ESCUELA POLITÉCNICA NACIONAL

FACULTAD DE INGENIERÍA DE SISTEMAS

DETECCIÓN DE ACTIVIDAD DE VOZ EN AMBIENTE NO RUIDOSO

TRABAJO DE TITULACIÓN PREVIO A LA OBTENCIÓN DEL TÍTULO DE INGENIERO EN SISTEMAS INFORMÁTICOS Y DE COMPUTACIÓN

EDGAR SANTIAGO YÉPEZ LÓPEZ
edgar.yepez@epn.edu.ec

DIRECTOR: PHD. JOSAFÁ DE JESÚS AGUIAR PONTES
josafa.aguiar@epn.edu.ec

Quito, Agosto 2018

AVAL

Certifico que el presente trabajo fue desarrollado por Edgar Santiago Yépez López bajo mi supervisión.

PHD. JOSAFÁ DE JESÚS AGUIAR PONTES
DIRECTOR DEL TRABAJO DE TITULACIÓN

DECLARACIÓN DE AUTORÍA

Yo, Edgar Santiago Yépez López, declaro bajo juramento que el trabajo aquí descrito, incluidas tablas e imágenes, es de mi autoría; que no ha sido previamente presentado para ningún grado o calificación profesional; y, que he consultado las referencias bibliográficas que se incluye en este documento.

A través de la presente declaración cedo mis derechos de propiedad intelectual correspondientes a este trabajo a la Escuela Politécnica Nacional, según lo establecido por su Reglamento, por la normatividad institucional vigente y por la Ley de Propiedad Intelectual.

EDGAR SANTIAGO YÉPEZ LÓPEZ

ÍNDICE

RESUMEN

El presente trabajo investiga la detección de actividad de voz en señales acústicas del habla humana en idioma francés que se producen en ambientes libres de ruido. Propone un método capaz de diferenciar eventos de habla y de pausa considerando que un evento de pausa que sucede por la pronunciación de una consonante oclusiva sorda debe tomarse como evento de habla. Para ello, emplea componentes especializados en detección de silencio, de inhalación y de consonantes oclusivas sordas basados en clasificadores de tipo Support Vector Machine (SVM), así como también un componente de toma de decisión basado en una Máquina de Estados Finitos. En este sentido, el presente trabajo muestra el drástico impacto que tiene la detección de consonantes oclusivas sordas sobre la predicción final de eventos de habla y pausa.

PALABRAS CLAVE: oclusiva, inhalación, silencio, habla, pausa

ABSTRACT

The current work investigates the detection of voice activity in acoustic signals of French human speech that occur in noise-free environments. It proposes a method capable of differentiating speech and pause events, considering that a pause event that occurs due to the pronunciation of a voiceless stop consonant must be taken as a speech event. For this, it uses components specialized in detection of silence, inhalation and voiceless stop consonants based on Support Vector Machine classifiers (SVM), as well as a component for decision making based on a Finite State Machine. In this sense, the current work shows the drastic impact that the detection of voiceless stop consonants has on the final prediction of speech and pause events.

KEYWORDS: stops, inhalation, silence, speech, pause

1 INTRODUCCIÓN

Un Sistema de Detección de Actividad de Voz (o VAD por las siglas en inglés para Voice Activity Detection) tiene como finalidad diferenciar entre eventos de habla y pausa dentro de una señal acústica [1, 2]. Las técnicas VAD se aplican en numerosas áreas científicas y técnicas, como por ejemplo, en el Procesamiento del Lenguaje Natural para la construcción de Sistemas de Reconocimiento Automático de la Voz (o ASR por las siglas en inglés para Automatic Speech Recognition) [3], o en las telecomunicaciones para la codificación y transmisión de señales de habla [4]. En este sentido, el requerimiento de la tasa de aciertos asociada a la predicción de un sistema VAD dependerá del área de aplicación. Sin embargo, por la naturaleza del habla, se sabe que un evento de habla no sucede al mismo tiempo que un evento de pausa, pues un locutor no puede hablar y callarse simultáneamente. No obstante, un evento de habla tiene en su composición algunos eventos de pausa, siendo el caso del corto silencio que suele ocurrir previo a pronunciar un fonema de consonante oclusiva sorda [5], tal como `/t/`, `/p/` o `/k/`. Así pues, esta composición de un evento de habla provoca una decaída en la tasa de aciertos asociada a la predicción de un sistema VAD que se rige estrictamente a asegurar que un evento de pausa no forma parte de un evento de habla. Consecuentemente, bajo la suposición de que la señal acústica sucede en un ambiente cuya presencia de ruido es nula o casi nula, el problema se plantea como la diferenciación entre eventos de habla y pausa considerando que un evento de pausa, sucedido por la pronunciación de una consonante oclusiva sorda, está contenido en uno de habla. En este sentido, como solución se propone un método VAD capaz de detectar la presencia de consonantes oclusivas sordas, para así filtrar y considerar como evento de habla a aquellos otrora considerados como pausa por efecto de la pronunciación de las consonantes en cuestión.

El fundamento teórico en el que se basa el presente trabajo establece que un método VAD está compuesto, generalmente, por dos etapas: una de extracción de características y una de decisión [6]. Se toma este fundamento ya que sugiere un enfoque de solución en el que el problema es tratado por partes. Así, al dar una solución a las partes se habrá dado solución al problema general.

La estructura del presente trabajo se desglosa de la siguiente manera: en la sección Metodología se expone el desarrollo del método VAD propuesto; en la sección Resultados se plantea las condiciones de experimentación, se presenta el conjunto de datos utilizado y se expone los resultados obtenidos; en la sección Discusión, dichos resultados son interpretados y comparados con aquellos obtenidos por el método VAD del proyecto WebRTC [7]; finalmente, en la sección Conclusión se expone las conclusiones obtenidas y el trabajo futuro.

1.1 Hipótesis

Es posible filtrar y considerar como evento de habla a un evento de pausa que sucede por la pronunciación de una consonante oclusiva sorda, para así elevar la tasa de aciertos de predicción de habla y pausa.

1.2 Objetivo General

Investigar la detección de actividad de voz en la señal acústica del habla para, mediante el uso de clasificadores automáticos y reglas definidas, diferenciar entre eventos de habla y de pausa considerando que uno de pausa, sucedido por la pronunciación de una consonante oclusiva sorda, está contenido en uno de habla.

1.3 Objetivos Específicos

- Extraer vectores de características de una señal acústica de habla.
- Analizar las características extraídas para identificar eventos de habla y de pausa.
- Identificar y marcar regiones que correspondan a consonantes oclusivas sordas dentro de la señal acústica.
- Filtrar los eventos de pausa que son parte de uno de habla según las marcas de consonantes oclusivas sordas.
- Construir un componente de comparación entre la diferenciación producida por el método VAD y la diferenciación de referencia producida por humanos, para evaluar la precisión, exhaustividad y Medida-F[1] de las predicciones.

1.4 Alcance

El método VAD propuesto se limita a analizar señales acústicas de habla producidas en ambientes con presencia nula o casi nula de ruido. Pretende diferenciar entre eventos de habla y de pausa, asumiendo que un evento de pausa corresponde únicamente a uno de dos tipos: silencio o inhalación. Además, considera que un evento de pausa, debido exclusivamente a la pronunciación de una consonante oclusiva sorda, está contenido en un evento de habla.

1.5 Marco Teórico

Como se mencionó en la sección Introducción, un sistema VAD está compuesto, generalmente, por dos etapas: una de extracción de características y una de decisión [6]. En la primera etapa se selecciona aquellas características acústicas más apropiadas para efecto de diferenciar un evento de habla y de pausa [2]. Posteriormente, se realiza un análisis del comportamiento de dichas características dentro de cada uno de los eventos en cuestión, para establecer reglas que permitan determinar la naturaleza del evento. En la literatura y en trabajos relacionados se sugiere que un conjunto apropiado de características acústicas está compuesto por: Llanura espectral [6, 9], Logaritmo de energía, Tasa de cambios de signo [2, 10, 11] y los Coeficientes Cepstrales en las Frecuencias de

[1] Medida en porcentaje que representa cuán correcta y completa es la predicción devuelta por un clasificador [8].

Mel (o MFCC por las siglas en inglés para Mel Frequency Cepstral Coeffcients) [12, 13]. En la segunda etapa, muchos métodos han sido propuestos para la toma de la decisión final acerca de la naturaleza de un evento en base a las características seleccionadas [14]. Entre algunos de estos métodos se puede rescatar el uso mayoritario de técnicas de umbralización [15], así como también el uso de modelos probabilísticos basados en clasificadores de tipo Support Vector Machine (SVM) [16, 17], redes neuronales [18, 19], modelos ocultos de Markov [19] y modelos de mezclas gaussianas [20]. En particular, el método VAD del proyecto WebRTC [7], que será utilizado para comparación de resultados con el método VAD propuesto, combina el uso del Logaritmo de energía más un modelo de mezclas gaussianas para producir la predicción final [21].

La variedad de los métodos propuestos para la segunda etapa se debe al reto que supone la detección de actividad de voz dentro de una señal acústica ruidosa, pues la literatura coincide en ello. Sin embargo, al tratar con señales acústicas no ruidosas surge un inconveniente debido a la presencia de fonemas correspondientes a consonantes oclusivas sordas en la señal. Un fonema de consonante oclusiva sorda se caracteriza por presentar un bloqueo sin vibración de cuerdas vocales y posterior liberación abrupta del aire en el tracto vocal [5, 22], como en los fonemas `/t/`, `/p/` y `/k/`. Dicho bloqueo del aire, de duración en el orden de los milisegundos, se refleja como una región de silencio en la señal acústica. Consecuentemente, un sistema VAD que pase por alto el hecho descrito y garantice que la región de silencio en cuestión corresponde a un evento de pausa producirá una baja tasa de aciertos en la predicción final de eventos de habla y de pausa. En este sentido, el presente trabajo busca solventar este inconveniente, siguiendo el fundamento teórico expuesto y como se detalla en la sección Metodología.

2 METODOLOGÍA

El método VAD propuesto desglosa su funcionamiento en cuatro partes: detección de silencio, detección de inhalación, detección de consonantes oclusivas sordas y toma de decisión. Cada parte está implementada por un componente, así, respectivamente, los componentes son: Componente de detección de silencio, Componente de detección de inhalación, Componente de detección de consonantes oclusivas sordas y Componente de toma de decisión (ver Figura 2.1). Cada uno se ejecuta en etapas (según lo dispuesto por el fundamento teórico), las cuales, a su vez, se dividen en fases.

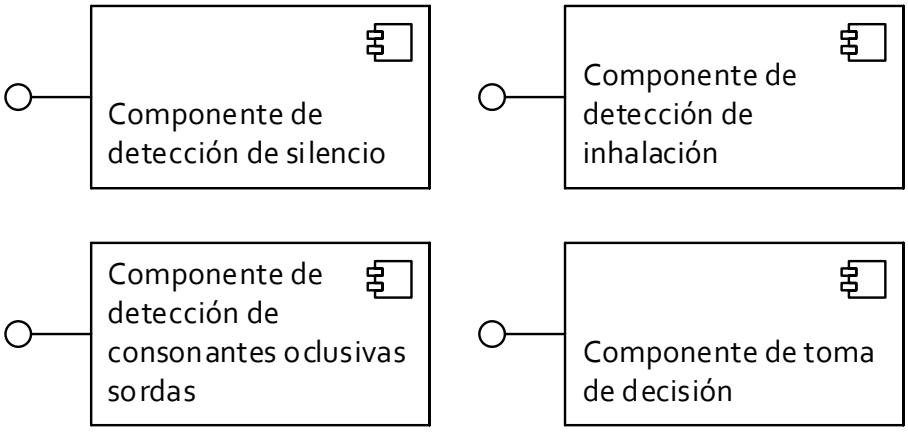


Figura 2.1. Componentes para detección de actividad de voz.

Cabe citar que la implementación práctica de los componentes hace uso de los siguientes paquetes de software:

- **openSMILE:** para extracción de características acústicas [23].
- **LibSVM:** para crear modelos de clasificación de tipo Support Vector Machine (SVM) [24].
- **RangeHandling:** para manejo de regiones y marcas de tiempo [25].
- **Praat:** para visualización de señales acústicas [26].

2.1 Componente de detección de silencio

Dada una muestra de señal acústica que contenga habla, inhalación y silencio (ver Figura 2.2); el objetivo es localizar regiones de silencio en dicha señal, lo cual se realiza en dos etapas: Etapa de identificación y Etapa de clasificación.

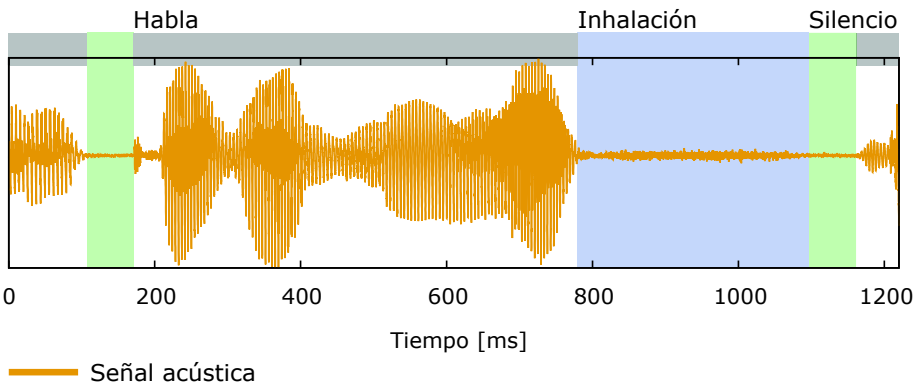


Figura 2.2. Muestra de señal acústica que contiene únicamente habla, inhalación y silencio.

2.1.1 Etapa de identificación

Esta etapa está compuesta por seis fases que, como se muestra en la Figura 2.3, incluyen: Fase de extracción y análisis de características acústicas, Fase de normalización, Fase de preparación, Fase de umbralización, Fase de limpieza y Fase de ajuste de fronteras.

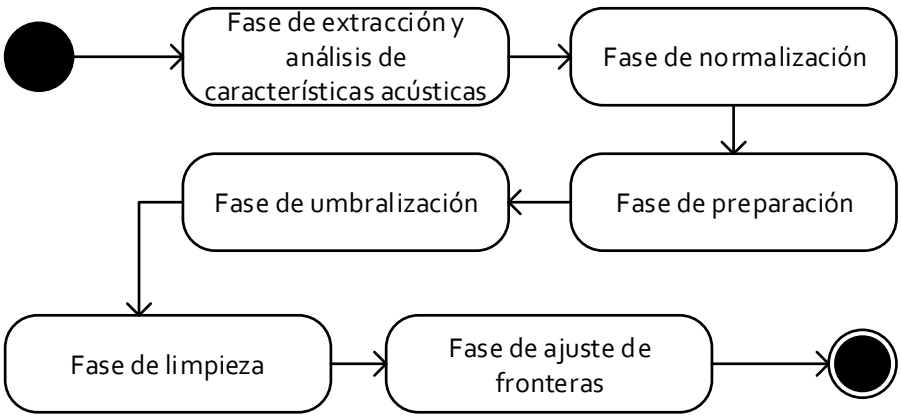


Figura 2.3. Fases de la Etapa de identificación.

Fase de extracción y análisis de características acústicas: De una muestra de señal acústica que contenga habla, inhalación y silencio se extrae características acústicas que sugieran la presencia de regiones de silencio en la muestra. Potenciales características que logran este objetivo son: Flujo espectral, Pendiente espectral y Logaritmo de energía. En particular, como se muestra en la Figura 2.4, las curvas de Flujo y Pendiente espectral muestran un comportamiento estable y constante a lo largo de regiones de silencio, mientras que la curva de Logaritmo de energía presenta un decrecimiento en estas mismas regiones.

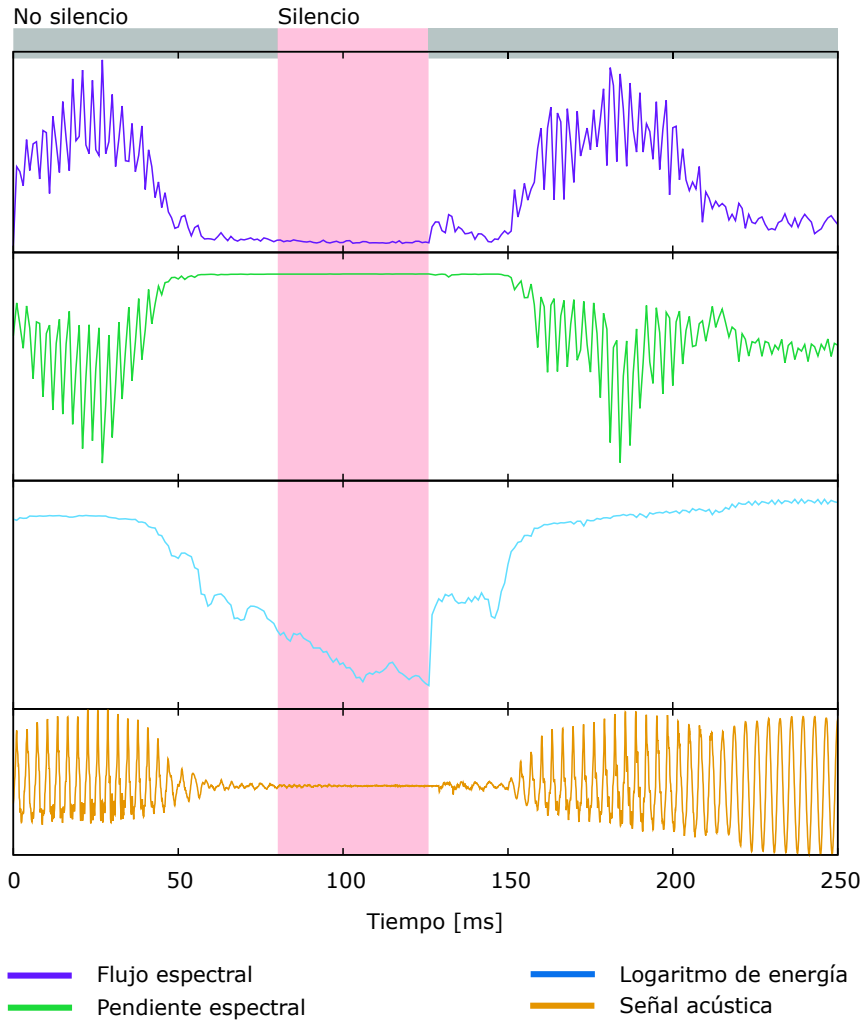
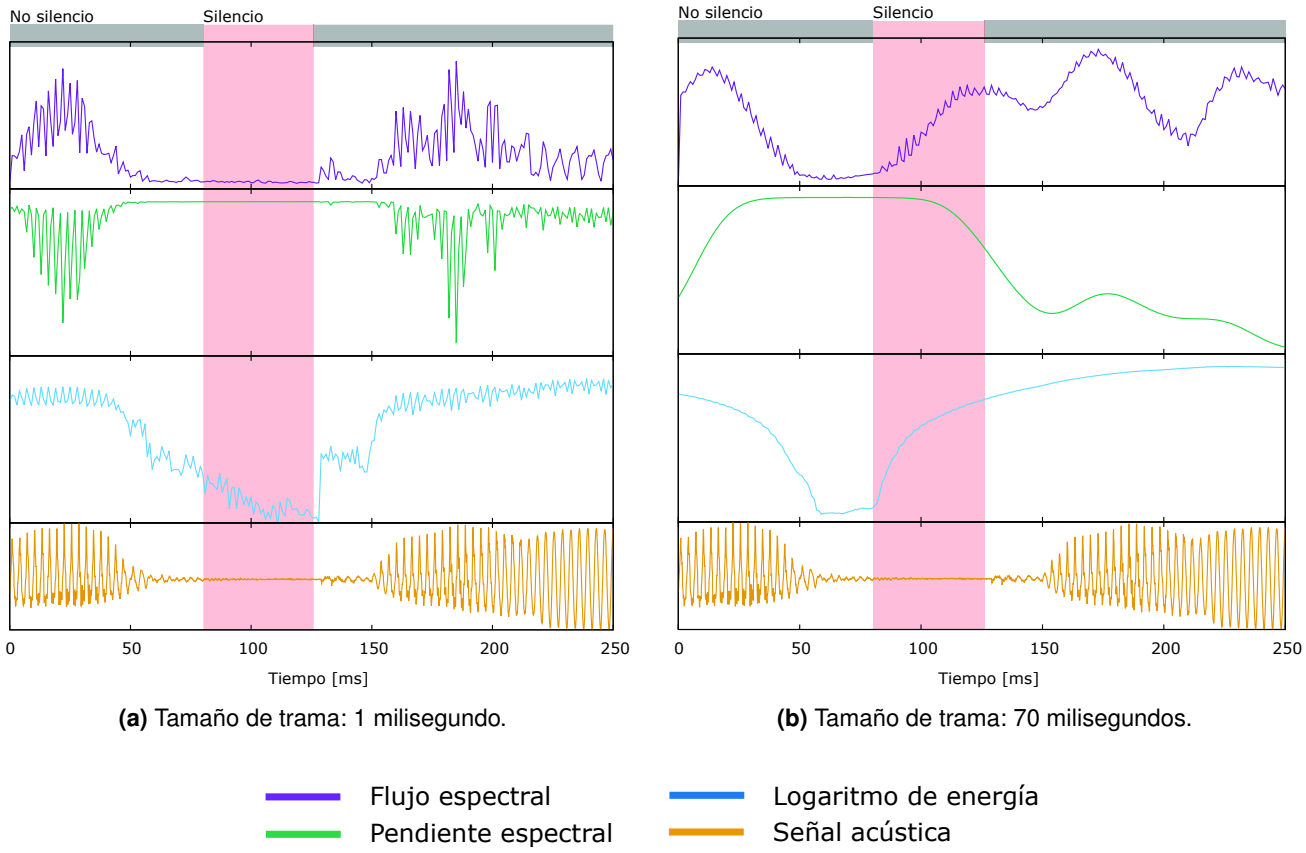


Figura 2.4. Comportamiento de características acústicas que sugieren la presencia de regiones de silencio.

La extracción de características se realiza por tramas de tamaño fijo [23], lo cual resulta en vectores que contienen los valores correspondientes a las tres características. El tamaño de trama establecido fue 10 milisegundos con salto de 1 milisegundo entre tramas adyacentes. Los experimentos realizados demostraron que, al disminuir el tamaño de trama, las curvas resultantes presentan fluctuaciones que dificultan la ejecución de las posteriores fases (ver Figura 2.5a). En contraste, al aumentarlo, el comportamiento de las características en regiones de no silencio podría ser abarcado (ver Figura 2.5b).



(a) Tamaño de trama: 1 milisegundo.

(b) Tamaño de trama: 70 milisegundos.

Figura 2.5. Comparación del comportamiento de las características a diferentes tamaños de trama.

Fase de normalización: Una vez obtenidos los vectores con las características acústicas, los valores en ellos son normalizados según un cálculo porcentual. Para ello, a partir de varias muestras de audio que contengan habla, inhalación y silencio se obtiene un valor aproximado de magnitud máxima para cada característica. Considerando este valor como el cien porciento, cualquier otro valor de magnitud se normaliza como el porcentaje correspondiente. La Tabla 2.1 muestra la magnitud máxima aproximada para cada característica a 10 milisegundos como tamaño de trama. Los vectores con las características normalizadas pasan a la Fase de preparación.

Característica acústica	**Magnitud máxima en tramas de 10[ms]**
Flujo espectral	$0{,}99$
Pendiente espectral	$0{,}54$
Logaritmo de energía	$25{,}53$

Tabla 2.1. Magnitud máxima aproximada de características acústicas en tramas de 10 milisegundos.

Fase de preparación: Seguidamente, se calcula el promedio simple de los valores en los vectores. El resultado de este promedio, comparado gráficamente con el Logaritmo de energía (normalizado como fue descrito en la

Fase de normalización), muestra un comportamiento que destaca la ubicación de regiones de no silencio y de silencio. Como se muestra en la Figura 2.6, las curvas del promedio de las características y el Logaritmo de energía normalizado describen el mismo patrón de fluctuación. Sin embargo, la curva de promedio tiende a separarse de la curva de Logaritmo de energía normalizado en regiones de no silencio. Por el contrario, las regiones donde estas curvas se mantienen cercanas corresponden principalmente a silencio.

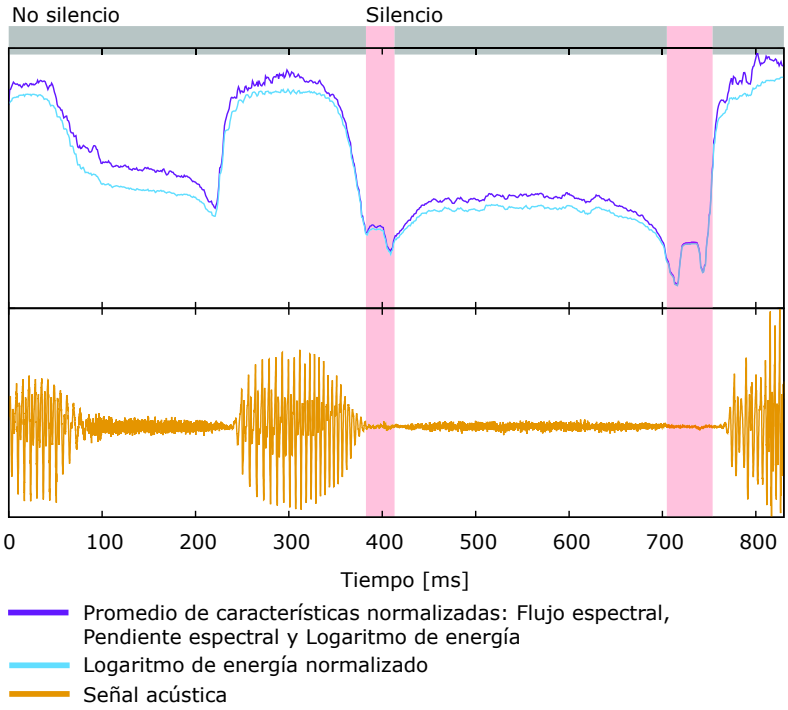


Figura 2.6. Comparación gráfica entre Logaritmo de energía normalizado y el promedio calculado de las características normalizadas.

Fase de umbralización: A continuación, sobre el promedio anteriormente calculado se aplica un umbral como se muestra en la Figura 2.7.

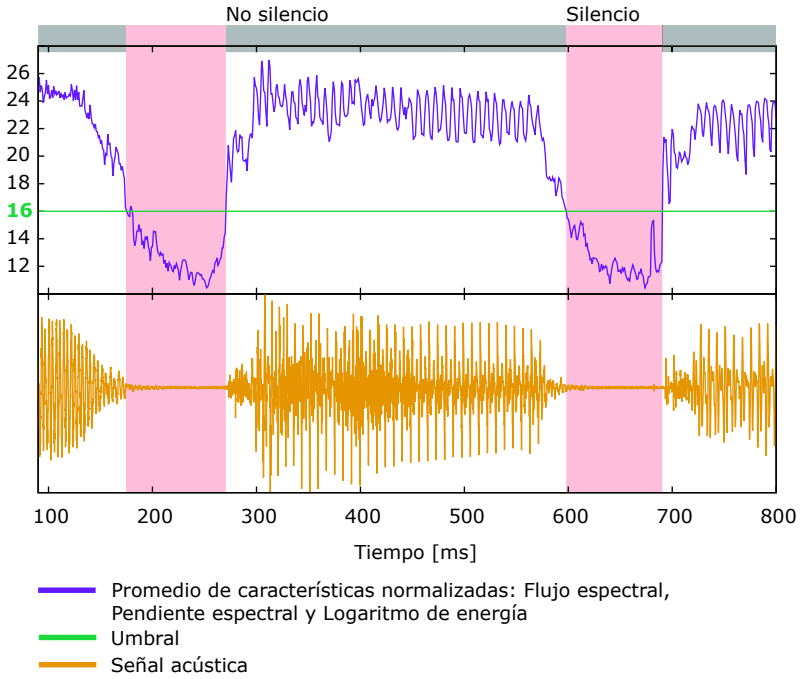


Figura 2.7. Aplicación de umbral sobre el promedio calculado de las características acústicas normalizadas.

Como resultado, valores debajo del umbral denotan la presencia de posibles regiones de silencio, y el resto, la presencia de regiones de no silencio. En este punto se obtiene una predicción inicial sobre las fronteras de dichas regiones, como se muestra en la Figura 2.8.

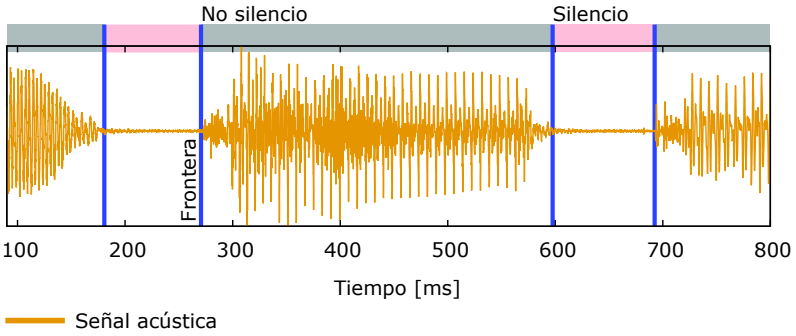
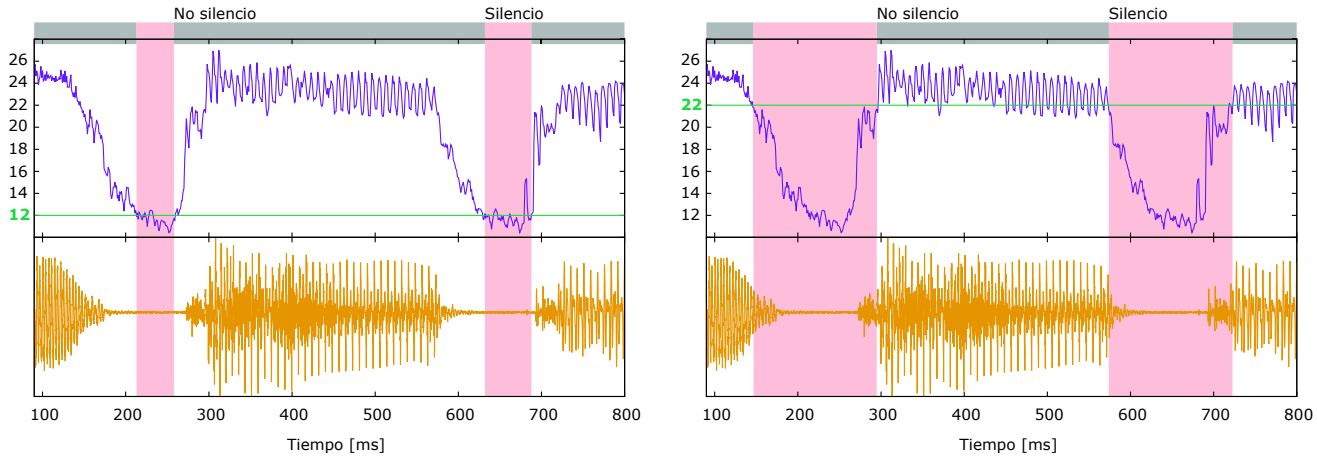


Figura 2.8. Predicción inicial sobre la ubicación de fronteras de regiones de silencio y de no silencio.

Por experimentación, el umbral establecido fue de 16 unidades. Los experimentos realizados demostraron que, al disminuir el umbral, regiones de silencio dejan de ser identificadas (ver Figura 2.9a). En contraste, al aumentarlo, regiones de no silencio son tomadas como de silencio (ver Figura 2.9b).



(a) Umbral: 12 unidades.

(b) Umbral: 22 unidades.

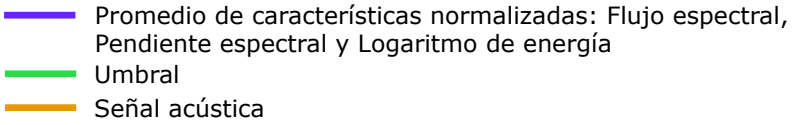


Figura 2.9. Comparación entre las regiones de silencio resultantes de variar el umbral.

Fase de limpieza: Sin embargo, la Fase de umbralización produce predicciones de fronteras que delimitan regiones cuya duración no proporciona información contundente acerca de la naturaleza de la región, sea ésta de silencio o de no silencio. A dichas regiones se las considera insignificantes. En este sentido, la fase actual intercambia entre silencio o no silencio, según corresponda, al valor de regiones insignificantes mediante un análisis de las regiones circundantes. Así, se tomará como regiones de no silencio a aquellas regiones insignificantes de silencio rodeadas por regiones de no silencio. La duración máxima definida para considerar que una región es insignificante fue 10 milisegundos.

Fase de ajuste de fronteras: No obstante, dado que el tamaño de trama utilizado para extraer características acústicas fue 10 milisegundos, las fronteras reales podrían encontrarse en un rango de $+10$ milisegundos de las fronteras predichas. Así, para obtener una predicción más exacta, las fases anteriormente descritas, desde extracción de características hasta limpieza, se ejecutan una vez más con diferentes parámetros. Para empezar, se hace necesario definir la región donde la frontera real podría estar presente. Dicha región inicia en cualquier frontera predicha y se extiende hasta 10 milisegundos a la derecha en la señal acústica, como se muestra en la Figura 2.10. Las fronteras reales no se ubican a la izquierda de las fronteras predichas debido a que el proceso de extracción de características analiza la señal acústica desde el lado izquierdo hacia el derecho.

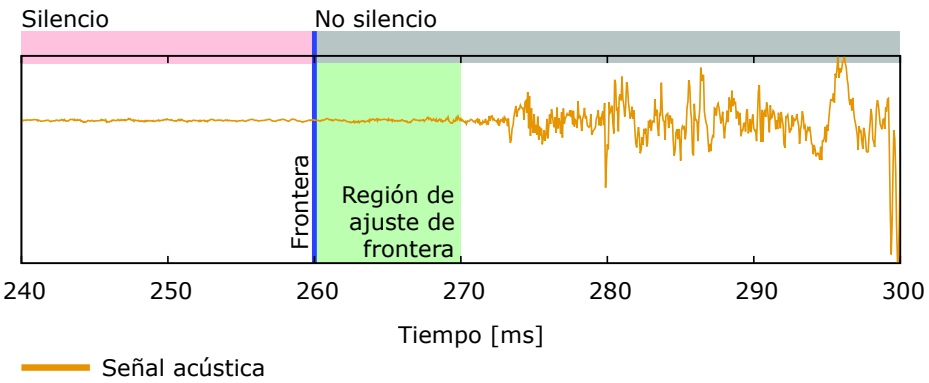


Figura 2.10. Definición de la región para ajuste de fronteras predichas.

Seguidamente, se establece un nuevo tamaño de trama para la extracción de características. Para el caso actual, 3 milisegundos fue el tamaño establecido. Finalmente, las fases mencionadas se vuelven a ejecutar utilizando el nuevo tamaño de trama y únicamente sobre la región recientemente definida. Las fronteras resultantes delimitan con mayor exactitud dónde inicia o termina una región de no silencio o de posible silencio. Cabe indicar que, para la re-ejecución de la Fase de normalización se utiliza los valores aproximados de magnitud máxima mostrados en la Tabla 2.2 para cada característica acústica.

Característica acústica	**Magnitud máxima en tramas de 3[ms]**
Flujo espectral	$0{,}52$
Pendiente espectral	$1{,}91$
Logaritmo de energía	$26{,}60$

Tabla 2.2. Magnitud máxima aproximada de características acústicas en tramas de 3 milisegundos.

2.1.2 Etapa de clasificación

Debido a la naturaleza de las características seleccionadas en la Etapa de identificación: Pendiente espectral, Flujo espectral y Logaritmo de energía; el proceso, como ha sido descrito hasta este punto, predice erróneamente como regiones de silencio a regiones débiles de no silencio. Una región débil de no silencio es aquella en la que existe habla, generalmente consonantes oclusivas [5], o inhalación y, sin embargo, el cálculo del promedio de sus características cae por debajo del umbral. Así, la etapa actual busca diferenciar, por medio de un clasificador de tipo SVM, entre regiones débiles de no silencio y regiones de silencio verdadero. Esta etapa está compuesta por cuatro fases que, como se muestra en la Figura 2.11, incluyen: Fase de preparación de datos, Fase de extracción de características acústicas, Fase de entrenamiento del clasificador y Fase de predicción.

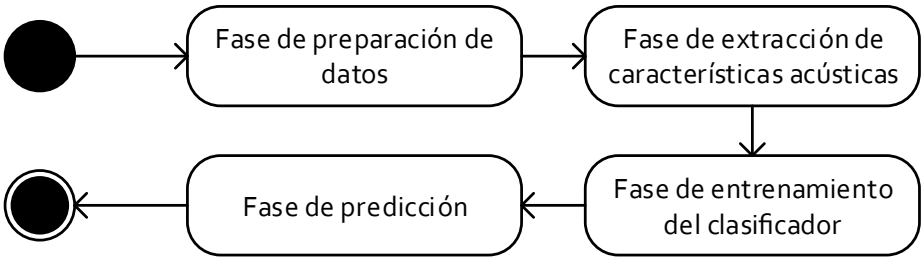


Figura 2.11. Fases de la Etapa de clasificación.

Fase de preparación de datos: Sobre un conjunto de señales acústicas que contengan habla, inhalación y silencio; la Etapa de identificación se ejecuta para obtener las regiones que conformarán el grupo de datos de entrenamiento para el clasificador. Las regiones de interés son únicamente aquellas predichas como silencio, ya que entre éstas existen las regiones erróneamente predichas (regiones débiles de no silencio). Así, cada región de interés es manualmente etiquetada como de silencio verdadero o como de no silencio, asimismo, las fronteras respectivas son manualmente ajustadas para abarcar la totalidad de dichas regiones. Consecuentemente, un clasificador de dos clases, silencio verdadero o no silencio, será posteriormente entrenado. Cabe indicar que el contexto, o alrededores, de las regiones de interés no es tomado en cuenta para conformar el conjunto de datos de entrenamiento, según lo descrito en la sección Resultados.

Fase de extracción de características acústicas: De las regiones de interés ya etiquetadas se extrae las características acústicas señaladas en la Lista 2.1 [2, 23]. El tamaño de trama para la extracción fue 10 milisegundos con salto de 1 milisegundo entre tramas adyacentes. Como resultado se obtiene un conjunto de vectores de características acústicas a cuyos valores se los normaliza en un rango de -1 a 1 según lo descrito por [24]. Los vectores normalizados, junto con la etiqueta correspondiente, conforman el conjunto de datos de entrenamiento para el clasificador.

- MFCC (12 elementos)
- Centroide espectral
- Entropía espectral
- Llanura espectral
- Flujo espectral
- Kurtosis espectral
- Nitidez espectral
- Oblicuidad espectral
- Pendiente espectral
- Varianza espectral
- Logaritmo de energía
- Tasa de cambios de signo
- Sonoridad

Lista 2.1. Características acústicas computadas para entrenamiento de clasificadores de tipo SVM y predicción.

Fase de entrenamiento del clasificador: Seguidamente, un clasificador de tipo SVM [24] es entrenado con los datos preparados. Como se dijo anteriormente, el clasificador predecirá una de dos clases: silencio verdadero o no silencio. Tanto el motivo de la elección de un clasificador de tipo SVM como los parámetros de entrenamiento son descritos en la sección Resultados.

Fase de predicción: Las anteriores fases, desde preparación de datos hasta entrenamiento del clasificador, se realizan una única vez. Una vez el clasificador haya sido entrenado, éste será utilizado en adelante para diferenciar entre regiones de silencio verdadero y regiones débiles de no silencio. Para ello, de cada región predicha como silencio, resultante de la Etapa de identificación, se extrae las características acústicas señaladas en la Lista 2.1 [2] a un tamaño de trama de 10 milisegundos y salto de 1 milisegundo, obteniendo así un conjunto de vectores de características acústicas. Luego, sobre estos vectores se aplica el mismo proceso de normalización aplicado al conjunto de vectores para entrenamiento del clasificador. Los vectores normalizados son luego entregados al clasificador, el cual asocia a ellos una etiqueta de silencio o de no silencio. Si el clasificador asocia la etiqueta de silencio a vectores continuos que representen una región cuya duración total sea mínimo 10 milisegundos, entonces dicha región se considera como de silencio, caso contrario, se considera de no silencio. En este punto se obtiene la predicción definitiva sobre regiones de silencio y de no silencio.

2.2 Componente de detección de inhalación

Dada una muestra de señal acústica que contenga habla, inhalación y silencio (ver Figura 2.2); el objetivo es localizar segmentos de inhalación en dicha señal, lo cual, de la misma manera que en el Componente de detección de silencio, se realiza en dos etapas: Etapa de identificación y Etapa de clasificación.

2.2.1 Etapa de identificación

Esta etapa está compuesta por seis fases que, como se muestra en la Figura 2.3, incluyen: Fase de extracción y análisis de características acústicas, Fase de normalización, Fase de preparación, Fase de umbralización, Fase de limpieza y Fase de ajuste de fronteras. La lógica de funcionamiento de cada fase es la misma que su homóloga en la Etapa de identificación del Componente de detección de silencio, sin embargo, se diferencian por sus parámetros de configuración.

Fase de extracción y análisis de características acústicas: De una muestra de señal acústica que contenga habla, inhalación y silencio se extrae características acústicas que sugieran la presencia de regiones de inhalación en la muestra. Potenciales características que logran este objetivo son: Logaritmo de energía, Sonoridad, Tasa de cambios de signo y Llanura espectral. En particular, como se muestra en la Figura 2.12, Logaritmo de energía y Sonoridad presentan un decrecimiento a lo largo de regiones de inhalación, mientras que Tasa de cambios de signo y Llanura espectral prestan un crecimiento a lo largo de estas mismas regiones. No obstante, este comportamiento de las características es también apreciable en otras regiones de la señal, mayormente, en regiones donde ocurren sonidos fricativos del habla [2].

[2] Un sonido fricativo sucede cuando el aire es expulsado del tracto vocal a través de un pequeño orificio formado por los labios, que puede o no estar acompañado por vibración de las cuerdas vocales. Algunos sonidos que corresponden a esta categoría son `/s/`, `/f/` o `/v/` [5].

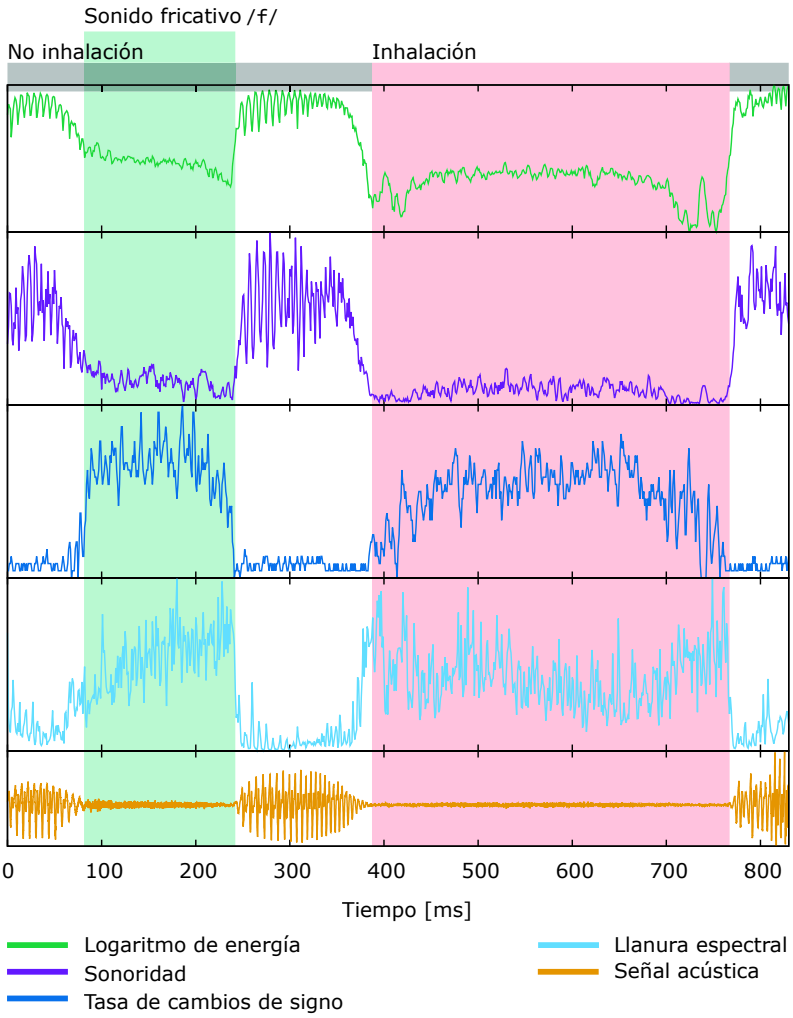


Figura 2.12. Comportamiento de características acústicas que sugieren la presencia tanto de regiones de inhalación como de sonidos fricativos del habla.

La extracción de características se realiza por tramas de tamaño fijo [23], lo cual resulta en vectores que contienen los valores correspondientes a las cuatro características. Tanto por experimentación como por las razones descritas en la Fase de extracción y análisis de características de la Etapa de identificación del Componente de detección de silencio, el tamaño de trama establecido fue 50 milisegundos con salto de 1 milisegundo entre tramas adyacentes.

Fase de normalización: Una vez obtenidos los vectores con las características acústicas, los valores en ellos son normalizados según el mismo proceso descrito en la Fase de normalización del Componente de detección de silencio. El cálculo porcentual se realiza a partir de los valores máximos de magnitud de cada característica. Dichos valores se muestran en la Tabla 2.3 para tramas de 50 milisegundos. Los vectores con las características normalizadas pasan a la Fase de preparación.

Característica acústica	Magnitud máxima en tramas de 50[ms]
Logaritmo de energía	$26{,}20$
Sonoridad	$6{,}31$
Tasa de cambios de signo	$0{,}92$
Llanura espectral	$0{,}66$

Tabla 2.3. Magnitud máxima de características acústicas en tramas de 50 milisegundos.

Fase de preparación: A continuación, sobre las características acústicas en los vectores se realiza la operación matemática descrita por la Ecuación 2.1. En dicha ecuación, n corresponde al número total de características, cuatro para el caso, y f_i corresponde a la i-ésima característica en el vector. La naturaleza de la ecuación hace relevante al orden de las características en los vectores, por lo tanto, la permutación que mejor logra destacar la presencia de posibles inhalaciones en la señal acústica es: Sonoridad, Logaritmo de energía, Llanura espectral y Tasa de cambios de signo. En este sentido, el sumatorio en la ecuación inicia con el valor de Sonoridad elevado a la mínima potencia, que es 1, y termina con el de Tasa de cambios de signo elevado a la máxima potencia, que es 4.

$$x = \sqrt[n]{\sum_{i=1}^{n} (f_i)^i}$$

Ecuación 2.1. Ecuación para preparación de características acústicas para detección de posibles regiones de inhalación.

Como se muestra en la Figura 2.13, la curva de la operación realizada y la de Tasa de cambios de signo normalizada describen el mismo patrón de fluctuación, sin embargo, la operación realizada provoca un incremento de los valores en los extremos de regiones de inhalación.

Fase de umbralización: A continuación, sobre el resultado del cálculo anteriormente realizado se aplica un umbral como se muestra en la Figura 2.14. Como resultado, valores arriba del umbral denotan la presencia de regiones de posible inhalación, y el resto, la presencia de regiones de no inhalación. Tanto por experimentación como por las razones descritas en la Fase de umbralización de la Etapa de identificación del Componente de detección de silencio, el umbral establecido fue de 19 unidades.
En este punto se obtiene una predicción inicial de las fronteras de regiones de posible inhalación y de no inhalación.

Fase de limpieza: De la misma manera que en su homóloga en la Etapa de identificación del Componente de detección de silencio, la Fase de umbralización produce predicciones de fronteras que delimitan regiones

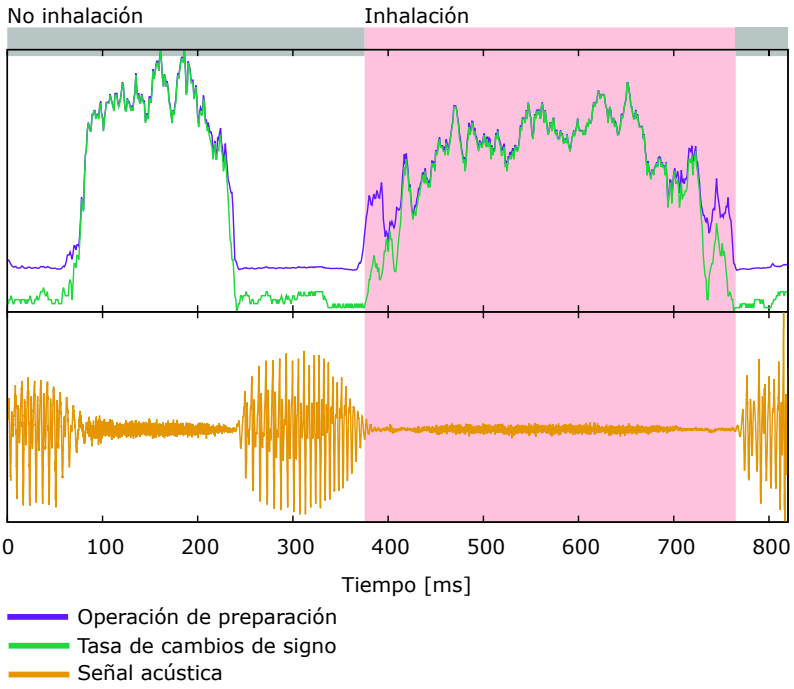


Figura 2.13. Comparación gráfica entre Tasa de cambios de signo normalizada y el resultado de la operación de preparación.

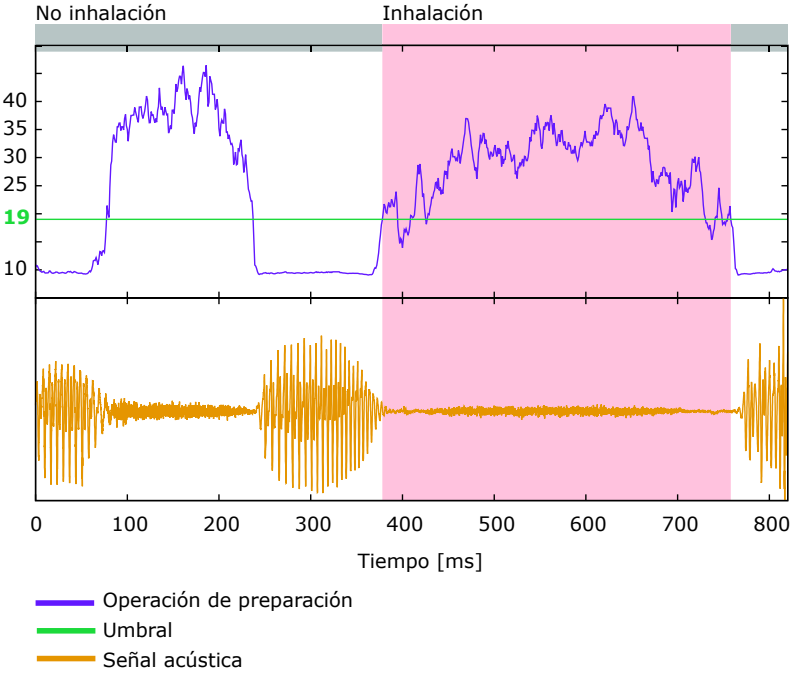


Figura 2.14. Aplicación de umbral sobre la operación de preparación de características acústicas normalizadas.

insignificantes. Así, es necesario realizar una limpieza de dichas regiones. Tanto el método de limpieza como la duración máxima definida para considerar que una región es insignificante son los mismos entre fases de limpieza homólogas.

Fase de ajuste de fronteras: Sin embargo, dado que el tamaño de trama utilizado para extraer características acústicas fue 50 milisegundos, las fronteras reales podrían encontrarse en un rango de $+50$ milisegundos de las fronteras predichas. Así, para obtener una predicción más exacta, el mismo proceso descrito en la Fase de ajuste de fronteras del Componente de detección de silencio se aplica sobre una región en la señal acústica que inicia en cualquier frontera predicha y se extiende 50 milisegundos a la derecha, con un tamaño de trama de 3 milisegundos. Durante la re-ejecución de la Fase de normalización se utiliza los valores aproximados de magnitud máxima mostrados en la Tabla 2.4 para cada característica acústica.

Característica acústica	**Magnitud máxima en tramas de 3[ms]**
Logaritmo de energía	$26{,}60$
Sonoridad	$5{,}60$
Tasa de cambios de signo	$0{,}88$
Llanura espectral	$0{,}66$

Tabla 2.4. Magnitud máxima de características acústicas en tramas de 3 milisegundos.

Las fronteras resultantes delimitan con mayor exactitud dónde inicia o termina una región de no inhalación o posible inhalación.

2.2.2 Etapa de clasificación

Como fue mencionado en la Fase de extracción y análisis de características acústicas, las características empleadas presentan el mismo comportamiento tanto en regiones de inhalación como en ciertas regiones de no inhalación, a saber, en aquellas donde ocurren sonidos fricativos de la voz. Así, en este punto, entre las regiones predichas como de inhalación por la Etapa de identificación existen tanto regiones de inhalación verdadera como de no inhalación. Por ello, la etapa actual busca diferenciar, por medio de un clasificador de tipo SVM, entre estas dos clases de regiones. Esta etapa está compuesta por cuatro fases que, como muestra la Figura 2.11, incluyen: Fase de preparación de datos, Fase de extracción de características acústicas, Fase de entrenamiento del clasificador y Fase de predicción.

Fase de preparación de datos: Sobre un conjunto de señales acústicas que contengan habla, inhalación y silencio; la Etapa de identificación se ejecuta para obtener las regiones que conformarán el grupo de datos de entrenamiento para el clasificador. Las regiones de interés son únicamente aquellas predichas como inhalación,

ya que entre éstas existen las regiones erróneamente predichas. Así, cada región de interés es manualmente etiquetada, según el caso, como de inhalación verdadera o de no inhalación, asimismo, las fronteras respectivas son manualmente ajustadas para abarcar la totalidad de dichas regiones. Consecuentemente, un clasificador de dos clases, inhalación verdadera o no inhalación, será posteriormente entrenado. Cabe indicar que el contexto, o alrededores, de las regiones de interés no es tomado en cuenta para conformar el conjunto de datos de entrenamiento, según lo descrito en la sección Resultados.

Fase de extracción de características acústicas: De las regiones de interés ya etiquetadas se extrae las características acústicas señaladas en la Lista 2.1 [2, 23]. El tamaño de trama para la extracción fue 50 milisegundos con salto de 1 milisegundo entre tramas adyacentes. Como resultado se obtiene un conjunto de vectores de características acústicas a cuyos valores se los normaliza en un rango de -1 a 1 según lo descrito por [24]. Los vectores normalizados, junto con la etiqueta correspondiente, conforman el conjunto de datos de entrenamiento para el clasificador.

Fase de entrenamiento del clasificador: Seguidamente, un clasificador de tipo SVM [24] es entrenado con los datos preparados. Dicho clasificador predecirá una de dos clases: inhalación verdadera o no inhalación.

Fase de predicción: Las anteriores fases, desde preparación de datos hasta entrenamiento del clasificador, se realizan una única vez. Una vez el clasificador haya sido entrenado, éste será utilizado en adelante para diferenciar entre regiones de inhalación verdadera y regiones de no inhalación. Esta fase se realiza de la misma manera que en su homóloga en la Etapa de clasificación del Componente de detección de silencio. No obstante, difiere en que el tamaño de trama para extraer característica acústicas se estableció en 50 milisegundos, y en que, para considerar a una región como de inhalación verdadera, la predicción devuelta por el clasificador debe sumar mínimo 100 milisegundos de tramas continuas de inhalación. En este punto se obtiene la predicción definitiva sobre regiones de inhalación y de no inhalación.

2.3 Componente de detección de consonantes oclusivas sordas

Dada una muestra de señal acústica que contenga únicamente habla, el objetivo es determinar la presencia de consonantes oclusivas sordas al inicio de dicha señal (ver Figura 2.15), lo cual se realiza en una única etapa: Etapa de clasificación.

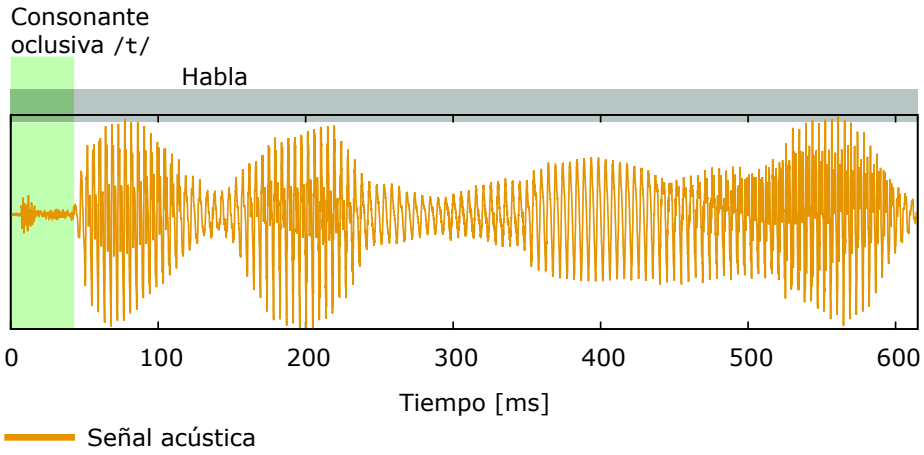


Figura 2.15. Muestra de señal acústica que contiene únicamente habla.

2.3.1 Etapa de clasificación

Esta etapa busca diferenciar, por medio de tres clasificadores de tipo SVM, entre regiones de consonante oclusiva sorda y de no consonante oclusiva sorda ubicadas al inicio de la señal acústica. Está compuesta por cuatro fases que, como muestra la Figura 2.11 incluyen: Fase de preparación de datos, Fase de extracción de características acústicas, Fase de entrenamiento de clasificadores y Fase de predicción.

Fase de preparación de datos: En un conjunto de señales acústicas que contengan únicamente habla, se etiqueta, manualmente y apoyado tanto en el método de conversión de grafema a fonema propuesto por [27] como en el método de segmentación fonética propuesto por [28], regiones de consonante oclusiva sorda (regiones de interés) y de no consonante oclusiva sorda para conformar el conjunto de datos de entrenamiento para los clasificadores. Consecuentemente, tres clasificadores de dos clases cada uno serán posteriormente entrenados. Cabe indicar que el contexto, o alrededores, de las regiones de interés no es tomado en cuenta para conformar el conjunto de datos de entrenamiento, según lo descrito en la sección Resultados.
Debido a la variedad en la duración de pronunciación de las consonantes oclusivas sordas, cada región etiquetada como tal es agrupada según su duración en uno de tres grupos. Como muestra la Tabla 2.5, el Grupo 1 abarca regiones de duración en un rango de 10 a 35 milisegundos, el 2, regiones de duración entre 35 y 50 milisegundos, y el 3, regiones de duración entre 50 y 150 milisegundos.

Grupo	Rango [ms]	
	Mínimo	Máximo
1	10	35
2	35	50
3	50	150

Tabla 2.5. Rangos de duración de regiones de consonantes oclusivas sordas distribuidos en tres grupos.

Fase de extracción de características acústicas: A cada grupo de regiones de interés le corresponderá un clasificador, por ello, la configuración para extracción de características es específica para cada grupo. Así, tanto de las regiones en un grupo como de las regiones etiquetadas como de no consonante oclusiva sorda se extrae las características acústicas señaladas en la Lista 2.1 [2, 23] a un tamaño de trama igual al valor mínimo del rango en el grupo y con salto de 1 milisegundo entre tramas adyacentes. Por lo tanto, para las regiones en el Grupo 1, el tamaño de trama será 10 milisegundos, para aquellas en el Grupo 2, 35 milisegundos, y para aquellas en el Grupo 3, 50 milisegundos. Como resultado se obtiene, por grupo, un conjunto de vectores de características acústicas a cuyos valores se los normaliza en un rango de -1 a 1 según lo descrito por [24]. Los vectores normalizados de un grupo, junto con la etiqueta correspondiente, conforman el conjunto de datos de entrenamiento para el clasificador respectivo.

Fase de entrenamiento de clasificadores: Seguidamente, tres clasificadores de tipo SVM [24] son entrenados con los datos del grupo respectivo. Cada clasificador predecirá una de dos clases: consonante oclusiva sorda y no consonante oclusiva sorda.

Fase de predicción: Las anteriores fases, desde preparación de datos hasta entrenamiento de clasificadores, se realizan una única vez. Una vez los clasificadores hayan sido entrenados, éstos serán utilizados para diferenciar entre regiones de consonante oclusiva sorda y regiones de no consonante oclusiva sorda. Para ello, inicialmente, se concatena las predicciones de regiones de silencio e inhalación resultantes de los componentes anteriores. Las regiones concatenadas son consideradas, desde este punto en adelante, como de pausa, mientras que el complemento, como de habla. Luego, una subregión de tamaño específico al inicio de cada región de habla es pasada por los tres clasificadores. Dicha subregión se determina por medio de las Ecuaciones 2.2 y 2.3. En ellas, $subregion_0$ y $subregion_f$ representan respectivamente el inicio y el final de la subregión en milisegundos. $region_0$ representa la frontera de inicio en milisegundos de una región de habla, mientras que $rango_{max}$ representa el valor máximo del rango de regiones del clasificador correspondiente conforme a la Tabla 2.5.

$$subregion_0 = region_0 - 20$$

Ecuación 2.2. Inicio de subregión para búsqueda de consonantes oclusivas sordas.

$$subregion_f = region_0 + rango_{max} + 10$$

Ecuación 2.3. Fin de subregión para búsqueda de consonantes oclusivas sordas.

Una vez determinadas las subregiones para búsqueda de consonantes oclusivas sordas, los clasificadores se ejecutan serialmente iniciando con el clasificador del menor rango. Cuando un clasificador devuelva una predicción afirmativa sobre la existencia de una consonante oclusiva sorda en la subregión, el proceso se detiene y los

siguientes clasificadores no analizan las subregiones respectivas. Una predicción de consonante oclusiva sorda se considera afirmativa cuando el clasificador prediga tramas continuas de consonante oclusiva sorda cuya duración total sea mayor o igual a la mitad del valor mínimo del rango de regiones del clasificador según lo descrito en la Tabla 2.5. Finalmente, la predicción de las fronteras de regiones de consonante oclusiva sorda al inicio de una región de habla se obtiene.

2.4 Componente de toma de decisión

Dada una secuencia de predicciones de regiones de habla, pausa y consonantes oclusivas sordas producidas por los tres componentes anteriores, el objetivo es determinar si las regiones de pausa corresponden al bloqueo del flujo del aire propio de la pronunciación de las consonantes oclusivas sordas [5]. El Componente de toma de decisión se encarga de este objetivo, para lo cual hace uso de una Máquina de Estados Finitos (MEF) cuyo diagrama se ilustra en la Figura 2.16. Inicialmente, la secuencia es invertida y pasada como entrada a la MEF, la cual busca subsecuencias de regiones de consonante oclusiva sorda seguidas de regiones de pausa. Cuando una subsecuencia de este tipo es encontrada, se analiza la duración de la región de pausa y, si resulta menor a cierto valor, entonces se la toma como parte del bloqueo del flujo del aire anteriormente mencionado y se la marca como región de habla. La duración máxima establecida para que una región de pausa sea tomada como de habla fue de 40 milisegundos.

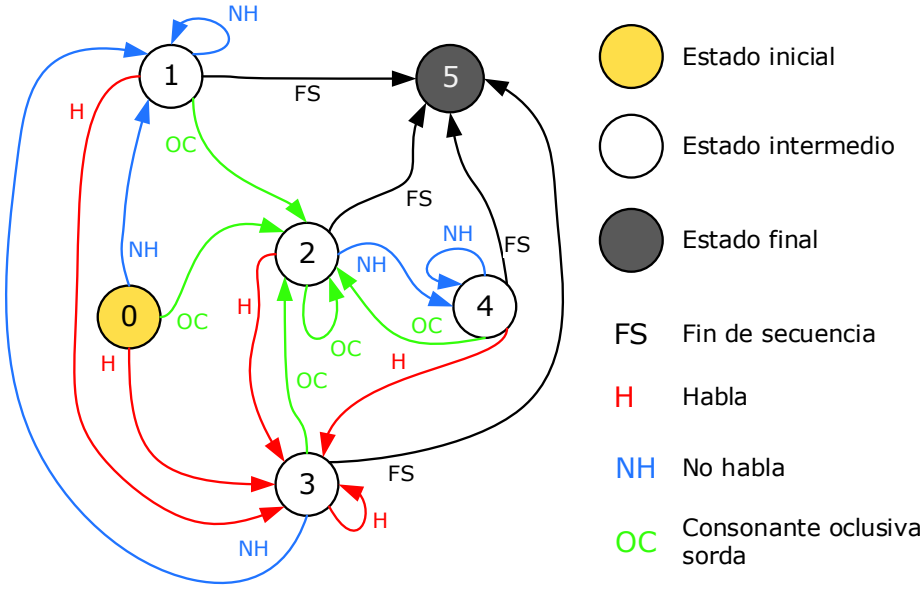


Figura 2.16. Máquina de Estados Finitos para análisis de secuencias de predicciones.

Luego, la secuencia de predicciones es invertida una vez más para, finalmente, representar las regiones de habla y pausa. Consecuentemente, se habrá obtenido los eventos de habla y pausa presentes en la señal acústica.

3 RESULTADOS Y DISCUSIÓN

En esta sección se describe la naturaleza del conjunto de datos empleado para la experimentación con el método VAD propuesto. Se presenta tanto los resultados obtenidos y su interpretación, así como una comparación de ellos con los devueltos por el método VAD propuesto por el proyecto WebRTC [7].

El conjunto de datos con el que se trabajó se compuso de doce noticieros en idioma francés [29] con las respectivas muestras de señales acústicas en formato `.WAV` cuyas duraciones totalizan 125 minutos. La Tabla 3.1 describe dichas señales acústicas. Cada señal estuvo compuesta, en promedio, por 123 ± 19 regiones de inhalación, 178 ± 24 regiones de habla, 855 ± 104 regiones de consonante oclusiva sorda y 47 ± 15 regiones de silencio producidas conscientemente por los locutores; las regiones de silencio debidas a la pronunciación de consonantes oclusivas sordas no fueron tomadas en consideración para el conteo mostrado.

Propiedades de las señales acústicas	
Canales	1
Frecuencia de muestreo	16 KHz
Precisión	16 bits
Velocidad de bits	256 kbps
Duración	10 ± 2 minutos
Libres de ruido y sonidos ajenos a la voz humana	Sí
Regiones presentes por señal acústica	
Regiones de silencio	47 ± 15
Regiones de Inhalación	123 ± 19
Regiones de habla	178 ± 24
Regiones de consonante oclusiva sorda	855 ± 104

Tabla 3.1. Descripción de las señales acústicas utilizadas para entrenamiento de clasificadores y pruebas.

El conjunto de datos descrito fue segmentado en dos partes, seis señales fueron utilizadas para entrenamiento de clasificadores y seis para pruebas. Además, para cada señal acústica, existió una referencia semi-manualmente etiquetada [27, 28] acerca de las regiones de silencio, inhalación y consonante oclusiva sorda.

3.1 Resultados

Los siguientes párrafos exponen los resultados del entrenamiento de los clasificadores empleados en los componentes expuestos en la sección Metodología, así como también los resultados de la operación de dichos componentes.

3.1.1 Resultado del entrenamiento de clasificadores

Tanto el Componente de detección de silencio, el Componente de detección de inhalación y el Componente de detección de consonantes oclusivas sordas emplean clasificadores SVM para su funcionamiento. Los datos para entrenar dichos clasificadores fueron obtenidos, según lo expuesto en la sección Metodología, a partir de las seis señales acústicas destinadas para entrenamiento. Es importante recordar que, durante la preparación de los mencionados datos, el contexto, o alrededores, de las regiones de interés respectivas no fue tomado en cuenta. Ello debido a que el conjunto de señales acústicas con el que se dispuso, según lo descrito anteriormente, no proveía la totalidad de las variantes de regiones que podrían rodear a las regiones de interés. Por esta razón, se optó por tomar a las regiones aisladas de su entorno de ocurrencia y representar a sus vectores de características como puntos independientes en un espacio geométrico cuyo número de dimensiones fue igual al número de características en el vector, 24 según la Lista 2.1. Así entonces, se adoptó un abordaje de clasificación basado en SVM debido tanto a su popularidad al tratarse de clasificación en espacios geométricos multidimensionales, a su simplicidad de uso y a que, además, ha sido empleado en otros trabajos de investigación relacionados con el presente [16, 17]. El entrenamiento de estos clasificadores se basó en la técnica de validación cruzada con cinco iteraciones. Para todos los clasificadores, las *funciones kernel* probadas fueron la función polinomial y la función de base radial (o RBF por las siglas en inglés para Radial Basis Function), siendo RBF la que devolvió una tasa de clasificación acertada mayor al $90\,\%$. Los parámetros de entrenamiento más óptimos fueron determinados mediante una búsqueda de cuadrícula. A continuación se presenta los resultados obtenidos.

Clasificador del Componente de detección de silencio: El clasificador de este componente fue entrenado con 800 vectores de silencio, 1266 vectores de no silencio y los siguientes parámetros:

- **Tipo de clasificador SVM:** nu-SVC
- **Tipo de kernel:** Función de base radial
- **Parámetro gamma:** $2{,}85$
- **Parámetro nu:** $0{,}2$

La tasa de vectores correctamente clasificados según validación cruzada de cinco iteraciones fue: $95{,}78\,\%$. La Figura 3.1 muestra la reducción a dos dimensiones, utilizando el método de análisis de componentes principales (o PCA por las siglas en inglés para Principal Component Analysis), del espacio de clasificación.

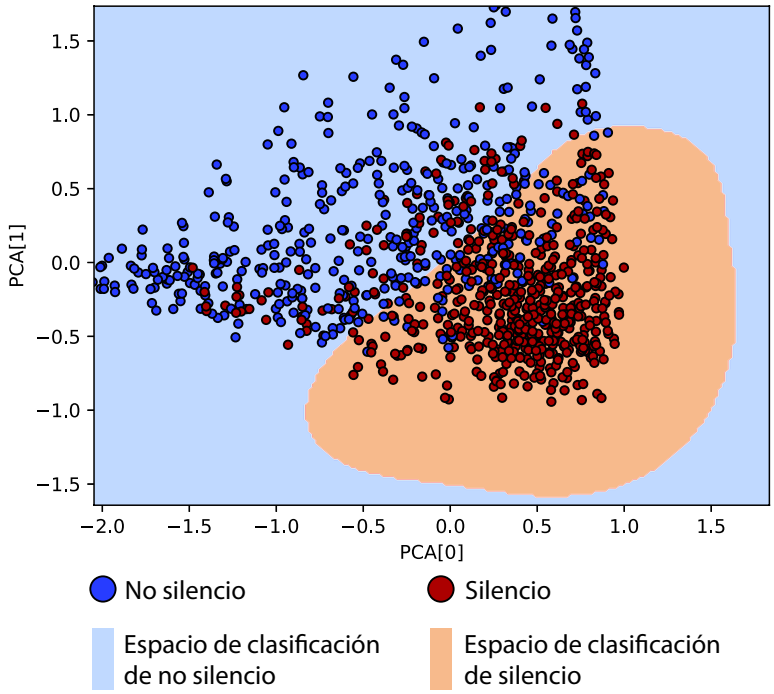


Figura 3.1. Reducción a dos dimensiones del espacio de clasificación para la detección de silencio.

Clasificador del Componente de detección de inhalación: El clasificador de este componente fue entrenado con 2267 vectores de inhalación, 2053 vectores de no inhalación y los siguientes parámetros:

- **Tipo de clasificador SVM:** nu-SVC
- **Tipo de kernel:** Función de base radial
- **Parámetro gamma:** $1{,}0$
- **Parámetro nu:** $0{,}06$

La tasa de vectores correctamente clasificados según validación cruzada de cinco iteraciones fue: $98{,}97\,\%$. La Figura 3.2 muestra la reducción a dos dimensiones del espacio de clasificación.

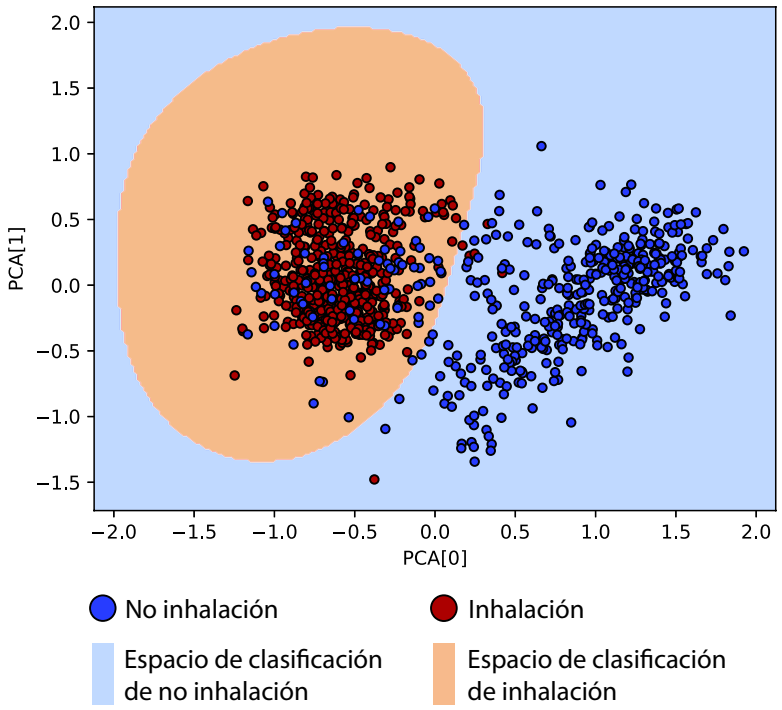


Figura 3.2. Reducción a dos dimensiones del espacio de clasificación para la detección de inhalación.

Clasificadores del Componente de detección de consonantes oclusivas sordas: Los tres clasificadores de este componente fueron entrenados con 920 vectores de no consonante oclusiva sorda y los siguientes parámetros:

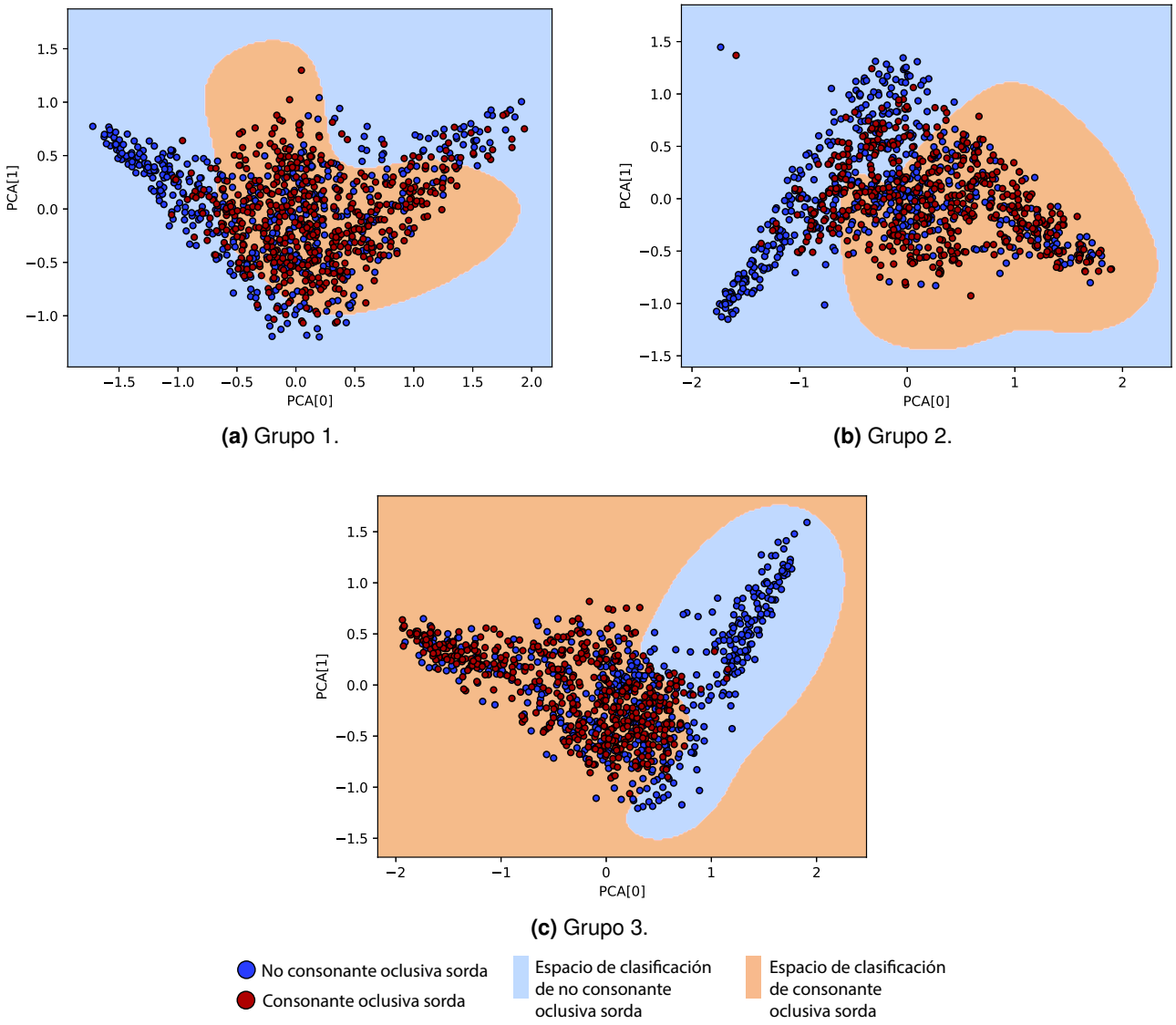
- **Tipo de clasificador SVM:** nu-SVC
- **Tipo de kernel:** Función de base radial
- **Parámetro gamma:** $3{,}7$
- **Parámetro nu:** $0{,}001$

No obstante, conforme a la Tabla 2.5, el clasificador del Grupo 1 fue entrenado con 537 vectores de consonante oclusiva sorda, el clasificador del Grupo 2, con 587 vectores, y el clasificador del Grupo 3, con 589 vectores. La Tabla 3.2 muestra la tasa de vectores correctamente clasificados según validación cruzada de cinco iteraciones para dichos clasificadores.

Clasificador	Tasa de vectores correctamente clasificados
Clasificador del Grupo 1	$98{,}20\,\%$
Clasificador del Grupo 2	$98{,}50\,\%$
Clasificador del Grupo 3	$98{,}46\,\%$

Tabla 3.2. Tasa de vectores correctamente clasificados según validación cruzada.

La Figura 3.3 muestra la reducción a dos dimensiones del espacio de clasificación para el grupo correspondiente.



(a) Grupo 1.

(b) Grupo 2.

(c) Grupo 3.

Figura 3.3. Reducción a dos dimensiones del espacio de clasificación para la detección de consonantes oclusivas sordas según los grupos de regiones.

3.1.2 Resultados de la operación de los componentes

Las seis señales acústicas destinadas para pruebas se utilizaron para evaluar la operación de los componentes según lo expuesto en la sección Metodología, obteniendo así, para cada una de ellas, la predicción de eventos de habla y de pausa. Apoyándose en las referencias semi-manualmente etiquetadas respectivas a cada señal acústica, los siguientes párrafos exponen los resultados obtenidos. Se incluye los resultados de la operación aislada del Componente de detección de inhalación. No se incluye resultados de la operación aislada del Componente de detección de silencio debido a que, en las mencionadas referencias, únicamente las regiones de silencio producidas conscientemente por los locutores se encontraron etiquetadas; no así con las regiones de silencio debidas

a la pronunciación de consonantes oclusivas sordas. Tampoco se incluye resultados de la operación aislada del Componente de detección de consonantes oclusivas sordas debido a que dicho componente se enfoca en determinar la presencia de consonantes oclusivas sordas al inicio de una región de habla, y no en determinar la totalidad de dichas consonantes en la señal acústica. Del mismo modo, no se incluye resultados de la operación aislada del Componente de toma de decisión ya que éste, para su funcionamiento, emplea reglas explícitamente enunciadas y no procesos de clasificación basados en modelos. Además, este componente depende fuertemente de las predicciones devueltas por los demás componentes.
Posteriormente, se expondrá los resultados de la operación serial conjunta de los componentes en dos escenarios:

- **Escenario 1:** Se excluye el uso del Componente de detección de consonantes oclusivas sordas.
- **Escenario 2:** Se incluye el uso del Componente de detección de consonantes oclusivas sordas.

Independientemente del escenario, los componentes operaron en el siguiente orden:

1. Componente de detección de silencio
2. Componente de detección de inhalación
3. Componente de detección de consonantes oclusivas sordas
4. Componente de toma de decisión

Las siguientes consideraciones fueron tomadas en cuenta para calcular los resultados:

1. Un Verdadero Positivo (VP), predicción acertada, corresponde a toda aquella región manualmente etiquetada como de pausa (sea silencio o inhalación) que fue correctamente clasificada como evento de pausa.

2. Un Verdadero Negativo (VN), predicción acertada, corresponde a toda aquella región manualmente etiquetada como de habla que fue correctamente clasificada como evento de habla.

3. Un Falso Positivo (FP), predicción errónea, corresponde a toda aquella región manualmente etiquetada como de habla que fue incorrectamente clasificada como evento de pausa.

4. Un Falso Negativo (FN), predicción errónea, corresponde a toda aquella región manualmente etiquetada como de pausa (sea silencio o inhalación) que fue incorrectamente clasificada como evento de habla.

Resultados de la operación aislada del Componente de detección de inhalación: La Tabla 3.3 muestra los resultados de la operación del Componente de detección de inhalación sobre las seis señales acústicas de prueba.

Señal acústica	VP	VN	FP	FN	Precisión	Exhaustividad	Medida-F
Señal 1	82	123	46	7	64,06 %	92,14 %	75,58 %
Señal 2	139	236	100	13	58,16 %	91,45 %	71,01 %
Señal 3	131	225	98	6	57,21 %	95,62 %	71,59 %
Señal 4	158	257	108	19	59,40 %	89,27 %	71,33 %
Señal 5	135	247	117	7	53,57 %	95,07 %	68,53 %
Señal 6	124	276	151	9	45,09 %	93,23 %	60,78 %

Tabla 3.3. Resultados del Componente de detección de inhalación sobre las señales acústicas de prueba.

En base a dichos resultados, el intervalo de confianza estadístico, calculado con la Distribución *t* de Student a $0{,}05$ de error estadístico, con respecto a la Medida-F es $70 \pm 5\,\%$. Ello muestra que, para cualquier señal acústica nueva que presente las propiedades descritas en la Tabla 3.1, el Componente de detección de inhalación presentará un desempeño de $70 \pm 5\,\%$ con respecto a la Medida-F.

Resultados de la operación conjunta de los componentes en el escenario 1: La Tabla 3.4 muestra los resultados de la operación conjunta de los componentes, excluido el Componente de detección de consonantes oclusivas sordas, sobre las seis señales acústicas de prueba.

Señal acústica	VP	VN	FP	FN	Precisión	Exhaustividad	Medida-F
Señal 1	118	584	482	20	19,67 %	85,51 %	31,98 %
Señal 2	209	955	784	50	21,05 %	80,70 %	33,39 %
Señal 3	226	1026	843	53	21,14 %	81,01 %	33,53 %
Señal 4	226	1160	957	29	19,10 %	88,63 %	31,43 %
Señal 5	196	1197	1013	16	16,21 %	92,45 %	27,57 %
Señal 6	188	1330	1153	19	14,02 %	90,82 %	24,29 %

Tabla 3.4. Resultados obtenidos de la operación conjunta de los componentes en el escenario 1.

En base a estos resultados, el intervalo de confianza estadístico, calculado con la Distribución *t* de Student a $0{,}05$ de error estadístico, con respecto a la Medida-F es $30 \pm 4\,\%$. Ello muestra que, para cualquier señal acústica nueva que presente las propiedades descritas en la Tabla 3.1, al no hacer uso del Componente de detección de consonantes oclusivas sordas, el método VAD propuesto presentará un desempeño de $30 \pm 4\,\%$ con respecto a la Medida-F.

Resultados de la operación conjunta de los componentes en el escenario 2: La Tabla 3.5 muestra los resultados de la operación conjunta de los componentes, incluido el Componente de detección de consonantes oclusivas sordas, sobre las seis señales acústicas de prueba.

Señal acústica	VP	VN	FP	FN	Precisión	Exhaustividad	Medida-F
Señal 1	93	128	26	27	78,15 %	77,50 %	77,82 %
Señal 2	153	220	49	24	75,74 %	86,44 %	80,74 %
Señal 3	168	265	82	37	67,20 %	81,95 %	73,85 %
Señal 4	171	268	65	52	72,46 %	76,68 %	74,51 %
Señal 5	170	268	84	30	66,93 %	85,00 %	74,89 %
Señal 6	150	323	146	35	50,68 %	81,08 %	62,37 %

Tabla 3.5. Resultados obtenidos de la operación conjunta de los componentes en el escenario 2.

En base a dichos resultados, el intervalo de confianza estadístico, calculado con la Distribución *t* de Student a $0{,}05$ de error estadístico, con respecto a la Medida-F es $74 \pm 6\,\%$.

La Tabla 3.6 muestra, para cada señal acústica de prueba, el conteo de predicciones erróneas FP y FN que cada componente produjo. Nótese que la sumatoria de los conteos individuales es siempre mayor o igual al conteo total. Esto se debe a que una misma predicción errónea puede ocurrir por acción de más de un componente.

Señal acústica	FP total	Conteo de predicciones erróneas FP por componente		
		Silencio	Inhalación	Consonantes Oclusivas Sordas
Señal 1	26	12	6	16
Señal 2	68	35	15	18
Señal 3	111	57	26	28
Señal 4	74	40	13	21
Señal 5	102	61	14	27
Señal 6	176	80	33	63

Señal acústica	FN total	Conteo de predicciones erróneas FN por componente		
		Silencio	Inhalación	Consonantes Oclusivas Sordas
Señal 1	27	13	17	21
Señal 2	24	14	7	16
Señal 3	37	13	10	30
Señal 4	52	13	25	42
Señal 5	30	11	12	24
Señal 6	35	10	10	27

Tabla 3.6. Conteo de predicciones erróneas FP y FN por componente.

Seguidamente, en la Tabla 3.7 se muestra, para cada señal acústica de prueba, la sumatoria total del conteo de predicciones erróneas de cada componente.

Señal acústica	FP + FN total	Sumatoria total de predicciones erróneas FP + FN		
		Silencio	Inhalación	Consonantes Oclusivas Sordas
Señal 1	53	25	23	37
Señal 2	92	49	22	34
Señal 3	148	70	36	58
Señal 4	126	53	38	63
Señal 5	132	72	26	51
Señal 6	211	90	43	90

Tabla 3.7. Sumatoria total de predicciones erróneas por componente.

A partir de estos valores, se calcula el porcentaje que el conteo de predicciones erróneas de cada componente representa con relación al total de predicciones erróneas para cada señal de acústica de prueba. Estos porcentajes se muestran en la Tabla 3.8.

Señal acústica	Silencio	Inhalación	Consonantes oclusivas sordas
Señal 1	47,17 %	43,40 %	69,81 %
Señal 2	53,26 %	23,91 %	36,96 %
Señal 3	47,30 %	24,32 %	39,19 %
Señal 4	42,06 %	30,16 %	50,00 %
Señal 5	54,55 %	19,70 %	38,64 %
Señal 6	42,65 %	20,38 %	42,65 %

Tabla 3.8. Porcentaje que el conteo de predicciones erróneas de cada componente representa con relación al total de predicciones erróneas.

Así, al calcular el intervalo de confianza estadístico de estos porcentajes, utilizando la Distribución *t* de Student a $0{,}05$ de error estadístico, se obtiene los resultados mostrados en la Tabla 3.9. Estos valores representan el porcentaje de error que produce cada componente con relación al total de predicciones erróneas.

	Silencio	Inhalación	Consonantes Oclusivas Sordas
Porcentaje de error	$48 \pm 5\,\%$	$27 \pm 9\,\%$	$46 \pm 12\,\%$

Tabla 3.9. Porcentaje de error que produce cada componente con relación al total de predicciones erróneas.

Los resultados expuestos muestran que, para cualquier señal acústica nueva que presente las propiedades descritas en la Tabla 3.1, el método VAD propuesto presentará un desempeño de $74 \pm 6\,\%$ con respecto a la Medida-F. Asimismo, al proyectar el $26 \pm 4\,\%$ de error resultante como el $100\,\%$, un $48 \pm 5\,\%$ de éste se deberá por causa del Componente de detección de silencio, un $27 \pm 9\,\%$, por causa del Componente de detección de inhalación, y un $46 \pm 12\,\%$, por causa del Componente de detección de consonantes oclusivas sordas. Estos últimos porcentajes no totalizan $100\,\%$ debido a que una predicción errónea puede darse por acción de varios componentes. Es más, el hecho de que tanto el Componente de detección de silencio y el Componente de detección de conso-

nantes oclusivas sordas produzcan similares porcentajes de error ($48 \pm 5\,\%$ y $46 \pm 12\,\%$ respectivamente) denota la relación de dependencia entre los dos componentes para la correcta predicción de eventos de habla y de pausa.

3.2 Discusión

Los siguientes párrafos discuten los resultados expuestos y presentan una comparación de éstos con los devueltos por el método VAD propuesto por el proyecto WebRTC [7].

3.2.1 Comparación de resultados con el método VAD del proyecto WebRTC

Las seis señales acústicas destinadas para pruebas se utilizaron también para evaluar el desempeño del método VAD propuesto del proyecto WebRTC [7]. Bajo las mismas consideraciones tomadas para evaluar la operación del método VAD propuesto en este trabajo y en base a las mismas referencias semi-manualmente etiquetadas [27, 28] respectivas a cada señal acústica, los resultados obtenidos se presentan en la Tabla 3.10. El método VAD en cuestión fue ejecutado con los parámetros listados a continuación que, por experimentación, resultaron producir los valores más altos de Medida-F.

- **Agresividad:** 3 (valor máximo)
- **Tamaño de trama:** 10 milisegundos
- **Tamaño de relleno:** 120 milisegundos

Señal acústica	VP	VN	FP	FN	Precisión	Exhaustividad	Medida-F
Señal 1	45	105	3	55	93,75 %	45,00 %	60,81 %
Señal 2	135	227	56	56	70,68 %	70,68 %	70,68 %
Señal 3	116	228	45	83	72,05 %	58,29 %	64,44 %
Señal 4	142	219	16	67	89,87 %	67,94 %	77,38 %
Señal 5	146	195	11	48	92,99 %	75,26 %	83,19 %
Señal 6	83	193	16	104	83,84 %	44,39 %	58,04 %

Tabla 3.10. Resultados obtenidos de la ejecución del método VAD propuesto por el proyecto WebRTC.

En base a estos resultados, el intervalo de confianza estadístico, calculado con la Distribución *t* de Student a $0{,}05$ de error estadístico, con respecto a la Medida-F es $69 \pm 9\,\%$. Ello muestra que, para cualquier señal acústica nueva que presente las propiedades descritas en la Tabla 3.1, el método VAD propuesto por el proyecto WebRTC presentará un desempeño de $69 \pm 9\,\%$ con respecto a la Medida-F.

Al comparar entre los resultados obtenidos por la operación en el escenario 2 del método VAD propuesto y los obtenidos por la propuesta del proyecto WebRTC, se puede notar que, como muestra la Figura 3.4, el método VAD

propuesto presenta un incremento de $5\,\%$ en cuanto a la Medida-F. Además, su intervalo de confianza estadístico garantiza una dispersión de ± 6 puntos porcentuales del valor de Medida-F resultante del análisis de una nueva señal acústica de prueba, comparado con los ± 9 puntos porcentuales de dispersión presentados por la propuesta VAD del proyecto WebRTC. Del mismo modo, al comparar los resultados de operación del método VAD obtenidos tanto en el escenario 1 como en el escenario 2, se puede apreciar el drástico impacto que tiene la detección de consonantes oclusivas sordas sobre el resultado final de predicción de eventos de habla y de pausa; pues en el escenario 2, los resultados mejoraron en un $44\,\%$.

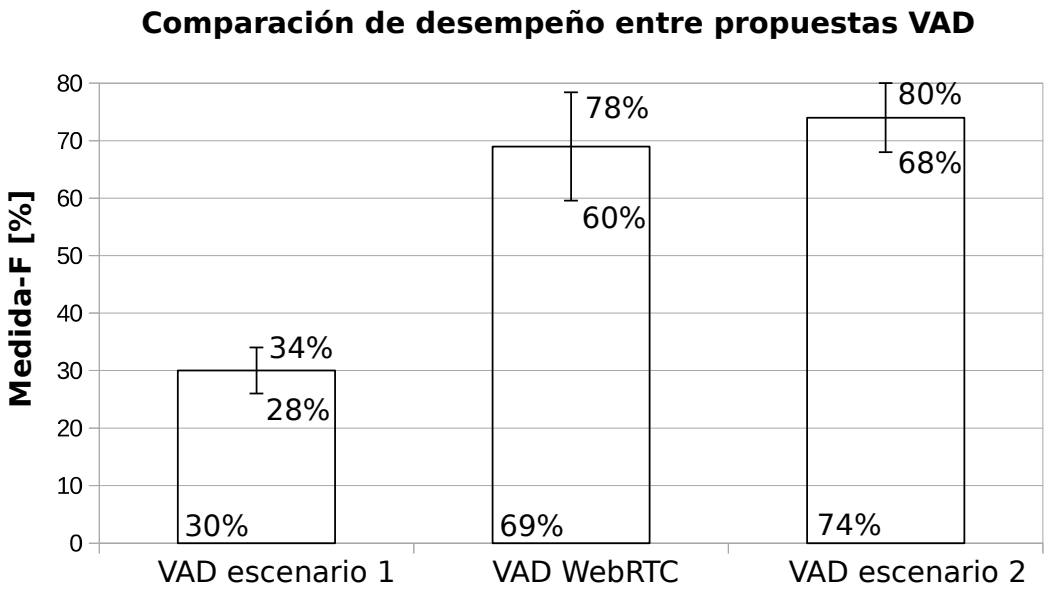


Figura 3.4. Comparación de resultados obtenidos por las diferentes propuestas VAD.

Esta comparativa demuestra que es posible filtrar y considerar como evento de habla a uno de pausa que sucede por la pronunciación de una consonante oclusiva sorda, logrando así elevar la tasa de aciertos de predicción de habla y pausa.

4 CONCLUSIONES

El trabajo presentado demostró la importancia que tiene la correcta detección de consonantes oclusivas sordas para incrementar la tasa de aciertos de predicción de eventos de habla y de pausa en una señal acústica libre de ruido. El método VAD propuesto utiliza clasificadores entrenados para detectar, respectivamente, regiones de silencio, de inhalación y de consonantes oclusivas sordas. Cada clasificador produjo una tasa de clasificación acertada mayor al $95\,\%$ durante su entrenamiento. No obstante, al ellos operar conjuntamente sobre las señales acústicas de prueba, el valor de Medida-F resultante fue $74 \pm 6\,\%$. Sin embargo, ello representa una mejora del $5\,\%$ con respecto a la Medida-F resultante de la propuesta VAD del proyecto WebRTC. Por otro lado, dado que un $46 \pm 12\,\%$ del total de las predicciones erróneas producidas por el método VAD se debió a la detección de consonantes oclusivas sordas, se hace necesario investigar otras maneras de detectar las consonantes mencionadas. En este sentido, el trabajo futuro se plantea en la línea de la detección de consonantes oclusivas sordas. De manera inicial, se supone que al tener en cuenta el contexto de pronunciación de una consonante oclusiva sorda durante el entrenamiento de clasificadores, la tasa final de aciertos de predicción de eventos de habla y de pausa se incrementará; pues dicho contexto no fue tomado en cuenta en el presente trabajo.

Por otro lado, la separación por componentes del proceso general de detección de actividad de voz posibilita extender las capacidades de detección de eventos de pausa en la señal, con lo cual, otro tipo de regiones aparte de silencio o inhalación podrían ser identificadas, por ejemplo: tos, risa, aplauso, etc. Sin embargo, la ejecución serial de los componentes evita que el método VAD pueda ser utilizado en tiempo real, es decir, en el momento en que la señal acústica se genera, además, incrementa el tiempo de computación requerido. No obstante, la separación en componentes propicia la ejecución de ellos de manera paralela.

REFERENCIAS

[1] J. Ramirez, J. M. Gorriz, and J. C. Segura, “Voice activity detection. fundamentals and speech recognition system robustness,” in *Robust Speech*, M. Grimm and K. Kroschel, Eds. Rijeka: IntechOpen, 2007, ch. 1. [Online]. Available: https://doi.org/10.5772/4740

[2] S. Graf, T. Herbig, M. Buck, and G. Schmidt, “Features for voice activity detection: a comparative analysis,” *EURASIP Journal on Advances in Signal Processing*, vol. 2015, no. 1, p. 91, Nov 2015. [Online]. Available: https://doi.org/10.1186/s13634-015-0277-z

[3] L. Karray and A. Martin, “Towards improving speech detection robustness for speech recognition in adverse conditions,” *Speech Communication*, vol. 40, no. 3, pp. 261–276, 2003. [Online]. Available: https://doi.org/10.1016/S0167-6393(02)00066-3

[4] D. K. Freeman, G. Cosier, C. B. Southcott, and I. Boyd, “The voice activity detector for the pan-european digital cellular mobile telephone service,” in *International Conference on Acoustics, Speech, and Signal Processing,*, vol. 1, May 1989, pp. 369–372. [Online]. Available: https://doi.org/10.1109/ICASSP.1989.266442

[5] P. Ladefoged, *Vowels and consonants. An Introduction to the Sounds of Languages*. Wiley-Blackwell, 2001.

[6] Y. Ma and A. Nishihara, “Efficient voice activity detection algorithm using long-term spectral flatness measure,” *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2013, no. 1, p. 87, Jul 2013. [Online]. Available: https://doi.org/10.1186/1687-4722-2013-21

[7] Various. Webrtc home. Software available at https://github.com/wiseman/py-webrtcvad. [Online]. Available: https://webrtc.org/

[8] Y. Sasaki, “The truth of the f-measure,” 2007. [Online]. Available: https://www.toyota-ti.ac.jp/Lab/Denshi/COIN/people/yutaka.sasaki/F-measure-YS-26Oct07.pdf

[9] S.-I. Kang and J.-H. Chang, “Voice activity detection based on discriminative weight training incorporating a spectral flatness measure,” *Circuits, Systems and Signal Processing*, vol. 29, no. 2, pp. 183–194, Apr 2010. [Online]. Available: https://doi.org/10.1007/s00034-009-9141-4

[10] J. Pang, “Spectrum energy based voice activity detection,” in *Computing and Communication Workshop and Conference (CCWC), 2017 IEEE 7th Annual*. IEEE, 2017, pp. 1–5.

[11] S.-H. Chen, R. C. Guido, T.-K. Truong, and Y. Chang, “Improved voice activity detection algorithm using wavelet and support vector machine,” *Computer Speech & Language*, vol. 24, no. 3, pp. 531–543, 2010.

[12] H. Wang, Y. Xu, and M. Li, “Study on the mfcc similarity-based voice activity detection algorithm,” in *2nd International Conference on Artificial Intelligence, Management Science and Electronic Commerce (AIMSEC)*, Aug 2011, pp. 4391–4394. [Online]. Available: https://doi.org/10.1109/AIMSEC.2011.6009945

[13] S. Joshi and S. Nagar, “Mfcc-based voice recognition system for home automation using dynamic programming,” *International Journal of Science and Research (IJSR)*, vol. 5, no. 2, pp. 494–498, 2016. [Online]. Available: https://doi.org/10.21275/v5i2.nov161160

[14] S. Tong, H. Gu, and K. Yu, “A comparative study of robustness of deep learning approaches for vad,” in *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*. IEEE, 2016, pp. 5695–5699.

[15] S.-S. Ahn and Y.-C. Lee, “An improved statistical model-based vad algorithm with an adaptive threshold,” *Journal of the Chinese Institute of Engineers*, vol. 29, no. 5, pp. 783–789, 2006.

[16] R. Johny Elton, P. Vasuki, and J. Mohanalin, “Voice activity detection using fuzzy entropy and support vector machine,” *Entropy*, vol. 18, no. 8, p. 298, 2016.

[17] Q.-H. Jo, J.-H. Chang, J. Shin, and N. Kim, “Statistical model-based voice activity detection using support vector machine,” *IET Signal Processing*, vol. 3, no. 3, pp. 205–210, 2009.

[18] X.-L. Zhang and J. Wu, “Deep belief networks based voice activity detection,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 4, pp. 697–710, 2013.

[19] M. Myllymäki and T. Virtanen, “Voice activity detection in the presence of breathing noise using neural network and hidden markov model,” in *Signal Processing Conference, 2008 16th European*. IEEE, 2008, pp. 1–5.

[20] Z. Shen, J. Wei, W. Lu, and J. Dang, “Voice activity detection based on sequential gaussian mixture model with maximum likelihood criterion,” in *10th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, Oct 2016, pp. 1–5.

[21] Various, “Vad-webrtc,” 2012. [Online]. Available: https://github.com/ideawu/rtc/blob/master/multimedia/webrtc/common_audio/vad/vad_core.c

[22] A. M. A. Ali, J. Van der Spiegel, and P. Mueller, “Acoustic-phonetic features for the automatic classification of stop consonants,” *IEEE transactions on speech and audio processing*, vol. 9, no. 8, pp. 833–841, 2001.

[23] F. Eyben, F. Weninger, F. Gross, and B. Schuller, “Recent developments in opensmile, the munich open-source multimedia feature extractor,” in *Proceedings of the 21st ACM International Conference on Multimedia*, ser. MM ’13. New York, NY, USA: ACM, 2013, pp. 835–838. [Online]. Available: http://doi.acm.org/10.1145/2502081.2502224

[24] C.-C. Chang and C.-J. Lin, “LIBSVM: A library for support vector machines,” *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 27:1–27:27, 2011, software available at http://www.csie.ntu.edu.tw/ cj-lin/libsvm.

[25] E. Yépez, “RangeHandling: Perl library for handling sequences of labels or data structures sorted by numerical stamps that represent ranges.” 2018, software available at https://gitlab.com/SantiagoYepez/RangeHandling.

[26] P. Boersma, “Praat, a system for doing phonetics by computer,” *Glot International*, pp. 341–345, 2001.

[27] J. de Jesus Aguiar Pontes and S. Furui, “Predicting the phonetic realizations of word-final consonants in context – a challenge for french grapheme-to-phoneme converters,” *Elsevier Speech Communication*, vol. 52, no. 10, pp. 847–862, 2010. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167639310001159

[28] B. Bigi, “Automatic Speech Segmentation of French: Corpus adaptation.” *2nd Asian Pacific Corpus Linguistics Conference*, p. 32, 2014.

[29] NHK. Nhk world-japan. [Online]. Available: https://www3.nhk.or.jp/nhkworld/

ORDEN DE EMPASTADO