OPTIMIZACIÓN DE VAD PARA VARIACIONES EN LA VELOCIDAD DEL HABLA DE AUDIOS EN ESPAÑOL ARGENTINO

Mateo Garcia Iacovelli[1]

[1]Ingeniería de Sonido, Universidad Nacional de Tres de Febrero
mateogi99@gmail.com

Matias Di Bernardo[2]

[2]Ingeniería de Sonido, Universidad Nacional de Tres de Febrero
matias.di.bernardo@hotmail.com

Gala Lucia Gonzalez Barrios[3]

[3]Doctoranda en Diseño Centrado en el Humano, Virginia Tech, USA
gala@vt.edu

Emmanuel Misley[4]

[4]Ingeniería de Sonido, Universidad Nacional de Tres de Febrero
emmanuel.misley@gmail.com

Resumen – La detección de actividad de voz constituye un paso en el procesamiento de audio que presenta dificultades cuando la velocidad del habla varía, como sucede en el español de Argentina. Los métodos que emplean parámetros estáticos tienden a presentar limitaciones al segmentar con precisión estos audios, lo que compromete la calidad de los conjuntos de datos utilizados en tareas como la conversión de texto a voz. Este trabajo propone una estrategia adaptativa que combina un clasificador de velocidad del habla —basado en transcripciones con marcas temporales— con un ajuste dinámico de los parámetros de un algoritmo de detección de voz. El sistema agrupa los fragmentos de audio según el índice de palabras por segundo y adapta los parámetros del algoritmo a cada grupo. Los resultados indican que esta propuesta genera una segmentación más precisa y detallada que los enfoques convencionales, especialmente en grabaciones de habla espontánea.

Palabras claves: Voice Activity Detection, Speech Rate, Adaptive VAD, Whisper, Speech Processing

1. INTRODUCCIÓN

El procesamiento de audio hablado es un área central en el desarrollo de sistemas de síntesis y reconocimiento automático del habla. Dentro de este campo, la detección de actividad de voz (VAD) constituye un procedimiento para segmentar el habla de manera precisa y eficiente, separándola de los silencios o el ruido de fondo. Sin embargo, la aplicación de algoritmos VAD presenta un desafío técnico en grabaciones extensas que contienen variaciones en la velocidad de emisión, una característica prosódica común en dialectos como el español argentino. Estas fluctuaciones en el ritmo del habla tienden a comprometer la sensibilidad de los detectores tradicionales, que operan con parámetros estáticos, afectando directamente la calidad y granularidad de los corpus de datos utilizados para el entrenamiento de modelos de conversión de texto a voz (TTS).

Para abordar la detección de voz, se han desarrollado sistemas VAD robustos basados en aprendizaje profundo que superan a los métodos clásicos en condiciones acústicas adversas [1]. No obstante, estos sistemas aún presentan limitaciones al enfrentar condiciones de habla no homogénea, donde la dinámica prosódica varía considerablemente dentro de una misma locución. Por otro lado, herramientas de transcripción a gran escala como Whisper [2] presentan una alta precisión en el reconocimiento del habla con marcas temporales, pero no abordan directamente el problema de la segmentación adaptativa. La limitación de los enfoques existentes reside, por tanto, en la falta de un mecanismo que ajuste dinámicamente la sensibilidad del detector de voz en función de las características prosódicas locales del hablante. Esta laguna da lugar a la necesidad de desarrollar estrategias que integren el análisis del ritmo del habla en el proceso de segmentación.

El objetivo de este estudio es proponer y evaluar una metodología de segmentación automática que combina un clasificador de velocidad del habla, basado en la transcripción temporizada de Whisper, con un proceso de optimización de hiperparámetros de un modelo VAD. Se busca evaluar si este enfoque adaptativo mejora la precisión de la segmentación en contextos de velocidad variable en comparación con métodos estáticos. Para ello, el presente artículo describe en primer lugar la metodología y el algoritmo implementado; a continuación, se presentan los resultados obtenidos a partir de un corpus de evaluación diseñado específicamente; finalmente, se discuten dichos resultados y se exponen las conclusiones derivadas del trabajo.

2. TRABAJOS RELACIONADOS

La evolución de los sistemas de detección de actividad de voz (VAD) ha estado marcada por una búsqueda constante de robustez frente a condiciones acústicas no ideales. Los enfoques iniciales, basados en métricas como la energía de la señal o la tasa de cruces por cero [3], fueron superados por métodos estadísticos [4, 5, 6] y, más recientemente, por arquitecturas de aprendizaje profundo [1, 7, 8]. Estos modelos modernos pueden discriminar la voz del ruido con alta precisión. Sin embargo, la mayoría de estos sistemas operan bajo la premisa de parámetros estáticos (umbrales, duraciones mínimas), lo que compromete su eficacia ante fluctuaciones prosódicas pronunciadas, como las variaciones en la velocidad del habla, características del discurso espontáneo [9, 10]. Esta rigidez inherente provoca errores de segmentación, ya sea omitiendo fragmentos de habla rápida o fragmentando incorrectamente pausas en el habla lenta.

Paralelamente, los avances en el reconocimiento automático del habla (ASR) han culminado en modelos a gran escala como Whisper [2], que ofrecen una transcripción de alta precisión en múltiples idiomas. Más allá de la conversión de audio a texto, la contribución fundamental de estos sistemas para la segmentación es su capacidad de generar información temporal a nivel de palabra mediante una supervisión débil. Variantes como WhisperX han refinado aún más esta capacidad, permitiendo una alineación forzada de alta precisión [11]. Estos modelos, evaluados exitosamente en diversos acentos y estilos de habla [12], proporcionan una fuente rica de metadatos sobre la estructura rítmica del discurso que los sistemas VAD tradicionales ignoran por completo.

El presente trabajo se posiciona en la intersección de estos dos campos. Mientras que los sistemas VAD se centran en el análisis de características acústicas de bajo nivel para detectar la presencia de voz, los modelos de ASR como Whisper operan a un nivel lingüístico para identificar palabras. La hipótesis fundamental de esta investigación es que la información temporal de alto nivel extraída del ASR puede ser utilizada para modular dinámicamente los parámetros de un VAD de bajo nivel. Al informar al detector sobre la velocidad del habla local, se puede ajustar su sensibilidad de manera adaptativa, superando así la limitación de los enfoques estáticos y logrando una segmentación más precisa y contextualmente consciente, especialmente en audios con alta variabilidad prosódica como los del español argentino.

3. CONCEPTOS Y TERMINOS

3.1 Optimización de hiperparámetros

En el contexto del aprendizaje automático, los hiperparámetros son parámetros de configuración externos al modelo, cuyo valor no se aprende durante el proceso de entrenamiento, pero que influyen directamente en su rendimiento. La optimización de hiperparámetros consiste en la búsqueda de la combinación de valores que maximiza o minimiza una métrica objetivo, como la precisión o el error. Existen diversas estrategias para esta búsqueda, entre las que se destacan:

- Grid Search: Consiste en definir un conjunto discreto de valores posibles para cada hiperparámetro y evaluar exhaustivamente todas las combinaciones. Aunque garantiza la exploración completa del espacio definido, su coste computacional es elevado y su eficiencia es baja si algunos parámetros tienen poco impacto.
- Random Search: Selecciona combinaciones de hiperparámetros al azar dentro de un espacio de búsqueda definido. Su ventaja radica en que puede encontrar configuraciones de alto rendimiento con un número significativamente menor de evaluaciones, especialmente cuando solo un subconjunto de los hiperparámetros es verdaderamente influyente.
- Tree-structured Parzen Estimator (TPE): Es una técnica de optimización bayesiana que construye modelos probabilísticos para estimar el rendimiento esperado en función de los valores de los hiperparámetros. A partir de esta estimación, el algoritmo prioriza la exploración de las combinaciones más prometedoras, resultando en una búsqueda más eficiente que las estrategias no informadas [13, 14].

3.2 Clasificación de la velocidad del habla

La tasa de elocución (speaking rate) se refiere a la velocidad con la que un hablante produce unidades lingüísticas. Aunque puede medirse en sílabas o fonemas por segundo, la métrica de palabras por segundo (WPS) es especialmente útil en contextos de análisis basados en transcripción automática. Formalmente, se define mediante la Ec. (1):

$$WPS = \frac{N_{Palabras}}{T} \quad (1)$$

Donde $N_{Palabras}$ es el número total de palabras pronunciadas en un intervalo temporal y T es la duración de dicho intervalo, medido en segundos. Esta métrica permite cuantificar objetivamente el ritmo del habla en diferentes segmentos de un audio.

3.3 Detector de Actividad de Voz: SileroVAD

Para la detección de actividad de voz, este trabajo emplea el modelo SileroVAD, un detector basado en redes neuronales ligeras, optimizado para su ejecución en tiempo real [15]. El modelo acepta entradas de audio con frecuencias de muestreo de 8000 Hz o 16000 Hz y es capaz de procesar fragmentos de 30 milisegundos en menos de un milisegundo en un único hilo de CPU. Su comportamiento puede ser ajustado mediante varios hiperparámetros, de los cuales los más relevantes para este estudio son:

- threshold: Un valor entre 0 y 1 que define la sensibilidad del modelo para distinguir entre voz y no-voz. Valores bajos aumentan la sensibilidad, lo que puede incrementar la detección de segmentos de habla pero también el riesgo de falsos positivos.
- min_speech_duration_ms: Establece la duración mínima, en milisegundos, que debe tener un segmento para ser clasificado como voz. Permite filtrar sonidos breves o esporádicos que no constituyen habla.
- min_silence_duration_ms: Define la duración mínima de silencio, en milisegundos, necesaria para considerar que un segmento de habla ha finalizado. Es útil para evitar la fragmentación prematura del discurso ante pausas breves.

4. DISPOSICIÓN EXPERIMENTAL, MÉTODOS APLICADOS

4.1 Arquitectura del Algoritmo Adaptativo

La metodología propuesta consiste en una arquitectura de segmentación automatizada, diseñada para ajustarse dinámicamente a las variaciones en la velocidad del habla dentro de un mismo archivo de audio. El sistema integra un modelo de transcripción de alto nivel (Whisper) para obtener información prosódica, un clasificador de velocidad del habla y un detector de actividad de voz (SileroVAD) cuyos hiperparámetros se optimizan de forma adaptativa. El objetivo es superar las limitaciones de los

sistemas VAD con parámetros estáticos, generando una segmentación más granular y precisa en contextos de habla heterogéneos. El flujo de trabajo completo del algoritmo se ilustra en la figura 1.

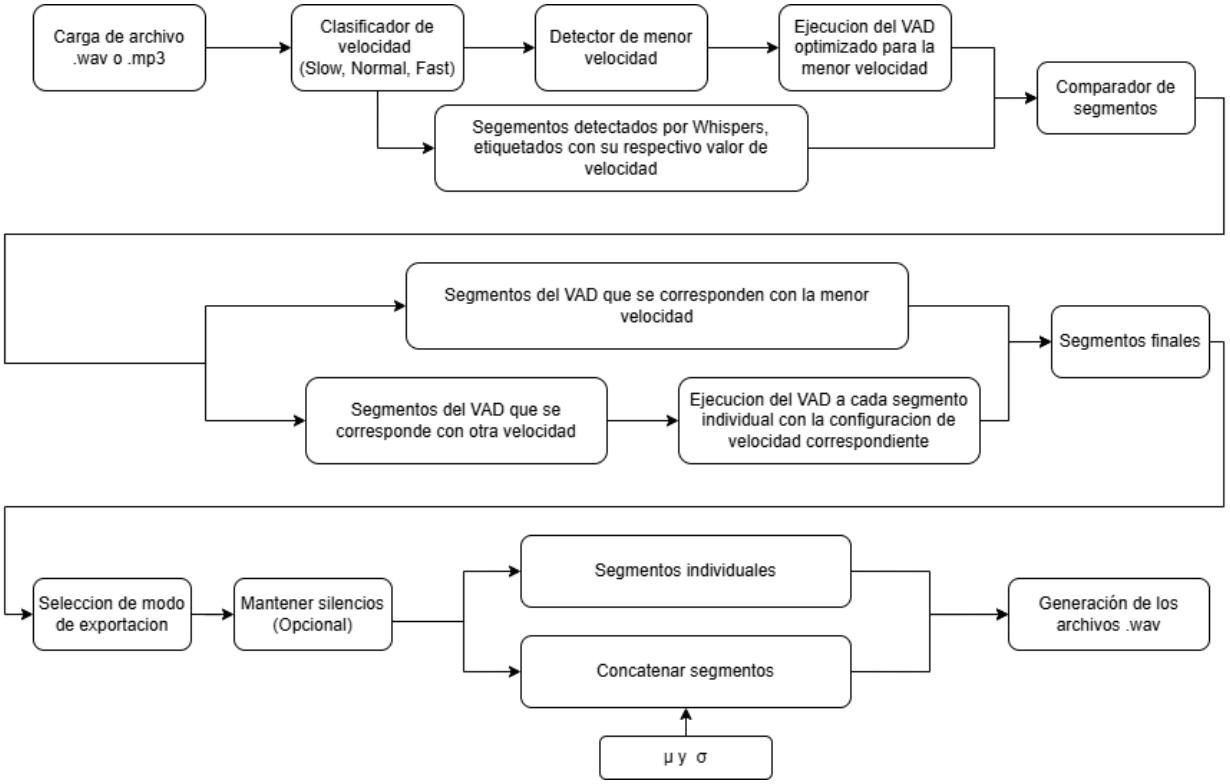

Figura 1 – Diagrama de flujo del algoritmo desarrollado.

4.2 Proceso de Segmentación y Optimización

1. Clasificación de la Velocidad del Habla: Inicialmente, el audio de entrada es procesado por el modelo Whisper para obtener una transcripción con marcas de tiempo a nivel de palabra. A partir de estos datos, se calcula el índice WPS para cada fragmento de habla detectado por Whisper. Los fragmentos se clasifican en tres categorías: Slow (WPS < 2), Normal (2 ≤ WPS ≤ 4) o Fast (WPS > 4).
2. Optimización Inicial: Se identifica la categoría de velocidad más lenta presente en el audio. Utilizando la biblioteca Optuna, se ejecuta un proceso de optimización (basado en TPE) para encontrar el conjunto de hiperparámetros de SileroVAD que maximiza la detección para esa categoría específica. El VAD se ejecuta sobre todo el audio con esta configuración inicial, generando una primera segmentación base.
3. Re-optimización Adaptativa: Se realiza una comparación temporal entre los segmentos obtenidos por el VAD y los fragmentos clasificados originalmente por Whisper. Si un

segmento detectado por el VAD se corresponde temporalmente con un fragmento que Whisper había clasificado en una categoría de velocidad diferente, el sistema aísla ese segmento y vuelve a ejecutar el VAD sobre él, pero esta vez utilizando los hiperparámetros optimizados para su categoría correcta. Este mecanismo iterativo permite que la sensibilidad del VAD se ajuste localmente a lo largo del audio.

4. Ajuste Fino y Exportación: Una vez completado el proceso de re-optimización, se realiza un ajuste final de los tiempos de inicio y fin de cada segmento para asegurar la coherencia con el audio original. El sistema permite exportar los fragmentos de forma individual o concatenada hasta alcanzar una longitud objetivo, con la opción de preservar los silencios originales entre ellos.

4.3 Corpus de Evaluación

Para evaluar el rendimiento del algoritmo, se seleccionó un corpus de cinco archivos de audio con características lingüísticas y prosódicas desafiantes. Estos audios fueron escogidos por presentar variaciones marcadas en la velocidad del habla, diversidad dialectal del español argentino y condiciones de grabación no controladas, factores que comprometen el desempeño de los sistemas VAD tradicionales. La descripción de cada audio se presenta en la tabla 1.

Tabla 1: Descripción de los audios utilizados para la evaluación del sistema.

Audio	Descripción
1	Programa de TV donde los participantes recitan trabalenguas. Contiene múltiples hablantes con cambios bruscos de velocidad.
2	Entrevistas a hinchas de fútbol de distintas provincias (Santiago del Estero, Córdoba, Mendoza y Tucumán), con diferencias dialectales.
3	Entrevista a un comediante de Tucumán que habla rápidamente y utiliza léxico coloquial intensivo.
4	Conversación telefónica donde una persona acelera el ritmo para confundir a la otra.
5	Fragmento de un audiolibro editado artificialmente con cambios de velocidad (x2 y x0.5) en diferentes momentos.

4.4 Métricas de Evaluación

El desempeño del algoritmo se evaluó mediante un enfoque complementario que combina análisis cuantitativo y validación cualitativa. Desde una perspectiva cuantitativa, se comparó el número total de segmentos de habla detectados por el método adaptativo propuesto frente a los obtenidos utilizando

la configuración por defecto de SileroVAD. Esta métrica sirve como un indicador primario del incremento en la sensibilidad y granularidad de la segmentación. Para la validación cualitativa, se realizó una inspección perceptual mediante la escucha crítica de todos los fragmentos generados. El objetivo de este análisis fue identificar errores de corte, definidos como aquellos casos en los que el sistema interrumpía una palabra a mitad de su articulación o realizaba una segmentación inadecuada dentro de una unidad lingüística continua, comprometiendo así la inteligibilidad del segmento resultante.

5. ANÁLISIS DE RESULTADOS

El rendimiento del sistema VAD adaptativo propuesto fue comparado cuantitativamente con la configuración por defecto del modelo SileroVAD, que emplea un conjunto estático de hiperparámetros. La tabla 2 resume el número total de segmentos de habla generados por cada enfoque para los cinco audios del corpus de evaluación.

Tabla 2: Comparación del número de segmentos de habla detectados.

	Cantidad de segmentos detectados	
Audio	Configuración por defecto	Configuración Optimizada
1	24	32
2	22	31
3	6	16
4	5	11
5	19	23

Los datos muestran que el método optimizado generó un mayor número de segmentos en todos los audios evaluados, lo que sugiere un incremento general en la sensibilidad del sistema. Este aumento fue más pronunciado en los audios que presentan una alta variabilidad prosódica natural. En el Audio 3, correspondiente a un hablante con un ritmo naturalmente rápido, el número de segmentos detectados aumentó un 167%. De manera similar, en el Audio 4, donde el hablante acelera deliberadamente el ritmo, el incremento fue del 120%. Se observaron también aumentos sostenidos en los audios con diversidad dialectal y cambios de ritmo (Audio 1 y 2, con un 33% y 41% respectivamente). En contraste, el Audio 5, un audiolibro con velocidad modificada artificialmente, presentó el incremento más moderado (21%).

Los resultados cuantitativos, validados por la escucha perceptual, indican que el método adaptativo propuesto produce una segmentación más granular y precisa que un enfoque estático. La discusión

no reside únicamente en el mayor número de segmentos, sino en la distribución de esta mejora. El hecho de que los incrementos más notables se observen en los audios de habla espontánea (Audio 3 y 4) y no en el audiolibro modificado artificialmente (Audio 5) es un hallazgo central.

Este patrón sugiere que los sistemas VAD estáticos no fallan únicamente ante cambios de velocidad, sino, de forma más crítica, ante las fluctuaciones rítmicas, la coarticulación y la irregularidad inherentes al habla natural. El Audio 5, aunque su velocidad fue alterada, parte de una locución de audiolibro caracterizada por una articulación clara y pausas bien definidas. Su estructura prosódica subyacente permanece regular, lo que facilita la tarea para un detector con parámetros fijos. Por el contrario, el habla espontánea y rápida de los Audios 3 y 4 presenta una dinámica más compleja que el método estático no segmenta eficazmente, resultando en una sub-segmentación severa.

El enfoque adaptativo, al ajustar su sensibilidad en función de la velocidad del habla local, segmenta con mayor eficacia estos contextos desafiantes. La validación cualitativa corroboró que los segmentos adicionales detectados por el método propuesto correspondían a unidades lingüísticas coherentes y no a falsos positivos, respetando la estructura del discurso. Por tanto, la mejora observada no es solo cuantitativa, sino que se traduce en una mayor fidelidad de la segmentación respecto al contenido hablado.

6. CONCLUSIÓNES

Este trabajo ha presentado una metodología para la detección de actividad de voz (VAD) que se adapta dinámicamente a las variaciones en la velocidad del habla. El método utiliza la transcripción temporizada de Whisper para clasificar segmentos de audio según su índice de palabras por segundo (WPS) y ajusta los hiperparámetros de un detector VAD para cada categoría. Los resultados indican que este enfoque adaptativo genera una segmentación significativamente más granular y precisa en comparación con un método estático, siendo este efecto más pronunciado en grabaciones de habla espontánea que en locuciones con velocidad modificada artificialmente.

La principal contribución de esta investigación es el desarrollo de un pipeline automatizado que permite generar corpus de habla con una segmentación más detallada y contextualmente consciente. Esto representa un avance para la creación de conjuntos de datos para sistemas de síntesis de voz

(TTS) que busquen modelar y reproducir de manera más fiel la variabilidad prosódica del habla humana.

Como línea de trabajo futuro, se propone la investigación de métodos para la definición automática de los umbrales de WPS. La determinación de estos límites de manera adaptativa, en lugar de empírica, podría mejorar la generalización del sistema a una mayor diversidad de hablantes, dialectos y estilos de habla.

REFERENCIAS

[1] S. Braun y I. Tashev. On training targets for noise-robust voice activity detection. Actas de la 29ª Conferencia Europea de Procesamiento de Señales (EUSIPCO), Dublín, Irlanda (2021) 421-425.

[2] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, y I. Sutskever. Robust Speech Recognition via Large-Scale Weak Supervision. arXiv:2212.04356 (2022).

[3] R. G. Bachu, S. Kopparthi, B. Adapa, y B. D. Barkana. Voiced/Unvoiced Decision for Speech Signals Based on Zero-Crossing Rate and Energy. En K. Elleithy (Ed.), Advanced Techniques in Computing Sciences and Software Engineering. Springer, Países Bajos (2010) 279-282.

[4] L. N. Tan, B. J. Borgstrom, y A. Alwan. Voice activity detection using harmonic frequency components in likelihood ratio test. Actas de la Conferencia Internacional de Acústica, Habla y Procesamiento de Señales (ICASSP), Dallas, TX, USA (2010) 4466-4469.

[5] K. Lee y D. P. Ellis. Voice activity detection in personal audio recordings using autocorrelogram compensation. Actas de INTERSPEECH, Pittsburgh, PA, USA (2006).

[6] J. Ramirez, J. M. Górriz, y J. C. Segura. Voice activity detection. Fundamentals and speech recognition system robustness. Robust speech recognition and understanding, 6(9) (2007) 1-22.

[7] S. Graf, T. Herbig, M. Buck, y G. Schmidt. Features for voice activity detection: A comparative analysis. EURASIP Journal on Advances in Signal Processing, 2015(1) (2015) 91.

[8] I. Tashev y S. Mirsamadi. DNN-based causal voice activity detector. Taller de Teoría de la Información y Aplicaciones (ITA), La Jolla, CA, USA (2016).

[9] S. Kumar, S. S. Buddi, U. O. Sarawgi, V. Garg, S. Ranjan, O. Rudovic, A. H. Abdelaziz, y S. Adya. Comparative Analysis of Personalized Voice Activity Detection Systems: Assessing Real-World Effectiveness. arXiv:2406.09443 (2024).

[10] B. Karan, J. J. van Vüren, F. de Wet, y T. Niesler. A Transformer-Based Voice Activity Detector. Actas de Interspeech 2024, Kos, Grecia (2024) 3819-3823.

[11] M. Bain, J. Huh, T. Han, y A. Zisserman. WhisperX: Time-Accurate Speech Transcription of Long-Form Audio. arXiv:2303.00747 (2023).

[12] C. Graham y N. Roll. Evaluating OpenAI's Whisper ASR: Performance analysis across diverse accents and speaker traits. JASA Express Letters, 4(2) (2024) 025206.

[13] T. Akiba, S. Sano, T. Yanase, T. Ohta, y M. Koyama. Optuna: A Next-generation Hyperparameter Optimization Framework. Actas de la 25ª Conferencia ACM SIGKDD sobre Descubrimiento de Conocimiento y Minería de Datos, Anchorage, AK, USA (2019) 2623-2631.

[14] S. Shekhar, A. Bansode, y A. Salim. A comparative study of hyper-parameter optimization tools. Actas de la Conferencia IEEE Asia-Pacífico sobre Ciencia de la Computación e Ingeniería de Datos (CSDE), (2021) 1-6.

[15] Silero Team. Silero VAD: Pre-trained enterprise-grade Voice Activity Detector (VAD), Number Detector and Language Classifier. Repositorio de software, GitHub (2024). https://github.com/snakers4/silero-vad