50 ms, buscando un equilibrio entre compresión de la información y preservación de la estructura latente. Valores más bajos permiten visualización directa y reducen el riesgo de sobreajuste, mientras que valores más altos pueden capturar relaciones complejas. Para el análisis con subsecuencias de 1 s, dada la mayor dimensionalidad del dataset (28 columnas frente a 14), se incorporaron valores adicionales de 15, 18 y 22.

- **min_cluster_size:** Este parámetro controla el tamaño mínimo que debe tener un grupo para ser considerado un clúster válido. Para el dataset de 50 ms, compuesto por más de 7 millones de ventanas, se fijó en 1000 para evitar clústeres espurios o ruidosos. En cambio, para el dataset de subsecuencias de 1 s (150 mil filas), se ajustó proporcionalmente a un valor mucho menor, fijándose en 50. Este escalamiento busca mantener una proporción similar respecto al tamaño total del conjunto de datos en cada caso.

Para asegurar la factibilidad computacional, el dataset de 7 millones de filas se reduce mediante un sub-muestreo estratificado de 200.000 ventanas por condición (Pre, Post y Control), obteniendo así una muestra representativa y balanceada del total de datos.

Se realiza la proyección UMAP seguida por el clustering de HDBSCAN para cada combinación de hiperparámetros en el dataset reducido, y luego se hace el cálculo de las métricas mencionadas anteriormente utilizando tan sólo 10.000 ventanas, debido al costo computacional de estos cálculos.

Una vez calculadas las métricas para cada combinación, se eligen las tres combinaciones que obtuvieron un mejor desempeño y se vuelven a calcular pero esta vez usando la totalidad de los datos. Dado que tanto UMAP como HDBSCAN permiten generalizar sus resultados a nuevos datos, se aprovecha esta propiedad para entrenar ambos algoritmos sobre el subconjunto muestreado y posteriormente extender las proyecciones y etiquetas de clúster al conjunto completo.

Finalmente, se escoge la combinación que muestre mejores métricas para el dataset completo.

3.4. Visualización

Visualizar los datos en dos dimensiones permite entender la estructura general del espacio de representación y la distribución global de los datos. Además, se puede evaluar visualmente la separación y cohesión de los clústeres obtenidos e identificar

posibles patrones temporales, transiciones o anomalías que no son evidentes en tablas numéricas.

Dado que los datos originales se encuentran en un espacio de alta dimensión, se emplea el algoritmo UMAP para proyectarlos en un espacio bidimensional interpretable. Esta técnica permite preservar tanto la topología local como global de los datos, facilitando la interpretación visual de las relaciones entre ventanas.

Para asegurar coherencia con la etapa de clusterización, se utilizan los mismos hiperparámetros de UMAP que resultaron óptimos durante el proceso de ajuste (maximizando DBCV y Silhouette Score). Esto garantiza que la representación visual respete las relaciones estructurales que dieron lugar a la agrupación obtenida.

Al igual que en el proceso de entrenamiento del modelo de clustering, se realiza una fase inicial de reducción de dimensión utilizando un submuestreo estratificado de 200.000 ventanas por condición (Control, Pre y Post). UMAP se ajusta sobre esta muestra y luego se extiende al conjunto completo de datos mediante su capacidad de transformar nuevos puntos en el mismo espacio latente.

El resultado son coordenadas (x, y) para cada ventana, que permiten visualizar gráficamente el conjunto completo en dos dimensiones. Estas proyecciones pueden colorearse de acuerdo con distintas variables (como etiquetas de clúster, timestamp, etc.), facilitando un análisis visual y revelando patrones complejos que serían difíciles de detectar en la representación original.

3.5. Cálculo de Métricas de Preservación Espacial y Temporal

Con el objetivo de evaluar en qué medida la proyección bidimensional generada por UMAP conserva las relaciones originales entre los datos, se calcularon las métricas descritas en la Sección 2.3: preservación de distancias, preservación de vecinos y estabilidad temporal. Estas métricas permiten cuantificar la fidelidad del espacio embebido tanto desde un punto de vista espacial como temporal.

El cálculo de estas métricas implica un costo computacional elevado, principal-

mente debido a la necesidad de estimar todas las distancias entre pares de muestras y determinar sus vecindarios en el espacio original y en el proyectado. Para reducir este costo, se adoptó una estrategia de muestreo que consiste en seleccionar un día aleatorio por sujeto. De esta forma, se analizan subsecciones contiguas del conjunto de datos ambulatorios, lo que preserva la continuidad temporal dentro del día y mantiene una representación representativa del comportamiento de cada individuo.

En el caso de las métricas de preservación de vecinos, es necesario definir el parámetro k , que indica el número de vecinos a considerar. Siguiendo el procedimiento propuesto en [40], se utiliza un rango de valores de k que varía entre el 1 % y el 20 % del número total de muestras disponibles para el sujeto en el día analizado. Esto permite evaluar la fidelidad de la proyección tanto en vecindades locales (pequeños k) como en vecindades más amplias (grandes k). Posteriormente, se promedian los valores obtenidos para cada k con el fin de obtener una medida global de preservación.

Finalmente, los valores de las métricas obtenidos para cada día se promedian por sujeto, obteniendo así una estimación del desempeño promedio de la reducción de dimensionalidad en términos de su capacidad para preservar la estructura espacial y la coherencia temporal de los datos originales.

Cabe destacar que no existen umbrales absolutos que determinen valores “buenos” o “malos” para cada métrica, ya que su interpretación depende del contexto del conjunto de datos y del método de proyección empleado. Por lo tanto, en este trabajo la evaluación de la calidad de la proyección se realiza de forma comparativa, considerando los valores obtenidos para cada ventana de tiempo utilizada, y comparando con los valores obtenidos en [40].

4. RESULTADOS Y ANÁLISIS

Esta sección de Resultados responde a dos ejes centrales: (i) determinar si la semana de fonación ambulatoria de pacientes fonotraumáticos y sus controles encierra patrones latentes consistentes, y (ii) establecer en qué medida dichos patrones discriminan estados clínicos (Pre, Post-terapia y Control) y describen matices de la producción vocal. Para ello, los registros multivariados de cada ventana se proyectan con UMAP. Sobre ese espacio latente, HDBSCAN identifica clústeres y señala el ruido potencial. La calidad de la proyección y del agrupamiento se cuantifica mediante DBCV, Silhouette y, posteriormente, se examinan diferencias entre condiciones mediante frecuencias relativas y tamaños de efecto de Cohen d. Finalmente, se contrasta la robustez del modelo frente a cambios en la granularidad temporal (ventanas de 50 ms vs 1 s) y se ilustran hallazgos clínicamente relevantes con estudios de caso individuales.

4.1. Resumen de la calidad del embedding y del clustering

n_comp	n_neigh	min_dist	clusters	DBCV	Silhouette	DB Index
9	50	0.05	3	0.823	0.635	0.387
9	100	0.05	3	0.819	0.629	0.384
9	30	0.05	3	0.811	0.630	0.400
9	50	0.10	3	0.772	0.596	0.434
9	30	0.10	3	0.768	0.600	0.437
9	100	0.10	3	0.761	0.587	0.431
9	10	0.01	4	0.761	0.508	0.499
9	10	0.01	4	0.761	0.508	0.499
9	30	0.01	4	0.742	0.524	0.521

Tab. 4.1: Top 9 combinaciones según DBCV (en negrita las 3 mejores).

Tras evaluar todas las combinaciones de hiperparámetros, la Tabla 4.1 muestra las nueve configuraciones con mejor desempeño según el índice DBCV. Se observa que las mejores métricas se alcanzan al proyectar los datos a 9 dimensiones, mientras que al reducir a un espacio de menor dimensión (por ejemplo, 5), la calidad del agrupamiento disminuye sistemáticamente.

Las tres configuraciones destacadas en negrita fueron re-evaluadas utilizando la totalidad del conjunto de datos. Como resultado, se determinó que la combinación óptima está compuesta por: **n_components = 9**, **n_neighbors = 50**, **min_dist = 0.05**. Esta configuración alcanzó un **DBCV de 0.856** y un **Silhouette Score de 0.671**, logrando identificar 3 clústeres y clasificando como ruido el **0.0006 %** de las ventanas, lo que sugiere una estructura latente densa y separable en los datos.

4.2. Visualización del espacio latente

La Figura 4.1 muestra el espacio latente obtenido al aplicar UMAP sobre los datos originales, configurado con los hiperparámetros que maximizaron el DBCV, pero reduciendo la dimensionalidad a dos componentes con fines visuales. Cada punto está coloreado según la etiqueta de clúster asignada por HDBSCAN.

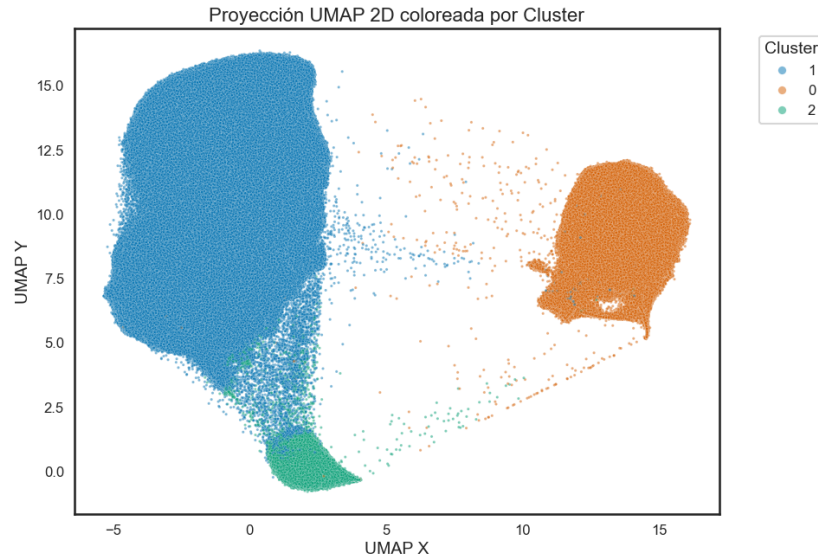


Fig. 4.1: Visualización UMAP de los datos, coloreado por clúster.

En el gráfico se observan tres nubes de puntos bien definidas, con distintas formas, tamaños y densidades, en su mayoría correspondientes a clústeres claramente diferenciados. Esto sugiere que el embedding captura una estructura latente coherente con la segmentación realizada.

Sin embargo, también se aprecian ciertos puntos cuyo color (es decir, su clúster) no coincide con la agrupación visual a la que parecen pertenecer en el plano 2D (como por ejemplo, ciertos puntos coloreados azul dentro de la nube de puntos del clúster 2). Este fenómeno se explica porque el algoritmo de clustering se aplica sobre el espacio latente original de **9 dimensiones**, donde las relaciones de proximidad entre puntos pueden diferir significativamente de las que se proyectan en dos dimensiones. La reducción a 2D introduce distorsiones locales debido a la compresión del espacio, lo

cual puede alterar las vecindades perceptibles.

Por lo tanto, esta visualización debe interpretarse como una representación simplificada que permite explorar la estructura global de los datos, pero no como una reproducción exacta de las fronteras de clúster.

4.3. Perfil de features por clúster

Antes de examinar las características específicas de cada clúster, es útil revisar su distribución relativa dentro del conjunto de datos. La Tabla 4.2 muestra el porcentaje de ventanas asignadas a cada uno de los tres clústeres identificados por HDBSCAN.

Clúster	Porcentaje (%)
1	77.02
0	19.06
2	3.92

Tab. 4.2: Distribución porcentual de ventanas por clúster.

Como puede apreciarse, el **clúster 1** concentra la mayoría de las ventanas, lo que sugiere que representa un patrón de fonación dominante o basal en la población estudiada. El **clúster 0** agrupa cerca del 19 % de las ventanas, mientras que el **clúster 2** es considerablemente más reducido, con menos del 4 %. Esta distribución es coherente con las proporciones observadas visualmente en el espacio latente proyectado (Figura 4.1), y permite contextualizar la relevancia relativa de cada agrupación al interpretar sus perfiles acústicos y aerodinámicos en el análisis que sigue.

Con el objetivo de visualizar cómo se distribuyen los valores de las distintas variables dentro de cada clúster, se construyeron diagramas de caja (boxplots) que permiten comparar la mediana, la dispersión y posibles asimetrías de cada feature entre clústers. En la Figura 4.2 se muestran las variables acústicas y aerodinámicas que evidencian los contrastes más marcados a nivel inter-clúster.

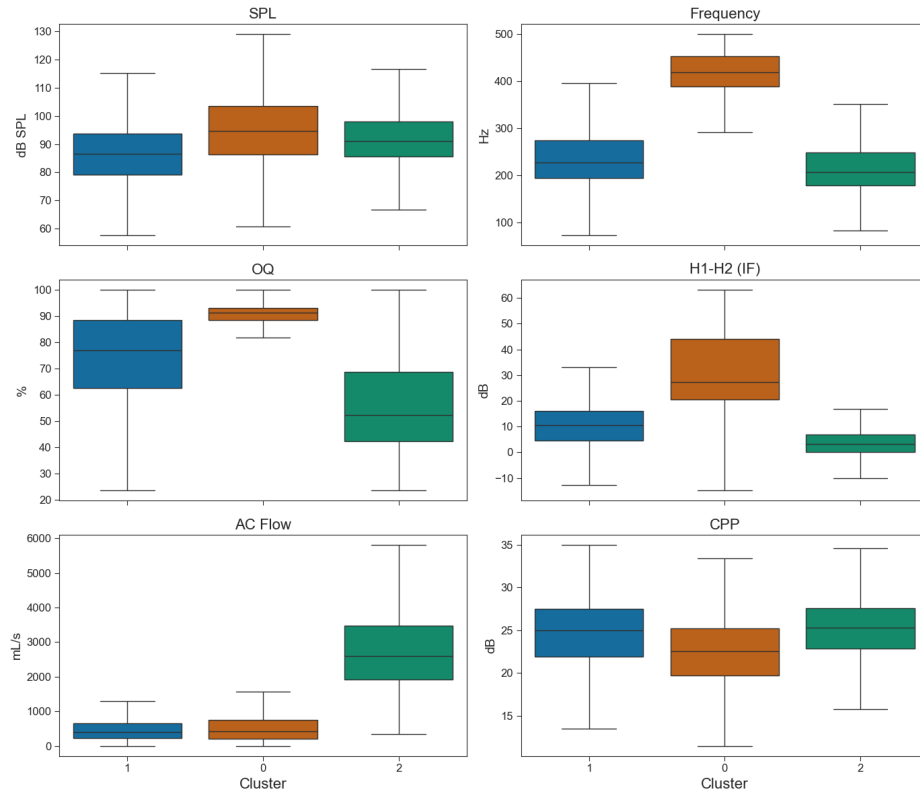


Fig. 4.2: Distribución de features por clúster.

La distribución de features para el **clúster 1** confirma que este corresponde a un comportamiento vocal basal, debido a sus valores ubicados dentro de rangos típicos del habla modal. Estas cifras se consideran fisiológicamente razonables, por lo que este clúster sirve como línea base de la fonación cotidiana.

Por otra parte, el **clúster 2** se caracteriza por presentar un AC Flow desproporcionadamente alto, con una mediana en torno a 2600 mL/s. Caudales superiores a 2 L/s resultan poco plausibles en condiciones de fonación normal [41] y suelen estar asociados a errores de calibración, ajustes inadecuados de ganancia en el modelo aerodinámico o artefactos de medición. Sin embargo, dado que el resto de las features de este clúster no presenta valores anómalos ni extremos, es menos probable que se trate de un error general en la adquisición, y más factible que el problema resida específicamente en la ganancia del modelo que estima el flujo aéreo.

Por último, el **clúster 0** concentra ventanas con frecuencias fundamentales muy

altas, acompañadas de valores elevados de OQ y H1-H2. Este patrón es coherente con una voz cantada, circunstancia esperable dado que una fracción de los participantes son cantantes. El nivel sonoro y un flujo moderado son compatibles con la hipótesis de un gesto vocal pensado para proyección más que para conversación cotidiana.

Para respaldar la hipótesis de que el **clúster 0** agrupa ventanas de voz cantada, se utilizó una variable booleana incluida en la base de datos que indica si una ventana corresponde efectivamente a canto. Los resultados muestran que un 83,5 % de las ventanas clasificadas en este clúster son de voz cantada, en marcado contraste con los clústers 1 y 2, donde sólo el 19 % y 5 % de las ventanas, respectivamente, presentan esta característica.

4.4. *Diferencias entre condiciones clínicas*

Para contrastar la distribución de los clústeres entre las distintas condiciones clínicas (Control, Pre-tratamiento y Post-tratamiento), la Figura 4.3 muestra el espacio latente separado por condición y coloreado por clúster.

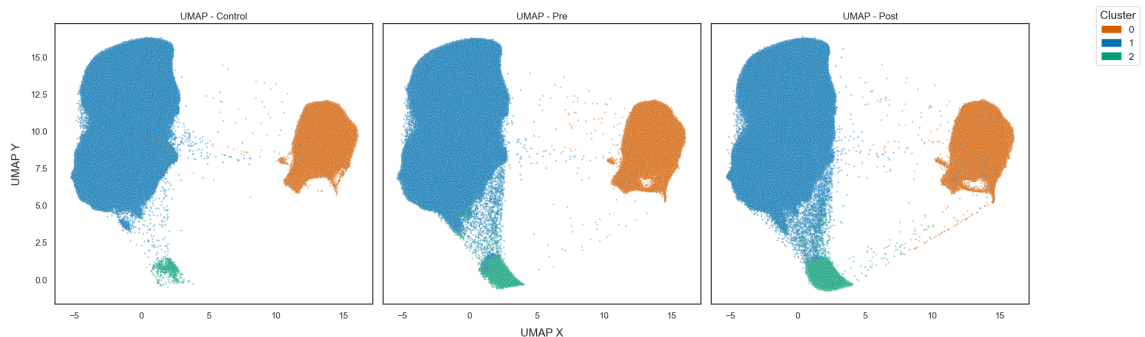


Fig. 4.3: Visualización UMAP de los datos separado por condición, coloreado por clúster.

A grandes rasgos, las tres proyecciones comparten una estructura similar. La principal diferencia entre estos radica en la zona del clúster 2, donde la densidad de puntos para este es menor para la condición Control. Para las condiciones patológicas, esta región se hace más densa y parece estar más interconectada con el clúster 1.

Para cuantificar este patrón, la Tabla 4.3 resume el porcentaje de ventanas por clúster y condición.

Clúster	Pre (%)	Post (%)	Control (%)
0	16.094	18.323	22.890
1	82.804	70.707	77.071
2	1.102	10.970	0.039

Tab. 4.3: Frecuencia relativa (%) de ventanas por condición en cada clúster

El **clúster 1** sigue siendo dominante (70-83 %), lo que sugiere un patrón vocal ampliamente compartido, independientemente del estado clínico. El **clúster 0** muestra un incremento gradual de Pre, Post a Control, posiblemente asociado a fonación de voz cantada o a una recuperación de la participación vocal típica.

El caso del **clúster 2** requiere un comentario adicional. En la sección anterior se demostró que las ventanas agrupadas en este clúster presentan valores de AC Flow que superan los 2 L/s, rango fisiológicamente improbable y, por tanto, compatible con un error de ganancia en la estimación aerodinámica. Sin embargo, su presencia se multiplica por diez tras la terapia (de 1.1 % a 10.9 %), esto podría deberse a un número mayor de ventanas afectadas por desajustes en la calibración del modelo aerodinámico durante las grabaciones Post. También cabe la posibilidad de que, tras la intervención terapéutica, los sujetos adopten un gesto vocal que eleve el flujo aéreo medio -o, al menos, la señal que el modelo interpreta como tal- y que dicho gesto quede capturado por el algoritmo como un clúster diferenciado.

Para evaluar la capacidad de los clústeres obtenidos para discriminar entre las distintas condiciones clínicas, se calculó el tamaño del efecto de Cohen's d sobre la proporción diaria de ventanas asignadas a cada clúster, agrupadas por sujeto y día. Este análisis permite cuantificar, para cada clúster, qué tan diferente es su presencia relativa entre dos condiciones, independientemente del tamaño muestral. Donde valores $|d|$ entre 0.3 y 0.5 son considerados efectos pequeños pero potencialmente relevantes,

entre 0.5 y 0.8 indican efectos moderados, y valores superiores a 0.8 se interpretan como efectos grandes.

Clúster	Pre-Post	Pre-Control	Post-Control
0	-0.182	0.316	0.112
1	0.404	-0.179	0.276
2	-0.332	-0.333	-0.442

Tab. 4.4: Tamaño del efecto (Cohen's d) entre condiciones para cada clúster. Valores positivos indican mayor presencia en la primera condición.

Se observa en la Tabla 4.4 que, si bien ningún clúster exhibe un tamaño de efecto alto para discriminar entre condiciones, hay algunos casos que presentan diferencias moderadas que podrían ser relevantes.

El **clúster 1** muestra un tamaño de efecto moderado en la comparación *Pre-Post* ($d = 0,40$), lo que sugiere que la terapia redistribuye parte de las ventanas hacia otros clústeres.

Asimismo, el **clúster 2** evidencia un tamaño de efecto también moderado en la comparación *Post-Control* ($d = -0,44$), dado el potencial artefacto del flujo, esta diferencia puede deberse a cambios genuinos en el habla, a errores de modelado, o a una combinación de ambos.

Estos resultados indican que, tras la intervención, los sujetos no solo reducen la proporción de ventanas en el clúster basal (1), sino que también incrementan notablemente la pertenencia al clúster 2 -cuyo significado fisiológico sigue abierto a interpretación. Futuros análisis deberán descartar o corregir posibles errores de ganancia para aclarar si se trata de un patrón fonatorio emergente o de un artefacto instrumental.

4.5. Efecto del tamaño de ventana

En esta sección se presentan los resultados más relevantes obtenidos al utilizar una segmentación con ventanas de un segundo de duración.

La configuración de hiperparámetros que maximizó el valor del índice DBCV fue: **n_components = 12**, **n_neighbors = 20**, y **min_dist = 0.01**. Bajo esta configuración se obtuvo un **DBCV de 0.791** y un **Silhouette Score de 0.613**. El modelo identificó un total de 6 clústers, clasificando únicamente un **0.002 %** de las ventanas como ruido.

La Figura 4.4 muestra la representación del espacio latente generado con UMAP, coloreado según el clúster asignado a cada punto.

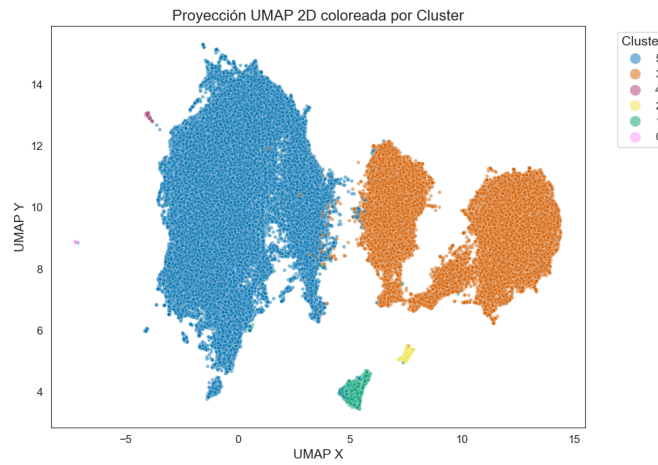


Fig. 4.4: Visualización UMAP de los datos en ventanas de 1 segundo, coloreada por clúster.

El aumento del tamaño de ventana provoca un cambio en la estructura del espacio latente, facilitando la emergencia de nuevas agrupaciones. En este caso, el algoritmo identificó tres clústers adicionales con respecto a la configuración con ventanas más pequeñas.

Los clústers emergentes presentan un tamaño reducido, en especial los **clústers 0 y 4**, por lo que deben interpretarse con cautela. Estos podrían corresponder a anomalías puntuales en ciertos sujetos o simplemente a ruido clasificado como clúster según el algoritmo. No obstante, debido al mayor tamaño de la ventana, cada punto representa más duración que en el escenario de 50ms; en consecuencia, un clúster reducido en número de puntos puede abarcar un tiempo total no despreciable.

Para analizar las características distintivas de cada clúster, a continuación se presentan los diagramas de caja (boxplots) por feature:

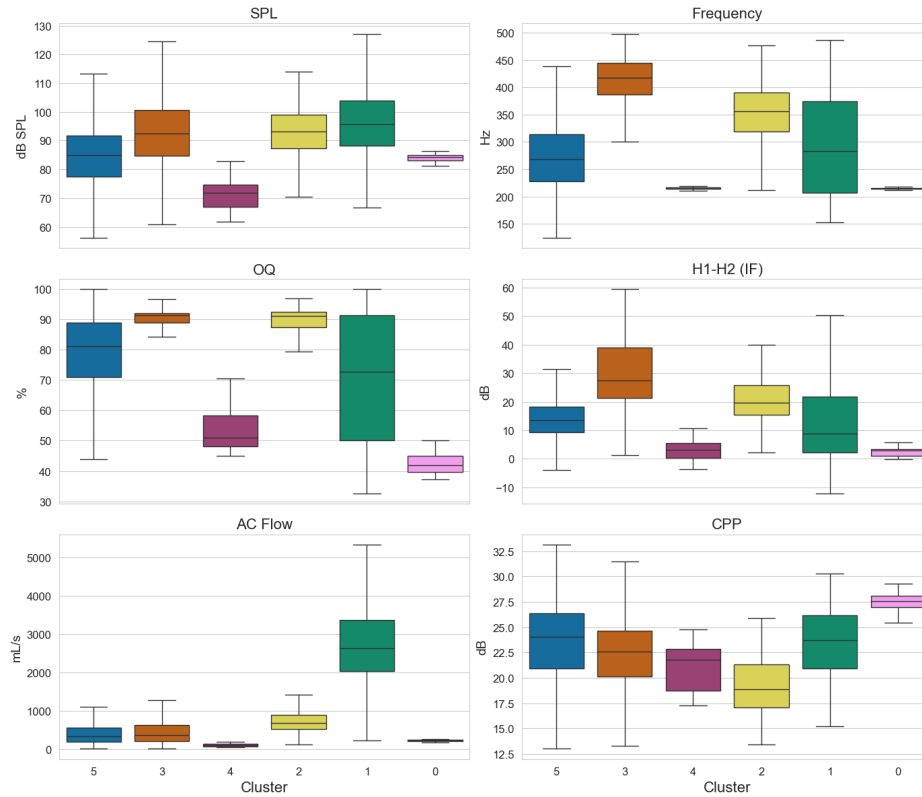


Fig. 4.5: Distribución de features por clúster, para datos en ventanas de 1s.

Según la distribución de las features en cada clúster, se puede concluir que el **clúster 5** corresponde al clúster de fonación cotidiana, debido a sus valores ubicados dentro de rangos típicos del habla modal. El **clúster 1** corresponde al que presenta anomalías dentro del parámetro AC Flow, mostrando valores desproporcionalmente altos y el **clúster 3** corresponde a la voz cantada, debido a la distribución de su frecuencia fundamental.

Por otra parte, el **clúster 0** exhibe la frecuencia fundamental y el nivel SPL más bajos de las voces habladas junto con un OQ reducido. Esta combinación sugiere una fonación con la glotis relativamente cerrada y un uso económico del flujo aéreo, propia de un habla tenue pero todavía modal.

El **clúster 2** combina una frecuencia alta y un OQ muy abierto, junto con AC Flow y H1-H2 intermedios. Puede interpretarse como una transición entre voz cantada y habla expresiva en registro alto.

Finalmente, el **clúster 4** presenta el SPL más bajo y el menor flujo aéreo, lo que podría reflejar una producción susurrada o interválica con esfuerzo mínimo.

Estos resultados muestran que ampliar la ventana temporal no altera los clústers fundamentales (fonación modal, voz cantada y flujo anómalo), pero permite aislar matices adicionales de la producción vocal.

4.6. *Preservación Espacial y Temporal*

Para complementar la evaluación visual del espacio latente, se calcularon las métricas de preservación espacial y temporal descritas en la Sección 2.3, tanto para el caso de ventanas de 50 ms como de 1 s. Para cada sujeto se seleccionó un día aleatorio y se promediaron los resultados obtenidos, con el fin de obtener un valor representativo del desempeño general. Además, se compararon los resultados entre ambas configuraciones de ventana temporal, con el objetivo de analizar cómo la escala de segmentación (50 ms vs. 1 s) afecta la preservación de la estructura espacial y la coherencia temporal del espacio embebido. Las expresiones matemáticas de cada métrica se detallan en el Anexo.

En la Tabla 4.5 se resume el resultado numérico de cada una de estas.

Categoría	Métrica	50 ms	1 s
Preservación de distancia	S_{Stress}	0.528	0.392
	S_{Pearson}	0.655	0.627
	S_{Spearman}	0.730	0.762
	S_{Kendall}	0.539	0.587
Preservación de vecinos	$\overline{S_{\text{trust}}}$	0.900	0.888
	$\overline{S_{\text{cont}}}$	0.908	0.961
Estabilidad temporal	T_{Stress}	0.636	0.437
	T_{Pearson}	0.595	0.635
	T_{Spearman}	0.749	0.754
	T_{Kendall}	0.559	0.565

Tab. 4.5: Comparación de métricas de preservación espacial y temporal para proyecciones UMAP con ventanas de 50 ms y 1 s.

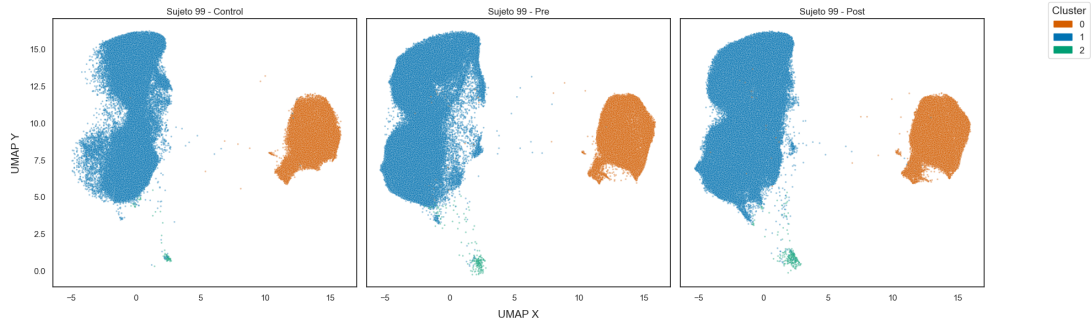
- **Preservación espacial.** El *Normalized Stress* (S_{Stress}) disminuyó de 0,528 a 0,392 al utilizar ventanas de 1s, lo que indica una mejor preservación global de las distancias en esta escala temporal (valores menores implican menor distorsión). Las correlaciones de Pearson, Spearman y Kendall se mantuvieron elevadas en ambos casos, con una leve mejora para las ventanas de 1s, especialmente en las métricas no lineales (Spearman 0,76 y Kendall 0,59), lo que sugiere una representación más coherente del orden relativo de las distancias.
- **Preservación de vecinos.** Los promedios de *Trustworthiness* y *Continuity* fueron altos en ambas configuraciones, evidenciando que las vecindades locales se preservan adecuadamente. No obstante, se observa un ligero descenso en la *Trustworthiness* y un aumento en la *Continuity* para la ventana de 1s, lo que refleja una proyección algo más coherente en términos de continuidad global, aunque con una mínima pérdida en la preservación estricta de los vecinos originales.
- **Estabilidad temporal.** El *Temporal Stress* (T_{Stress}) se redujo de 0,636 a 0,437, indicando una menor deformación en la magnitud de los desplazamientos entre instantes consecutivos para la ventana de 1s. Las correlaciones temporales (Pearson, Spearman y Kendall) aumentaron levemente, lo que sugiere una mejora en la consistencia del orden temporal y en la estabilidad global del embedding al emplear intervalos más largos.

Estas métricas nos muestran que la proyección mantiene una buena fidelidad tanto espacial como temporal, con valores de preservación elevados y deformaciones moderadas. El uso de ventanas más largas mejora ligeramente la consistencia temporal y la preservación global de las distancias, a costa de una mínima pérdida en la preservación local de vecinos. En comparación con los valores reportados en [40], los resultados obtenidos se encuentran dentro de un rango considerado de buena calidad, lo que confirma que las proyecciones generadas logran preservar de manera adecuada la estructura espacial y la coherencia temporal del espacio original.

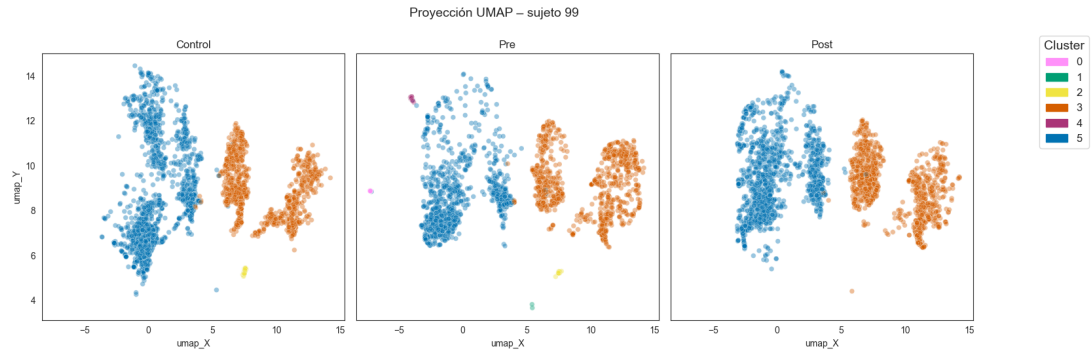
4.7. Análisis de caso individual

Para ilustrar cómo se distribuyen los clústeres a nivel individual y explorar posibles eventos vocales atípicos, se analiza el caso del sujeto 99, comparado con su control emparejado. Este análisis no busca generalizar resultados, sino funcionar como una prueba de concepto que permita observar dinámicas temporales o patrones particulares que podrían perderse en los análisis agregados.

A continuación, se presenta la visualización UMAP de este sujeto, utilizando dos resoluciones temporales distintas:



(a) Visualización UMAP con ventanas de 50 ms.



(b) Visualización UMAP con ventanas de 1 s.

Fig. 4.6: Visualización UMAP del sujeto 99 y su control, separados por condición y coloreados según clúster.

En la Figura 4.6a, correspondiente a ventanas de 50 ms, no se observan diferencias sustanciales entre condiciones, más allá de una mayor proporción del clúster 2 en una de ellas. En cambio, en la visualización con ventanas de 1 s (Figura 4.6b), se evidencia

que durante la condición pretratamiento aparecen los clústers 0 y 4, los cuales están ausentes en las otras condiciones.

Dado el tamaño reducido de estos clústers, podrían representar eventos atípicos asociados a la condición patológica del sujeto. Para explorar su aparición en el tiempo, se grafica la proporción horaria de cada clúster a lo largo de la semana para el sujeto 99 en la etapa pretratamiento.

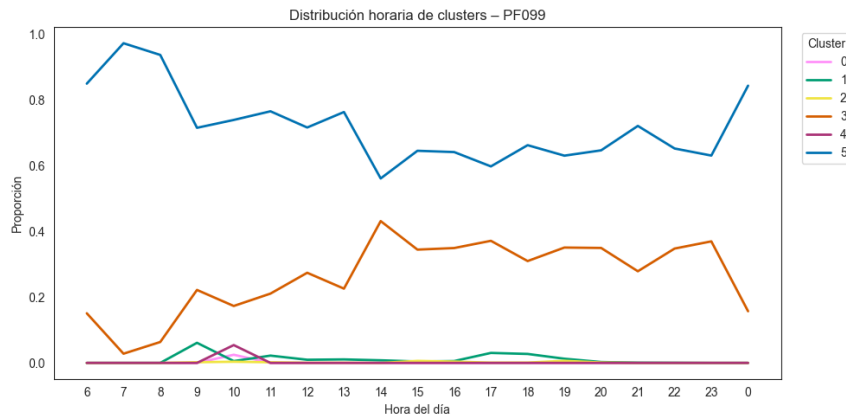


Fig. 4.7: Distribución horaria de clústers para el sujeto 99 en la condición pretratamiento.

Se observa que los clústers 0 y 4 emergen de manera puntual durante la mañana, alrededor de las 10:00 horas. Ambos se caracterizan por presentar niveles bajos de SPL, lo que sugiere que podrían corresponder a una voz matutina afectada por la patología, reflejando un posible estado de fatiga vocal al inicio del día. Si bien estas observaciones no permiten extraer conclusiones definitivas, respaldan la utilidad del enfoque para detectar patrones transitorios de interés clínico en sujetos individuales.

5. CONCLUSIONES Y TRABAJO FUTURO

Este trabajo presentó un pipeline no supervisado para caracterizar el comportamiento vocal ambulatorio a partir de señales captadas por acelerómetro, combinando UMAP con HDBSCAN. En conjunto, ambas técnicas permitieron descubrir una estructura latente clara y estable en los datos, además de habilitar una visualización interpretable y coherente espacial y temporalmente. Esto respaldado por los resultados tanto de los índices de calidad de agrupamiento como de preservación temporal y espacial.

Los valores obtenidos para las métricas de preservación espacial y temporal indican que la proyección conserva de forma satisfactoria la estructura original de los datos, con correlaciones elevadas y niveles de distorsión moderados. En particular, al comparar las escalas temporales de 50ms y 1s, se observó que las ventanas más largas mejoran la coherencia temporal y la preservación global de las distancias, sin comprometer significativamente la fidelidad local.

La distribución de features por clúster muestra un clúster "basal" que concentra la fonación modal cotidiana, un clúster asociado fuertemente a voz cantada, coherente con la presencia de cantantes en la muestra, y un clúster minoritario con AC Flow anómalamente alto, probablemente vinculado a artefactos de ganancia/modelado más que a un gesto fisiológico plausible. Este análisis apoya la utilidad del método para separar distintos patrones y casos potencialmente espurios asociados al habla.

Adicionalmente, el clúster basal domina en Pre, Post y Control. Sin embargo, se observó un aumento marcado del clúster con AC Flow elevado en Post respecto de Pre y Control, así como variaciones moderadas en la proporción diaria de ventanas por clúster entre condiciones. Estos resultados sugieren que parte de los cambios Post-

tratamiento podrían estar influidos por cambios de calibración/estimación o por ajustes fonatorios que el modelo interpreta como mayor flujo; ambos escenarios ameritan verificación específica.

Al pasar de ventanas atómicas de 50 ms a subsecuencias de 1 s, el método resolvió más matices sin perder los patrones esenciales. La mayor ventana capturó configuraciones más largas (p.ej., habla tenue modal con SPL y OQ bajos; transiciones hacia registros altos) que no emergían con la escala fina. Esto refuerza que la escala temporal impacta la granularidad del fenómeno observado, preservando lo esencial y revelando subestados adicionales.

El análisis del sujeto 99 ilustró cómo el enfoque permite ubicar temporalmente clústeres poco frecuentes (p.ej., estados de baja intensidad matinal) que podrían vincularse con fatiga o condición particular del día. Aun cuando no busca generalización, esta observación por sujeto-día muestra el potencial del método para detectar episodios transitorios de interés clínico.

UMAP + HDBSCAN demuestra ser una herramienta robusta y escalable para condensar millones de ventanas en un espacio latente interpretable, perfilar clústers diferenciados, comparar condiciones clínicas y escalas temporales, y explorar casos individuales y eventos puntuales. Este marco aporta bases objetivas para monitoreo terapéutico y detección temprana de comportamientos hiperfuncionales en la vida diaria.

Como trabajo futuro, se propone repetir el análisis restringiendo a contextos controlados (mañana vs. tarde; días laborales vs. fin de semana, etc.) para ver si emergen clústers adicionales y si los tamaños de efecto entre condiciones se amplifican. Además, extender la exploración por franjas horarias para caracterizar fatiga/recuperación diaria.

Finalmente, se sugiere priorizar protocolos de verificación de ganancia/calibración para IBIF, como puede ser chequeos diarios automáticos, límites fisiológicos adaptativos o detección de outliers por sujeto.

ANEXO

Fórmulas

Métricas espaciales

- **Normalized Stress:**

$$S_{\text{Stress}} = \frac{\sum_{i,j} (d_{ij} - d'_{ij})^2}{\sum_{i,j} d_{ij}^2}$$

donde d_{ij} es la distancia euclidiana entre los puntos x_i y x_j en el espacio original, y d'_{ij} es la distancia entre sus proyecciones $P(x_i)$ y $P(x_j)$.

- **Shepard diagram correlations:** miden la relación entre d_{ij} y d'_{ij} mediante coeficientes de correlación:

$$S_{\text{Pearson}} = \rho_{\text{Pearson}}(d_{ij}, d'_{ij})$$

$$S_{\text{Spearman}} = \rho_{\text{Spearman}}(d_{ij}, d'_{ij})$$

$$S_{\text{Kendall}} = \rho_{\text{Kendall}}(d_{ij}, d'_{ij})$$

- **Trustworthiness:**

$$S_{\text{Trust}} = 1 - \frac{2}{nk(2n - 3k - 1)} \sum_{i=1}^n \sum_{j \in U_k(i)} (r_{ij} - k)$$

donde $U_k(i)$ es el conjunto de vecinos que están en la proyección pero no en el espacio original, y r_{ij} es el rango de j respecto a i en el espacio original.

- **Continuity:**

$$S_{\text{Cont}} = 1 - \frac{2}{nk(2n - 3k - 1)} \sum_{i=1}^n \sum_{j \in V_k(i)} (\hat{r}_{ij} - k)$$

donde $V_k(i)$ es el conjunto de vecinos que están en el espacio original pero no en la proyección, y $\hat{r}_{ij}$ es el rango de j respecto a i en el espacio proyectado.

Métricas temporales

- **Temporal Stress:**

$$T_{\text{Stress}} = \frac{\sum_i (\Delta d_i - \Delta d'_i)^2}{\sum_i (\Delta d_i)^2}$$

donde $\Delta d_i = \|x_{i+1} - x_i\|$ representa el cambio consecutivo en el espacio original, y $\Delta d'_i = \|P(x_{i+1}) - P(x_i)\|$ el cambio correspondiente en el espacio proyectado. Valores bajos de T_{Stress} indican que la proyección preserva adecuadamente la dinámica temporal.

- **Temporal Shepard diagram correlations:**

$$T_{\text{Pearson}} = \rho_{\text{Pearson}}(\Delta d_i, \Delta d'_i)$$

$$T_{\text{Spearman}} = \rho_{\text{Spearman}}(\Delta d_i, \Delta d'_i)$$

$$T_{\text{Kendall}} = \rho_{\text{Kendall}}(\Delta d_i, \Delta d'_i)$$

Bibliografía

- [1] R. E. Hillman, C. E. Stepp, J. H. V. Stan, M. Zañartu, and D. D. Mehta, “An updated theoretical framework for vocal hyperfunction,” *American Journal of Speech-Language Pathology*, vol. 29, no. 4, pp. 2254–2260, 2020, accessed: 2024-05-01. [Online]. Available: https://pubs.asha.org/doi/abs/10.1044/2020_AJSLP-20-00104
- [2] D. Mehta, J. Van Stan, M. Zañartu, M. Ghassemi, J. Gutttag, V. Espinoza Catalán, J. Cortés, H. Cheyne, and R. Hillman, “Using ambulatory voice monitoring to investigate common voice disorders: Research update,” *Frontiers in Bioengineering and Biotechnology*, vol. 3, p. 155, 10 2015.
- [3] S. D. Gray, I. R. Titze, R. W. Chan, and T. H. Hammond, “Vocal fold pathology associated with hyperfunctional voice use: experimental evidence,” *Annals of Otolaryngology, Rhinology & Laryngology*, vol. 104, no. 1, pp. 13–18, 1987.
- [4] R. E. Hillman, E. B. Holmberg, J. S. Perkell, M. Walsh, and C. Vaughan, “Objective assessment of vocal hyperfunction: an experimental framework and initial results,” *Journal of Speech and Hearing Research*, vol. 32, no. 2, pp. 373–392, 1989.
- [5] E. Van Houtte, K. Van Lierde, E. D’haeseleer, and S. Claeys, “The prevalence of laryngeal pathology in a treatment-seeking population with dysphonia,” *The Laryngoscope*, vol. 120, no. 2, pp. 306–312, 2010.
- [6] M. El Uali Abeida, K. R. Scherer, A. Höhne, W. Seidner, U. Eysholdt, and F. Rosanowski, “Study of the influence of psychological factors in the etiology of vo-

- cal nodules in women,” *Journal of Voice*, vol. 27, no. 1, pp. 129.e15–129.e20, 2013.
- [7] M. Ghassemi, Z. Syed, D. D. Mehta, J. H. Van Stan, R. E. Hillman, and J. V. Guttag, “Uncovering voice misuse using symbolic mismatch,” *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1634–1643, 2016.
- [8] J. H. Van Stan, D. D. Mehta, and R. E. Hillman, “The effect of voice ambulatory biofeedback on the daily performance and retention of a less hyperfunctional vocal motor pattern in patients with phonotraumatic vocal hyperfunction,” *Journal of Speech, Language, and Hearing Research*, vol. 58, no. 6, pp. 1624–1636, 2015, estudio que muestra la utilidad clínica del monitoreo ambulatorio para intervención (biofeedback).
- [9] M. Södersten, S. Granqvist, B. Hammarberg, and A. Szabo, “Long-term average spectrum (ltas) and voice range profile (vrp) of patient and non-patient children, women and men,” *Folia Phoniatica et Logopaedica*, vol. 54, no. 4, pp. 206–215, 2002, trabajo asociado al desarrollo y uso del sistema de monitoreo ambulatorio VoxLog.
- [10] A. Szabo, B. Hammarberg, A. Häggqvist, and M. Södersten, “A voice accumulator device for evaluation of voice use,” *Logopedics Phoniatrics Vocology*, vol. 26, no. 3, pp. 102–117, 2001, describe el “Voice Accumulator”, un dispositivo temprano para dosificar el uso vocal.
- [11] D. D. Mehta, M. Zañartu, S. W. Feng, H. A. Cheyne II, and R. E. Hillman, “Mobile voice health monitoring using a wearable accelerometer sensor and a smartphone platform,” *IEEE Transactions on Biomedical Engineering*, vol. 59, no. 11, pp. 3090–3096, 2012.
- [12] M. Zañartu, J. C. Ho, D. D. Mehta, R. E. Hillman, and G. R. Wodicka, “Sub-

- glottal impedance-based inverse filtering of voiced sounds using neck surface acceleration,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 9, pp. 1929–1939, 2013.
- [13] J. Van Stan, D. Mehta, A. Ortiz, J. Burns, L. Toles, K. Marks, M. Vangel, T. Hron, S. Zeitels, and R. Hillman, “Differences in weeklong ambulatory vocal behavior between female patients with phonotraumatic lesions and matched controls,” *Journal of Speech, Language, and Hearing Research*, vol. 63, pp. 372–384, 01 2020.
- [14] V. Espinoza Catalán, M. Zañartu, J. Van Stan, D. Mehta, and R. Hillman, “Glottal aerodynamic measures in women with phonotraumatic and nonphonotraumatic vocal hyperfunction,” *Journal of Speech, Language, and Hearing Research*, vol. 60, pp. 2159–2169, 08 2017.
- [15] J. Cortés, V. Espinoza Catalán, M. Ghassemi, D. Mehta, J. Van Stan, R. Hillman, J. Guttag, and M. Zañartu, “Ambulatory assessment of phonotraumatic vocal hyperfunction using glottal airflow measures estimated from neck-surface acceleration,” *PLOS ONE*, vol. 13, p. e0209017, 12 2018.
- [16] J. P. Cortés, V. M. Espinoza, J. H. Van Stan, A. Ghasemzadeh, D. D. Mehta, R. E. Hillman, and M. Zañartu, “Ambulatory monitoring of subglottal pressure estimated from neck-surface vibration in individuals with and without voice disorders,” *IEEE Transactions on Biomedical Engineering*, vol. 69, no. 12, pp. 3707–3718, 2022.
- [17] A. Ghasemzadeh, R. E. Hillman, and D. D. Mehta, “Consistency of the signature of phonotraumatic vocal hyperfunction across different ambulatory voice measures,” *Journal of Speech, Language, and Hearing Research*, vol. 67, no. 6, pp. 1997–2014, 2024.
- [18] H. Huang, Y. Wang, C. Rudin, and E. P. Browne, “Towards a comprehensive

- evaluation of dimension reduction methods for transcriptomic data visualization,” *Communications Biology*, vol. 5, no. 1, p. 719, 2022. [Online]. Available: <https://doi.org/10.1038/s42003-022-03628-x>
- [19] I. T. Jolliffe and J. Cadima, “Principal component analysis: a review and recent developments,” *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 374, no. 2065, p. 20150202, 2016.
- [20] L. van der Maaten and G. Hinton, “Visualizing data using t-sne,” *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, 11 2008.
- [21] L. McInnes, J. Healy, and J. Melville, “Umap: Uniform manifold approximation and projection for dimension reduction,” *arXiv preprint arXiv:1802.03426*, 2018. [Online]. Available: <https://arxiv.org/abs/1802.03426>
- [22] P. Nazari, S. Damrich, and F. A. Hamprecht, “Geometric autoencoders – what you see is what you decode,” *arXiv preprint arXiv:2306.17638*, 2023. [Online]. Available: <https://arxiv.org/abs/2306.17638>
- [23] A. Mazher, “Visualization framework for high-dimensional spatio-temporal hydrological gridded datasets using machine-learning techniques,” *Water*, vol. 12, no. 2, 2020. [Online]. Available: <https://www.mdpi.com/2073-4441/12/2/590>
- [24] M. Ali, M. W. Jones, X. Xie, and M. Williams, “Timecluster: dimension reduction applied to temporal data for visual analytics,” *The Visual Computer*, vol. 35, no. 6, pp. 1013–1026, 2019. [Online]. Available: <https://doi.org/10.1007/s00371-019-01673-y>
- [25] V. V. Baligodugula and F. Amsaad, “Unsupervised learning: Comparative analysis of clustering techniques on high-dimensional data,” 2025. [Online]. Available: <https://arxiv.org/abs/2503.23215>

- [26] J. H. Leland McInnes and J. Melville, “UMAP: Uniform manifold approximation and projection for dimension reduction - documentation,” <https://umap-learn.readthedocs.io>, 2023, accessed: 2025-05-20.
- [27] L. McInnes, J. Healy, and J. Melville, “Umap: Uniform manifold approximation and projection for dimension reduction,” 2020. [Online]. Available: <https://arxiv.org/abs/1802.03426>
- [28] A. K. Jain, M. N. Murty, and P. J. Flynn, “Data clustering: A review,” *ACM Computing Surveys (CSUR)*, vol. 31, no. 3, pp. 264–323, 1999.
- [29] N. Tavakoli, S. Siامي-Namini, M. A. Khanghah, F. M. Soltani, and A. S. Namin, “Clustering time series data through autoencoder-based deep learning models,” *CoRR*, vol. abs/2004.07296, 2020. [Online]. Available: <https://arxiv.org/abs/2004.07296>
- [30] T. W. Liao, “Clustering of time series data? a survey,” *Pattern Recognition*, vol. 38, no. 11, pp. 1857–1874, 2005.
- [31] M. Ali, A. Alqahtani, M. W. Jones, and X. Xie, “Clustering and classification for time series data in visual analytics: A survey,” *IEEE Access*, vol. 7, pp. 181 314–181 338, 2019.
- [32] C. Pealat, G. Bouleux, and V. Cheutet, “Improved time-series clustering with umap dimension reduction method,” in *2020 25th International Conference on Pattern Recognition (ICPR)*, 2021, pp. 5658–5665.
- [33] L. McInnes, J. Healy, and S. Astels, “hdbscan: Hierarchical density based clustering,” *The Journal of Open Source Software*, vol. 2, 03 2017.
- [34] J. H. Leland McInnes and S. Astels, “HDBSCAN: Hierarchical density-based spatial clustering of applications with noise - documentation,” <https://hdbscan.readthedocs.io/en/latest/>, 2024, accessed: 2025-05-23.

- [35] P. J. Rousseeuw, “Silhouettes: A graphical aid to the interpretation and validation of cluster analysis,” *Journal of Computational and Applied Mathematics*, vol. 20, pp. 53–65, 1987.
- [36] Number Analytics, “Silhouette score clustering evaluation,” <https://www.numberanalytics.com/blog/silhouette-score-clustering-evaluation>, 2022, accessed: 2025-07-02.
- [37] D. Moulavi, P. Andretta Jaskowiak, R. Campello, A. Zimek, and J. Sander, “Density-based clustering validation,” 04 2014.
- [38] J. Venna and S. Kaski, “Neighborhood preservation in nonlinear projection methods: An experimental study,” in *International Conference on Artificial Neural Networks*. Springer, 2001, pp. 485–491.
- [39] J. B. Kruskal, “Multidimensional scaling by optimizing goodness of fit to a non-metric hypothesis,” *Psychometrika*, vol. 29, no. 1, pp. 1–27, 1964.
- [40] E. F. Vernier, R. Garcia, I. P. da Silva, J. L. D. Comba, and A. C. Telea, “Quantitative evaluation of time-dependent multidimensional projection techniques,” *CoRR*, vol. abs/2002.07481, 2020. [Online]. Available: <https://arxiv.org/abs/2002.07481>
- [41] R. I. Zraick, L. Smith-Olinde, and L. L. Shotts, “Adult normative data for the kaypentax phonatory aerodynamic system model 6600,” *Journal of Voice*, vol. 26, no. 2, pp. 164–176, 2012. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0892199711000075>