Universidad de Buenos Aires
Facultad de Ciencias Exactas y Naturales

Medición de la velocidad del habla con aprendizaje profundo

Tesis de Licenciatura en Ciencias de Datos

Pedro Ortiz

Director: Pablo Riera

Buenos Aires, 2025

MEDICIÓN DE LA VELOCIDAD DEL HABLA CON APRENDIZAJE PROFUNDO

En esta tesis se aborda el problema de estimar la velocidad del habla —medida en fonemas por segundo (FPS) o sílabas por segundo (SPS)— a partir de señales acústicas. Esta tarea resulta relevante en contextos educativos, clínicos y tecnológicos, donde el ritmo del habla puede aportar información clave sobre la producción y comprensión del lenguaje. A diferencia de métodos tradicionales que requieren transcripciones explícitas, aquí se propone una aproximación que trabaja directamente sobre representaciones acústicas derivadas del audio.

Se utilizan posteriogramas, matrices temporales que reflejan la probabilidad de activación de cada fonema a lo largo del tiempo, como entrada para modelos de regresión. Se exploran tanto modelos lineales con atributos fonéticos diseñados manualmente como arquitecturas neuronales convolucionales que operan sobre estas representaciones, en particular modelos 1D y 2D con distintas estrategias de convolución y resumen temporal.

Los resultados muestran que los modelos que preservan la estructura fonética de las representaciones alcanzan el mejor desempeño en ambas tareas. También se presentan experimentos controlados con oraciones en español e inglés, que permiten observar cómo los modelos responden a cambios en el ritmo del habla, incluso fuera del dominio original de entrenamiento. Estos hallazgos refuerzan la idea de que un diseño de modelo alineado con la estructura del lenguaje puede mejorar significativamente la estimación de la velocidad del habla.

Palabras clave: Velocidad del habla, Posteriogramas, Fonemas por segundo, Sílabas por segundo, Redes neuronales convolucionales, Reconocimiento del habla, Representaciones acústicas.

MEDICIÓN DE LA VELOCIDAD DEL HABLA CON APRENDIZAJE PROFUNDO

This thesis addresses the problem of estimating speech rate —measured in phonemes per second (FPS) or syllables per second (SPS)— directly from acoustic signals. This task is relevant in educational, clinical, and technological settings, where the rhythm of speech can provide valuable information about language production and comprehension. Unlike traditional methods that rely on explicit transcriptions, this approach operates directly on acoustic representations extracted from audio.

We use posteriograms, temporal matrices that encode the probability of phoneme activation over time, as input to regression models. Both linear models with handcrafted phonetic features and convolutional neural network architectures are explored, including 1D and 2D models with different convolution and temporal pooling strategies.

Results show that models preserving the phonetic structure of the input yield the best performance in both FPS and SPS prediction tasks. Additionally, controlled experiments with utterances in both Spanish and English demonstrate the models' ability to respond to speech rate variations, even outside the training domain. These findings support the idea that model designs aligned with the structure of language can lead to substantial improvements in estimating speech rate.

Keywords: Speech rate, Posteriograms, Phonemes per second, Syllables per second, Convolutional neural networks, Speech recognition, Acoustic representations.

AGRADECIMIENTOS

A mi familia. Porque aunque esto sea un logro personal y existieron momentos de duda, siempre tuve el soporte que necesite, no todos entramos a la universidad en las mismas condiciones y si hoy entrego esta tesis es porque siempre tuve una gran educación y también porque tuve la posibilidad de elegir entre estudiar o trabajar. En casa sobran libros, sobran lugares para estudiar y jamas dudaron de mi capacidad. Gracias a mis padres por este privilegio y por siempre hacer todo lo posible, y en algunos momentos hasta lo imposible, para que yo pueda seguir adelante.

A mis abuelos, que siempre confían ciegamente en mí y me apoyan en cada paso, aun sin saber demasiado bien qué estaba estudiando o haciendo. Gracias Chicho por tus llamaditos para saber como estaba el clima acá.

A mi padrino, que en todo momento está presente, desde llevarme a Agronomía para inscribirme al CBC, hasta mandarme un mensajito de suerte o una llamada para ponerse al día con todo.

A Juan, por bancarse la convivencia y mis rispideces diariamente durante más de dos años. Y a su familia también, por estar en los más mínimos detalles y hacer que la adaptación a Capital haya sido un mero trámite.

A Nacho, mis tíos y primos, que cada uno a su manera están siempre apoyando y presentes.

A Agus, Delfi, Goico, Joaco, Juani, Luz, Martín, Sol y Valen, por alegrar cursadas, viajes, tardes de estudio; por hacer llevadera la carrera en toda ocasión, y sobre todo, por hacerme tomar mate.

A mis amigos de Bolívar, por las comidas, asados, juntadas, innumerables momentos de risa y por ayudarme a despejar la cabeza de la facultad.

A Pony por alegrarme cada día de cursada virtual por allá por 2021, y recibirme con una alegría inconmensurable cada vez que me ve.

A Pablo, por aceptar dirigir esta tesis y acompañarme durante todo el proceso, siempre dispuesto a aclarar dudas y orientar el trabajo.

Y por último, pero no menos importante, a la UBA por la educación pública, y a la Facultad, por su excelencia académica y sobre todo por la calidad humana que encontré en cada etapa del camino.

A toda persona que se tome el tiempo para leer este trabajo.

Índice general

1. INTRODUCCIÓN

1.1. Motivación

La **velocidad del habla** o speech rate —definida como la cantidad de palabras, sílabas o fonemas que una persona pronuncia en un intervalo de tiempo determinado— es un aspecto central en la caracterización del lenguaje oral. Su medición suele expresarse en palabras por minuto, sílabas por segundo o fonemas por segundo. Esta variable no es constante, sino que depende de múltiples factores, incluyendo el idioma, el contexto discursivo, así como características individuales del hablante como edad, género, dialecto o condiciones psicológicas y fisiológicas.

Además, distintos estudios han mostrado que la velocidad promedio de habla también varía según el idioma. Por ejemplo, investigaciones translingüísticas [1] [2] han demostrado que lenguas como el japonés y el español tienden a presentar mayores tasas de sílabas por segundo que otras como el inglés o el alemán. Sin embargo, estas diferencias de velocidad superficial se compensan con la densidad informativa de cada lengua, de modo que la cantidad de información transmitida por segundo tiende a ser constante entre idiomas, expuesto en el trabajo de Coupé et al.[2]. Este hallazgo refuerza la idea de que la velocidad del habla debe entenderse no sólo en términos de rapidez articulatoria, sino también considerando la eficiencia comunicativa y las propiedades estructurales del idioma.

Desde una perspectiva funcional, la velocidad del habla incide directamente sobre la **claridad, inteligibilidad y eficacia de la expresión**, siendo un aspecto de vital importancia en la comunicación oral.

Su correcta medición resulta de gran utilidad en distintos campos. Un ejemplo de esto es el estudio de Cohn et al. [3] perteneciente al terreno de la **lingüística computacional**, donde se discute el efecto que tienen los ajustes en la velocidad del habla que realizan los usuarios del socialbot Alexa y las implicaciones que esto posee para la interacción humano-computadora y las teorías de adaptación y acomodación lingüística.

En el **ámbito educativo**, por ejemplo, existen estudios [4] sobre la comprensión auditiva de adultos aprendiendo un segundo idioma, demostrando que velocidades de habla moderadamente rápidas reducen significativamente la comprensión, mientras que ritmos lentos no mostraron mejoras relevantes frente a velocidades promedio. Por su parte, otro trabajo [5] encontró que los estudiantes comprendían mejor y consideraban más importantes los temas presentados a una velocidad de habla más lenta. Simonds et al. [6] en cambio, analizaron cómo la velocidad del habla de un instructor afectaba la percepción de los estudiantes sobre su claridad, credibilidad y la retención de información. Los resultados mostraron que una velocidad moderada o rápida mejoraba la percepción de claridad y credibilidad, así como el aprendizaje afectivo. Finalmente, otras investigaciones [7] exploraron cómo los cambios en la velocidad del habla de madres hacia sus hijos pequeños influían en su desarrollo lingüístico, de esta forma se muestra el amplio rango de estudios relacionados al tópico educación donde el speech rate posee gran relevancia.

En el **campo de la salud**, la velocidad del habla ha mostrado ser un **biomarcador relevante** para el diagnóstico temprano y monitoreo de enfermedades neurológicas como Parkinson, Alzheimer o deterioro cognitivo leve. Esto lo podemos ver en distintos estudios [8] donde se concluyó que características como la velocidad del habla, la cantidad y dura-

ción de pausas son indicadores sensibles en las etapas iniciales del Alzheimer, permitiendo un tamizaje lingüístico accesible. También en otros trabajos [9] [10] que mostraron cómo en pacientes con enfermedad de Parkinson se observa un patrón característico de disfluencias y reducción progresiva en la velocidad del habla. Estos estudios refuerzan la hipótesis de que las alteraciones en la velocidad articulatoria y su regularidad pueden servir como marcadores de progresión de dicha enfermedad.

En este contexto, el desarrollo de métodos automáticos y robustos para estimar la velocidad del habla cobra especial relevancia, tanto en aplicaciones clínicas como educativas y tecnológicas. Esta tesis se enmarca dentro de ese objetivo, proponiendo técnicas basadas en aprendizaje profundo para abordar esta tarea desde una perspectiva moderna y generalizable.

1.2. Trabajos anteriores y objetivos

El análisis de la velocidad del habla ha sido un área de interés en procesamiento de señales y lingüística computacional desde hace décadas. Inicialmente, los enfoques tradicionales se basaban en **técnicas de procesamiento de señales** [11], utilizando transformaciones en el dominio temporal y de frecuencia para extraer características acústicas relevantes. Entre estas, la energía, la frecuencia fundamental y la modulación temporal fueron utilizadas para inferir de manera indirecta la velocidad del habla, sin necesidad de transcripciones o segmentaciones fonéticas. Por ejemplo, la detección de picos de energía o de modulación permitió estimar el ritmo del habla en condiciones controladas [12]. Sin embargo, estos métodos suelen ser altamente sensibles al ruido ambiental y tienen dificultades para generalizar entre diferentes hablantes, dialectos o estilos de habla.

Con la expansión del **aprendizaje automático**, surgieron alternativas más robustas como el uso de redes neuronales profundas [13]. Sin embargo, fueron las redes neuronales convolucionales [14], las que han demostrado ser especialmente efectivas en el modelado del habla, permitiendo extraer patrones relevantes directamente a partir de representaciones acústicas como espectrogramas o posteriogramas, sin depender de una segmentación textual previa. En particular, estas arquitecturas permiten modelar el ritmo, la fluidez y otras propiedades del habla, superando en muchos casos a los métodos tradicionales tanto en precisión como en capacidad de generalización.

En el marco de esta tesis, se continúa una línea de trabajo previamente desarrollada en el laboratorio por Palazzo [15]. Allí se exploró la estimación de la velocidad del habla en términos de fonemas por segundo mediante **modelos de regresión lineal** sobre estadísticas fonéticas derivadas de posteriogramas. Si bien ese enfoque resultó útil como baseline, presenta limitaciones en su capacidad para capturar relaciones no lineales complejas y para adaptarse a nuevos datos. Además, si bien la métrica de fonemas por segundo es informativa, en contextos educativos o clínicos puede resultar más interpretable la velocidad expresada en sílabas por segundo.

El presente trabajo busca extender esta línea mediante el uso de redes neuronales convolucionales, con el objetivo principal de estimar de manera precisa la velocidad del habla -medida en fonemas o en sílabas por segundo- directamente a partir de posteriogramas derivados de señales de audio.

Los objetivos específicos de esta tesis son:

- **Desarrollar un pipeline experimental** que permita procesar señales de audio y

transformarlas en representaciones temporales adecuadas para el entrenamiento de modelos.

- **Implementar modelos de regresión lineal** como baseline, utilizando estadísticas acústicas derivadas.
- **Diseñar, entrenar y evaluar distintas arquitecturas de redes convolucionales 1D y 2D** para estimar la velocidad del habla sin necesidad de segmentación textual.
- **Evaluar los modelos predictivos** utilizando métricas estándar (MSE, R^2, CCC, coeficiente de Pearson) con intervalos de confianza mediante bootstrapping.
- **Comparar el rendimiento de los modelos** para las tareas de estimación de fonemas por segundo y sílabas por segundo, identificando ventajas, desafíos y limitaciones de cada enfoque.

1.3. Estructura de la tesis

La presente tesis se organiza en seis secciones, incluyendo este capítulo introductorio, con una progresión lógica que intenta reflejar el proceso de investigación y desarrollo del trabajo.

En el **Capítulo 1**, se introduce el problema de estimar la velocidad del habla en unidades como sílabas y fonemas, destacando su relevancia en campos como la lingüística computacional, la educación y la salud. Se enuncian los objetivos del trabajo y se justifica su enfoque.

El **Capítulo 2** está dedicado al marco teórico y metodológico. Se introducen las unidades del habla consideradas, las representaciones acústicas utilizadas como entrada (posteriogramas), los modelos de regresión y redes neuronales empleados, así como las métricas de evaluación.

En el **Capítulo 3** se lleva a cabo un análisis exploratorio del conjunto de datos, con una descripción detallada del corpus TIMIT y un estudio de correlaciones y tendencias entre variables relevantes como duración, pausas y velocidad del habla.

El **Capítulo 4** describe la metodología experimental, incluyendo el criterio de partición de los datos, la extracción de atributos y el proceso de entrenamiento y evaluación de modelos.

El **Capítulo 5** concentra todos los experimentos realizados, organizados de forma progresiva desde enfoques simples hasta arquitecturas más complejas. Se comienza con modelos de regresión lineal como línea base, luego se incorporan atributos fonéticos temporales y estrategias de combinación de características. A continuación, se exploran distintas arquitecturas neuronales basadas en convoluciones 1D y 2D. Por último, se presentan análisis comparativos de métricas de desempeño con intervalos de confianza, así como una evaluación cualitativa mediante oraciones propias grabadas ad hoc.

Finalmente, el **Capítulo 6** sintetiza las conclusiones del trabajo y propone posibles líneas de trabajo futuro. Los últimos capítulos contienen la bibliografía utilizada y apéndices con información complementaria.

2. MARCO TEÓRICO Y HERRAMIENTAS

2.1. Unidades del habla

En el estudio del lenguaje oral, existen múltiples niveles de análisis que permiten descomponer el flujo continuo del habla en unidades discretas y significativas. Dos de las unidades más relevantes son los **fonemas** y las **sílaba**s.

Fonemas

Los **fonemas** son las unidades sonoras mínimas que permiten distinguir significados en una lengua. Aunque los alfabetos escritos suelen tener un número limitado de letras (por ejemplo, 26 en inglés), los sistemas fonológicos pueden incluir entre 40 y 60 fonemas, dependiendo del idioma y del alfabeto fonético utilizado.

Por ejemplo, la palabra *fox* (zorro) se transcribe fonéticamente como /f ɑ k s/ lo que aporta información sobre su pronunciación más allá de la ortografía. Asimismo, palabras como *flower* (flor) y *flour* (harina) se pronuncian de forma idéntica (/ˈflaʊ.ər/), pero tienen significados diferentes; estos casos se conocen como **homófonos**.

Cada fonema representa una categoría abstracta de sonidos que puede realizarse de distintas formas físicas. Las realizaciones acústicas específicas se denominan **fonos** y se escriben entre corchetes (por ejemplo, [d], [ð]). Cuando distintos fonos corresponden al mismo fonema según el contexto fonológico, se denominan **alófonos**. Así, el fonema /d/ puede realizarse como [d], [ð] según la posición en la palabra y el idioma. Formalmente, sea p_0 un fonema del lenguaje $L1$, entonces p_0 puede pronunciarse como cualquier fono dentro de los alófonos de p_0.

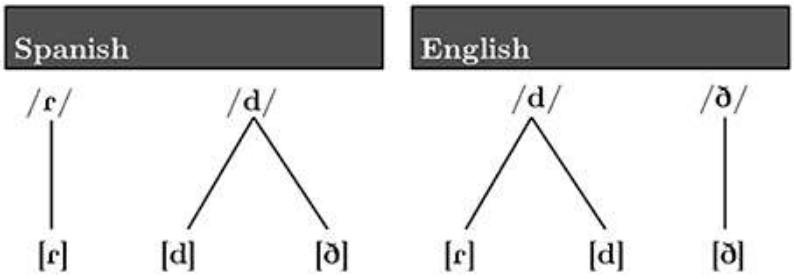

Fig. 2.1: Relación entre fonos y fonemas en español e inglés. Adaptado de [16].

En la Figura 2.1 podemos ver la relación entre fonemas en español e inglés con sus respectivos alófonos. Podemos observar que los fonos [d], [ð] y [r] son los alófonos de distintos fonemas dependiendo del lenguaje. Los fonos [d] y [ð] distinguen significados de palabras en inglés (e.g., [ðei] “they” y [dei] “day”).

Sílabas

Las **sílabas** son unidades prosódicas que agrupan uno o más fonemas en torno a un núcleo vocálico. A diferencia de los fonemas, que son unidades abstractas fonológicas, las

sílabas tienen un rol central en la estructura rítmica del habla y en su percepción por parte de los oyentes.

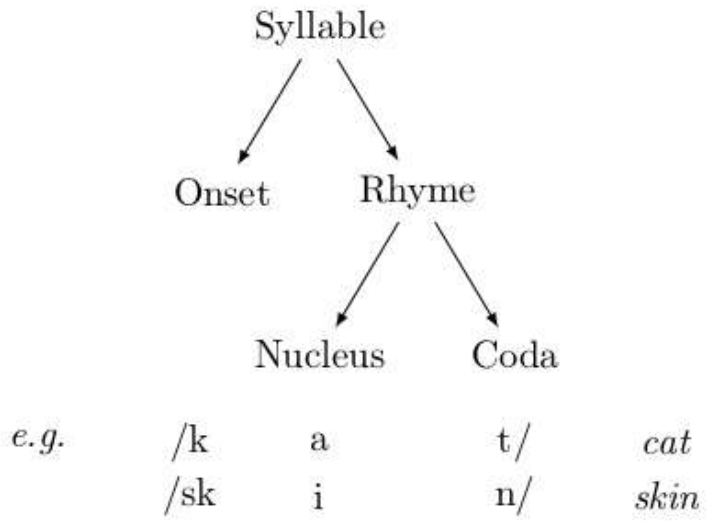

Fig. 2.2: Estructura de una sílaba. Adaptado de [17].

Desde una perspectiva perceptiva, las sílabas representan bloques más fácilmente fragmentables y reconocibles por los hablantes, lo que las convierte en una unidad particularmente relevante en contextos educativos, clínicos o de análisis del ritmo del habla.

Estimar la **velocidad del habla en sílabas por segundo** es especialmente útil en aplicaciones donde la inteligibilidad o el procesamiento auditivo juegan un papel central, como en la enseñanza de idiomas o en la evaluación de patologías del habla. Además, métricas basadas en sílabas suelen ser más interpretables que aquellas basadas exclusivamente en fonemas.

Comparación y elección

Tanto los fonemas como las sílabas aportan información valiosa para el análisis del habla. Los fonemas ofrecen una segmentación fina y precisa, útil para el análisis detallado de patrones articulatorios. Por su parte, las sílabas permiten una evaluación más global y comprensible del ritmo prosódico.

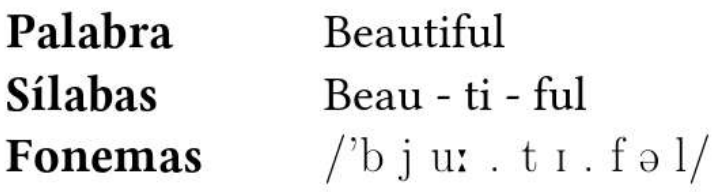

Fig. 2.3: Ejemplo de división de la palabra “Beautiful” en fonemas y en sílabas.

En este trabajo se propone abordar ambos niveles: se estimará la velocidad del habla tanto en fonemas como en sílabas por segundo, permitiendo así comparar su utilidad y robustez en distintos contextos de aplicación.

2.2. Representaciones acústicas

Las representaciones acústicas son un componente fundamental en tareas de análisis del habla, ya que permiten transformar la señal de audio pura en estructuras informativas y manejables para modelos computacionales. En el contexto de esta tesis, se utilizan **posteriogramas**, que son representaciones temporales donde cada punto en el tiempo indica la probabilidad de que se esté pronunciando un determinado fonema.

Para generar estos posteriogramas, se emplea el alineador forzado **Charsiu** [18], una herramienta originalmente diseñada para alinear automáticamente transcripciones escritas con sus correspondientes segmentos temporales de audio. Sin embargo, en esta tesis no se utiliza su modo de alineamiento tradicional, sino su *modo predictivo*, el cual permite estimar la probabilidad de fonemas directamente a partir del audio, sin necesidad de transcripción previa. Esto convierte a Charsiu en una herramienta sumamente flexible para el preprocesamiento de grandes volúmenes de datos no anotados.

Charsiu se basa en el modelo `en_w2v2_fc_10ms`, una versión adaptada de **wav2vec2** [19], el cual fue preentrenado sobre grandes corpus de habla en inglés. Este modelo utiliza una combinación de redes convolucionales y arquitectura Transformer [20] para extraer características relevantes del audio y generar predicciones fonéticas cada 10 milisegundos. El resultado es una matriz temporal, donde cada fila corresponde a un fonema del inventario fonético y cada columna representa una ventana temporal llamada *frame*. Los valores en esta matriz indican la probabilidad de activación de cada fonema en cada instante, incluyendo un fonema especial ($/SIL/$) que representa silencios.

La Figura 2.4 muestra un ejemplo de posteriograma normalizado correspondiente a una oración en inglés. Esta representación permite observar claramente cómo se distribuyen los fonemas a lo largo del tiempo, lo cual resulta esencial para capturar el ritmo, la duración y la fluidez del habla —factores clave para estimar su velocidad.

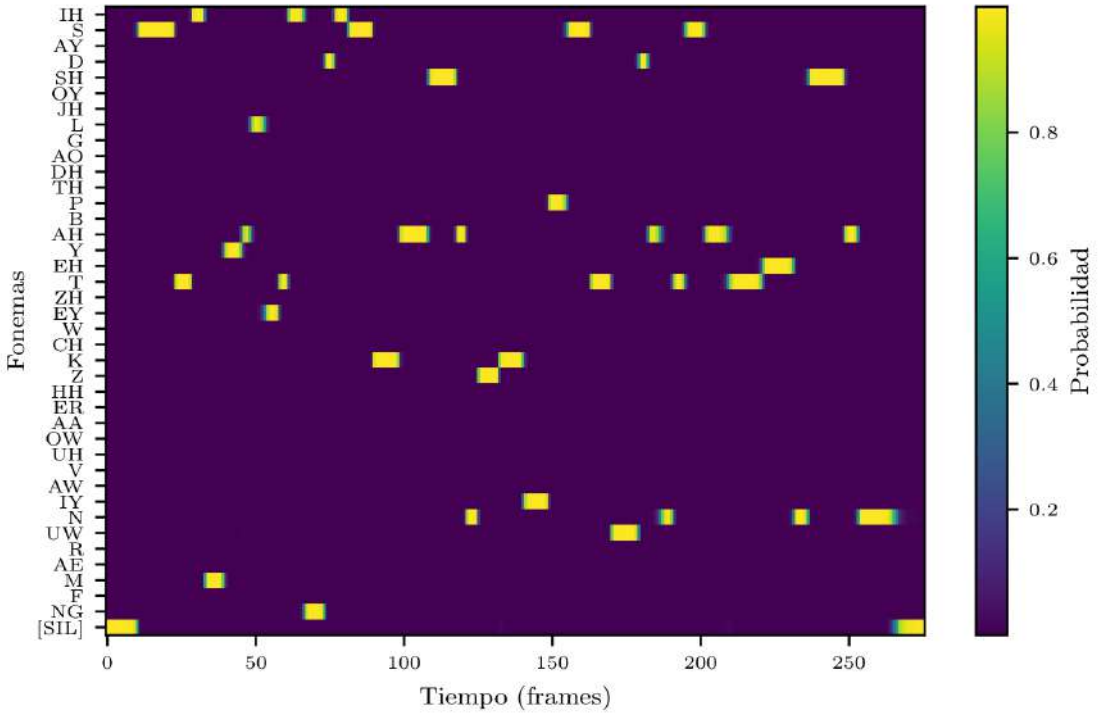

Fig. 2.4: Posteriograma normalizado correspondiente a la oración "Stimulating discussions keep students' attention". La escala de colores representa la probabilidad de activación de cada fonema a lo largo del tiempo.

A diferencia de representaciones como los espectrogramas, que codifican información espectral de bajo nivel, los posteriogramas permiten trabajar directamente con unidades lingüísticas como los fonemas. Esto los convierte en una representación mucho más alineada con el objetivo de estimar la velocidad del habla en términos de fonemas o sílabas por segundo.

2.3. Medición de la velocidad del habla

Existen distintas formas de cuantificar la velocidad del habla. En este trabajo el foco está puesto en la **velocidad media**, definida como la cantidad de unidades fonéticas (fonemas o sílabas) pronunciadas por segundo a lo largo de un segmento de audio. Esta medida permite capturar de forma global el ritmo de una locución.

Sea a un audio con duración total $T(a)$, y $N(a)$ la cantidad de fonemas (o sílabas) pronunciadas en ese fragmento. La velocidad media del habla se calcula entonces como:

$$\bar{v}(a) = \frac{N(a)}{T(a)}$$

Este enfoque tiene la ventaja de ser simple, interpretable y robusto frente a variaciones locales en la señal. En contextos educativos o clínicos, donde interesa evaluar de forma

general el ritmo del habla de una persona, esta medida global resulta particularmente informativa.

Cabe mencionar que existen otras formas más detalladas de medir el ritmo del habla, como la **velocidad instantánea**, que estima cuántas unidades de habla se producen en intervalos temporales acotados, proporcionando una visión más dinámica. Sin embargo, este trabajo está centrado exclusivamente en la velocidad media como variable objetivo para modelar.

2.4. Modelos de regresión

Como primer enfoque para la estimación de la velocidad del habla, se optó por utilizar modelos de regresión lineal como técnica base. Estos modelos permiten establecer un benchmark inicial sobre el cual comparar enfoques más complejos, como las redes neuronales. La regresión lineal ofrece una interpretación clara de las relaciones entre variables, facilitando el análisis de qué características extraídas del habla contribuyen en mayor medida a la predicción.

La idea fundamental consiste en modelar la velocidad del habla como una combinación lineal de un conjunto de atributos acústicos extraídos a partir de posteriogramas. Para ello, se planteó el problema como un caso de aprendizaje supervisado, donde las muestras de entrada son características obtenidas desde las representaciones acústicas y las etiquetas son las velocidades de habla previamente calculadas en el preprocesamiento (ya sea en fonemas o sílabas por segundo).

El modelo de regresión lineal utilizado se expresa mediante la siguiente ecuación:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \varepsilon$$

donde:

- y representa la variable dependiente, es decir, la velocidad del habla.
- x_i son las variables predictoras (features) calculadas a partir de los posteriogramas.
- β_i son los coeficientes del modelo, con la restricción $\beta_i \geq 0$ para todos $i = 1, 2, \ldots, n$, lo que asegura que cada atributo tenga un efecto positivo o nulo sobre la variable de salida.
- ε es el término de error, modelado como una variable aleatoria con esperanza nula ($E[\varepsilon] = 0$) y varianza finita, que representa el componente de variación no explicado por el modelo.

Para este trabajo, se empleó regresión lineal con restricción de positividad en los coeficientes (también conocida como *regresión lineal positiva*), permitiendo así interpretar cada atributo como una contribución directa, nunca reductiva, a la estimación de la velocidad.

Si bien los modelos de regresión lineal presentan limitaciones al modelar relaciones no lineales, su implementación sirve como baseline sólido sobre el cual evaluar el aporte de técnicas más avanzadas, como las redes neuronales convolucionales que se exploran en la sección siguiente.

2.5. Redes neuronales convolucionales

Las **redes neuronales convolucionales** (CNN, por sus siglas en inglés) son un tipo de arquitectura de redes profundas que han demostrado una enorme efectividad en el procesamiento de datos con estructura espacial o temporal, como imágenes, video o señales de audio. Inicialmente introducidas para reconocimiento de patrones en imágenes [21], estas redes han sido adaptadas y ampliamente utilizadas en tareas de reconocimiento de habla, clasificación de sonidos y análisis de series temporales.

Estructura básica de una CNN

Una CNN se compone de capas convolucionales que aprenden filtros capaces de detectar patrones locales en los datos. En el caso de señales de audio, estos patrones pueden corresponder a transiciones acústicas, presencia de determinados fonemas, o incluso características prosódicas.

En su forma más simple, una CNN incluye:

- **Capas convolucionales**: aplican filtros sobre fragmentos locales de la señal (ej: ventanas temporales).
- **Funciones de activación no lineales**: como ReLU o Tanh, que introducen aspectos no lineales en el modelo.
- **Operaciones de pooling**: reducen la dimensionalidad y ayudan a la red a capturar invariancias.
- **Capas densas (fully connected)**: transforman los mapas de activación en predicciones finales.

Aplicaciones en señales de audio

El uso de CNNs en procesamiento de audio ha sido especialmente relevante gracias a su capacidad para obtener características jerárquicas directamente de representaciones como espectrogramas o posteriogramas. A diferencia de los métodos tradicionales basados en la extracción manual de características (por ejemplo, detección de picos o estadísticas globales), las CNNs permiten el aprendizaje automático de representaciones relevantes directamente desde los datos.

En este trabajo, las CNNs fueron utilizadas para predecir la velocidad del habla a partir de los posteriogramas. Esta tarea requiere modelar patrones acústicos distribuidos en el tiempo, por lo que las CNNs son particularmente adecuadas debido a:

- su capacidad para capturar dependencias locales en el tiempo;
- su eficiencia computacional, especialmente en arquitecturas 1D y 2D reducidas;
- su robustez frente a variaciones en la longitud de las señales.

Variantes exploradas en esta tesis

En el desarrollo experimental de esta tesis se implementaron y compararon diferentes variantes de CNNs:

- **CNNs 1D**: utilizadas sobre series temporales fonéticas (como [canales, tiempo]), agregando capas convolucionales seguidas de pooling y capas lineales.
- **CNNs 2D**: aplicadas a representaciones tipo imagen (como [canales, fonemas, tiempo]), inspiradas en el tratamiento de estadísticos en la regresión lineal.
- **Arquitecturas con pooling explícito**: como *TemporalPooling*, que únicamente realiza un promedio sobre el tiempo antes de la capa fully connected y posterior predicción.

Motivación de su uso

El uso de CNNs en este contexto está motivado por su eficacia en aprender representaciones discriminativas desde datos crudos, sin necesidad de segmentaciones manuales. A diferencia de los modelos lineales utilizados como baseline, las CNNs permiten capturar relaciones no lineales complejas entre las secuencias de fonemas y las medidas de velocidad del habla. Esto las convierte en una herramienta poderosa y flexible para tareas de predicción en señales acústicas.

Este enfoque se apoya en diversos trabajos previos que demostraron el éxito de las CNNs en tareas de detección de eventos acústicos y reconocimiento de habla, como los presentados por Abdel-Hamid et al. [22], y Sainath et al. [23].

2.6. Métricas a utilizar

A la hora de evaluar la calidad de los modelos desarrollados para predecir la velocidad del habla, se utilizaron diversas métricas estándar en problemas de regresión. Estas métricas permiten cuantificar la diferencia entre los valores predichos por los modelos y los valores reales medidos.

Error cuadrático medio (MSE - Mean Squared Error)

El MSE fue la métrica utilizada para entrenar tanto los modelos de regresión como las redes neuronales convolucionales, buscando en ambos casos minimizarla. Penaliza fuertemente los errores grandes debido al término cuadrático, lo cual la vuelve útil en contextos donde es importante minimizar desviaciones grandes. Sin embargo, su sensibilidad a outliers puede ser una desventaja si los datos presentan valores atípicos. Está definido por:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Coeficiente de determinación (R^2)

El R^2 mide la proporción de la varianza total de la variable dependiente que es explicada por el modelo. Toma valores entre 0 y 1, donde 1 indica una predicción perfecta, mientras que un valor cercano a 0 indica un modelo que predice únicamente la media. Una de sus ventajas es que tiene una interpretación intuitiva, pero puede ser engañosa en modelos con pocos datos o con overfitting. Se define como:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

Utiliza el promedio real $\bar{y}$ como valor de referencia.

Coeficiente de determinación ajustado (R^2_{adj})

El coeficiente de determinación ajustado, denotado como R^2_{adj}, se define como:

$$R^2_{\text{adj}} = 1 - \left(\frac{(1 - R^2)(n-1)}{n - p - 1}\right)$$

donde n es el número de observaciones, p la cantidad de predictores en el modelo e $\bar{y}$ el promedio de los valores reales.

A diferencia del R^2 tradicional, que siempre tiende a aumentar con la incorporación de nuevas variables, el R^2_{adj} penaliza la complejidad del modelo. Sólo aumenta si los nuevos predictores mejoran efectivamente la capacidad explicativa, lo cual lo convierte en una métrica más conservadora y adecuada para evitar sobreajuste, especialmente en contextos de alta dimensionalidad.

Coeficiente de correlación de Pearson (ρ)

El coeficiente de correlación de Pearson, utiliza el promedio de los valores reales $\bar{y}$, y está definido como:

$$\rho = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}\sqrt{\sum_{i=1}^{n}(\hat{y}_i - \bar{\hat{y}})^2}}$$

Esta métrica evalúa la relación lineal entre las predicciones y los valores reales. A diferencia de MSE, no considera la escala absoluta de los errores, sino la tendencia conjunta entre ambas variables. Una correlación alta no garantiza necesariamente una predicción precisa en valor absoluto, pero sí indica una buena alineación.

Coeficiente de Concordancia (CCC)

El CCC combina la correlación lineal de Pearson con la concordancia en la media y la varianza entre las predicciones y los valores reales. En esta expresión, ρ representa el coeficiente de Pearson entre y y $\hat{y}$, mientras que σ_y, $\sigma_{\hat{y}}$ son las desviaciones estándar, y μ_y, $\mu_{\hat{y}}$ las medias de las variables reales y predichas, respectivamente.

$$\rho_c = \frac{2\rho\sigma_y\sigma_{\hat{y}}}{\sigma_y^2 + \sigma_{\hat{y}}^2 + (\mu_y - \mu_{\hat{y}})^2}$$

A diferencia del R^2 o el MSE, el CCC penaliza tanto los desplazamientos sistemáticos como las diferencias en escala entre ambas distribuciones. Es particularmente útil cuando se

desea una coincidencia no sólo en la tendencia, sino también en la localización y dispersión de las predicciones.

Comparación general

Cada métrica utilizada aporta una perspectiva complementaria sobre el rendimiento de los modelos. El error cuadrático medio (MSE) ofrece una visión sobre la magnitud absoluta del error, penalizando con mayor severidad las desviaciones grandes. El coeficiente de determinación clásico (R^2) mide la proporción de la varianza explicada por el modelo, aunque tiende a aumentar con la inclusión de predictores adicionales, incluso si estos no aportan valor explicativo real. Por este motivo, se incorporó también el coeficiente de determinación ajustado (R^2_{adj}), que introduce una penalización basada en la cantidad de predictores, brindando una evaluación más conservadora y robusta en contextos donde se exploran múltiples atributos.

A su vez, el coeficiente de correlación de concordancia (CCC) cuantifica la similitud estructural entre las predicciones y los valores reales, evaluando tanto la correlación como la cercanía a la identidad. Finalmente, el coeficiente de Pearson refleja la fuerza y dirección de la relación lineal entre las variables, aunque puede resultar insensible a errores sistemáticos o sesgos en escala.

Dado que cada métrica captura diferentes aspectos del desempeño, en este trabajo se adoptó una estrategia de evaluación múltiple, considerando todas ellas de forma complementaria. Esto permite obtener una visión más completa y equilibrada del rendimiento de los modelos, tanto en términos de precisión absoluta como de capacidad explicativa y concordancia estructural.

3. ANÁLISIS EXPLORATORIO DE DATOS

3.1. Descripción del dataset

Para la evaluación de los modelos, se utilizará la base de datos **TIMIT** [24], un corpus de **habla inglesa** ampliamente reconocido en la investigación del procesamiento automático del habla. Este conjunto fue desarrollado mediante una colaboración entre el *Massachusetts Institute of Technology* (MIT), *SRI International* y *Texas Instruments* (TI), con el objetivo de proporcionar datos estandarizados para el entrenamiento y evaluación de sistemas de reconocimiento de voz.

El corpus contiene 6300 grabaciones de audio correspondientes a 630 hablantes provenientes de ocho regiones dialectales distintas de los Estados Unidos. La distribución de género incluye aproximadamente un 70 % de hablantes masculinos y un 30 % femeninos. Cada hablante aportó 10 frases: dos frases de calibración acústica, cinco oraciones fonéticamente ricas, y tres seleccionadas aleatoriamente. En total, el corpus abarca 2342 frases únicas, organizadas para capturar variabilidad fonética, dialectal y contextual.

Una característica distintiva de TIMIT es su **anotación fonética detallada**, que incluye marcas de tiempo precisas para fonemas, palabras, pausas internas y silencios iniciales o finales. Además, el corpus proporciona metadatos lingüísticos y demográficos sobre los hablantes. Estas anotaciones fueron validadas manualmente por expertos, lo cual permite utilizar el corpus como referencia confiable para estudios que analizan características temporales del habla.

Esta granularidad en las anotaciones convierte a TIMIT en un recurso ideal para tareas como la estimación de la *velocidad del habla*, ya que permite calcular con precisión duraciones fonéticas, identificar pausas e inferir ritmo y fluidez. Por estas razones, TIMIT es frecuentemente utilizado como benchmark en investigaciones que emplean técnicas de aprendizaje automático para modelar las dinámicas temporales del habla.

Nota: Para todos los análisis y experimentos realizados en esta tesis se utilizaron únicamente los subconjuntos de `train` y `dev` del corpus TIMIT. El conjunto `test` se reservó para las evaluaciones finales y no fue empleado durante el proceso de entrenamiento, validación ni análisis exploratorio.

3.1.1. Tipos de oraciones en TIMIT

El corpus TIMIT está compuesto por tres tipos principales de oraciones, con diferentes propiedades fonéticas y propósitos experimentales:

- **Oraciones SA (Dialect Sentences):** Son dos frases fijas (SA1 y SA2) que todos los hablantes del corpus deben leer. Tienen un contenido fonético idéntico y están diseñadas para facilitar comparaciones directas entre hablantes, ya que controlan completamente la variabilidad lingüística.
- **Oraciones SX (Phonetically Compact Sentences):** Estas oraciones son seleccionadas de un subconjunto diseñado para maximizar la cobertura fonética con una cantidad reducida de palabras. Cada hablante lee un conjunto distinto de frases SX, lo que introduce cierta variabilidad estructural.

- **Oraciones SI (Phonetically Diverse Sentences):** Son frases únicas asignadas individualmente a cada hablante, tomadas del conjunto completo del corpus. Su objetivo es aumentar la diversidad fonética global del conjunto de datos.

Esta clasificación es especialmente relevante para el análisis exploratorio de la velocidad del habla. En particular, las oraciones `SA`, al ser comunes a todos los hablantes, permiten examinar variaciones de velocidad eliminando el efecto del contenido lingüístico. Esta propiedad será aprovechada en el análisis posterior (ver Figura 3.6) para evaluar hasta qué punto la estructura del enunciado contribuye a la variabilidad observada en las métricas de velocidad.

3.2. Correlaciones y tendencias

Uno de los objetivos principales del análisis exploratorio de datos fue comprender cómo se comportan las medidas de velocidad del habla —fonemas por segundo (FPS) y sílabas por segundo (SPS)— a lo largo del corpus, y cómo se relacionan entre sí y con otras variables lingüísticas o características del enunciado.

Antes de profundizar en la caracterización de FPS y SPS, se analizó el posible impacto de las pausas en su medición. Para ello, se compararon las velocidades calculadas considerando los segmentos silenciosos y sin tenerlos en cuenta.

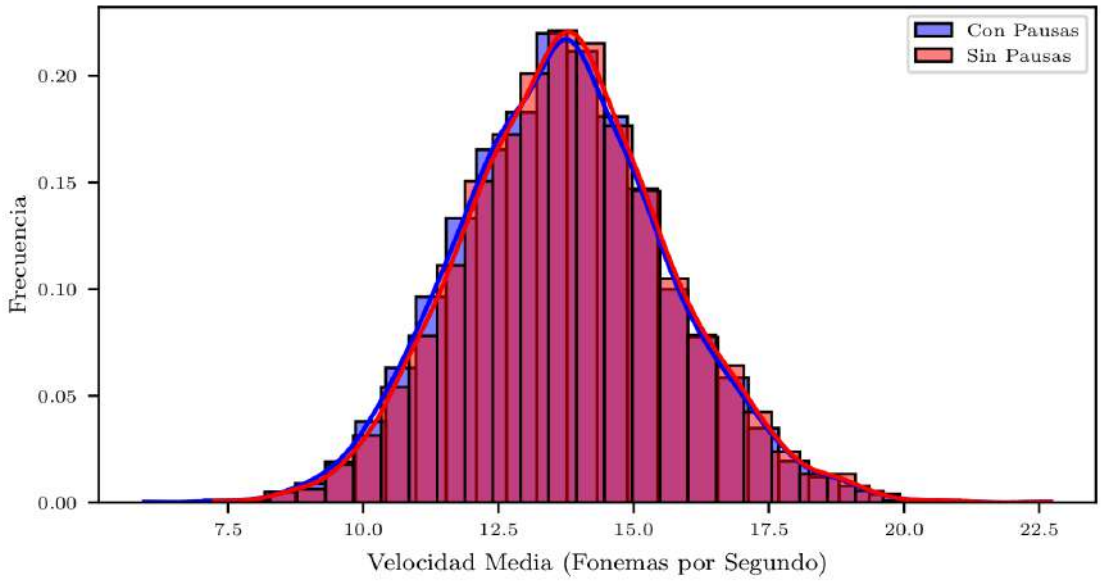

Fig. 3.1: Comparación entre velocidades con y sin pausas para FPS.

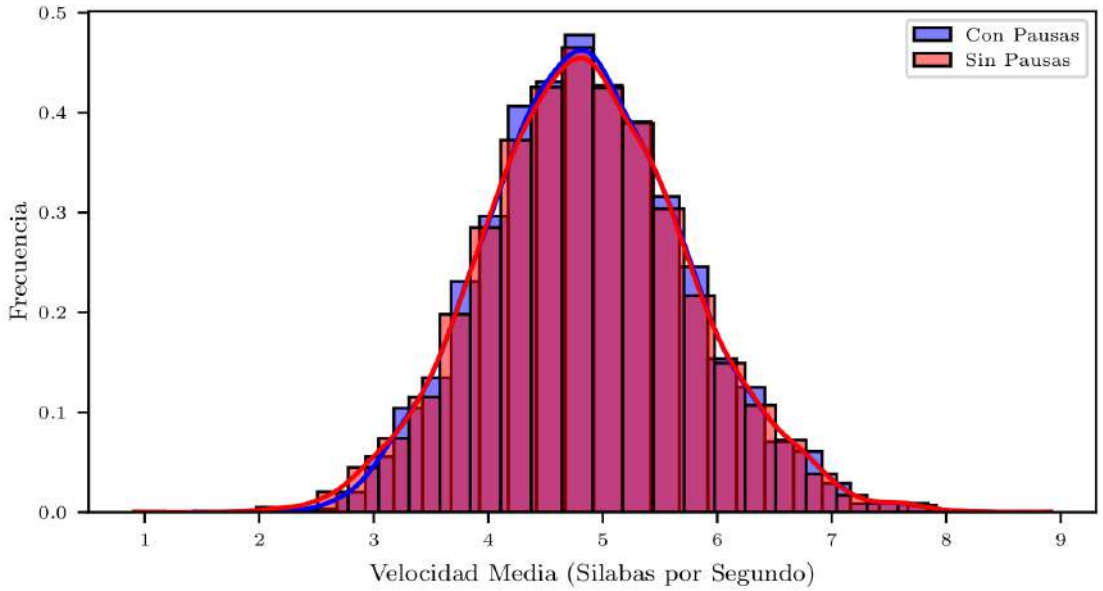

Fig. 3.2: Comparación entre velocidades con y sin pausas para SPS.

Como se muestra en las Figuras 3.1 y 3.2, las diferencias son mínimas. Esto se debe a las características del corpus TIMIT, donde las oraciones son breves y pronunciadas de forma controlada. La duración típica de los audios ronda los tres segundos, y las pausas detectadas suelen ser inferiores a una décima de segundo, como se nota en la Figura 3.3.

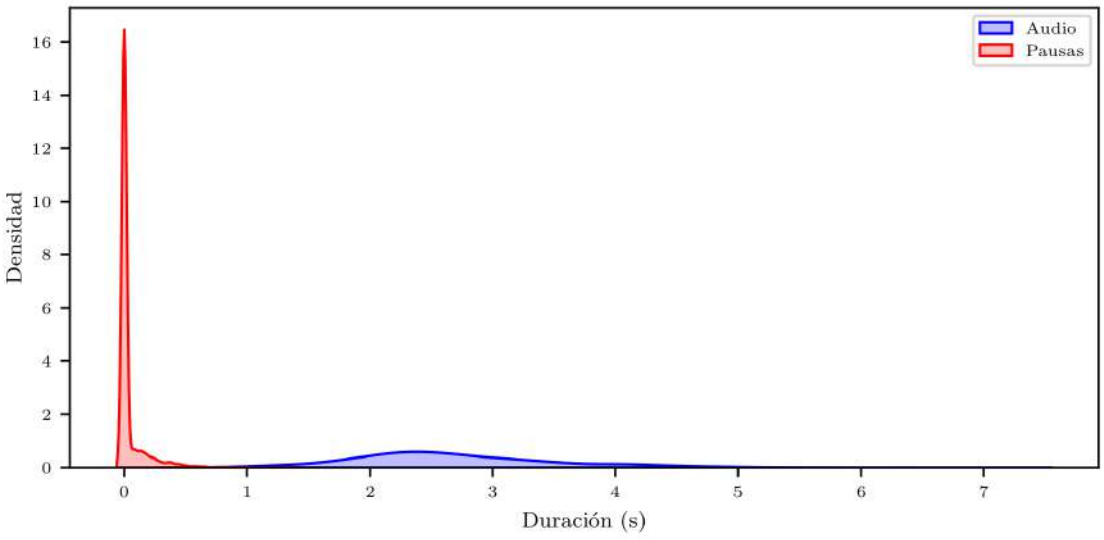

Fig. 3.3: Distribución general de duración de los audios y de las pausas, en segundos.

Dado que en contextos reales como la educación o la salud los silencios pueden ser informativos, se optó por trabajar con la velocidad total —incluyendo pausas— en todo

el análisis posterior, aunque en este corpus no haya grandes diferencias.

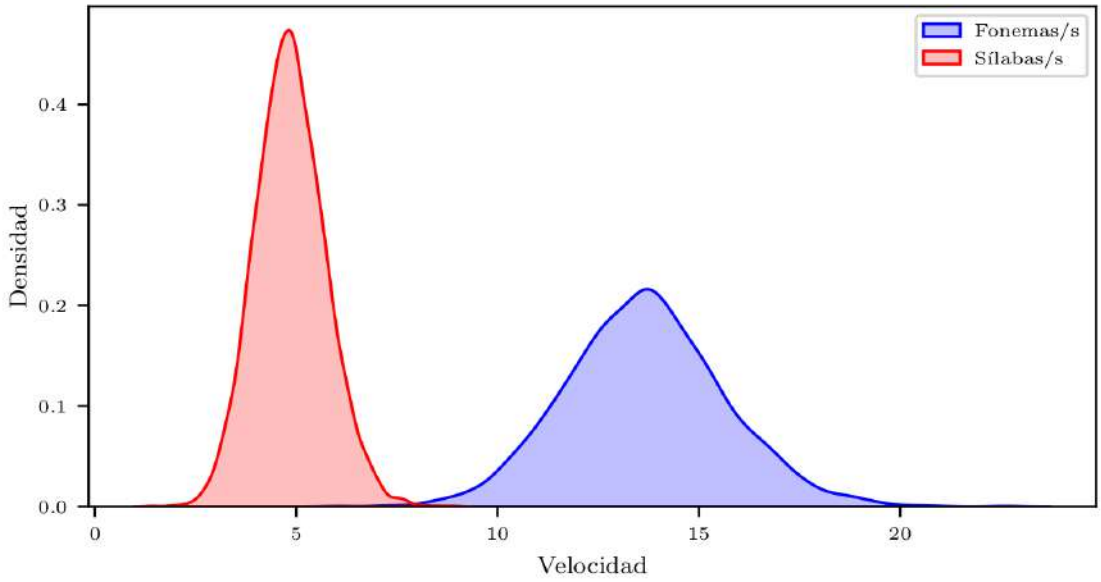

Fig. 3.4: Distribución general de la velocidad del habla medida en fonemas por segundo (FPS) y sílabas por segundo (SPS).

La Figura 3.4 muestra que ambas métricas presentan distribuciones unimodales, con FPS desplazado hacia valores más altos, como es esperable dado que una sílaba puede contener múltiples fonemas. Además, se observa que FPS tiene una mayor dispersión que SPS, lo cual sugiere una mayor sensibilidad a variaciones fonéticas locales y anticipa que los modelos tendrán más dificultad para predecir esta métrica (por ejemplo, en términos de error cuadrático medio).

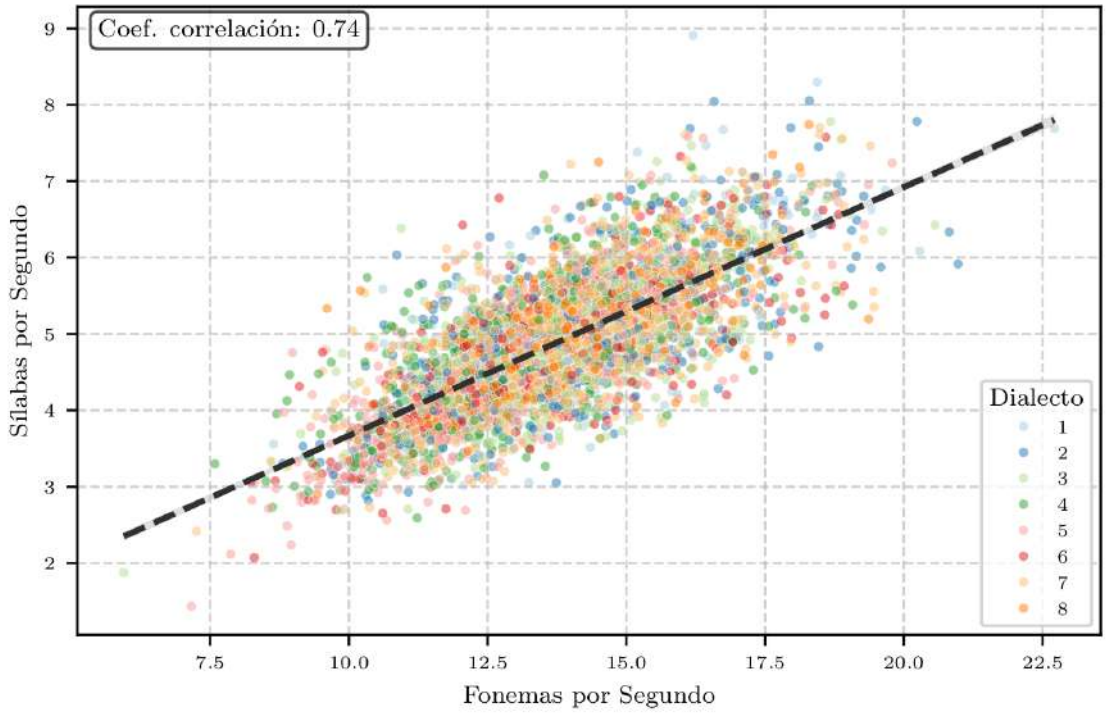

Fig. 3.5: Relación entre fonemas por segundo (FPS) y sílabas por segundo (SPS), coloreado por dialecto.

Como muestra la Figura 3.5, existe una fuerte relación lineal entre FPS y SPS. Esta correlación positiva confirma que ambas métricas reflejan un mismo fenómeno subyacente. Sin embargo, la pendiente de la recta es distinta de uno, lo que sugiere que la cantidad promedio de fonemas por sílaba varía entre hablantes, introduciendo cierta variabilidad en la relación entre ambas medidas.

Los puntos en la figura están coloreados según el dialecto del hablante, lo que permite explorar visualmente si existen patrones regionales asociados a la velocidad del habla. Aunque no se observan agrupamientos claramente diferenciados, algunas regiones parecen ocupar zonas ligeramente desplazadas, lo que podría indicar efectos dialectales sutiles. Sin embargo, estas diferencias no son suficientemente marcadas como para extraer conclusiones sólidas, por lo que se requeriría un análisis más detallado para poder obtener distinciones concretas.

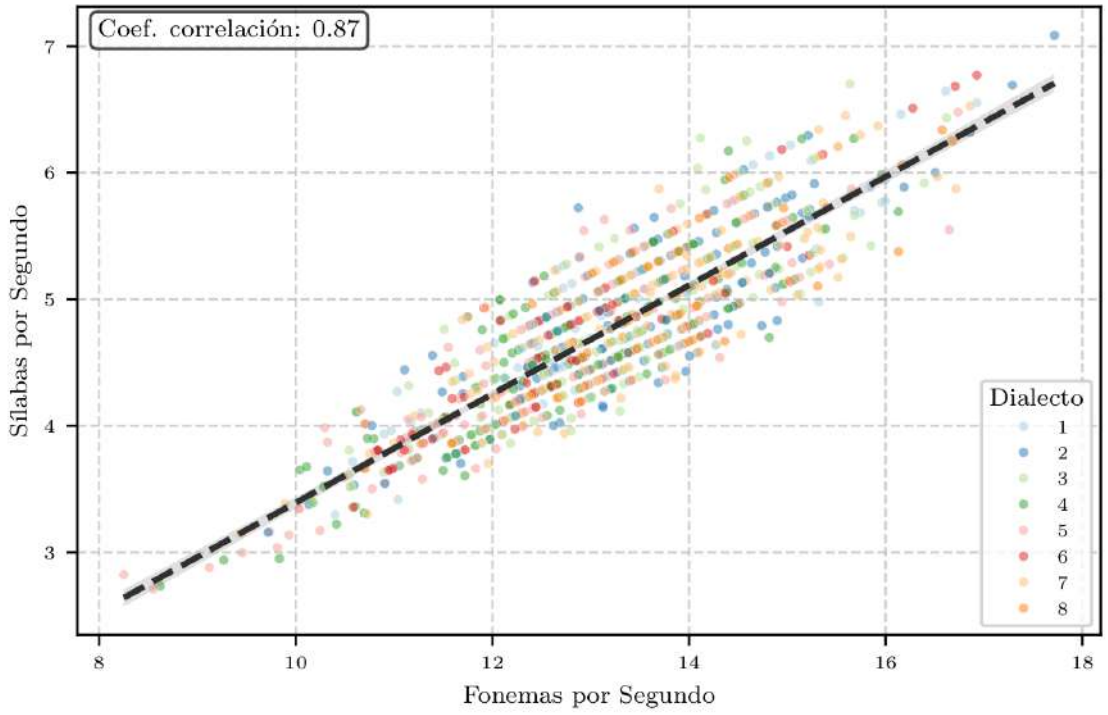

Fig. 3.6: Relación entre FPS y SPS para oraciones tipo `SA`, coloreado por dialecto.

En la Figura 3.6 se restringe el análisis a las oraciones `SA`, que son compartidas por todos los hablantes y tienen un contenido fonético idéntico. Esta visualización permite observar que, al controlar el contenido lingüístico, la variabilidad entre hablantes se reduce. El agrupamiento más compacto de los puntos sugiere que estas oraciones podrían ser útiles como referencia o control en estudios de velocidad del habla, y que parte de la variabilidad global observada en el corpus podría deberse a diferencias estructurales entre oraciones.

En conjunto, este análisis exploratorio respalda el uso de FPS y SPS como variables objetivo relevantes, y sugiere que características como la presencia de pausas, la estructura del enunciado y la variación interhablante pueden influir en la velocidad del habla. Si bien no se identificaron patrones categóricos fuertes asociados a dialectos o tipos de oración, las tendencias observadas permiten delinear hipótesis y orientar el diseño posterior de los modelos predictivos.

4. METODOLOGÍA

4.1. Obtención de variables objetivo

El dataset original TIMIT no incluye información explícita sobre la velocidad del habla. Por este motivo, fue necesario generar manualmente las columnas objetivo (targets) que luego se utilizaron para entrenar los modelos supervisados. Se definieron dos métricas principales: *fonemas por segundo* (FPS) y *sílabas por segundo* (SPS).

Fonemas por segundo (FPS).

La base TIMIT incluye alineamientos fonéticos temporales por cada audio, lo que permitió contar con relativa facilidad la cantidad de fonemas por segmento. A partir de esto, la métrica se definió de forma directa como:

$$FPS(a) = \frac{N_f(a)}{T(a)}$$

donde N_f representa el número total de fonemas en el audio, y T su duración total en segundos.

Sílabas por segundo (SPS).

La estimación de la cantidad de sílabas no estaba presente en el dataset y requirió una etapa de procesamiento adicional. Para esto, se utilizaron los archivos `.WRD` del corpus, que contienen la transcripción ortográfica de cada oración. Se desarrolló un script que recorre cada oración palabra por palabra y estima la cantidad de sílabas utilizando dos métodos complementarios:

- Se consultó el diccionario fonético de CMU[1] (Carnegie Mellon Pronouncing Dictionary), que provee para cada palabra su representación fonética, a partir de la cual es posible contar las sílabas según los núcleos vocálicos.
- En los casos en que la palabra no se encontraba en CMU, se utilizó la librería `syllapy`[2] como método de respaldo.

A continuación, se sumaron las sílabas por oración y se dividió por su duración total en segundos:

$$SPS(a) = \frac{N_s(a)}{T(a)}$$

[1] http://www.speech.cs.cmu.edu/cgi-bin/cmudict
[2] https://pypi.org/project/syllapy/

4.2. División de conjuntos

El corpus TIMIT utilizado en este trabajo ya provee una partición oficial entre conjuntos de *entrenamiento* y *test*, lo que permite una evaluación estandarizada y reproducible. Siguiendo esta división, todas las decisiones experimentales respetan esta separación, evitando cualquier tipo de filtración de información entre etapas.

Además, para evaluar de manera rigurosa tanto los modelos lineales como las redes neuronales, se aplicaron estrategias de validación diferentes pero complementarias, adaptadas a las necesidades particulares de cada enfoque.

Regresiones lineales: validación cruzada por hablante

Para los modelos de regresión lineal, se implementó una validación cruzada de 6 folds (*6-fold cross-validation*) sobre el conjunto de entrenamiento oficial. Esta partición se realizó a nivel de hablante, asegurando que ningún locutor aparezca en más de un fold. De esta manera, se evita el *data leakage* —una situación en la que el modelo tiene acceso durante el entrenamiento a información del conjunto de validación o test, ya sea de forma directa o indirecta—, lo cual puede conducir a una sobreestimación ficticia del desempeño. Al evitar esta fuga de información, se garantiza que la evaluación refleje con mayor fidelidad la capacidad del modelo para generalizar a nuevos hablantes.

Este esquema de validación se utilizó para seleccionar un hiperparámetro fundamental: el umbral de probabilidad a partir del cual se considera que un fonema ha sido realmente pronunciado (ver Sección 4.3). Este umbral afecta el cálculo de las estadísticas de entrada para la regresión y, por lo tanto, tiene impacto directo en la precisión del modelo. Para ello, se evaluaron múltiples valores posibles y se seleccionó aquel que generó el menor error promedio entre folds.

Durante cada iteración, se entrenó el modelo (con un valor fijo de umbral) sobre cinco sextos de los datos y se lo evaluó, con la métrica de entrenamiento (MSE) sobre el fold restante. Finalmente, los resultados se promediaron a lo largo de los folds.

Redes neuronales: partición explícita de validación

En el caso de los modelos neuronales, se utilizó una división explícita del conjunto de entrenamiento en dos subconjuntos: uno para entrenamiento y otro para validación. Esta división también se realizó por hablante, manteniendo la independencia total entre los subconjuntos.

El conjunto de validación se utilizó para implementar *early stopping*, seleccionando el modelo con mejor desempeño en validación y evitando el sobreajuste. Finalmente, el conjunto de test original provisto por TIMIT se mantuvo reservado y fue utilizado exclusivamente para evaluar el rendimiento final de los modelos una vez completado el entrenamiento.

Este diseño permite comparar de manera justa y controlada la capacidad predictiva de los modelos lineales y los basados en redes neuronales, asegurando tanto la robustez como la validez experimental de los resultados presentados.

4.3. Extracción de features

En esta etapa, se describen los atributos utilizados como entrada para los modelos de regresión lineal, los cuales fueron calculados directamente a partir de los posteriogramas sin normalizar, denotados P, obtenidos mediante el alineador Charsiu. Dichos atributos intentan capturar características relevantes de la señal de habla, como patrones de activación fonética o dinámica temporal, sin necesidad de transcripciones explícitas.

Los atributos seleccionados fueron inspirados y adaptados de la tesis de Palazzo [15], priorizando aquellos que demostraron mejor capacidad predictiva en su estudio, manteniendo así la línea de trabajo del laboratorio y coherencia con investigaciones anteriores. A continuación se detallan los utilizados:

Estadísticas por fonema.

Se computaron medidas de promedio y variabilidad sobre cada curva fonética $p_i(t)$, correspondiente a la probabilidad del fonema i en el tiempo. En particular, denotando para cada fonema, $p_i(t)$ la señal original, $\frac{dp_i}{dt}$ la derivada temporal de primer orden y $\frac{d^2p_i}{dt^2}$ la derivada temporal de segundo orden. Se consideraron las siguientes estadísticas:

- $\overline{p_i}$, $\overline{\frac{dp_i}{dt}}$, $\overline{\frac{d^2p_i}{dt^2}}$: medias de cada señal.
- $\sigma(p_i)$, $\sigma\left(\frac{dp_i}{dt}\right)$, $\sigma\left(\frac{d^2p_i}{dt^2}\right)$: desvíos estándar de cada una.

El conjunto de *features* se puede describir como:

$$B = \bigcup_{i=1}^{|P|} \left\{ \overline{p_i}, \overline{\frac{dp_i}{dt}}, \overline{\frac{d^2p_i}{dt^2}}, \sigma(p_i), \sigma\left(\frac{dp_i}{dt}\right), \sigma\left(\frac{d^2p_i}{dt^2}\right) \right\}$$

Fonemas probables.

Otra fuente de atributos fue el conteo de activaciones fonéticas con alta probabilidad. Se computó la cantidad de veces que cada fonema alcanza una probabilidad superior a un cierto umbral, considerando únicamente los inicios de activación —es decir, el momento en el que la probabilidad supera el umbral por primera vez—.

Formalmente, definimos el siguiente conjunto de características:

$$C = \bigcup_{i=0}^{|P|} g(p_i), \quad \text{donde} \quad g(p_i) = \frac{1}{T} \sum_{t=1}^{T} \text{máx}\left(0, \frac{d\left(\mathbb{1}\left\{\text{Softmax}(P)_{t,i} > \theta\right\}\right)}{dt}\right),$$

donde:

- T es la cantidad de frames temporales;
- θ representa el umbral de probabilidad;
- $\mathbb{1}\{\cdot\}$ es la función indicadora que vale 1 si la condición se cumple y 0 en caso contrario;
- La derivada temporal d/dt se utiliza para detectar las transiciones de 0 a 1 (activaciones);

- La función máx$(0, \cdot)$ descarta las caídas y se enfoca en los momentos de encendido de la probabilidad fonética.

El objetivo de esta fórmula es capturar el número de veces que la activación del fonema p_i supera el umbral de manera significativa en el tiempo.

El valor del umbral θ fue determinado mediante validación cruzada con 6 particiones, evaluando un rango de valores entre 0.05 y 0.95 en pasos de 0.05, como fue explicado en la Sección 4.2.

4.4. Uso de SylNet como sistema de referencia

Con el objetivo de contar con una referencia externa en la tarea de estimación de velocidad del habla en sílabas por segundo (SPS), se utilizó el sistema **SylNet** [25]. Este modelo, desarrollado para la detección automática de sílabas en señal de audio, fue empleado en esta tesis exclusivamente como contador de sílabas, sin modificar su estructura ni analizar internamente sus activaciones o salidas probabilísticas.

Cálculo de la velocidad del habla con SylNet

Para cada archivo de audio, se utilizó SylNet para estimar la cantidad total de sílabas presentes. Esta estimación se dividió luego por la duración total del audio en segundos, generando así una medida de **sílabas por segundo**. Esta métrica fue utilizada para evaluar el rendimiento del sistema externo frente a los modelos entrenados específicamente para esta tesis.

Resultados

La Tabla 4.1 muestra los resultados obtenidos por SylNet en el conjunto de test, empleando las métricas estándar de evaluación utilizadas en todo el trabajo: error cuadrático medio (MSE), coeficiente de determinación (R^2), coeficiente de correlación de concordancia (CCC) y correlación de Pearson.

Modelo	**MSE**	R^2	**CCC**	**Pearson**
SylNet	0.592	0.211	0.630	0.685

Tab. 4.1: Rendimiento de SylNet como baseline externo para la predicción de SPS.

Discusión

Si bien los resultados de SylNet son razonables como punto de partida, quedan muy por debajo del rendimiento alcanzado tanto por las regresiones lineales como por las redes convolucionales entrenadas en esta tesis, como se analiza en la Sección 5.7. Sin embargo, su inclusión permite contar con un **baseline externo automático**, independiente de los procesos desarrollados ad hoc, y refuerza la validez de las mejoras obtenidas por los modelos diseñados específicamente para esta tarea.

5. RESULTADOS

5.1. Baseline simple

Como punto de partida, se entrenaron modelos de regresión lineal restringidos a coeficientes positivos, utilizando únicamente estadísticas básicas por fonema, como la media, medias de las derivadas de primer y segundo orden, y desviaciones estándar, tal como fue descrito en la Sección 4.3. Esto dio lugar a vectores de entrada de dimensión 235.

Los modelos fueron entrenados utilizando mínimos cuadrados ordinarios, y se evaluaron dos variantes: una para predecir fonemas por segundo (FPS) y otra para sílabas por segundo (SPS). Para evaluar su rendimiento se utilizaron las métricas MSE, R^2, coeficiente de correlación concordante (CCC) y coeficiente de Pearson.

Target	**MSE**	R^2	**CCC**	**Pearson**
FPS	2.343	0.390	0.563	0.629
SPS	0.453	0.329	0.534	0.586

Tab. 5.1: Resultados del modelo de regresión lineal con estadísticas fonéticas básicas.

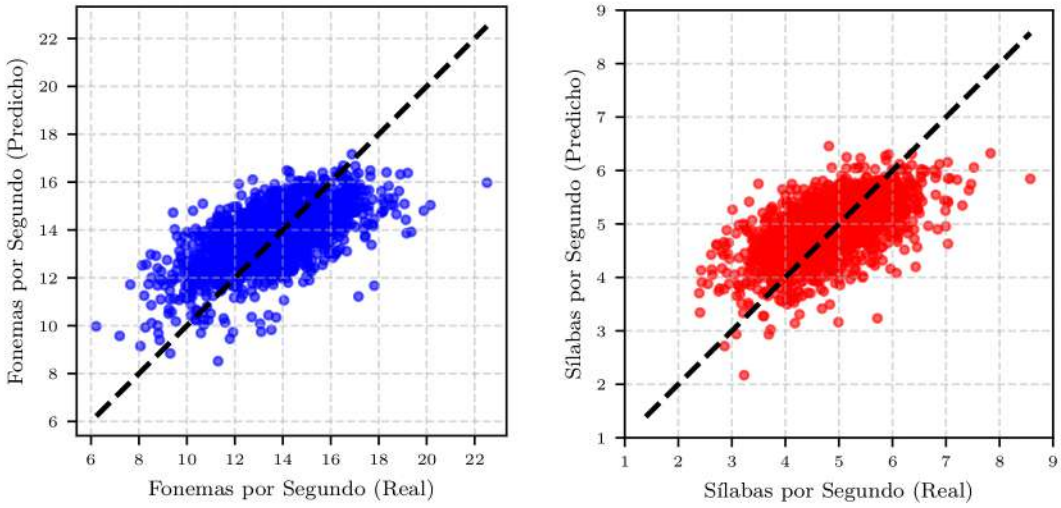

Fig. 5.1: Relación entre valores reales y predichos para FPS (izquierda) y SPS (derecha) usando el modelo lineal simple.

Los resultados observados en la Tabla 5.1 muestran que, a pesar de su simplicidad, estos modelos logran capturar una parte limitada de la variabilidad en la velocidad del habla. El desempeño es levemente mejor para la predicción de FPS, y sienta una línea base clara a superar por las estrategias más complejas desarrolladas en las siguientes secciones.

5.2. Incorporación de umbrales

Con el objetivo de capturar información más directamente relacionada con la presencia fonética efectiva en el tiempo, se propuso una nueva familia de atributos basada en el conteo de apariciones de fonemas con alta probabilidad en los posteriogramas. A diferencia de las estadísticas clásicas utilizadas en la Sección 5.1, esta estrategia incorpora información temporal discretizada sobre la ocurrencia de fonemas, filtrando únicamente aquellos cuya probabilidad supera cierto umbral.

Formalmente, se utilizó la salida del posteriograma $P \in \mathbb{R}^{T \times F}$, donde T representa los marcos temporales y $F = 39$ el número de fonemas. Luego de aplicar una función *softmax* en cada instante de tiempo, se contabilizó la cantidad de veces que cada fonema excedía un umbral de activación $u \in [0, 1]$, produciendo vectores de características de dimensión 40.

Para determinar el umbral óptimo, se exploró un rango de valores entre 0.05 y 0.95 con incrementos de 0.05, como fue explicado en las secciones 4.2 y 4.3. El valor seleccionado fue $u = 0.5$ para el modelo predictor de fonemas por segundo y $u = 0.45$ para el modelo predictor de sílabas por segundo.

Target	**MSE**	R^2	**CCC**	**Pearson**	**Umbral**
FPS	1.307	0.660	0.799	0.814	0.5
SPS	0.206	0.695	0.827	0.837	0.45

Tab. 5.2: Resultados del modelo de regresión lineal con atributos de conteo de fonemas con alta probabilidad.

Los resultados, presentados en la Tabla 5.2, muestran mejoras sistemáticas con respecto al baseline para ambos targets, en particular para SPS. El patrón de correlación se vuelve más pronunciado, como se puede observar en las Figuras 5.2, lo que sugiere que el patrón de activación discreta de fonemas es informativo para modelar la velocidad del habla.

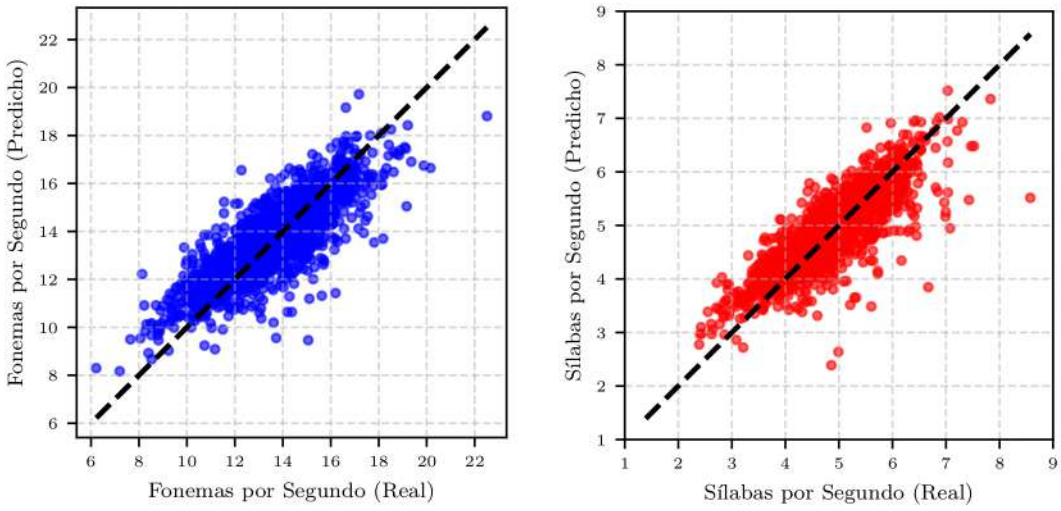

Fig. 5.2: Relación entre valores reales y predichos para FPS (izquierda) y SPS (derecha) usando atributos de conteo fonético con umbral óptimo.

Este resultado destaca el valor de incorporar información umbralizada proveniente de los posteriogramas, lo cual permite capturar mejor la dinámica temporal del habla sin recurrir a segmentaciones explícitas ni transcripciones.

5.3. Mezcla de atributos

En esta última etapa de experimentación con modelos lineales, se propuso una estrategia de combinación de atributos que integrara tanto las estadísticas fonéticas básicas como los conteos umbralizados de apariciones fonémicas descritos en las secciones anteriores. Esta mezcla tiene como objetivo aprovechar tanto la información continua derivada de la forma de las curvas de activación fonética como los patrones de aparición discretos en el tiempo.

El vector final de características resultante alcanza una dimensión de 274, siendo el más grande entre los experimentados hasta el momento. Esta mayor cantidad de atributos plantea el riesgo de sobreajuste, por lo que se reportaron tanto las métricas estándar como el coeficiente de determinación ajustado (R^2_{adj}), que penaliza la complejidad excesiva del modelo.

Los resultados se presentan en la Tabla 5.3, donde se observa que esta configuración alcanza el mejor rendimiento general en todos los targets, incluyendo mejoras en el error cuadrático medio (MSE), el coeficiente de correlación de concordancia (CCC) y la correlación de Pearson.

Target	**MSE**	R^2	**CCC**	**Pearson**
FPS	1.008	0.738	0.851	0.860
SPS	0.156	0.769	0.875	0.879

Tab. 5.3: Resultados del modelo de regresión lineal con combinación de atributos fonéticos continuos y discretos.

Particularmente, se visualiza en la Tabla 5.4 como el R^2_{adj} se mantiene superior al obtenido en configuraciones previas, justificando así el uso del conjunto ampliado de características.

Modelo	**FPS** (R^2_{adj})	**SPS** (R^2_{adj})
Regresión simple	0.358	0.293
Con umbral (probabilidad)	0.657	0.693
Combinado	**0.721**	**0.754**

Tab. 5.4: Comparación del coeficiente de determinación ajustado (R^2_{adj}) para los distintos modelos de regresión lineal.

A pesar del aumento en el número de parámetros, el modelo de combinación de atributos logra generalizar de forma más efectiva que las variantes anteriores, como podemos observar en las Figuras 5.3.

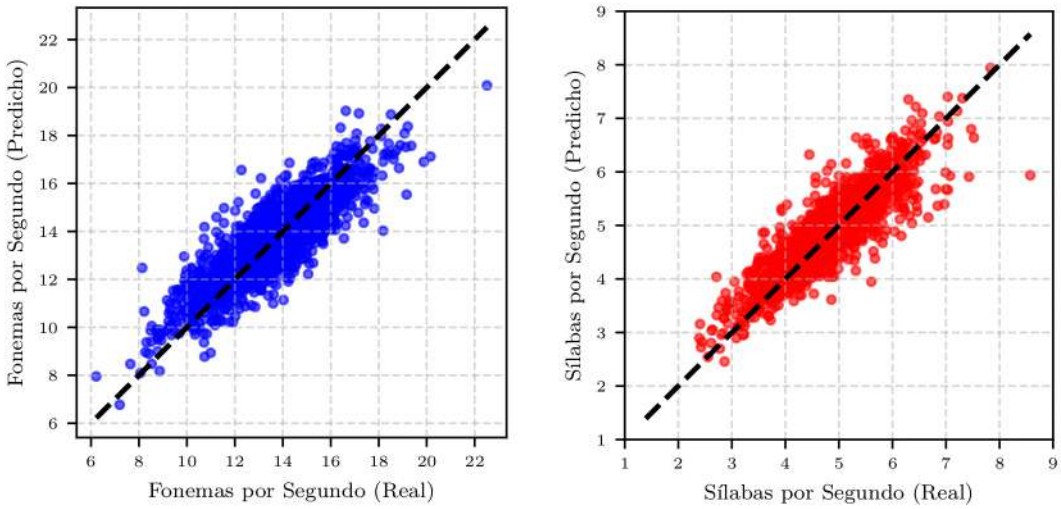

Fig. 5.3: Relación entre valores reales y predichos para FPS (izquierda) y SPS (derecha) con el modelo lineal combinado.

Esto sugiere que ambas fuentes de información (forma de activación y frecuencia de aparición) capturan aspectos complementarios del ritmo del habla y, cuando se integran, permiten una mejor estimación de la velocidad en fonemas y sílabas por segundo.

En resumen, la combinación de descriptores continuos y discretos no sólo incrementa la capacidad explicativa del modelo, sino que establece un nuevo umbral de rendimiento dentro de los enfoques lineales, sirviendo como base sólida de comparación frente a los modelos no lineales desarrollados en la siguiente sección.

5.4. Red simple con pooling

En la exploración de modelos neuronales, se propuso como punto de partida una arquitectura extremadamente simple basada en una operación de *pooling* global que resume la dimensión temporal de los posteriogramas. Esta arquitectura, denominada `TemporalPooling`, tiene como objetivo capturar el nivel general de activación fonética a lo largo del tiempo y generar una estimación de la velocidad del habla a partir de dicha información agregada.

La entrada al modelo es un tensor de forma $[T, F]$, donde $T = 768$ corresponde a la cantidad máxima de cuadros temporales utilizada (cada cuadro representa un fragmento de 10 ms de audio) y $F = 39$ es la cantidad de fonemas brindados por el alineador. El tamaño fijo se definió tras un análisis de longitud de todas las muestras disponibles, donde se observó que únicamente una de ellas superaba los 768 cuadros, y lo hacía exclusivamente por presencia de silencio. Este caso se analiza en el Apéndice 6.2, donde se incluye una visualización del posteriograma y se justifica el recorte sin pérdida de contenido fonético relevante.

Para evitar que los cuadros agregados por padding afecten el resultado, se aplica una máscara que anula su contribución en el *pooling*. La operación se realiza únicamente sobre los cuadros válidos, definidos por la longitud real de cada muestra. Se implementaron dos variantes:

- **Pooling promedio**: suma de activaciones normalizada por la cantidad de cuadros

válidos.

- **Pooling máximo**: se toma el valor máximo de activación para cada fonema, ignorando el padding.

El resultado del *pooling* es un vector de tamaño F, que luego se pasa por una capa lineal que produce la predicción final de velocidad.

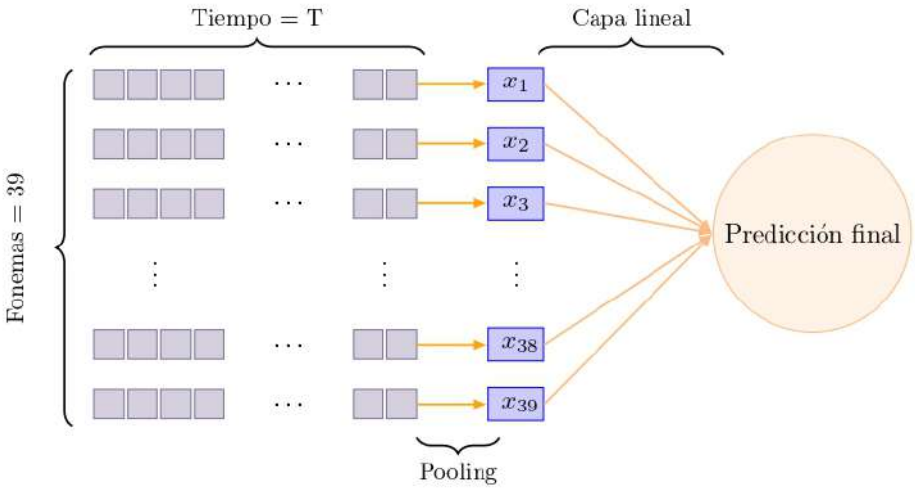

Fig. 5.4: Esquema de la arquitectura TemporalPooling.

El modelo fue entrenado por separado para cada tarea (FPS y SPS), y evaluado sobre el conjunto de validación mediante las métricas MSE, R^2, CCC y coeficiente de correlación de Pearson. Los resultados se muestran en la Tabla 5.5.

Target	**MSE**	R^2	**CCC**	**Pearson**
FPS	3.380	0.120	0.411	0.437
SPS	0.570	0.156	0.430	0.461

Tab. 5.5: Resultados del modelo `TemporalPooling` con pooling promedio.

En la Figura 5.5 se puede observar la relación entre los valores reales y las predicciones obtenidas por el modelo.

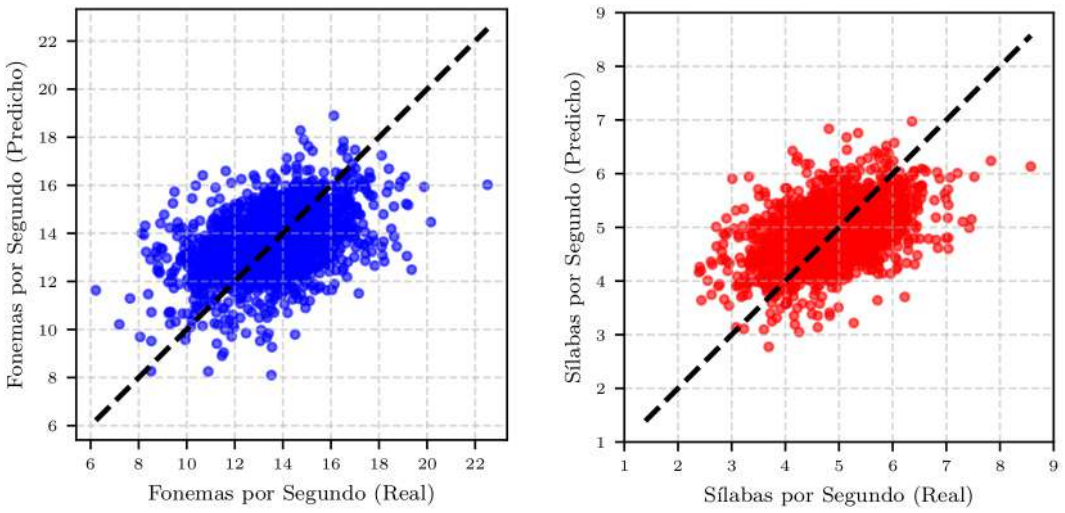

Fig. 5.5: Predicción vs. valores reales para FPS (izquierda) y SPS (derecha) usando `TemporalPooling`.

Debido a su simplicidad estructural y del bajo número de parámetros, el modelo `TemporalPooling` no logra superar el rendimiento de los modelos lineales discutidos en las Secciones 5.1, 5.2 y 5.3. Por lo que este modelo sirve únicamente como punto de partida para la exploración de arquitecturas más complejas que exploren interacciones temporales más finas.

5.5. CNN1D

Con el objetivo de explorar arquitecturas ligeramente más expresivas sin incrementar significativamente la complejidad del modelo, se propuso una red neuronal convolucional simple, denominada `CNN1D`. Esta arquitectura busca aprender patrones temporales en las activaciones fonéticas utilizando convoluciones 1D, seguidas de una agregación temporal y una capa lineal final que produce la predicción de velocidad del habla.

El modelo toma como entrada un tensor de forma $[T, F]$, donde $T = 768$ es la cantidad de cuadros temporales y $F = 39$ es la cantidad de fonemas. A diferencia de la arquitectura de `TemporalPooling`, aquí se aplica una capa de convolución 1D sobre la dimensión temporal. En concreto, se utiliza una capa convolucional con F canales de entrada, un tamaño de kernel k y C canales de salida, seguida por una activación ReLU y una operación de *pooling* manual mediante promedio enmascarado.

Vemos la arquitectura en la Figura 5.6, que puede resumirse como:

- **Convolución 1D:** $\text{Conv1d}(F, \text{out\_channels} = C, \text{kernel\_size} = k, \text{padding} = 1)$
- **ReLU:** Activación no lineal.
- **Máscara de longitud:** Se anulan posiciones fuera de la duración real del audio.
- **Pooling promedio manual:** Promedio sobre los cuadros válidos.
- **Perceptrón final:** $\text{Linear}(C, 1)$

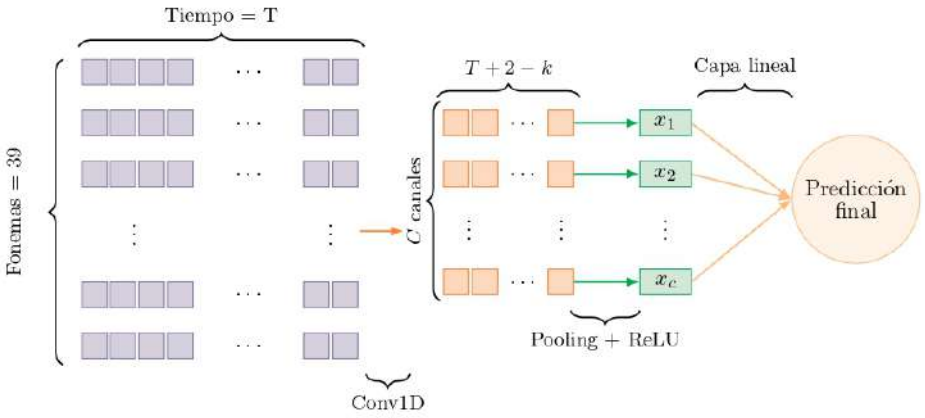

Fig. 5.6: Esquema de la arquitectura CNN1D

El modelo fue evaluado con diferentes combinaciones de hiperparámetros, en particular:

- Tamaños de kernel: $k = \{3, 5\}$
- Canales de salida: $C = \{10, 50\}$

La Tabla 5.6 resume el desempeño del modelo `CNN1D` para la mejor configuración de hiperparámetros, medido sobre el conjunto de evaluación para ambas tareas (FPS y SPS)[1]. Claramente se observa una mejora respecto de modelos más simples, confirmando que la capacidad adicional de la convolución permite capturar patrones fonéticos más complejos.

Configuración	**MSE**	R^2	**CCC**	**Pearson**
FPS **(k=3, C=50)**	1.024	0.734	0.845	0.857
SPS **(k=5, C=50)**	0.170	0.748	0.858	0.865

Tab. 5.6: Resultados de `CNN1D` la mejor configuración de hiperparámetros.

La Figura 5.7 presenta la relación entre valores reales y predichos para las mejores configuraciones encontradas según el target. Se observa una fuerte alineación, especialmente en SPS, lo que confirma el buen ajuste del modelo a los datos.

[1] Todos los resultados se encuentran en el Apéndice 6.2

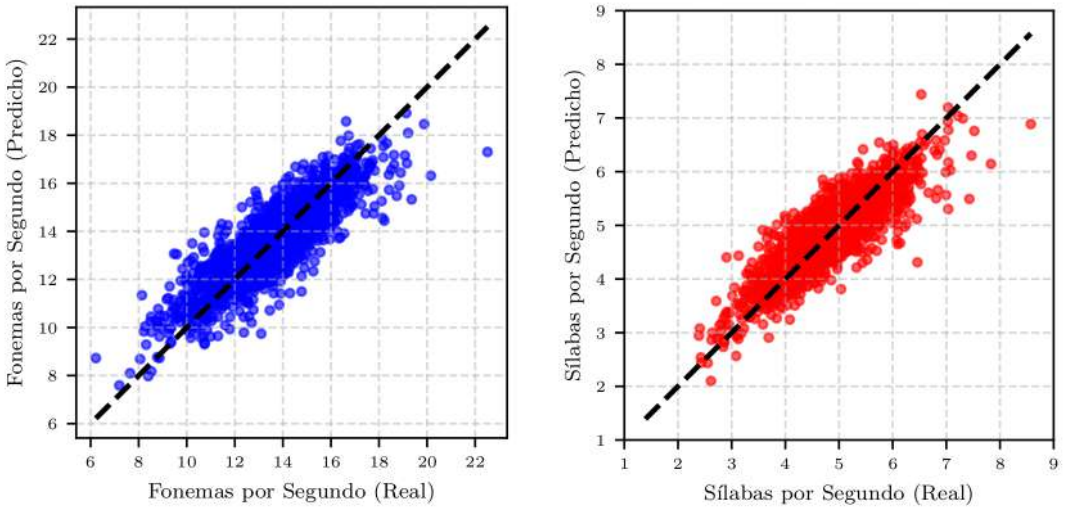

Fig. 5.7: Predicción vs. valores reales para FPS (izquierda) y SPS (derecha). Se utilizó `CNN1D` con $k = 3$ y $C = 50$ para FPS y `CNN1D` con $k = 5$ y $C = 50$ para SPS.

Los resultados obtenidos con `CNN1D` reflejan que una arquitectura convolucional simple es capaz de capturar patrones temporales relevantes para estimar la velocidad del habla. La combinación de convoluciones locales, activaciones no lineales y agregación por promedio resulta eficaz, sin incurrir en una gran cantidad de parámetros. Esta arquitectura sirve como transición natural hacia modelos más complejos que se exploran en secciones posteriores.

5.6. CNN2D

La arquitectura `CNN2D` fue diseñada con el objetivo de replicar, mediante convoluciones, la forma en que los modelos lineales anteriores procesaban las curvas de activación de cada fonema. A diferencia de las redes convolucionales unidimensionales vistas en la sección anterior, aquí se emplean convoluciones bidimensionales (`Conv2d`) con un **kernel de forma** $(1, k)$, es decir, convoluciones que actúan **sobre el tiempo** pero compartiendo parámetros **a lo largo de todos los fonemas**. Esta estrategia permite extraer características temporales de forma independiente para cada canal fonético, como si cada filtro aprendiera a calcular la media, pendiente o curvatura de un fonema específico, al igual que los atributos derivados de las estadísticas fonéticas discutidos en las secciones anteriores.

La entrada al modelo es un tensor de forma $[1, T, F]$, donde $T = 768$ cuadros temporales y $F = 39$ fonemas. Este tensor se reorganiza como $[1, 1, F, T]$ para aplicar las convoluciones bidimensionales. La arquitectura básica incluye:

- **Conv2D (1):** $1 \to C$ canales, kernel $(1, k)$, padding adecuado para mantener la longitud.
- **ReLU**
- **Conv2D (2):** $C \to C$, kernel $(1, k)$, mismo padding.
- **ReLU**

- **Máscara de longitudes**: se enmascaran los cuadros fuera de los T_i reales.
- **Pooling temporal**: promedio manual de las activaciones válidas.
- **Perceptrón final:** Linear$(C \cdot F, 1)$

La Figura 5.8 ilustra el flujo general de esta arquitectura.

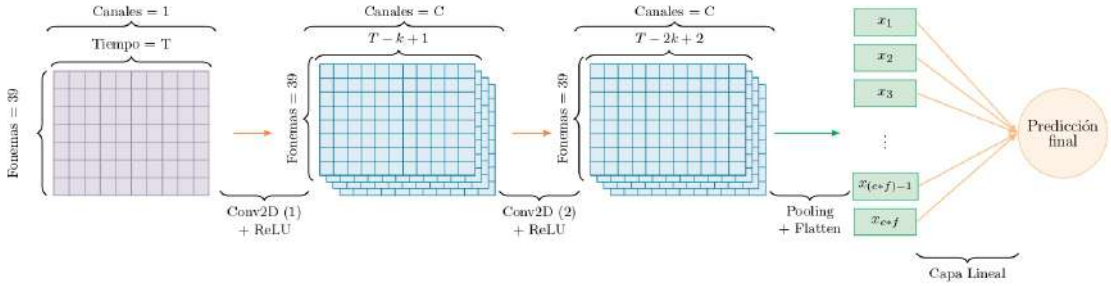

Fig. 5.8: Diagrama esquemático de la arquitectura `CNN2D`. Las convoluciones actúan sobre el tiempo, compartiendo filtros a lo largo de todos los fonemas.

El uso de convoluciones 2D con forma $(1, k)$ permite preservar la identidad fonética de cada canal, permitiendo que cada filtro convolucional actúe de manera similar sobre las curvas de activación individuales. Este diseño contrasta con las `Conv1D` de la sección anterior, que mezclan información entre canales (fonemas), dificultando una interpretación más estructurada. En cambio, `CNN2D` está más alineado con la lógica de los atributos de la regresión lineal, donde cada fonema se analiza individualmente a lo largo del tiempo.

Se exploraron distintas configuraciones de hiperparámetros, variando el tamaño del kernel k y la cantidad de canales C, en particular:

- Tamaños de kernel: $k = \{1, 3, 5\}$
- Canales de salida: $C = \{1, 5, 10, 50, 100\}$

La Tabla 5.7 muestra los resultados obtenidos en las tareas de predicción de FPS y SPS respectivamente por las mejores configuraciones de hiperparámetros obtenidas.

Configuración	**MSE**	R^2	**CCC**	**Pearson**
FPS **(k=5, C=50)**	0.744	0.806	0.896	0.899
SPS **(k=5, C=100)**	0.126	0.813	0.900	0.904

Tab. 5.7: Rendimiento del modelo `CNN2D` para predicción de fonemas por segundo (FPS) y sílabas por segundo (SPS). Se muestran las métricas obtenidas para las mejores configuraciones de hiperparámetros en cada tarea.

En la Figura 5.9, se muestran los gráficos de dispersión entre valores reales y predichos para la mejor configuración encontrada según el target. El ajuste es superior al de todos los modelos estudiados previamente, especialmente en SPS, indicando que esta arquitectura logra capturar mejor las variaciones fonéticas.

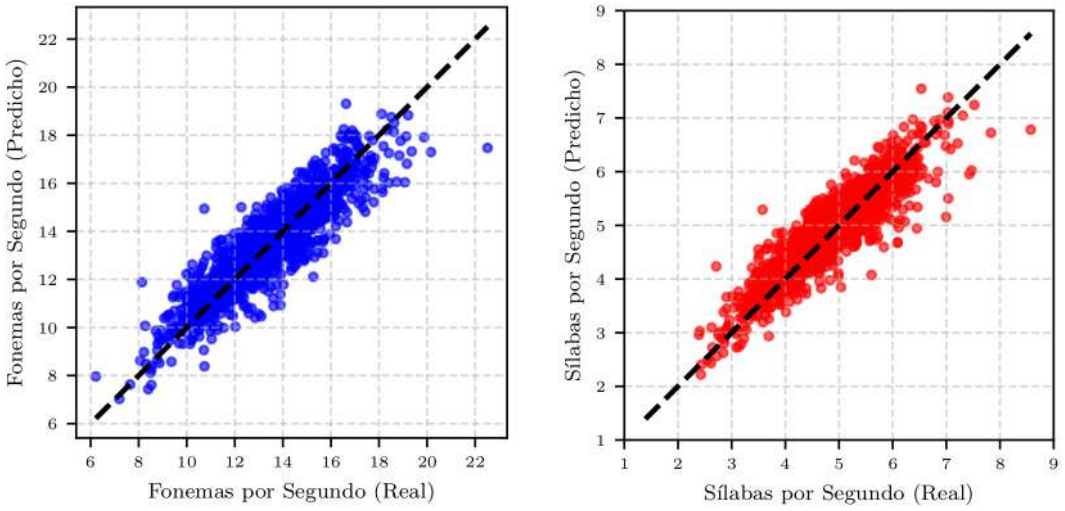

Fig. 5.9: Relación entre valores reales y predichos para FPS (izquierda) y SPS (derecha) utilizando la arquitectura `CNN2D`. Las configuraciones correspondieron a $k = 5$, $C = 50$ para FPS, y $k = 5$, $C = 100$ para SPS.

`CNN2D` presenta una forma de modelado que incorpora inductivamente la estructura del problema: trata cada fonema como una serie temporal independiente, permitiendo a la red aprender patrones temporales especializados. Su rendimiento superior en comparación con `CNN1D` y `TemporalPooling` sugiere que esta estructura inductiva resulta ventajosa para capturar regularidades en el ritmo del habla. Para interpretar adecuadamente estas diferencias de desempeño, es importante considerar también la complejidad de cada arquitectura. La comparación entre rendimiento y cantidad de parámetros puede encontrarse en la Subsección 5.7, donde se analizan estos aspectos en detalle.

Convergencia en entrenamiento

Finalmente, se analizó el comportamiento del entrenamiento de los modelos neuronales mediante sus curvas de pérdida. La Figura 5.10 muestra, en una única visualización, la evolución del error cuadrático medio (MSE) sobre el conjunto de entrenamiento y validación a lo largo de las épocas —es decir, los ciclos completos en los que el modelo recorre todo el conjunto de entrenamiento— para las tres arquitecturas evaluadas.

Se observa que los modelos basados en CNN1D y CNN2D convergen rápidamente y detienen el entrenamiento gracias al mecanismo de *early stopping*, generalmente entre las 90 y 130 épocas. En cambio, el modelo `TemporalPooling` completó todas las épocas permitidas, lo que indica una convergencia más lenta y una posible menor capacidad de representación, entendida como la habilidad del modelo para capturar patrones complejos en los datos de entrada.

Cabe destacar que todos los modelos fueron optimizados bajo las mismas condiciones de entrenamiento: se utilizó el mismo tamaño de *batch*, tasa de aprendizaje y función de pérdida. Esto permite atribuir las diferencias observadas exclusivamente a la arquitectura del modelo y no a factores externos. Las configuraciones detalladas de cada experimento pueden consultarse en el Apéndice 6.2, donde se listan los archivos `.yaml` utilizados.

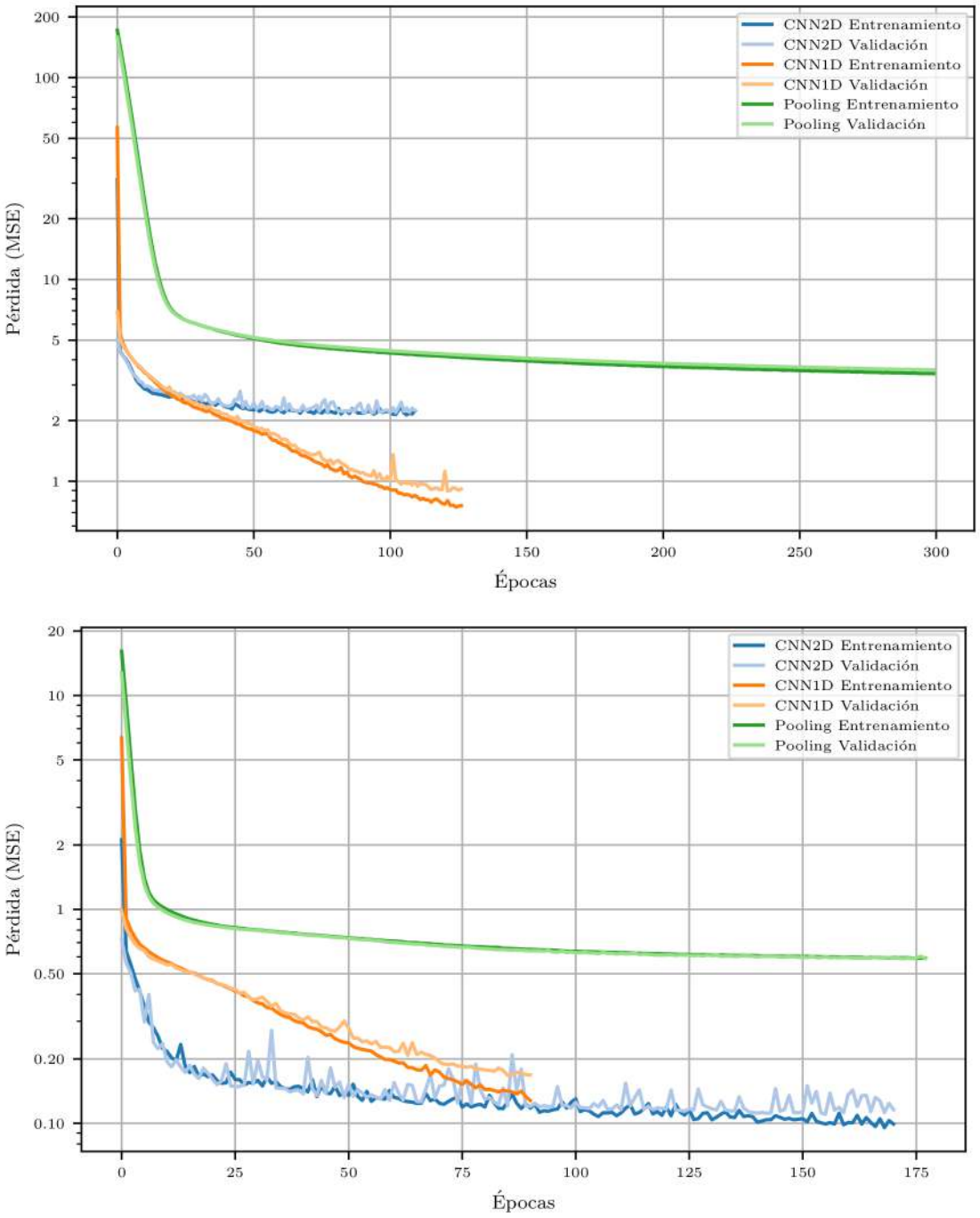

Fig. 5.10: Curvas de entrenamiento y validación (MSE) para los modelos `TemporalPooling`, `CNN1D` y `CNN2D`, para fonemas (arriba) y para sílabas (abajo).

5.7. Comparación de métricas

En esta sección se presentan de forma sistemática los resultados obtenidos por todos los modelos evaluados, tanto lineales como neuronales, en las tareas de estimación de fonemas por segundo (FPS) y sílabas por segundo (SPS). Cada modelo fue evaluado

utilizando múltiples métricas: error cuadrático medio (MSE), coeficiente de determinación (R^2), coeficiente de correlación de concordancia (CCC) y coeficiente de Pearson.

Resultados tabulados

Las Tablas 5.8 y 5.9 resumen las métricas obtenidas para modelos seleccionados según las tareas. Los resultados completos se pueden observar en el Apéndice 6.2

Modelo	**MSE**	**Pearson**	**CCC**	R^2
Regresión combinada	1.008	0.860	0.851	0.738
Pooling - promedio	3.380	0.437	0.411	0.120
CNN1D ($k = 3, c = 50$)	1.024	0.857	0.845	0.734
CNN2D ($k = 1, c = 1$)	2.672	0.566	0.523	0.305
CNN2D ($k = 5, c = 50$)	**0.744**	**0.899**	**0.896**	**0.806**

Tab. 5.8: Comparación de modelos seleccionados para la tarea de predicción de **FPS**.

Modelo	**MSE**	**Pearson**	**CCC**	R^2
SylNet	0.592	0.685	0.630	0.211
Regresión combinada	0.156	0.880	0.875	0.769
Pooling - promedio	0.570	0.461	0.430	0.156
CNN1D ($k = 5, c = 50$)	0.170	0.865	0.858	0.748
CNN2D ($k = 1, c = 1$)	0.545	0.480	0.433	0.193
CNN2D ($k = 3, c = 100$)	**0.126**	**0.904**	**0.900**	**0.813**
CNN2D ($k = 5, c = 100$)	**0.126**	**0.904**	**0.900**	**0.813**

Tab. 5.9: Comparación de modelos seleccionados para la tarea de predicción de **SPS**.

El análisis comparativo permite identificar patrones consistentes en el desempeño de las distintas arquitecturas evaluadas. Para ambas tareas —predicción de fonemas por segundo (FPS) y sílabas por segundo (SPS)—, los modelos basados en convoluciones 2D (`CNN2D`) con kernels más anchos ($k = 3$ o $k = 5$) y mayor cantidad de canales (`C=100`) alcanzan los mejores valores en todas las métricas.

En particular, para la tarea de predicción de **FPS**, el modelo `CNN2D` ($k = 5, c = 100$) logra el menor error cuadrático medio (MSE = 0.744) y los mayores coeficientes de correlación y determinación. En el caso de **SPS**, los modelos `CNN2D` ($k = 5, c = 100$) y `CNN2D` ($k = 3, c = 100$) se posicionan como los más precisos (MSE = 0.126), con valores de Pearson y CCC superiores a 0.90.

Los modelos de regresión lineal también ofrecen desempeños competitivos, cuando combinan estadísticas fonéticas continuas y atributos derivados de apariciones umbralizadas, con valores cercanos a los mejores modelos neuronales en ambas tareas, lo cual refuerza la idea de que parte de la estructura del problema puede ser capturada eficientemente mediante modelos lineales bien diseñados.

Por otro lado, los modelos más simples como `Pooling - Promedio` o `CNN2D` ($k = 1, c = 1$) muestran desempeños limitados, particularmente en métricas de correlación, lo que indica que estrategias que no modelan adecuadamente la estructura temporal del habla no son suficientes para capturar las variaciones rítmicas presentes en los datos.

En conjunto, estos resultados confirman que la incorporación de inductivas estructurales —ya sea a través de arquitecturas convolucionales especializadas o de atributos fonéticos diseñados— permite mejorar significativamente la estimación de la velocidad del habla.

Intervalos de confianza sobre el MSE

Para complementar la comparación de desempeño, se calcularon intervalos de confianza [2] del MSE para los modelos seleccionados en cada tarea. Esta estimación se realizó mediante *bootstrapping*, una técnica de remuestreo que permite obtener una distribución empírica del error a partir de los datos originales, sin asumir supuestos fuertes sobre la distribución. De esta forma, se obtiene un rango de valores dentro del cual es esperable que se encuentre el MSE real del modelo con cierta confianza (en este caso, 95 %).

En la parte superior de la Figura 5.11 se observa que el modelo `CNN2D` ($k = 5, c = 50$) presenta el menor MSE en la tarea de FPS, y su intervalo de confianza no se solapa con el de otros modelos como la regresión combinada o `CNN1D`, indicando que la diferencia es estadísticamente significativa. De forma análoga, en la parte inferior, los modelos `CNN2D` ($k = 3, c = 100$) y `CNN2D` ($k = 5, c = 100$) se destacan como los más precisos para SPS, también con intervalos separados del resto.

[2] https://github.com/luferrer/ConfidenceIntervals

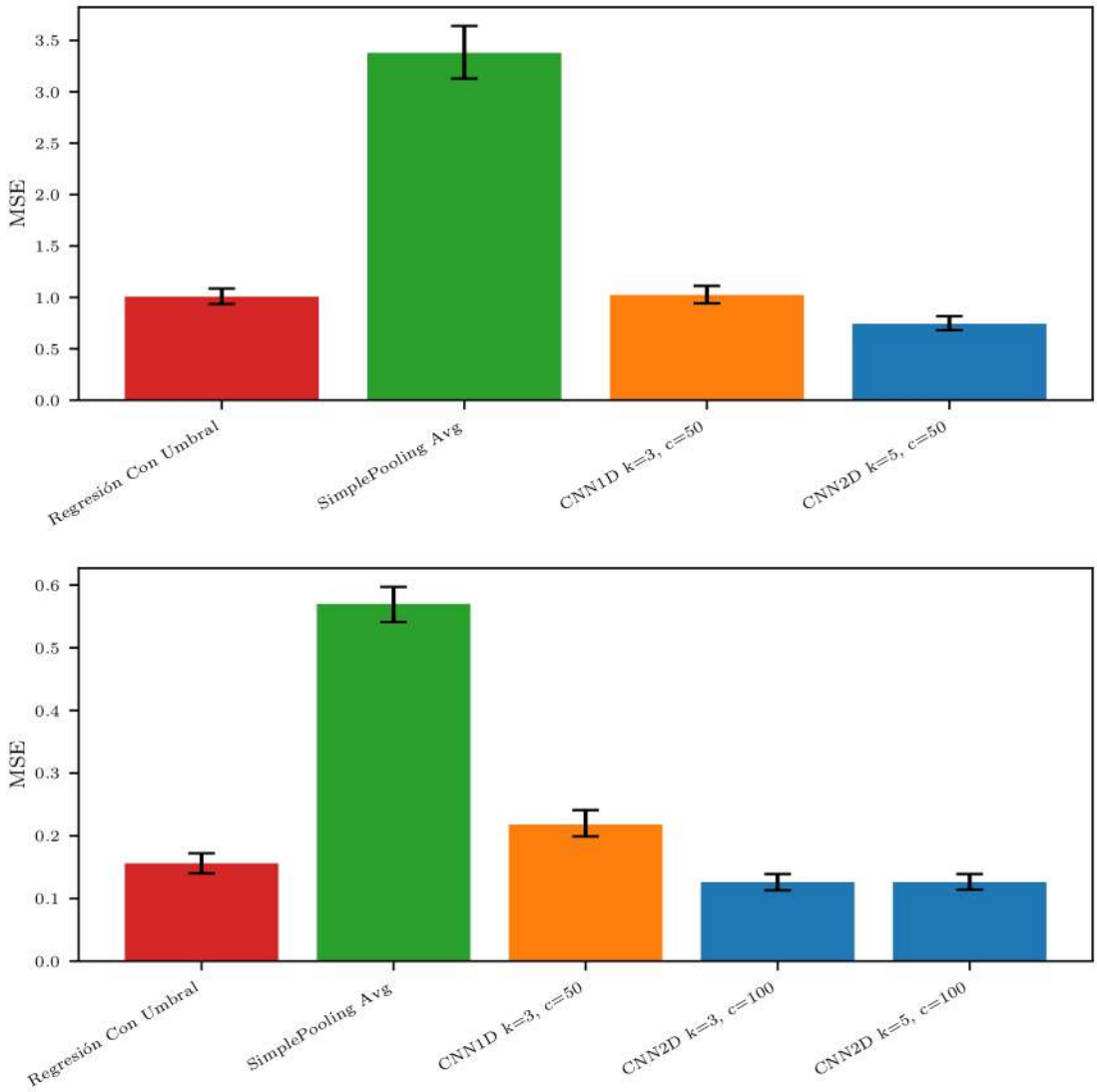

Fig. 5.11: Intervalos de confianza del MSE para modelos seleccionados. Para FPS arriba, y abajo para SPS.

Esto sugiere que, aunque algunos modelos alternativos ofrecen buenos resultados, no alcanzan el mismo nivel de precisión que las arquitecturas convolucionales más expresivas. La inclusión de inductivas estructurales en estas redes parece ser clave para capturar de forma más robusta la dinámica temporal del habla.

Comparación por métrica

En esta sección se comparan los modelos seleccionados exclusivamente en términos de error cuadrático medio (MSE), acompañado de sus respectivos intervalos de confianza[3].

[3] Todos los intervalos fueron estimados mediante *bootstraping* sobre las predicciones del conjunto de test, utilizando 1000 muestras de remuestreo con reemplazo y un nivel de confianza del 95 %.

Se incluyen únicamente aquellos modelos cuyo MSE no se diferencia estadísticamente del mejor modelo observado.

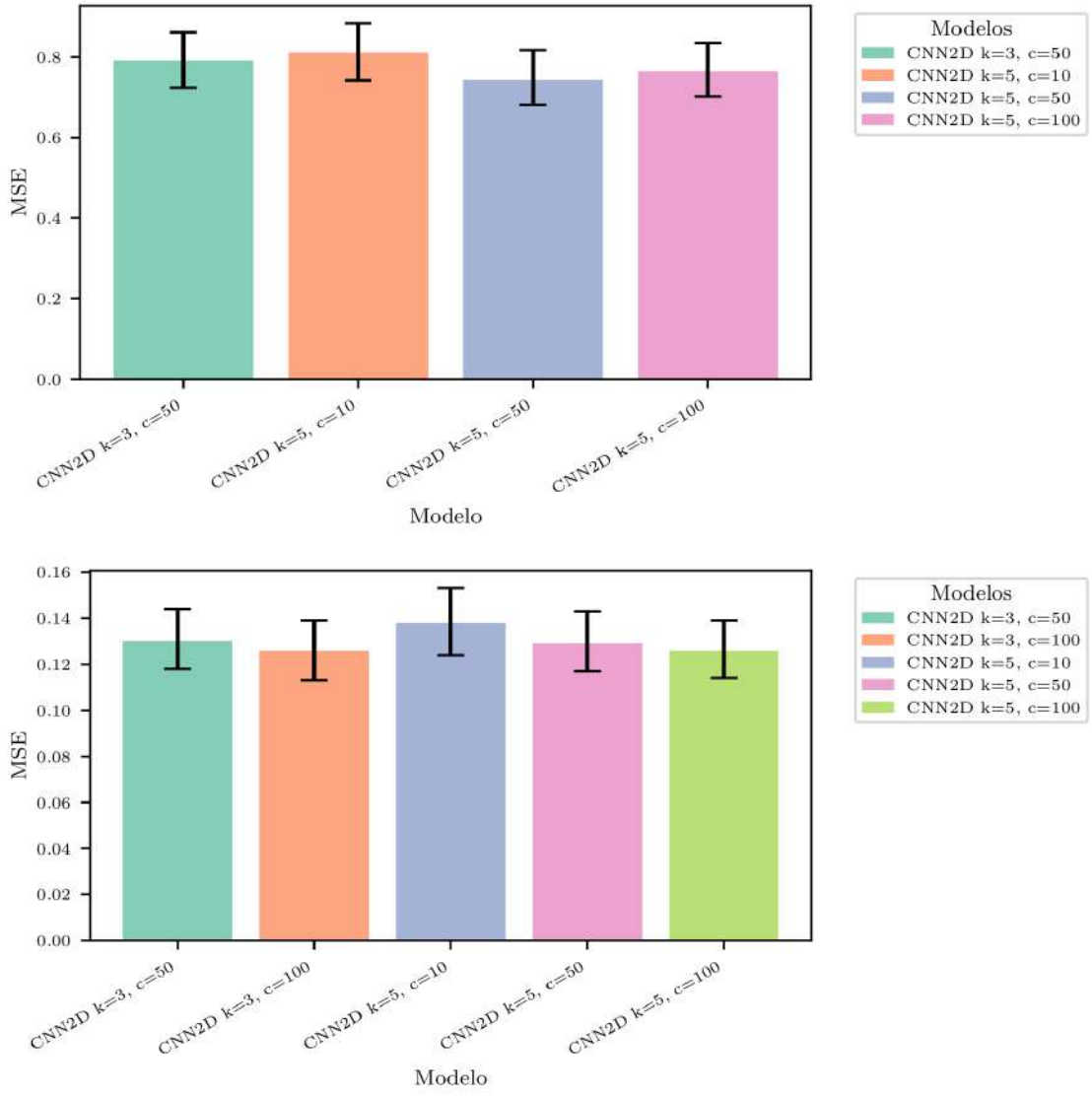

Fig. 5.12: Comparación de modelos CNN2D seleccionados para FPS (arriba) y SPS (abajo) en términos de MSE, con intervalos de confianza.

Como puede observarse en la Figura 5.12, los modelos seleccionados —todas variantes de la arquitectura `CNN2D` con diferentes configuraciones de kernel y cantidad de filtros— presentan valores de MSE muy similares. Las diferencias no son estadísticamente significativas, lo que sugiere que múltiples configuraciones dentro de esta familia arquitectónica logran un desempeño comparable en términos de error absoluto.

Esta flexibilidad puede aprovecharse para priorizar otros aspectos del modelo, como la eficiencia computacional o la cantidad de parámetros. El detalle completo de las métricas restantes (Pearson, CCC y R^2) para estos mismos modelos se presenta en la Tabla 5.10.

Tarea	Modelo	Pearson	CCC	R^2
FPS	CNN2D ($k = 3, c = 50$)	0.891	0.886	0.794
FPS	CNN2D ($k = 5, c = 10$)	0.888	0.884	0.789
FPS	CNN2D ($k = 5, c = 50$)	0.899	0.896	0.806
FPS	CNN2D ($k = 5, c = 100$)	0.895	0.892	0.801
SPS	CNN2D ($k = 3, c = 50$)	0.900	0.896	0.807
SPS	CNN2D ($k = 3, c = 100$)	0.904	0.900	0.813
SPS	CNN2D ($k = 5, c = 10$)	0.895	0.890	0.796
SPS	CNN2D ($k = 5, c = 50$)	0.903	0.898	0.809
SPS	CNN2D ($k = 5, c = 100$)	0.904	0.900	0.813

Tab. 5.10: Métricas complementarias para los modelos CNN2D seleccionados en FPS y SPS.

Estos resultados refuerzan la robustez de la arquitectura `CNN2D`, mostrando que, aún con variaciones en los hiperparámetros, es posible obtener predictores confiables sin comprometer el rendimiento predictivo.

Rendimiento vs. complejidad

Para entender cómo se relaciona la complejidad de los modelos con su rendimiento, en la Figura 5.13 se graficó el MSE en función de la cantidad de parámetros. Se observa una tendencia general a la mejora con arquitecturas más grandes, aunque existen excepciones destacables, como los `CNN1D` o la `Regresión Combinada`, que logran un buen desempeño con una baja cantidad de parámetros.

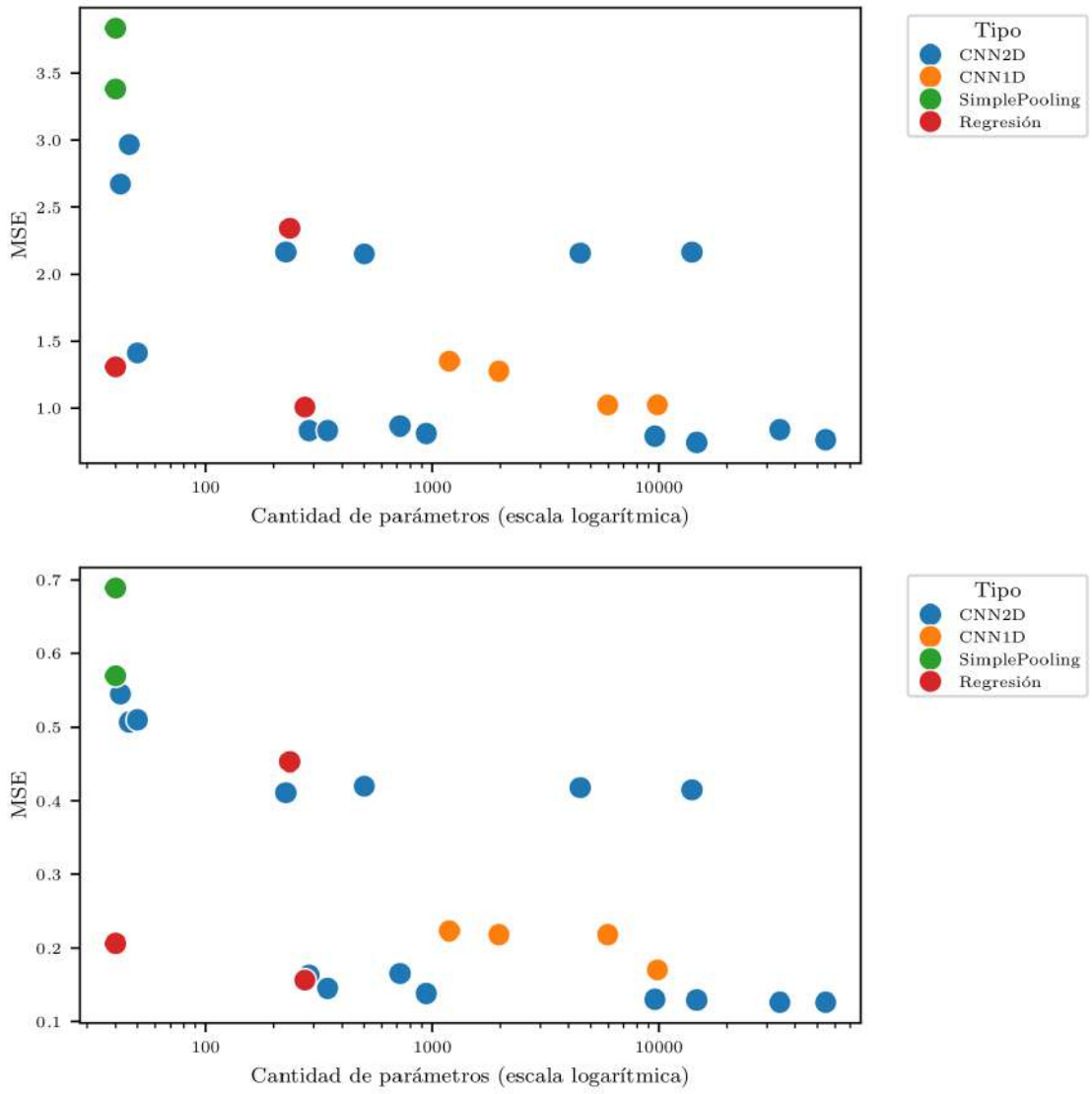

Fig. 5.13: Relación entre cantidad de parámetros y MSE para FPS (arriba) y SPS (abajo). El color indica el tipo de modelo.

Estos resultados permiten identificar modelos adecuados bajo diferentes restricciones de complejidad. En escenarios donde se requiere mantener el número de parámetros por debajo de 10.000, la mejor opción es `CNN2D` ($k = 3, c = 50$), que ofrece un rendimiento competitivo con una arquitectura relativamente compacta. Si la restricción es aún más severa (menos de 1.000 parámetros), la configuración `CNN2D` ($k = 5, c = 10$) resulta la más adecuada para ambas tareas, lo que permite lograr un buen balance entre eficiencia y precisión. Esto refuerza la utilidad de modelos convolucionales diseñados con criterio fonético, incluso en entornos con recursos limitados.

Discusión

En conjunto, estos resultados muestran que el modelo `CNN2D` no sólo alcanza el mejor desempeño general, sino que lo hace manteniendo coherencia entre métricas distintas. Sin embargo, otras alternativas como `CNN1D` o incluso regresiones lineales enriquecidas ofrecen una excelente relación entre precisión y complejidad, lo cual resulta útil para aplicaciones con restricciones de recursos, como implementaciones en navegadores web, procesamiento en servidores de baja capacidad o el despliegue en sistemas de asistencia vocal en tiempo real con múltiples usuarios concurrentes.

5.8. Evaluación controlada con oraciones propias

Para validar la capacidad de generalización de los modelos entrenados en el corpus TIMIT, se propuso una estrategia de evaluación controlada mediante la grabación de oraciones propias. Este enfoque busca analizar la capacidad de los modelos para capturar correctamente variaciones de velocidad del habla en escenarios fuera del dominio de entrenamiento, incluyendo otro idioma (español) y estilos de pronunciación no presentes en TIMIT.

Diseño del experimento

Se seleccionó una oración breve y simple: *"The cat is under the table"*, junto con su equivalente en español: *"El gato está debajo de la mesa"*. Para cada oración, se generaron cinco versiones de audio:

1. Versión hablada a velocidad normal.
2. Versión hablada de forma lenta.
3. Versión hablada de forma rápida.
4. Versión original acelerada a **2x** con un software de edición de audio (sin cambiar el tono).
5. Versión original ralentizada a **0.5x** con un software de edición de audio (sin cambiar el tono).

Esto da un total de **10 archivos de audio**, cinco por idioma. Las grabaciones se realizaron en condiciones acústicas razonables pero no controladas, usando un micrófono estándar. Posteriormente, se calcularon los posteriogramas y se procesaron con los mejores modelos seleccionados para las tareas de predicción de FPS y SPS.

Resultados obtenidos

Se presenta a continuación la comparación entre las velocidades estimadas por los modelos[4] y una referencia de velocidad real, calculada de forma aproximada como el cociente entre la cantidad de fonemas o sílabas esperadas y la duración total del archivo de audio.

[4] Se utilizaron las mejores configuraciones de `CNN2D` según el target y la regresión lineal combinada para las estimaciones.

Para realizar la estimación real, sin tender a subestimar la velocidad de pronunciación, se quitaron los segmentos de silencio tanto al principio como al final del audio lo cual amplía la duración sin modificar el contenido fonético. Permitiendo así una inferencia más limpia y alineada con los datos de entrenamiento del modelo. Estos silencios se aprecian claramente en las figuras de la subsección 5.8.

Idioma	Audio	CNN2D	Reg. lineal	Real
Español	Normal	13.71	10.13	8.90
	Lento	11.32	7.55	4.66
	Rápido	15.41	13.18	12.06
	Normal ×2	25.22	22.87	17.71
	Normal ×0.5	15.12	11.95	4.44
Inglés	Normal	14.01	12.04	9.83
	Lento	12.20	11.93	6.20
	Rápido	17.15	15.29	13.60
	Normal ×2	20.56	19.05	17.35
	Normal ×0.5	13.29	11.49	4.86

Tab. 5.11: Comparación de FPS estimado por el modelo CNN2D y la regresión lineal.

Idioma	Audio	CNN2D	Reg. Lineal	SylNet	Real
Español	Normal	6.20	4.62	3.65	6.28
	Lento	4.38	2.29	2.36	3.29
	Rápido	7.69	6.10	4.55	8.51
	Normal ×2	10.31	10.75	8.03	12.50
	Normal ×0.5	6.83	4.74	2.93	3.43
Inglés	Normal	5.07	4.40	4.20	4.62
	Lento	4.19	4.31	2.89	2.92
	Rápido	6.27	5.81	6.17	6.40
	Normal ×2	7.93	7.84	6.98	8.16
	Normal ×0.5	4.94	4.37	3.37	2.29

Tab. 5.12: Comparación de SPS estimado por el modelo CNN2D, la regresión lineal y SylNet.

A partir de las Tablas 5.11 y 5.12 se observa que los modelos tienden a sobreestimar sistemáticamente la velocidad del habla, aunque con diferente intensidad según el tipo de métrica. Esta tendencia puede explicarse parcialmente por la distribución de velocidades presente en el corpus de entrenamiento.

En el caso de fonemas por segundo (FPS), la sobreestimación es más marcada, especialmente en las condiciones de habla lenta (`Normal ×0.5` y `Lento`). En estas configuraciones, el modelo CNN2D llega a predecir más del triple de la velocidad real. Este fenómeno sugiere que el modelo ha aprendido una representación sesgada hacia velocidades más rápidas, probablemente debido a que en el conjunto de entrenamiento la gran mayoría de

los ejemplos se ubican en un rango de entre 10 y 17 fonemas por segundo (percentiles 5–95 %). En contraste, los audios ralentizados que se utilizaron para esta evaluación caen por debajo de ese umbral, saliéndose del rango de exposición del modelo. Por esta razón, el modelo tiende a sobreestimar sistemáticamente en estos casos extremos, ya que no ha sido entrenado con suficientes ejemplos de velocidad lenta.

Por su parte, en las estimaciones de sílabas por segundo (SPS), la sobreestimación es menos pronunciada e incluso más ajustada en varios casos. La CNN2D logra valores cercanos a los reales en condiciones de habla normal o rápida, y las desviaciones se mantienen en márgenes moderados para la mayoría de los ejemplos. Esto puede deberse a que la variabilidad en SPS es menor que en FPS y que el rango típico de velocidades en el conjunto de entrenamiento para SPS se ubica aproximadamente entre 3.5 y 6.5 sílabas por segundo (percentiles 5–95 %). Como las velocidades reales de la mayoría de los audios evaluados se encuentran dentro de ese rango o cerca de sus extremos, el modelo muestra una mejor capacidad de generalización en este caso.

Otro aspecto interesante es que, en comparación con SylNet, el modelo CNN2D mejora sus predicciones en la mayoría de las condiciones. Esto sugiere que, a pesar de las limitaciones del corpus y la arquitectura, las técnicas utilizadas para estimar SPS en esta tesis resultan competitivas frente a alternativas previas.

En consecuencia, un posible eje de mejora para trabajos futuros consiste en incrementar la diversidad de velocidades de habla en el conjunto de entrenamiento. Una estrategia viable sería aplicar técnicas de *data augmentation* por escalamiento temporal, es decir, generar versiones sintéticas de los audios que simulen velocidades más lentas o extremas. Esto permitiría al modelo calibrar mejor su respuesta en estos casos fuera de distribución, reduciendo el sesgo hacia velocidades medias o rápidas observado en esta evaluación.

Visualización de posteriogramas

Para ilustrar la diferencia entre las señales de entrada que recibe el modelo, se muestran los posteriogramas generados para las versiones habladas a velocidad normal de ambas oraciones.

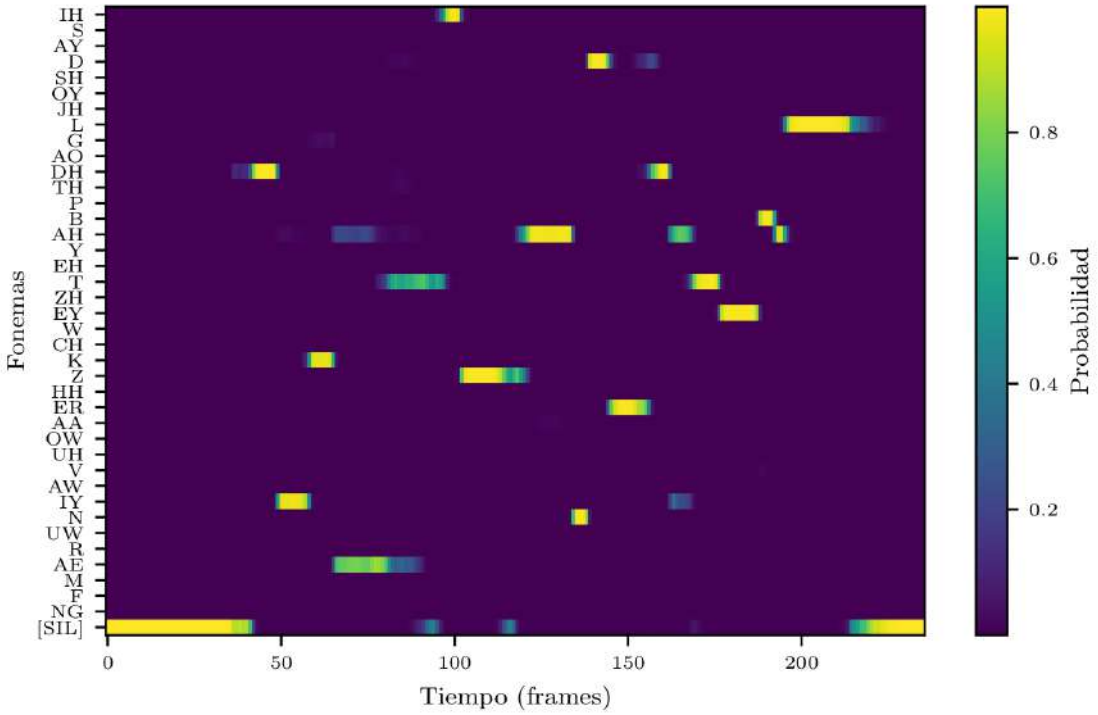

Fig. 5.14: Posteriograma generado con Charsiu para el audio normal de la oración en inglés.

Como puede observarse, el posteriograma de la oración en inglés (Figura 5.14) presenta activaciones más definidas y localizadas en el tiempo, con picos claros de probabilidad para fonemas individuales. Esto contrasta con la representación más difusa y menos marcada del posteriograma correspondiente al audio en español (Figura 5.15), donde las probabilidades aparecen distribuidas y con menor contraste.

Esta diferencia es coherente con el hecho de que el modelo utilizado, `en_w2v2_fc_10ms`, fue entrenado sobre datos en inglés. La degradación en la definición de los fonemas en español podría estar asociada a una menor confianza del modelo al inferir fonemas no presentes en su dominio original, lo cual explicaría parcialmente el menor desempeño observado en las predicciones de velocidad en este idioma.

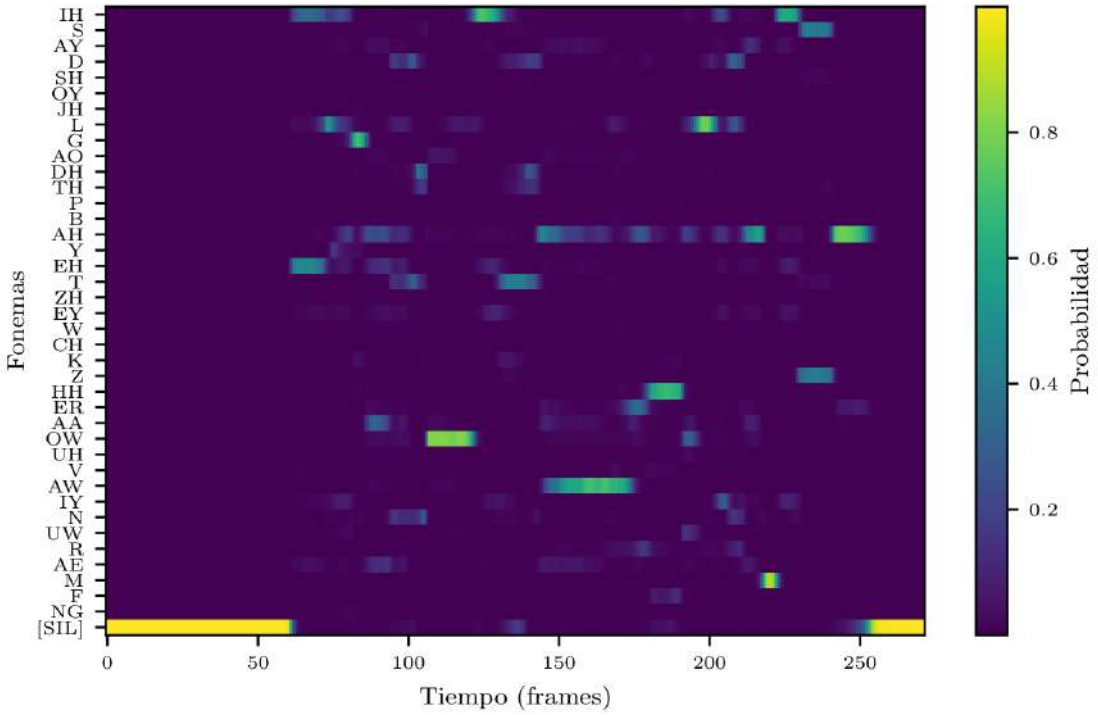

Fig. 5.15: Posteriograma generado con Charsiu para el audio normal de la oración en español.

Discusión

Los resultados permiten comparar no sólo el comportamiento del modelo neuronal frente a cambios controlados de velocidad, sino también contrastarlo con el de una regresión lineal basada en atributos fonéticos.

En términos generales, ambos modelos responden correctamente a los cambios de velocidad: los valores predichos aumentan en las versiones aceleradas y disminuyen en las ralentizadas. Sin embargo, se observan diferencias sistemáticas en la magnitud de las predicciones. En particular, el modelo basado en redes neuronales (`CNN2D`) tiende a sobreestimar las velocidades con respecto a los valores reales, especialmente en los audios más lentos. Esto puede deberse a la presencia de silencios al principio y al final de las grabaciones, los cuales fueron incluidos en el cálculo del "ground truth", pero no siempre son relevantes para el modelo.

La regresión lineal, en cambio, produce estimaciones generalmente más conservadoras, que en muchos casos se aproximan mejor a las velocidades reales. Esto podría deberse a que esta clase de modelos se basa directamente en promedios y frecuencias de activación fonética, lo cual le permite capturar de manera más estable ciertos patrones generales, aunque con menor capacidad de adaptación a variaciones más complejas en la señal.

Esta diferencia entre ambos enfoques subraya la importancia de considerar no sólo el desempeño en corpus controlados como TIMIT, sino también el comportamiento fuera de dominio. Además, refuerza la necesidad de incorporar mecanismos de detección de silencios reales o segmentación precisa si se busca una estimación más fiel de la velocidad de habla en contextos prácticos.

6. CONCLUSIONES Y TRABAJO FUTURO

6.1. Conclusiones

A lo largo de este trabajo se exploraron diferentes estrategias para la estimación automática de la velocidad del habla, medida tanto en fonemas por segundo (FPS) como en sílabas por segundo (SPS), a partir de representaciones acústicas derivadas de posteriogramas. Se desarrollaron modelos de regresión lineal y redes neuronales convolucionales diseñadas con distintos grados de complejidad y capacidad de representación.

Los resultados experimentales permiten extraer varias conclusiones relevantes. En primer lugar, se constató que la predicción de **SPS resulta más sencilla que la de FPS**, reflejando su menor variabilidad y mayor estabilidad ante factores como la densidad fonética o la segmentación. Esto se evidenció tanto en las métricas de error (menores valores de MSE para SPS) como en la estabilidad del entrenamiento de los modelos. Esta observación también refuerza la idea de que, para muchas aplicaciones prácticas —especialmente en contextos educativos o clínicos—, trabajar con sílabas por segundo no sólo es más interpretable, sino también más robusto desde el punto de vista computacional.

En segundo lugar, se comprobó que los modelos lineales logran capturar parte significativa de la variabilidad en los datos cuando se diseñan cuidadosamente los atributos, especialmente al combinar estadísticas fonéticas con conteos umbralizados. Sin embargo, fueron las arquitecturas basadas en **convoluciones 2D con kernels temporales y separación por fonema** las que ofrecieron los mejores resultados. Estas redes, al tratar cada canal fonético como una serie temporal independiente, lograron modelar de forma más efectiva la dinámica del habla.

Estos hallazgos refuerzan la idea de que el **diseño de arquitecturas que respeten las propiedades lingüísticas subyacentes** —como la independencia entre fonemas o la relación temporal dentro de una unidad de habla— puede llevar a mejoras sustanciales en tareas de análisis del habla. Incorporar de manera explícita supuestos estructurales se mostró clave para obtener buenos desempeños, tanto en términos de error como de concordancia entre predicción y realidad.

En suma, este trabajo no sólo valida las hipótesis planteadas sobre la viabilidad de estimar la velocidad del habla a partir de posteriogramas sin segmentación textual, sino que también destaca la importancia de alinear las decisiones arquitectónicas con el conocimiento lingüístico disponible.

6.2. Trabajo futuro

Este trabajo abre múltiples líneas posibles de continuación y expansión. Una de las más directas es la incorporación de nuevos corpus que extiendan la diversidad lingüística y prosódica del material de entrenamiento. Si bien los modelos evaluados demostraron cierta capacidad de generalización a oraciones en español, todos fueron entrenados exclusivamente sobre el corpus TIMIT, compuesto por habla en inglés. Explorar datasets de habla en español permitiría validar los hallazgos actuales en contextos lingüísticos distintos y evaluar la robustez real de los modelos frente al cambio de idioma.

Además, se podrían desarrollar variantes multilingües entrenadas directamente sobre

corpus que integren distintas lenguas, lo cual abriría la puerta a sistemas verdaderamente generalizables. En ese sentido, incorporar representaciones fonéticas más abstractas —como las utilizadas por modelos como `Phonet` [26]— podría facilitar la transferencia entre idiomas al reducir la dependencia de los fonemas específicos de cada lengua.

Desde el punto de vista metodológico, una posible mejora consiste en integrar modelos preentrenados más avanzados como `Whisper` [27] o `wav2vec-XLS-R` [28], que han sido entrenados sobre grandes cantidades de datos multilingües y podrían proporcionar posteriogramas más informativos para tareas de estimación del ritmo del habla.

Finalmente, a nivel arquitectónico, podría explorarse la incorporación de mecanismos de atención o modelos recurrentes que capturen dependencias de largo plazo, así como estrategias de entrenamiento multitarea que permitan predecir simultáneamente otras variables del habla (como duración total, número de pausas, velocidad con pausas excluidas, o velocidad media instantánea).

En conjunto, estas extensiones permitirían avanzar hacia modelos más precisos, robustos y versátiles, capaces de operar en condiciones de mayor variabilidad lingüística, discursiva y acústica.

Bibliografía

[1] François Pellegrino, Christophe Coupé, and Egidio Marsico. A cross-language perspective on speech information rate. *Language*, 87(3):539–558, 2011.

[2] Christophe Coupé, Yoon Mi Oh, Dan Dediu, and François Pellegrino. Different languages, similar encoding efficiency: Comparable information rates across the human communicative niche. *Science Advances*, 5(9):eaaw2594, 2019.

[3] Michelle Cohn, Kai-Hui Liang, Melina Sarian, Georgia Zellou, and Zhou Yu. Speech rate adjustments in conversations with an amazon alexa socialbot. *Frontiers in Communication*, Volume 6 - 2021, 2021.

[4] Roger Griffiths. Speech rate and nns comprehension: A preliminary study in time-benefit analysis. *Language Learning*, 40(3):311–336, 1990.

[5] Sheri Robinson, Heather Sterling, Christopher Skinner, and Daniel Robinson. Effects of lecture rate on students' comprehension and ratings of topic importance. *Contemporary Educational Psychology - CONTEMP EDUC PSYCHOL*, 22:260–267, 04 1997.

[6] Brent Simonds, Kevin Meyer, Margaret Quinlan, and Stephen Hunt. Effects of instructor speech rate on student affective learning, recall, and perceptions of nonverbal immediacy, credibility, and clarity. *Communication Research Reports*, 23:187–197, 09 2006.

[7] Barry Guitar and Lisa Marchinkoski. Influence of mothers' slower speech on their children's speech rate. *Journal of Speech, Language, and Hearing Research*, 44:853–861, 08 2001.

[8] Gréta Szatlóczki, Ildiko Hoffmann, Veronika Vincze, Kálmán János, and Magdolna Pakaski. Speaking in alzheimer's disease, is that an early sign? importance of changes in language abilities in alzheimer's disease. *Frontiers in Aging Neuroscience*, 7, 09 2015.

[9] Francisco Martínez-Sánchez, Juan Meilán, Juan Carro, Consolación Iñiguez, Lymarie Millian Morell, Isabel M Valverde, José Alburquerque, and Dolores López. Speech rate in parkinson's disease: A controlled study. *Neurologia (Barcelona, Spain)*, 31, 02 2015.

[10] Sabine Skodda. Aspects of speech rate and regularity in parkinson's disease. *Journal of the neurological sciences*, 310:231–6, 08 2011.

[11] Steven Greenberg, Hannah Carvey, Leah Hitchcock, and Shuangyu Chang. Temporal properties of spontaneous speech—a syllable-centric perspective. *Journal of Phonetics*, 31(3):465–485, 2003. Temporal Integration in the Perception of Speech.

[12] Dagen Wang and Shrikanth S. Narayanan. Robust speech rate estimation for spontaneous speech. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(8):2190–2201, 2007.

[13] Xiangyu Zeng, Shi Yin, and Dong Wang. Learning speech rate in speech recognition. *CoRR*, abs/1506.00799, 2015.

[14] Natalia Tomashenko and Yuri Khokhlov. Speaking rate estimation based on deep neural networks. 10 2014.

[15] Tomas Palazzo. Predicción de la velocidad del habla con modelos preentrenados. 2024.

[16] Shannon L. Barrios, Anna M. Namyst, Ellen F. Lau, Naomi H. Feldman, and William J. Idsardi. Establishing new mappings between familiar phones: Neural and behavioral evidence for early automatic processing of nonnative contrasts. *Frontiers in Psychology*, Volume 7 - 2016, 2016.

[17] Gianpaolo Coro. *A Step Forward in Multi-granular Automatic Speech Recognition.* 08 2013.

[18] Jian Zhu, Cong Zhang, and David Jurgens. Phone-to-audio alignment without text: A semi-supervised approach. *CoRR*, abs/2110.03876, 2021.

[19] Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *CoRR*, abs/2006.11477, 2020.

[20] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *CoRR*, abs/1706.03762, 2017.

[21] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[22] Ossama Abdel-Hamid, Abdel-rahman Mohamed, Hui Jiang, Li Deng, Gerald Penn, and Dong Yu. Convolutional neural networks for speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 22(10):1533–1545, 2014.

[23] Tara N. Sainath, Ron J. Weiss, Andrew Senior, Kevin W. Wilson, and Oriol Vinyals. Learning the speech front-end with raw waveform cldnns. In *Interspeech 2015*, pages 1–5, 2015.

[24] J. Garofolo, Lori Lamel, W. Fisher, Jonathan Fiscus, D. Pallett, N. Dahlgren, and V. Zue. Timit acoustic-phonetic continuous speech corpus. *Linguistic Data Consortium*, 11 1992.

[25] Shreyas Seshadri and Okko Räsänen. Sylnet: An adaptable end-to-end syllable count estimator for speech. *CoRR*, abs/1906.09825, 2019.

[26] J. C. Vásquez-Correa, P. Klumpp, J. R. Orozco-Arroyave, and E. N.ºth. Phonet: A Tool Based on Gated Recurrent Neural Networks to Extract Phonological Posteriors from Speech. In *Proc. Interspeech 2019*, pages 549–553, 2019.

[27] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision, 2022.

[28] Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, Alexei Baevski, Alexis Conneau, and Michael Auli. XLS-R: self-supervised cross-lingual speech representation learning at scale. *CoRR*, abs/2111.09296, 2021.

7. APÉNDICES

Justificación del truncamiento de posteriogramas

En el procesamiento de los datos para el entrenamiento de los modelos neuronales, fue necesario unificar la longitud temporal de todos los posteriogramas. Dado que los modelos no trabajan directamente con secuencias de longitud variable, se optó por truncar todos los posteriogramas a una longitud fija de **768 cuadros temporales** (lo que equivale a aproximadamente 7.68 segundos, dado el paso temporal de 10ms).

Esta decisión se basa en un análisis previo de la duración de los audios en el corpus TIMIT. Como se muestra en la Figura 7.1, **sólo un archivo de audio superaba la longitud de 768 cuadros**. En dicho caso, se verificó visualmente que el segmento eliminado correspondía únicamente a una porción final de silencio (es decir, no contenía activaciones relevantes de fonemas).

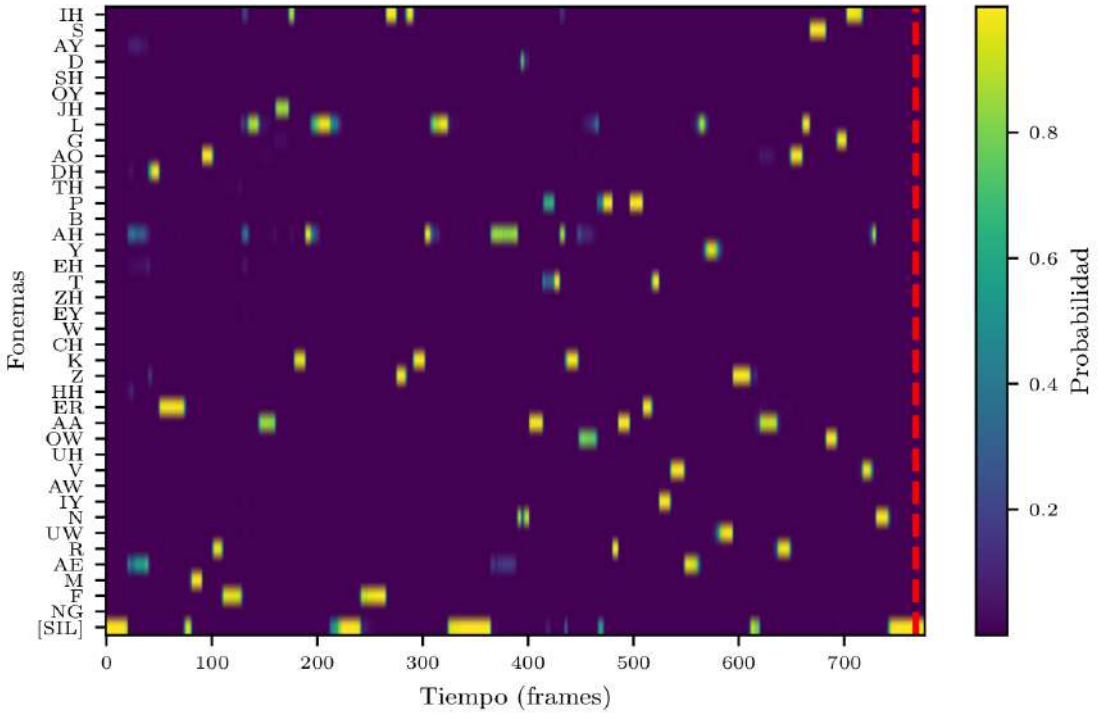

Fig. 7.1: Posteriograma completo del único audio truncado. La línea roja punteada marca el cuadro 768.

La línea roja punteada en la Figura 7.1 indica el punto exacto donde se aplica el truncamiento. Como puede observarse, la región posterior a dicho punto está dominada por activaciones del fonema de silencio `/SIL/`, lo que confirma que el recorte no afecta información fonética significativa.

Por esta razón, el truncamiento no introduce sesgo relevante ni pérdida de contenido

lingüístico útil, y permite mantener un preprocesamiento uniforme y compatible con arquitecturas convolucionales y batch training. Para el resto de los audios, cuya longitud era inferior a 768 cuadros, se aplicó padding con ceros sin afectar el contenido efectivo.

Configuraciones de entrenamiento

Este apéndice documenta las configuraciones utilizadas para entrenar los distintos modelos neuronales desarrollados en esta tesis. Cada bloque YAML representa un archivo de configuración utilizado durante el entrenamiento y evaluación.

Modelo `SimpleTemporalPooling`

```yaml
task: fonemas
model:
    name: "SimpleTemporalPooling"
    params:
        num_phonemes: 39
        pooling_type: avg
data:
    target_col: Fonemas por Segundo
    batch_size: 64
    num_workers: 4
    fixed_size: 768
training:
    epochs: 300
    lr: 0.001
    loss: MSELoss
    optimizer: Adam
    early_stopping_patience: 15
logging:
    tensorboard: True
    log_interval: 10
```

Listing 7.1: Configuración para SimpleTemporalPooling

Modelo `CNN1D`

```yaml
task: fonemas
model:
    name: "TinyCNN1D"
    params:
        in_channels: 39
        out_channels: 10
        kernel_size: 5
data:
    target_col: Fonemas por Segundo
    batch_size: 64
    num_workers: 4
    fixed_size: 768
training:
    epochs: 300
```

```
    lr: 0.001
    loss: MSELoss
    optimizer: Adam
    early_stopping_patience: 15
logging:
    tensorboard: True
    log_interval: 10
output: outputs/fonemas_tinycnn_k5_c10
```

Listing 7.2: Configuración para CNN1D

Modelo `CNNFeatureConv`

```
task: fonemas
model:
    name: "CNNFeatureConv"
    params:
        kernel_size: 5
        num_fonemes: 39
        out_channels: 10
data:
    batch_size: 64
    fixed_size: 768
    num_workers: 4
    target_col: Fonemas por Segundo
training:
    early_stopping_patience: 15
    epochs: 300
    loss: MSELoss
    lr: 0.001
    optimizer: Adam
logging:
    tensorboard: True
    log_interval: 10
output: outputs/cnnfeatures_fonemas_k5_c10
```

Listing 7.3: Configuración para CNNFeatureConv

Resultados totales

Se muestran los resultados de todos los modelos para las métricas: error cuadrático medio (MSE), coeficiente de determinación (R^2), coeficiente de correlación de concordancia (CCC) y coeficiente de Pearson.

Modelo	**MSE**	**Pearson**	**CCC**	R^2
Regresión simple	2.343	0.629	0.563	0.390
Regresión con umbral	1.307	0.814	0.799	0.660
Regresión combinada	**1.008**	**0.860**	**0.851**	**0.738**
Pooling - promedio	**3.380**	**0.437**	**0.411**	**0.120**
Pooling - máximo	3.834	0.386	0.372	0.002
CNN1D ($k = 3, c = 10$)	1.350	0.807	0.789	0.649
CNN1D ($k = 3, c = 50$)	**1.024**	**0.857**	**0.845**	**0.734**
CNN1D ($k = 5, c = 10$)	1.276	0.819	0.807	0.668
CNN1D ($k = 5, c = 50$)	1.025	0.857	0.847	0.733
CNN2D ($k = 1, c = 1$)	2.672	0.566	0.523	0.305
CNN2D ($k = 1, c = 5$)	2.164	0.662	0.613	0.437
CNN2D ($k = 1, c = 10$)	2.150	0.667	0.619	0.440
CNN2D ($k = 1, c = 50$)	2.156	0.666	0.615	0.439
CNN2D ($k = 1, c = 100$)	2.163	0.664	0.615	0.437
CNN2D ($k = 3, c = 1$)	2.966	0.488	0.426	0.228
CNN2D ($k = 3, c = 5$)	0.833	0.885	0.880	0.783
CNN2D ($k = 3, c = 10$)	0.867	0.881	0.874	0.774
CNN2D ($k = 3, c = 50$)	0.791	0.891	0.886	0.794
CNN2D ($k = 3, c = 100$)	0.840	0.884	0.876	0.781
CNN2D ($k = 5, c = 1$)	1.412	0.799	0.765	0.633
CNN2D ($k = 5, c = 5$)	0.833	0.885	0.881	0.783
CNN2D ($k = 5, c = 10$)	0.811	0.888	0.884	0.789
CNN2D ($k = 5, c = 50$)	**0.744**	**0.899**	**0.896**	**0.806**
CNN2D ($k = 5, c = 100$)	0.764	0.895	0.892	0.801

Tab. 7.1: Comparación de modelos para la tarea de predicción de **FPS**.

Modelo	**MSE**	**Pearson**	**CCC**	R^2
SylNet	**0.592**	**0.685**	**0.630**	**0.211**
Regresión simple	0.453	0.586	0.534	0.329
Regresión con umbral	0.206	0.837	0.827	0.695
Regresión combinada	**0.156**	**0.880**	**0.875**	**0.769**
Pooling - promedio	**0.570**	**0.461**	**0.430**	**0.156**
Pooling - máximo	0.689	0.300	0.256	-0.020
CNN1D ($k = 3, c = 10$)	0.223	0.819	0.806	0.669
CNN1D ($k = 3, c = 50$)	0.218	0.824	0.812	0.677
CNN1D ($k = 5, c = 10$)	0.218	0.825	0.819	0.678
CNN1D ($k = 5, c = 50$)	**0.170**	**0.865**	**0.858**	**0.748**
CNN2D ($k = 1, c = 1$)	0.545	0.480	0.433	0.193
CNN2D ($k = 1, c = 5$)	0.411	0.634	0.595	0.391
CNN2D ($k = 1, c = 10$)	0.420	0.624	0.583	0.378
CNN2D ($k = 1, c = 50$)	0.418	0.630	0.592	0.382
CNN2D ($k = 1, c = 100$)	0.415	0.631	0.588	0.386
CNN2D ($k = 3, c = 1$)	0.507	0.519	0.476	0.249
CNN2D ($k = 3, c = 5$)	0.163	0.875	0.869	0.759
CNN2D ($k = 3, c = 10$)	0.165	0.875	0.866	0.756
CNN2D ($k = 3, c = 50$)	0.130	0.900	0.896	0.807
CNN2D ($k = 3, c = 100$)	**0.126**	**0.904**	**0.900**	**0.813**
CNN2D ($k = 5, c = 1$)	0.510	0.522	0.485	0.245
CNN2D ($k = 5, c = 5$)	0.145	0.889	0.883	0.786
CNN2D ($k = 5, c = 10$)	0.138	0.895	0.890	0.796
CNN2D ($k = 5, c = 50$)	0.129	0.903	0.898	0.809
CNN2D ($k = 5, c = 100$)	**0.126**	**0.904**	**0.900**	**0.813**

Tab. 7.2: Comparación de modelos para la tarea de predicción de **SPS**.