UNTREF UNIVERSIDAD NACIONAL DE TRES DE FEBRERO

INGENIERÍA EN SONIDO

Evaluación de cadenas de procesameintos en la creación de conjuntos de datos del habla

Tesis final presentada para obtener el título de Ingeniero de Sonido de la

Universidad Nacional de Tres de Febrero (UNTREF)

TESISTA: Matías Di Bernardo (42.229.438)
TUTOR/A: Gala Lucía Gonzáles Barrios (Lic.)

Fecha de defensa: mes y año | Locación (Ej. Saenz Peña), Argentina

AGRADECIMIENTOS

Se propone incluir este apartado, donde se debe agradecer primeramente a las autoridades de la Universidad, al coordinador de la carrera, al tutor y a los docentes implicados en el desarrollo de la investigación. Seguidamente agradecer a familiares o a aquellas personas que se quiera. También puede incluirse en la siguiente hoja una dedicatoria personal. A modo de ejemplo el contenido podría ser:

"En primer lugar dar gracias a la Universidad Nacional de Tres de Febrero (UNTREF), a su Rector Lic. Anibal Jozami, a todo su personal docente y no docente. Por promover un espacio ideal para el desarrollo de ideas y nuevos pensamientos y brindar a todos y cada uno de los alumnos, de esta casa de altos estudios, todos los recursos que esta institución dispone. Esta investigación no hubiera sido posible sin una formación académica acorde, por este motivo debo extender mi agradecimiento a los docentes de la carrera de Ingeniería de Sonido de la UNTREF, a su coordinador Ing. Alejandro Bibondo, que siendo la primera carrera de estas características del país, es muy importante contar con un cuerpo docente afín a las exigencias que este desafío propone, prestando su dedicación y vocación de enseñar. Un especial agradecimiento por la participación de esta tesis a la tutora Ing. Nombre Apellido, que supo transmitirme sus conocimientos y ayudarme a organizarme y fijarme un rumbo concreto y delineado, disponiendo desmedidamente de su tiempo. Por otra parte, quisiera hacer una mención especial al Ing. Hernan San Martin, que permitió el uso de las instalaciones de su laboratorio para poder trabajar y la disposición de todos sus recursos para que dicha investigación se realizara en tiempo y forma. Por último y no menos importante, quiero dar un afectuoso y cálido agradecimiento a mi familia…"

DEDICATORIA

Elige a quién o a qué quieres dedicárselo.

Elegir el motivo de la dedicatoria (orientativo).

ÍNDICE DE CONTENIDOS

Índice de figuras

Índice de tablas

RESUMEN

Su contenido no debe superar una página. Se indicarán los objetivos del trabajo, los métodos y resultados principales. A dos espacios debajo del resumen, en la misma página, se colocarán hasta 5 palabras clave que identifican los contenidos del trabajo.

Palabras Clave:

ABSTRACT

Ídem que para castellano.

Keywords:

1. INTRODUCCIÓN

1.1. FUNDAMENTACIÓN

Los modelos de texto a habla (TTS, por sus siglas en inglés) experimentaron un avance tecnológico exponencial en los últimos años: mediante redes neuronales profundas se alcanzaron resultados de elevada calidad sonora e inteligibilidad (Tan et al., 2021). No obstante, la fuerte dependencia de estos sistemas respecto a los datos de entrenamiento dificulta la obtención de voces sintetizadas con naturalidad para la gran diversidad de hablantes. Esta dificultad es especialmente notable en regiones con escasez de conjuntos de datos extensos, como ocurre en distintas provincias de Argentina.

En este marco, se han desarrollado sistemas de TTS en español rioplatense (Ortega Riera et al., 2023) que alcanzan resultados aceptables, pero se enfrentan a la limitada cantidad de datos específicos de los diferentes dialectos de Argentina, lo cual impide lograr sistemas más robustos y naturales. Tradicionalmente, la generación de bases de datos para entrenar modelos de TTS se orienta a recopilar grandes volúmenes de grabaciones de alta calidad (realizadas en estudios profesionales) y a emplear hablantes con características específicas (por ejemplo, locutores), lo que da lugar a un corpus homogéneo en sus características acústicas y prosódicas. Este enfoque fue crucial para la convergencia de modelos basados en aprendizaje profundo, pero representa una barrera de entrada para numerosos idiomas y variedades dialectales que no disponen de recursos para producir dichos datasets.

La literatura denomina “idiomas de bajos recursos” (low-resource languages) a estos casos; dentro de ellos se incluyen dialectos específicos de una lengua, como sería el español rioplatense o las variantes propias de determinadas provincias argentinas. Para entrenar modelos de TTS en lenguajes de bajos recursos se ha explorado la utilización de datos recolectados en Internet (Cooper, 2019), conformando conjuntos heterogéneos procedentes de diversas fuentes y de calidad de audio variable. Estos corpus suelen denominarse datos salvajes (ITW, “in-the-wild” por sus siglas en inglés). Además, con el avance de la inteligencia artificial generativa, han surgido diferentes mejoras en la arquitecturas de los sistemas de TTS mas actuales (Xie et al., 2024), lo que hace que los conjuntos de datos ITW sean una fuente especialmente atractiva para capturar la gran diversidad del fenómeno del habla.

El principal problema de entrenar modelos de TTS con conjuntos ITW es la elevada variabilidad en la calidad de las grabaciones, lo que incide directamente en la capacidad de los modelos neuronales para aprender los patrones subyacentes y, en muchos casos, impide la convergencia hacia resultados satisfactorios. Para abordar esta limitación, recientemente se han propuesto cadenas de preprocesamiento que extraen, a partir de un gran conjunto de datos, subgrupos con mejor calidad de audio (Yu et al., 2024). Si bien existen distintas variantes de estas cadenas en la literatura, no se ha llevado a cabo una caracterización acústica exhaustiva de la variabilidad que generan los conjuntos resultantes tras su aplicación. La validación suele basarse en el entrenamiento de modelos TTS y en la evaluación de su convergencia; sin embargo, no se suele caracterizar toda la cadena mediante parámetros acústicos que permitan comparar diferentes implementaciones bajo criterios comunes, ni definir configuraciones óptimas según objetivos distintos (por ejemplo, maximizar la calidad del audio frente a maximizar la cantidad de horas del corpus). El impacto de la calidad de los datos en el entrenamiento de modelos de TTS a sida profundamente estudiado (Ayllón et al., 2019), pero no se ha analizado las diferencias entre los dataset ITW y los dataset profesionales mediante un análisis objetivo.

La investigación propuesta en este trabajo tiene como objetivo determinar si es posible cuantificar la eficacia de estas cadenas de procesamiento mediante parámetros acústicos. Este tipo de análisis no solo facilita la iteración y la optimización de los procesos de filtrado de datos, sino que también abre la posibilidad de desarrollar con mayor facilidad bases de datos para lenguajes de bajos recursos, contribuyendo así a disponer de sistemas TTS de mayor calidad para una amplia variedad de idiomas y acentos locales.

1.2. OBJETIVOS

1.2.1. Objetivo general

El objetivo de la investigación es evaluar con parámetros objetivos y subjetivos, el impacto de cadenas de procesamiento de conjuntos de datos *in-the-wild* para el entrenamiento de modelos de texto a voz basados en redes neuronales profundas.

1.2.2. Objetivo especifico

Los objetivos específicos son:

- Crear un dataset *in-the-wild* en español de Argentina. Recopilar datasets de voces profesionales en español (grabaciones de alta calidad realizadas por hablantes profesionales).
- Desarrollar una cadena automática de preprocesamiento modular para la generación de conjuntos de datos de habla, y procesar el conjunto de datos ITW con la cadena bajo diferentes configuraciones operativas.
- Evaluar métricas acústicas en los distintos conjuntos de datos generados y comparar dichos resultados con los obtenidos en datasets tradicionales y determinar, según criterios acústicos, cuál de los conjuntos generados puede considerarse óptimo (comparando media y desvío de los diferentes conjuntos).
- Entrenar un modelo de estimación de distribuciones y comparar la similitud entre los diferentes conjuntos en el espacio latente. Determinar el conjunto de datos óptimo según criterios de similitud basados en estimación de densidad.
- Comparar los resultados del análisis acústico con los derivados del análisis por estimación de densidad. Analizar de forma estadística la relevancia de las diferencias observadas en los distintos parámetros.
- Validar los resultados en el contexto de clonación de voz mediante modelos TTS zero-shot.

1.3. ESTRUCTURA DE LA INVESTIGACIÓN

El trabajo propuesto corresponde a una investigación de carácter tecnológico orientada al desarrollo y evaluación de una herramienta de software para la selección automática de audios, destinada a la generación de conjuntos de datos de habla. El objetivo principal es crear un dataset en español con los diferentes acentos de Argentina, contribuyendo al avance de las tecnologías del habla en el país y, en consecuencia, a la soberanía tecnológica nacional. El desarrollo de esta tesis se enmarca en el proyecto Archivoz del grupo de investigación Intercambios Transorgánicos, radicado en el MUNTREF.

Organización del documento:

En el capítulo 2 se presenta el marco teórico: se exponen los fundamentos de la inteligencia artificial y se describen las arquitecturas aplicables a los modelos modernos de TTS, incluyendo tanto modelos secuenciales como modelos generativos. Además, se detallan las métricas acústicas seleccionadas para la caracterización de los datos.

El capítulo 3 ofrece una recapitulación de los modelos de TTS actuales y de las cadenas de procesamiento que han surgido en los últimos años.

En el capítulo 4 se describen con detalle las etapas del desarrollo: recopilación de datos, diseño y construcción del software, metodología de comparación propuesta y el entrenamiento de modelos mediante redes neuronales.

El capítulo 5 presenta los resultados y el análisis de los experimentos descritos en la sección anterior.

Finalmente, el capítulo 6 expone las conclusiones generales de la tesis, y el capítulo 7 propone líneas de investigación futuras y posibles aplicaciones no exploradas en el presente trabajo.

2. MARCO TEÓRICO

2.1. DESCRIPTORES DE CALIDAD DE AUDIO

2.1.1. Métricas de degradación de la señal

Explicar: PESQ, POLQA, SNR, SI-SDR, STOI

2.1.2. Métricas de entorno

Explicar: T30, C50, C80, D50

2.1.3. Métricas del habla

Explicar: F0, Speaker Rate, MCD

2.2. TEXT-TO-SPEECH (TTS)

Los sistemas de text-to-speech (TTS) convierten texto en señal de voz (Tan et al., 2021). Históricamente pueden agruparse en tres grandes enfoques:

- Enfoque concatenativo: Ensamblan fragmentos pre grabados de voz (unidades) para formar enunciados. Ofrecen alta naturalidad cuando el corpus es homogéneo y extenso, pero presentan baja flexibilidad y alto coste de recopilación (Hunt y Black, 1996).
- Enfoque paramétrico: Modelan parámetros acústicos (por ejemplo, mediante HMM) y luego sintetizan la señal a partir de los parámetros predichos. Tienen mayor flexibilidad y requieren un menor tamaño de corpus, aunque su calidad perceptual suele ser inferior a la voz grabada (Tokuda et al., 2013).
- Enfoque neuronal: Emplean redes neuronales para mapear texto a representaciones intermedias (p. ej. mel-espectrogramas) y vocoders neuronales para generar la forma de onda. Dentro de este grupo hay variantes auto regresivas (mayor fidelidad pero más lentas) y no-autoregresivas (más rápidas y escalables). Los sistemas actuales de mayor calidad combinan un modelo de predicción de espectrogramas, como pueden

ser Tacotron2 (Shen et al., 2018) o FastSpeech (Ren et al., 2021), con un vocoder neural, como pueden ser WaveNet (van den Oord et al., 2016) o HiFi-GAN (Kong et al., 2020).

2.3. REDES NEURONALES

Las redes neuronales son modelos parametrizados por capas de neuronas artificiales que aprenden funciones complejas a partir de datos (Goodfellow et al., 2016). En TTS y procesamiento de audio se emplean arquitecturas diversas: redes convolucionales (CNN) para extracción de características tiempo-frecuencia; redes recurrentes y Transformers (Vaswani et al., 2017) para modelado secuencial; y mecanismos de *attention* en tareas seq2seq.

Las redes permiten aprender mapeos directos (texto $\rightarrow$ espectrograma) y modelos generativos (vocoder, modelos de densidad). Su flexibilidad explica el salto cualitativo en TTS, pero también la fuerte dependencia de la cantidad y calidad de los datos de entrenamiento.

2.4. INTELIGENCIA ARTIFICIAL GENERATIVA

Completar

2.5. MODELOS DE DIFUSIÓN

Completar

3. ESTADO DEL ARTE

3.1. MODELOS DE TTS

Describir la parte de TTS basados en modelos de difusión hasta llegar a F5 TTS.

3.2. CADENAS DE PRE PROCESAMIENTO

En los últimos años se han desarrollado numerosas cadena de procesamiento, todas con diferentes configuraciones y particularidades. Para ilustrar las diferencias en las diferentes etapas se conforma la Tabla 1, donde se comparan diferencias de modelos, configuraciones y criterios en el desarrollo de cadena de pre procesamiento automático para la creación de datasets.

Tabla 1. Comparación entre diferentes etapas en una cadena de pre procesamiento para TTS.

Nombre del estudio	Algoritmo de Denoising	Voice Activity Detection	Estimador MOS y umbral	Sistema TTS evaluados
AutoPrep - (Yu et al., 2024)	BSRRN	TDNN	DNS MOS: 2.4	DurIAN TTS
Text-to-Speech in the wild - (Jung et al., 2025)	Demucs	Whisper X Pipeline	Nisqa: 3	GradTTS y VITS
WeNeetSpeech - (Ma et al., 2024)	MBTFNet	Rezamblyzer	DNS MOS: 3.6, 3.8, 4	VALL-E y NS2
SCEP - (Sabra et al., 2024)	U-Net	Casual DNN	Usan SNR y PESQ	No evalúa
Muyan TTS - (Li et al., 2024)	FRCRN y VoiceFixer	No usa	Nisqa: 3.8	FireRedTTS y CozyVoice2
Emilia - (He et al., 2024)	UVR-MDX-N et Inst	Silero VAD	DNS MOS: 3	VoiceBox

4. DESARROLLO

4.1. RECOPILACIÓN DE BASES DE DATOS

4.1.1. Datos *in-the-wild*

Para evaluar las cadenas de procesamiento sobre conjuntos de datos de habla se emplea el corpus en español de Argentina recopilado por el grupo de investigación Intercambios Transorgánicos. Este corpus consta de 24 horas de grabaciones realizadas en condiciones heterogéneas —tanto en calidad de audio como en diversidad de hablantes— y proviene mayoritariamente de fuentes públicas en internet. Por su variabilidad y carácter no controlado, este conjunto *in-the-wild* resulta idóneo para validar procedimientos de preprocesado de audio. En lo sucesivo, se hará referencia a esta colección como el conjunto de datos *original* (versión sin procesamientos).

Con el objetivo de captar diferencias dialectales relevantes para Argentina, la selección de hablantes sigue la clasificación regional propuesta por de Weinberg y de Mirande, 2004. En concreto, el corpus incluye 32 hablantes con acento bonaerense y 27 con acento centro, para un total de 59 hablantes. La estrategia a futuro consiste en ampliar la cobertura dialectal para incorporar todas las variedades representativas del país; sin embargo, en esta tesis se valida inicialmente la cadena de preprocesamiento sobre estas dos variantes representativas.

4.1.2. Datos profesionales

Para complementar el corpus *in-the-wild* y disponer de material de mayor calidad contra el cual contrastar el análisis objetivo, se incorporan conjuntos de datos profesionales disponibles en trabajos previos. Dado que no existe una colección pública extensa exclusivamente de español de Argentina, se incluyen también recursos con variantes dialectales cercanas cuando procede.

- ELRA Dataset (Guevara-Rukoz et al., 2020): aproximadamente 8 horas de audio con dialecto bonaerense, 44 hablantes.
- Emilia (Torres et al., 2019): aproximadamente 4 horas de audio (dialecto bonaerense).

- HaCASpa Dataset (): aproximadamente 10 horas de audio con dialecto bonaerense, 50 hablantes.

  El conjunto profesional suma (determinar si agrego más data o no)

4.2. DESARROLLO DE LA CADENA DE PRE PROCESAMIENTO

La cadena de preprocesamiento diseñada en esta Tesis es una secuencia modular y reproducible de etapas destinadas a transformar material *in-the-wild* en subconjuntos utilizables para entrenamiento de TTS. Las etapas principales son las siguientes:

1. Post-procesado y metadatos: Normalización de niveles, etiquetado de metadatos (origen, duración, dialecto, condiciones de captura) y generación de los subconjuntos finales para evaluación y para alimentación a modelos TTS.

2. Voice Activity Detection (VAD): Eliminación de segmentos sin voz y segmentación inicial en enunciados. En la implementación se usa Silero VAD (Team, 2024) con una estrategia adaptativa de optimización de hiperparámetros basada en clasificar el ritmo de habla (lento/normal/rápido) mediante los timestamps de Whisper y optimizar los parámetros del VAD por categoría. Además, se controla la longitud final de las unidades (concatenación / recorte) para ajustarlas a la distribución requerida por modelos TTS downstream.

3. Denoising / Speech enhancement: Aplicación opcional de la cadena para mitigar ruido y artefactos de grabación. Se evalúan modelos prácticos y eficientes en CPU, además de la opción sin denosing, considerando el balance entre mejora perceptual y preservación de la identidad vocal. Los modelos seleccionados son Demucs (Défossez et al., 2020) y DeepFilterNet (Schroter et al., 2022).

4. Filtrado de calidad no intrusivo: Evaluación y filtrado mediante modelos de calidad perceptual no intrusiva para aceptar o rechazar fragmentos según puntuación mínima. Se utiliza modelos predictivos de MOS (Mean-Opinion-Score) como métrica conjunta de calidad. Se evaluan los modelos NISQA (Mittag et al., 2021) y DNS MOS (Reddy et al., 2021).

5. Transcripción (STT): Obtención de transcripciones automáticas necesario para obte-

ner los pares texto-audio que se requieren para el entrenamiento; en esta etapa se prioriza la precisión a costa de un mayor tiempo de cómputo, dado el impacto de errores de transcripción en la calidad final del corpus. Se utiliza el modelo de Whisper Large para la transcripción (Radford et al., 2023).

La implementación enfatiza portabilidad y bajo costo computacional, de modo que grupos con recursos limitados puedan reproducir la cadena y comparar variantes de configuración sin necesidad de utilizar hardware de alto rendimiento como serían GPUs. El diagrama de flujo de toda la cadena propuesta se presenta en la Figura 1.

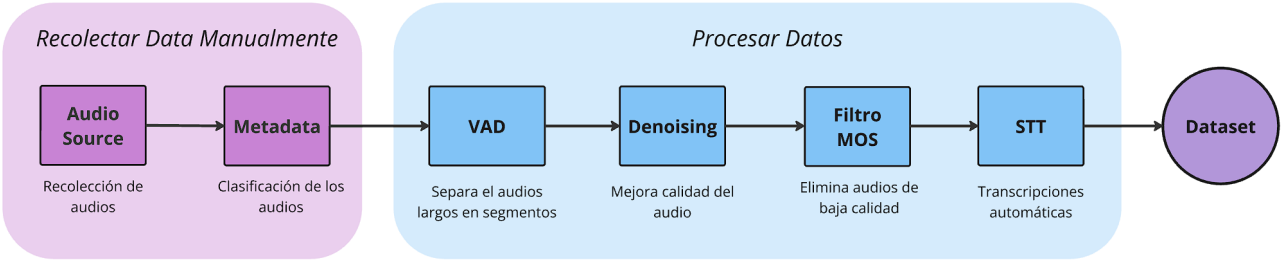

Figura 1. Diagrama de flujo del proceso completo para generar el dataset.

4.2.1. Diferentes configuraciones

Para evaluar el impacto de decisiones de diseño en la confección de cadenas de pre procesamiento, se proponen las siguientes variantes:

- Condiciones de denoising: DeepFilterNet (DFN), Demucs, y *no-denoising*. Estas alternativas representan puntos intermedios entre eficiencia, mejora perceptual y preservación de timbre.
- Modelos de calidad no intrusiva: NISQA y DNSMOS, seleccionados por su uso extendido y su diferente sensibilidad a condiciones de ruido. Estos modelos son las mas usados en la literatura, donde se utiliza uno u el otro pero no se han contrastado para determinar el modelo mas optimo.
- Umbrales de filtrado: Para NISQA se evaluaron (3.0, 3.5, 3.8, 4.2) y para DNSMOS (2.7, 3.0, 3.2, 3.4). Los umbrales se eligieron de forma empírica analizando el nivel predicho

de MOS para todo el conjunto de datos original.

La selección de umbrales parte de un análisis empírico presentado en la Figura 2, donde se calcula el valor de MOS resultante de NISQA y DNS MOS para todos los segmentos del dataset original. Es interesante notar como DNS MOS presenta un valor medio inferior y menos varianza. En consecuencia, para hacer una comparación justa, se seleccionaron los umbrales para garantizar un filtrado equitativo para los dos modelos.

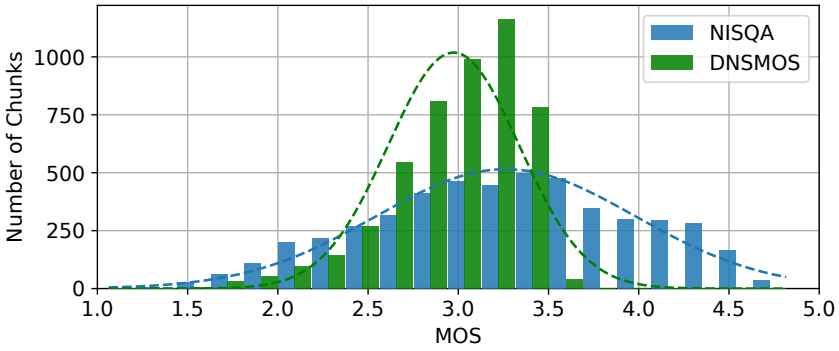

Figura 2. Resultado de MOS por cantidad de segmentos del dataset original para los 2 modelos a comparar.

En la Figura 3 se presentan de manera gráfica las diferentes variantes a analizar. Cada configuración genera un subconjunto procesado sobre el que se calculan las métricas objetivas (ver sección siguiente) para permitir una comparación reproducible y dirigida por métrica sin necesidad de entrenar modelos TTS para cada variante. A estas variantes de procesamiento se las denomina sub dataset.

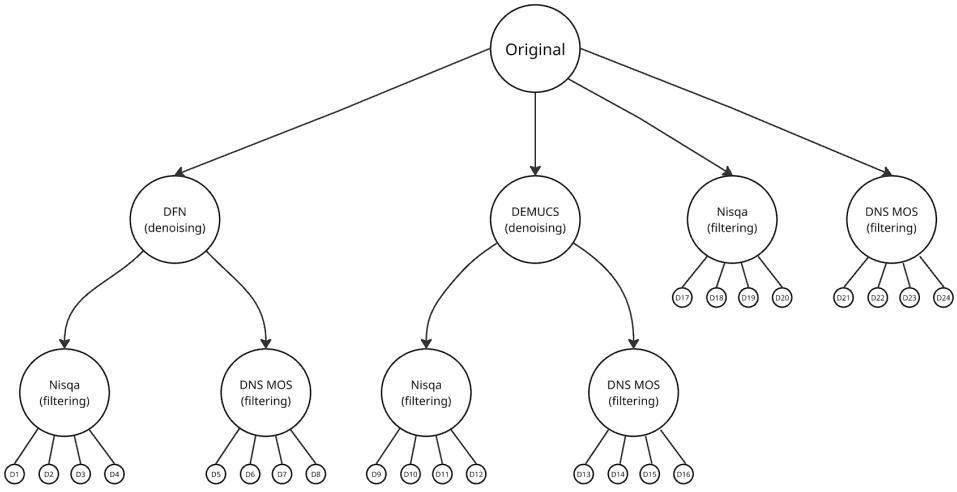

Figura 3. Variantes propuestas para evaluar diferentes configuración de la cadena de procesamientos.

4.3. EVALUACIÓN DE LOS CONJUNTOS DE DATOS

La evaluación se articula alrededor de cuatro bloques de métricas complementarias que capturan cantidad, calidad de señal, condiciones acústicas y preservación de características del hablante. Estas métricas se combinan en una métrica compuesta que permite ordenar y seleccionar la mejor configuración según criterios definidos (ver formulación matemática al final).

Todas las métricas buscan cuantificar diferencias relativas entre el dataset original y los sub datasets resultantes de las diferentes variantes; de esta forma, se pueden aplicar esta misma metodología de procesamiento para caracterizar el funcionamiento de la cadena de procesamientos, independientemente de las características del conjunto de datos de partida.

4.3.1. Reducción del corpus

El tamaño del dataset se mide en duración total (horas) y la *reducción de datos* (RD) se cuantifica según Ecuación 1, donde el subíndice P hace referencia a las diferentes variantes del dataset procesado, y el subíndice O hace referencia a la cantidad de horas del dataset original.

$$RD_P = 1 - \frac{\text{HORAS}_P}{\text{HORAS}_O} \tag{1}$$

Un valor de RD más pequeño es preferible (menos pérdida de datos). Esta medida captura el trade-off básico entre condiciones del filtrado agresivas y cantidad de material utilizable.

4.3.2. Calidad de la grabación

La evaluación de la calidad de la grabación tiene por objetivo cuantificar, de forma objetiva y reproducible, las mejoras perceptuales y la fidelidad de la señal obtenidas tras aplicar las distintas etapas de la cadena (por ejemplo, denoising y filtrado). Para ello se emplean medidas complementarias: PESQ, SI-SDR y SNR; en un principio se contempla la

posibilidad de utilizar también el STOI como un parámetro para cuantificar inteligibilidad, pero empíricamente no se encontraron diferencias significativas de este valor, con lo cual fue descartado.

PESQ (Perceptual Evaluation of Speech Quality) se utiliza como un indicador aproximado de la calidad subjetiva de la señal comparando la versión procesada con la referencia original; SI-SDR (Scale-Invariant Signal-to-Distortion Ratio) mide la fidelidad de la señal de forma robusta frente a cambios de escala y permite evaluar cuánto contenido de la señal original se preserva tras el procesamiento; la definición teórica para calcular estos parámetros requiere de tener una versión original de la señal y una versión degradada, como en este caso se quiere comparar ambos dataset de forma independiente, se utiliza el modulo Pytorch Squim (Kumar et al., 2023), que permite predecir los valores de PESQ, SI-SDR de forma no intrusiva (estimación ciega).

El SNR se estima mediante WADA-SNR (Kim y Stern, 2008) para obtener una cifra robusta de relación señal-ruido basada en la distribución de la amplitud de la onda. Todas estas métricas se computan a nivel de segmento y luego se resumen mediante la media y la desviación estándar para cada subconjunto, de modo que sea posible comparar distribuciones antes y después del procesamiento. Finalmente, las puntuaciones individuales se integran en el bloque *Calidad señal* (CS) donde se suma la influencia de todos estos valores Ecuación 2, teniendo en cuenta que se espera que los valores de PESQ, SI-SDR y SNR suban (mejoría) al aplicar la cadena de procesamientos. Los subindices respetan la condiciones anterior donde P es por procesado y O es original, estos subíndices se mantiene consistentes para todas las métricas.

$$CS_P = \frac{\text{PESQ}_O}{\text{PESQ}_P} + \frac{\text{SI-SDR}_O}{\text{SI-SDR}_P} + \frac{\text{SNR}_O}{\text{SNR}_P} \tag{2}$$

4.3.3. Condiciones acústicas

Las métricas acústicas buscan describir las condiciones de sala y la presencia de reverberación o de energía tardía en las grabaciones, aspectos que afectan la utilidad de los audios para entrenamiento de TTS, especialmente la alineación temporal entre texto y mel-spectrograma. En este trabajo se emplean descriptores clásicos como T_{30} (tiempo de rever-

beración aproximado) y medidas de claridad como C_{50} y D_{50}, que resumen la proporción de energía inicial frente a la energía reverberada y permiten detectar grabaciones con exceso de reverberación o mala claridad.

Nuevamente, es necesario calcular estos parámetros de forma ciega y sin referencia de las condiciones del entorno original,con estas limitaciones, estas métricas se estiman mediante un modelo CNN que fue entrenado para calcular parámetros acústicos de forma ciega, y fue validado para voces del español argentino, de manera que la estimación sea práctica sobre material *in-the-wild* sin requerir respuestas impulsivas de sala (Ortiz, 2023). Las mediciones se calculan por segmento y se agregan mediante estadísticos (media y desvío) para cada subconjunto; en la métrica compuesta se incorporan las mejores relativas para evaluar si una configuración reduce la reverberación y mejora la claridad respecto del conjunto original.

En este caso, estas dos métricas se suman para contabilizar la influencia de las *Condiciones acústicas* (CA) en la Ecuación 3, donde el T_{30} mejora si disminuye, pero el C_{50} mejora si se incrementa. Se descarta el D_{50} ya que los resultados empíricos fueron muy similares al análisis del C_{50}, con lo cual no se estaría agregando información redundante y se decide en consecuencia eliminar el aporte de esta métrica. El análisis de estas variables permite discriminar configuraciones que mejoran la «condición de sala» del corpus sin sacrificar en exceso otros atributos.

$$CA_P = \frac{T_{30,P}}{T_{30,O}} + \frac{C_{50,O}}{C_{50,P}} \tag{3}$$

4.3.4. Diferencias del habla

Las medidas de diferencias del habla están pensadas para garantizar que las etapas de preprocesamiento, en particular los algoritmos de denoising, no alteren indebidamente la identidad del hablante ni la variabilidad prosódica del corpus, aspectos críticos para síntesis con preservación de timbre y estilo. En la tesis se consideran dos indicadores principales: la desviación estándar del F0 (F0-STD) y la distorsión mel-cepstral media (MCD).

El F0-STD captura la variabilidad prosódica de los hablantes; su cálculo se realiza a nivel de segmento usando el estimador PESTO (Riou et al., 2023) para extracción robusta de pitch,

y se compara entre versión original y procesada para detectar reducciones anómalas de variabilidad que indiquen pérdida de naturalidad o sesgo en la selección de hablantes. El MCD se calcula entre los coeficientes mel-cepstrales de la señal original y de la señal procesada (especialmente relevante en audios sometidos a denoising) para cuantificar cambios tímbricos (Kominek et al., 2008); en la formulación adoptada el MCD se expresa como incremento porcentual o normalizada respecto a un valor de referencia aceptable de 5 dB según (Xie et al., 2024), de esta forma se penalizan las configuraciones que alteran significativamente el timbre.

Estas dos componentes se combinan en el bloque *Diferencias del habla* (DH), donde cualquier diferencia relativa de F0-STD es penalizada, y lo mismo para el MCD respecto al valor de referencia Ecuación 4. Cabe aclarar que el MCD solo se calcula para las variantes con denoising, esto favorece de alguna forma a las variantes sin denoising que no modifican las características del hablante.

$$SD_P = \left| 1 - \frac{\text{F0std}_P}{\text{F0std}_O} \right| \; + \; \frac{\text{MCD}_P}{5} \tag{4}$$

4.3.5. Métrica conjunta

Para ordenar configuraciones se propone una métrica compuesta que suma los cuatro bloques anteriores: reducción del dataset (RD), calidad de señal (CS), condiciones acústicos (CA) y diferencias de habla (DH). Esta métrica compuesta, plantea de forma matemática las relaciones de compromiso que se asumen con la cadena de procesamientos.

Se plantea la métrica compuesta como un objetivo de optimización Ecuación 5, donde se busca minimizar el aporte negativo de cada variable para poder identificar la mejor variante P de la cadena de procesamientos.

$$\min_{P \in \text{Conf}} \left\{ RD_P \; + \; CS_P \; + \; CA_P \; + \; DH_P \right\} \tag{5}$$

Con esta formulación, configuraciones que mantienen mayor cantidad de horas, mejoran la calidad de señal, mejoran (o mantienen) condiciones acústicas y preservan características del hablante obtendrán puntuaciones más bajas (mejor). Se propone una configuración

inicial donde todos los bloques tienen el mismo peso; la elección de pesos distintos permite priorizar, por ejemplo, la preservación vocal frente a la cantidad de horas si el objetivo es síntesis con alta fidelidad de identidad.

4.4. ENTRENAMIENTO DEL MODELO DE ESTIMACIÓN DE DENSIDAD

Explicar porque necesito hacer lo del modelo de estimación de densidad

4.4.1. Validación con medelo zero-shot

Definir modelo y explicar la justificación de este experimento

4.5. DESCRIPCIÓN DE PRUEBAS ESTADÍSTICAS

Detallar un poco las validaciones estadísticas que se van a realizar

4.6. MODELO DE TTS ZERO-SHOT

Definir el modelo de TTS a usar, el porque de la selección y la descripción del último experimento

5. RESULTADOS Y ANÁLISIS

5.1. RESULTADOS DE LAS DIFERENTES VARIANTES DE LA CADENA

5.1.1. Métricas separadas

Primero se evalúa el impacto de las diferentes métricas por separado según las diferentes configuraciones. Al comparar la mejora de PESQ respecto a la reducción del dataset Figura 4, se observa que

Primero es compara la relación entre calidad de audio y cantidad de horas del dataset. La Figura 4 muestra, en el eje horizontal, el porcentaje de reducción del dataset producido por el filtrado (más a la derecha significa pérdida de más horas) y, en el eje vertical, el incremento relativo de la puntuación PESQ expresado en porcentaje. Sobre la misma gráfica se comparan las tres variantes de la etapa de realce consideradas: DeepFilterNet, Demucs y la condición sin denoise. El comportamiento observado evidencia un trade-off claro entre cantidad y calidad: umbrales de filtrado más estrictos producen mejoras mayores en PESQ pero sacrifican un mayor número de horas de grabación. En concreto, las variantes con denoising aportan ganancias moderadas en PESQ (típicamente inferiores al 8 % según el umbral), mientras que la variante sin realce logra aumentos mayores en PESQ (superiores al 10 %) a costa de una reducción de dataset mucho más agresiva. Entre los métodos de realce usados, Demucs tiende a ofrecer las mayores mejoras de PESQ a lo largo de los umbrales evaluados, y DeepFilterNet presenta ganancias más pequeñas para filtros equivalentes; esta diferencia implica que la elección del algoritmo de realce y del umbral de NISQA debe hacerse en función del balance deseado entre calidad perceptual y preservación de horas útiles.

La Figura 3 relaciona la reducción porcentual de horas con la mejora relativa en T_{30} (tiempo de reverberación estimado) cuando se emplea DNSMOS como métrica de filtrado. En la gráfica, un aumento en la mejora de T_{30} implica una reducción de la energía tardía y, por ende, una menor reverberación aparente en los segmentos retenidos. Los resultados muestran que el filtrado aporta ganancias en condiciones sin denoise de forma especialmente notable —es decir, cuando no se aplica realce previo el filtrado selecciona segmentos con menor reverberación y la mejora en T_{30} es sustancial—; en las condiciones ya denoised las ganancias en T_{30} son más moderadas (por debajo aproximadamente del 5 % en los umbrales

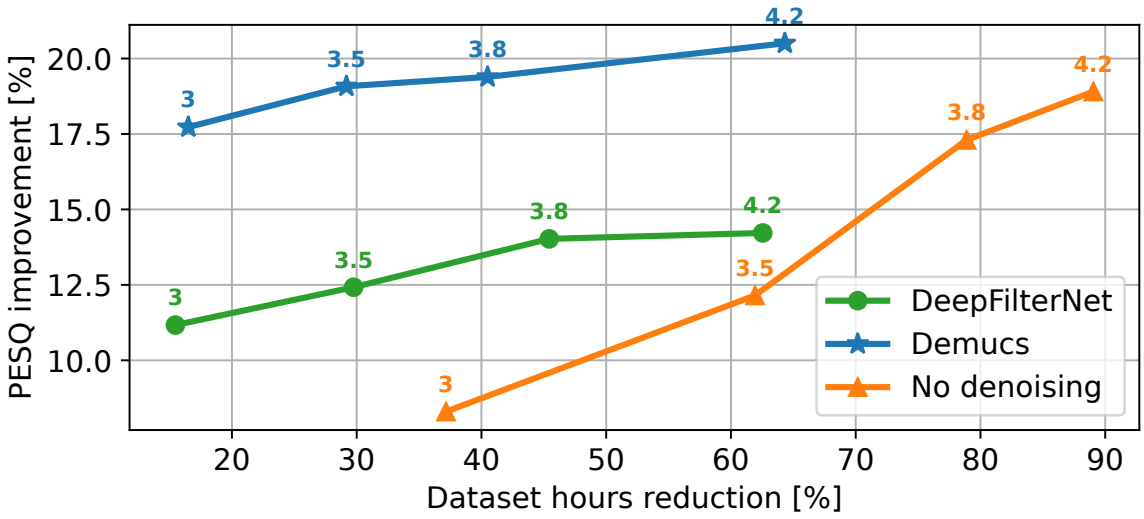

Figura 4. Comparación entre PESQ y cantidad de horas de las variantes.

considerados). Además, no se detecta una diferencia significativa en T_{30} entre DeepFilterNet y Demucs, aunque Demucs presenta una ventaja apreciable (alrededor de 5 %) en medidas de claridad como C_{50}. En conjunto, la figura subraya que el filtrado basado en DNSMOS puede reducir la reverberación del subconjunto resultante, pero el beneficio marginal depende fuertemente de si previamente se aplicó un algoritmo de realce y de la agresividad del umbral. :contentReference[oaicite:1]index=1

La Figura 4 ilustra la relación entre la mejora relativa de PESQ (eje horizontal) y la diferencia porcentual en la desviación estándar de $F0$ ($F0\text{-}STD$, eje vertical) para las variantes evaluadas con DNSMOS. Esta representación revela un efecto colateral relevante: a medida que aumenta la mejora perceptual (PESQ) por filtrado más estricto, suele observarse una reducción de la variabilidad prosódica medida por $F0\text{-}STD$. La explicación práctica es que el filtrado agresivo tiende a eliminar segmentos y hablantes de peor calidad, lo cual reduce la heterogeneidad prosódica del subconjunto retenido. No obstante, los métodos de realce que mejor preservan el timbre de voz muestran cambios menores en $F0\text{-}STD$ ante el mismo grado de filtrado, lo que indica que ciertos denoisers permiten mejorar la calidad percibida sin sacrificar tanto la variabilidad prosódica. En la práctica, la figura pone de manifiesto el compromiso entre mejorar la calidad objetiva del audio y mantener la diversidad de patrones de entonación: selección excesiva orientada únicamente a PESQ puede empobrecer la variabilidad del corpus, con posible impacto en tareas de síntesis que requieran conservar rasgos prosódicos y de identidad. Además, en el paper se reporta que la MCD varía poco con el filtrado y que Demucs típicamente obtiene una MCD algo menor que

DeepFilterNet, lo que refuerza la recomendación de evaluar simultáneamente métricas de calidad y métricas de preservación vocal antes de fijar una configuración final. :contentReference[oaicite:2]index=2

5.1.2. Métricas conjunta

Hacer el ranking de todas las variantes según el resultado de optimización.

5.2. COMPARACIÓN ENTRE CONJUNTOS DE DATOS

Comparar entre conjuntos de datos profesionales vs datos ITW (por parametros acústicos)

5.3. COMPARACIÓN POR ESTIMACIÓN DE DENSIDAD

Comparar entre conjuntos de datos profesionales vs datos ITW (por estimación de densidad)

5.4. VALIDACIÓN MODELO DE TTS ZERO-SHOT

Compara calidad de clonación respecto a la calidad o subset del audio original

6. CONCLUSIONES

En las conclusiones del Plan de Investigación, debe plantearse cómo será la exposición de los resultados y qué es lo que se espera obtener en resumen de las pruebas que se realicen.

7. LÍNEAS FUTURAS DE INVESTIGACIÓN

Este trabajo pretende contribuir a la unificación de criterios en el diseño y evaluación de cadenas de preprocesado, lo que facilitará la identificación de las configuraciones más adecuadas para distintos casos de uso. Entre las líneas futuras de investigación se destacan, en particular, el desarrollo de modelos de denoising o speech enhancement personalizados: mediante técnicas de adaptación, dichos modelos buscarían aproximar conjuntos de grabaciones diversas hacia el dominio acústico del corpus con el que fue entrenado el modelo TTS, aumentando así la robustez y la fidelidad de la síntesis en condiciones reales.

Bibliografia

Ayllón, D., Sánchez-Hevia, H., Figueroa, C., & Lanchantin, P. (2019). Investigating the Effects of Noisy and Reverberant Speech in Text-to-Speech Systems. *Proc. Interspeech*, 1511-1515. https://doi.org/10.21437/Interspeech.2019-3104

Cooper, E. (2019,). *Text-to-Speech Synthesis Using Found Data for Low-Resource Languages* [PhD. thesis]. Columbia University.

Défossez, A., Synnaeve, G., & Adi, Y. (2020). Real Time Speech Enhancement in the Waveform Domain. *Proc. Interspeech*, 3291-3295. https://doi.org/10.21437/Interspeech.2020-2409

de Weinberg, M., & de Mirande, N. (2004). *El español de la Argentina y sus variedades regionales*. Asociación Bernardino Rivadavia, Proyecto Cultural Weinberg/Fontanella. https://books.google.com.ar/books?id=fpxiAAAAMAAJ

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press.

Guevara-Rukoz, A., Demirsahin, I., He, F., Chu, S.-H. C., Sarin, S., Pipatsrisawat, K., Gutkin, A., Butryna, A., & Kjartansson, O. (2020). Crowdsourcing Latin American Spanish for Low-Resource Text-to-Speech. *Language Resources and Evaluation Conference (LREC)*, 6504-6513. https://aclanthology.org/2020.lrec-1.801/

He, H., Shang, Z., Wang, C., Li, X., Gu, Y., Hua, H., Liu, L., Yang, C., Li, J., Shi, P., Wang, Y., Chen, K., Zhang, P., & Wu, Z. (2024). Emilia: An Extensive, Multilingual, and Diverse Speech Dataset For Large-Scale Speech Generation. *IEEE Spoken Language Technology Workshop (SLT)*, 885-890. https://doi.org/10.1109/SLT61566.2024.10832365

Hunt, A. J., & Black, A. W. (1996). Unit selection in a concatenative speech synthesis system using a large speech database. *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 373-376.

Jung, J., Zhang, W., Maiti, S., Wu, Y., Wang, X., Kim, J.-H., Matsunaga, Y., Um, S., Tian, J., Shim, H.-j., Evans, N., Chung, J. S., Takamichi, S., & Watanabe, S. (2025). The Text-to-speech in the Wild (TITW) Database. *Proc. Interspeech*, 4798-4802. https://doi.org/10.21437/Interspeech.2025-2536

Kim, C., & Stern, R. (2008). Robust signal-to-noise ratio estimation based on waveform amplitude distribution analysis. *Proc. Interspeech*, 2598-2601.

Kominek, J., Schultz, T., & Black, A. W. (2008). Synthesizer voice quality of new languages calibrated with mean mel cepstral distortion. *Proc. Speech Technology Under-Resourced Languages*, 63-68.

Kong, J., Kim, J., & Bae, J. (2020). HiFi-GAN: generative adversarial networks for efficient and high fidelity speech synthesis. *Proceedings of the 34th International Conference on Neural Information Processing Systems*.

Kumar, A., Tan, K., Ni, Z., Manocha, P., Zhang, X., Henderson, E., & Xu, B. (2023). Torchaudio-Squim: Reference-Less Speech Quality and Intelligibility Measures in Torchaudio. *IEEE International Conference of Acoustics, Speech and Signal Processing (ICASSP)*, 1-5. https://doi.org/10.1109/ICASSP49357.2023.10096680

Li, X., Jia, K., Sun, H., Dai, J., & Jiang, Z. (2024). Muyan-TTS: A Trainable Text-to-Speech Model Optimized for Podcast Scenarios with a $50K Budget. *arXiv preprint arXiv:2504.19146*. https://arxiv.org/abs/2504.19146

Ma, L., Guo, D., Song, K., Jiang, Y., Wang, S., Xue, L., Xu, W., Zhao, H., Zhang, B., & Xie, L. (2024). WenetSpeech4TTS: A 12,800-hour Mandarin TTS Corpus for Large Speech Generation Model Benchmark. *Proc. Interspeech*, 1840-1844. https://doi.org/10.21437/Interspeech.2024-2343

Mittag, G., Naderi, B., Chehadi, A., & Möller, S. (2021). NISQA: A Deep CNN-Self-Attention Model for Multidimensional Speech Quality Prediction with Crowdsourced Datasets. *Proc. Interspeech*, 2127-2131. https://doi.org/10.21437/Interspeech.2021-299

Ortega Riera, P., Passano, N., Paez, D., Bach, F., Pupkin, I., Sacerdoti, E., Yommi, M., & Martín, H. (2023). Implementación y Evaluación de un Sistema de Clonación de Voz Rioplatense para Asistencia en la Comunicación Oral. *Jornadas de Acústica, Audio y Sonido*.

Ortiz, M. (2023). Estimación ciega de parámetros acústicos de un recinto. *Master Thesis*.

Radford, A., Kim, J., Xu, T., Brockman, G., Mcleavey, C., & Sutskever, I. (2023). Robust Speech Recognition via Large-Scale Weak Supervision. *Proceedings of the 40th International Conference on Machine Learning*, *202*, 28492-28518.

Reddy, C. K. A., Gopal, V., & Cutler, R. (2021). Dnsmos: A Non-Intrusive Perceptual Objective Speech Quality Metric to Evaluate Noise Suppressors. *IEEE International Conference of Acoustics, Speech and Signal Processing (ICASSP)*, 6493-6497. https://doi.org/10.1109/ICASSP39728.2021.9414878

Ren, Y., Hu, C., Tan, X., Qin, T., Zhao, S., Zhao, Z., & Liu, T.-Y. (2021). FastSpeech 2: Fast and High-Quality End-to-End Text to Speech. *International Conference on Learning Representations (ICLR)*.

Riou, A., Lattner, S., Hadjeres, G., & Peeters, G. (2023). PESTO: Pitch Estimation with Self-supervised Transposition-equivariant Objective. *Proceedings of the 24th International Society for Music Information Retrieval Conference (ISMIR)*.

Sabra, A., Wronka, C., Mao, M., & Hijazi, S. (2024). SECP: A Speech Enhancement-Based Curation Pipeline for Scalable Acquisition of Clean Speech. *IEEE International Conference of Acoustics, Speech and Signal Processing (ICASSP)*, 11981-11985. https://doi.org/10.1109/ICASSP48485.2024.10446973

Schroter, H., Escalante-B, A. N., Rosenkranz, T., & Maier, A. (2022). Deepfilternet: A Low Complexity Speech Enhancement Framework for Full-Band Audio Based On Deep Filtering. *IEEE International Conference of Acoustics, Speech and Signal Processing (ICASSP)*, 7407-7411. https://doi.org/10.1109/ICASSP43922.2022.9747055

Shen, J., Pang, R., Weiss, R. J., Schuster, M., Jaitly, N., Yang, Z., Chen, Z., Zhang, Y., Wang, Y., Skerry-Ryan, R. J., Saurous, R. A., Agiomyrgiannakis, Y., & Wu, Y. (2018). Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 4779-4783. https://doi.org/10.1109/ICASSP.2018.8461368

Tan, X., Qin, T., Soong, F. K., & Liu, T.-Y. (2021). A Survey on Neural Speech Synthesis. *arXiv preprint arXiv:2106.15561*. https://arxiv.org/abs/2106.15561

Team, S. (2024). Silero VAD: pre-trained enterprise-grade Voice Activity Detector (VAD), Number Detector and Language Classifier.

Tokuda, K., Nankaku, Y., Toda, T., Zen, H., Yamagishi, J., & Oura, K. (2013). Speech synthesis based on hidden Markov models. *Proceedings of the IEEE*, *101*(5), 1234-1252. https://doi.org/10.1109/JPROC.2013.2251852

Torres, H. M., Gurlekian, J. A., Evin, D. A., & Cossio Mercado, C. G. (2019). Emilia: a speech corpus for Argentine Spanish text to speech synthesis. *Language Resources and Evaluation*, *53*(3), 419-447.

van den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., & Kavukcuoglu, K. (2016). WaveNet: A Generative Model for Raw Audio. *9th ISCA Workshop on Speech Synthesis Workshop (SSW 9)*, 125.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention Is All You Need [NeurIPS 2017]. *Advances in Neural Information Processing Systems*.

Xie, T., Rong, Y., Zhang, P., & Liu, L. (2024). Towards Controllable Speech Synthesis in the Era of Large Language Models: A Survey. *arXiv preprint arXiv:2412.06602*. https://arxiv.org/abs/2412.06602

Yu, J., Chen, H., Bian, Y., Li, X., Luo, Y., Tian, J., Liu, M., Jiang, J., & Wang, S. (2024). AutoPrep: An Automatic Preprocessing Framework for In-The-Wild Speech Data. *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1136-1140. https://doi.org/10.1109/ICASSP48485.2024.10447759

ANEXO I. FORMATO INTERNO

AI 1. *Numeración*

Las páginas serán enumeradas a partir del Índice de Contenidos, con números romanos colocados en la parte media inferior de cada página. A partir de la Introducción, todas las páginas serán enumeradas con números arábigos ubicados en la parte inferior derecha. No usar la palabra “página” antes de la numeración de las páginas.