

Evaluación de cadenas de procesamiento 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: Guillermo Marzik (Ing.)

COTUTOR/A: Gala Lucía Gonzalez Barrios (Lic.)

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

RESUMEN	IX
ABSTRACT	X
1 INTRODUCCIÓN	1
1.1 FUNDAMENTACIÓN	1
1.2 OBJETIVOS	2
1.2.1 Objetivo general	2
1.2.2 Objetivos específicos	3
1.3 ESTRUCTURA DE LA INVESTIGACIÓN	3
2 MARCO TEÓRICO	5
2.1 DESCRIPTORES DE CALIDAD DE AUDIO	5
2.1.1 Métricas de degradación de la señal	5
2.1.2 Métricas de entorno	6
2.1.3 Métricas del habla	7
2.2 TEXT-TO-SPEECH (TTS)	8
2.3 REDES NEURONALES	9
2.4 INTELIGENCIA ARTIFICIAL GENERATIVA	9
2.5 MODELOS DE DIFUSIÓN	10
3 ESTADO DEL ARTE	12
3.1 MODELOS DE TTS	12
3.2 CADENAS DE PRE PROCESAMIENTO	12
3.3 CONJUNTOS DE DATOS DEL HABLA EN ESPAÑOL	12
4 DESARROLLO	14
4.1 RECOPIACIÓN DE BASES DE DATOS	14
4.1.1 Datos <i>in-the-wild</i>	14
4.1.2 Datos profesionales	14
4.2 DESARROLLO DE LA CADENA DE PRE PROCESAMIENTO	15
4.2.1 Diferentes configuraciones	16
4.3 EVALUACIÓN DE LOS CONJUNTOS DE DATOS	18

4.3.1	Reducción del corpus	18
4.3.2	Calidad de la grabación	19
4.3.3	Condiciones acústicas	20
4.3.4	Diferencias del habla	21
4.3.5	Métrica conjunta	22
4.4	ENTRENAMIENTO DEL MODELO DE ESTIMACIÓN DE DENSIDAD	22
4.4.1	Validación con medelo zero-shot	22
4.5	DESCRIPCIÓN DE PRUEBAS ESTADÍSTICAS	22
4.6	MODELO DE TTS ZERO-SHOT	23
5	RESULTADOS Y ANÁLISIS	24
5.1	RESULTADOS DE LAS DIFERENTES VARIANTES DE LA CADENA	24
5.1.1	Ejemplo de cálculo	24
5.1.2	Métricas separadas	25
5.1.3	Métricas compuestas	27
5.1.4	Optimización de la cadena	28
5.1.5	Análisis del peso de cada métrica	29
5.2	COMPARACIÓN ENTRE CONJUNTOS DE DATOS	29
5.3	COMPARACIÓN POR ESTIMACIÓN DE DENSIDAD	30
5.4	VALIDACIÓN MODELO DE TTS ZERO-SHOT	30
6	CONCLUSIONES	31
7	LÍNEAS FUTURAS DE INVESTIGACIÓN	32
	BIBLIOGRAFÍA	33

Índice de Figuras

Figura 1. Diagrama de flujo del proceso completo para generar el dataset.	16
Figura 2. Resultado de MOS por cantidad de segmentos del dataset original para los 2 modelos a comparar.	17
Figura 3. Variantes propuestas para evaluar diferentes configuración de la cadena de procesamientos.	18
Figura 4. Comparación entre PESQ y cantidad de horas de diferentes variantes con NISQA.	25
Figura 5. Comparación entre T30 y cantidad de horas de diferentes variantes con DNS MOS.	26
Figura 6. Comparación entre F0-std y PESQ para diferentes variantes con DNS MOS.	27
Figura 7. Comparación entre reducción de datos y calidad de señal para diferentes variantes con Nisqa.	28
Figura 8. Comparación entre calidad de señal y condiciones acústicas para diferentes variantes con DNSMOS.	28

Índice de Tablas

Tabla 1. Comparación entre diferentes etapas en una cadena de pre procesamiento para TTS.	12
Tabla 2. Resumen de datasets	13
Tabla 3. Mejora SNR por métrica de calidad y algoritmo de denoising.	24
Tabla 4. Métricas compuestas y total para todas las configuraciones.	29

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 sido 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 esta tesis tiene como objetivo determinar 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 audios de habla, 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. Objetivos específicos

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

Esta tesis 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

Las métricas de degradación de la señal cuantifican, a nivel objetivo, la diferencia entre una señal de referencia (habitualmente la señal limpia o de alta calidad) y una señal procesada o degradada (por ejemplo, una señal comprimida o con ruido de fondo). A continuación se describen las más empleadas en evaluación de voz y TTS.

PESQ es un índice objetivo diseñado para predecir la calidad percibida de la voz en telefonía y sistemas de comunicación. Compara la señal de referencia y la señal degradada mediante un modelo perceptual que incluye etapas de alineamiento temporal, modelado perceptual y mapeo a una escala MOS-LQO (Mean Opinion Score — Listening Quality Objective). PESQ fue normalizado originalmente como la recomendación ITU-T P.862 (ITU-T, 2001) y se utiliza ampliamente para evaluación de códecs y transmisiones telefónicas.

Esta métrica tiene la limitación del rango de frecuencia que evalúa (va entre 200 Hz a 3500 Hz), para un análisis mas completo surge la métrica denominada POLQA, que es la tercera generación de métricas ITU para evaluación de calidad de voz end-to-end ITU-T P.863 (ITU-T, 2018). Mejora y extiende PESQ en ancho de banda (incluye super-wideband y fullband), es más robusta frente a ciertas distorsiones modernas (ecualización, delays, codificación amplia banda) y proporciona predicciones MOS más fiables en escenarios actuales.

Desde un lado objetivo de la calidad de la señal se tienen métricas como el SNR y el SI-SDR. El SNR (relación señal a ruido por sus siglas en inglés) es la relación entre la potencia de la señal útil y la potencia del ruido de fondo, normalmente expresada en decibelios. Es una medida física simple, muy útil para caracterizar condiciones de captura o transmisión, pero su correlación con la percepción humana puede ser pobre cuando las distorsiones son no-gaussianas, no-aditivas o hay alteraciones de fase y tiempo; por ello suele utilizarse junto con métricas perceptuales.

El SI-SDR (relación invariante de distorsión a señal por sus siglas en inglés) es una versión modificada y más robusta del SDR tradicional diseñada para evitar penalizaciones por

escalado de amplitud entre referencia y estimado (Le Roux et al., 2019). En la práctica, SI-SDR proyecta la señal estimada sobre la referencia (elimina la diferencia de escala) y calcula la relación señal/residuo resultante; por su formulación, es ampliamente usada en separación de fuentes y evaluación de redes de denoising o separación en el dominio temporal. Su definición y discusión de ventajas frente a SDR clásico están recogidas en la literatura sobre evaluación objetiva de separación.

Por último, entre las métricas de calidad de señales del habla se puede incluir el STOI (inteligibilidad objetiva de tiempo corto por sus siglas en inglés), que estima inteligibilidad de habla en condiciones ruidosas mediante una medida de correlación entre segmentos de tiempo-frecuencia de la señal limpia y la señal degradada (Taal et al., 2011). Fue propuesta para predecir la inteligibilidad de señal procesada (p. ej. filtrado espectral, separación) y ha mostrado alta correlación con experimentos de inteligibilidad humana en muchos escenarios. No mide naturalidad, sino inteligibilidad.

Por su definición, para calcular todas estas métricas se necesita tener una referencia limpia (o la señal sin ruido para SNR). En la práctica no siempre se puede tener esta señal de referencia, por eso han surgido métodos de estimación ciega de todos estos parámetros, que permiten estimar la calidad de un audio sin tener la señal de referencia.

Hay que destacar, que ninguna métrica única captura todos los aspectos perceptuales (naturalidad, inteligibilidad, artefactos, coloración). En evaluación de TTS y sistemas de procesamiento de voz es habitual combinar métricas intrusivas (PESQ/POLQA, SI-SDR, MCD) con medidas de inteligibilidad (STOI) y, cuando es posible, pruebas subjetivas (MOS).

2.1.2. Métricas de entorno

Las métricas de entorno (o de acústica de salas) describen propiedades de la sala o del campo acústico que afectan la percepción de la voz y la música.

Uno de los parámetros principales, es el tiempo de reverberación (T), que se define clásicamente como el tiempo que tarda el nivel sonoro en decaer 60 dB tras cortar la fuente (Sabine). Por razones prácticas se calculan estimadores como (T_{30}) (pendiente del decaimiento entre -5 dB y -35 dB extrapolada a -60 dB) o (T_{20}) (entre -5 y -25 dB). El (T_{30}) es de uso corriente en caracterización objetiva de salas y está normalizado por ISO 3382

(«ISO 3382-2:2008 Acoustics – Measurement of room acoustic parameters – Part 2: Reverberation time in ordinary rooms», 2008). Los valores de (T) afectan inteligibilidad, claridad y sensación de espacialidad acústica.

Otro parámetros que se usa para evaluar la percepción de audios en auditorios es la claridad (C_{50}) (para voz) y (C_{80}) (para música), ya que miden la relación energía temprana/energía tardía. Este parámetro se define matemáticamente como:

$$C_{50} = 10 \log_{10} \frac{\int_0^{50,ms} p^2(t) dt}{\int_{50,ms}^{\infty} p^2(t) dt} \quad (1)$$

donde ($p(t)$) es la respuesta al impulso en la posición de escucha. Valores altos de (C_{50}) indican mayor proporción de energía temprana (mejor inteligibilidad para voz).

El D_{50} (a menudo expresado en porcentaje) es la fracción de energía que llega en los primeros 50 ms respecto a la energía total; es otra forma de cuantificar claridad/inteligibilidad para voz (relacionado con (C_{50})). Se calcula a partir de la respuesta al impulso y es relevante para evaluación de locales docentes, salas de conferencia y condiciones de grabación.

2.1.3. Métricas del habla

Las métricas del habla describen propiedades prosódicas y espectrales de la voz que influyen en la naturalidad y en la identificación del hablante. Dentro de esta categoría, una de las métricas principales es la frecuencia fundamental de la voz (F_0), cuyo valor representa la frecuencia de vibración principal de las cuerdas vocales de los diferentes hablantes (Bäckström et al., 2022) y constituye la base de la percepción del *pitch*. Se estima mediante algoritmos de detección de pitch (autocorrelación, cepstrum, algoritmos basados en modelos probabilísticos) y se usa para análisis prosódico, control de entonación en TTS y evaluación de naturalidad. La distribución estadística de F_0 , su variación temporal (contorno), y su relación con la energía son indicadores usados tanto en evaluación objetiva como perceptual.

Otra métrica muy utilizada en el análisis de sistemas de TTS es la distorsión media cepstral (MCD por sus siglas en ingles). MCD (a veces llamado Mel-Cepstral Distance) cuantifica la distancia entre dos secuencias de coeficientes mel-cepstales (por ejemplo, señal

natural vs. señal sintetizada) y se expresa en dB (Kominek et al., 2008). La expresión de esta métrica es la siguiente:

Se calcula comúnmente como una raíz cuadrada de la suma de cuadrados normalizada entre vectores de coeficientes y es ampliamente utilizada para evaluar la calidad espectral en TTS y vocoders; sin embargo, su correlación con la calidad percibida no es perfecta y debe complementarse con pruebas perceptuales. La métrica fue propuesta en trabajos clásicos sobre evaluación objetiva de síntesis.

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 (J. 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

La inteligencia artificial generativa (IAG) comprende técnicas cuyo objetivo es modelar la distribución de datos para generar muestras nuevas que sean plausibles, ya sea para imágenes, texto, o audio. La generación de estos *samples* nuevos puede ser condicionadas o no por información adicional (*prompts*). Los principales paradigmas contemporáneos son:

- Modelos autoregresivos: modelan la probabilidad conjunta como producto de condicionales (p. ej. PixelRNN en imágenes (Van Den Oord et al., 2016) o WaveNet en audio (van den Oord et al., 2016)). Son conceptualmente simples y producen alta calidad pero pueden ser lentos en inferencia por su naturaleza secuencial.
- VAE (Variational Autoencoders): modelos latentes que optimizan una cota variacional de la verosimilitud. Buen control del espacio latente y entrenamiento estable, pero a ofrecen menor fidelidad en muestras crudas (Kingma y Welling, 2013).
- GANs (Generative Adversarial Networks): enfrenta dos redes neuronales, un generador vs un discriminador que compiten en un juego adverso, generan muestras de alta fidelidad rápida y eficientemente (Goodfellow et al., 2020). Aunque el entrenamiento puede ser inestable, se han aplicado a vocoders y síntesis auditiva (ej. HiFi-GAN por J. Kong et al., 2020).
- Modelos basados en score / difusión: incluyen los modelos de difusión y score-matching (DDPM, score-based models) que adicionan ruido progresivamente a datos reales y

aprenden a invertir ese proceso para muestrear (Ho et al., 2020). Actualmente han mostrado resultados competitivos o superiores en calidad de muestras y estabilidad de entrenamiento en imagen y audio (ej. DiffWave por Z. Kong et al., 2021).

En audio/TTS, cada paradigma tiene ventajas: los autoregresivos (WaveNet) alcanzan alta naturalidad; los GANs (HiFi-GAN) ofrecen inferencia rápida y alta fidelidad; y los modelos de difusión (DiffWave) combinan calidad con mayor estabilidad y flexibilidad condicional. La elección depende de trade-offs entre calidad, velocidad y control.

2.5. MODELOS DE DIFUSIÓN

Los modelos de difusión son una clase de generadores probabilísticos basados en procesos estocásticos de adición y eliminación de ruido. Su formulación moderna se fundamenta en dos ideas principales: (1) definir un proceso directo (forward) que corrompe los datos añadiendo ruido gaussiano progresivamente hasta obtener ruido casi puro; (2) aprender el proceso inverso (reverse) —un modelo condicional que predice pasos de denoising— para transformar ruido en datos sintéticos.

La definición matemática plantea una muestra de datos x_0 , donde el proceso directo añade ruido en (T) pasos:

$$q(x_t | x_{t-1}) = \mathcal{N}(x_t; \sqrt{1 - \beta_t} x_{t-1}, \beta_t), \quad (2)$$

con una escala de varianzas (β_t) . El proceso inverso se parametriza como:

$$p_\theta(x_{t-1} | x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)), \quad (3)$$

y se entrena minimizando una variación de la evidencia (ELBO) o una pérdida de denoising equivalente. El trabajo de Ho et al., 2020 formaliza esta familia (DDPM) y conecta la aproximación por denoising con score-matching (Song et al., 2021).

Los modelos de difusión proporcionan calidad competitiva sin entrenamiento adversario y permiten control condicional (por ejemplo, condicionamiento en mel-espectrogramas para vocoders). La principal limitación práctica es el coste de muestreo (múltiples pasos),

aunque variantes (DDIM, sampling acelerado, distillation) reducen el número de pasos sin pérdida sustancial de calidad.

3. ESTADO DEL ARTE

3.1. MODELOS DE TTS

FALTA: 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	DurlIAN 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

La gran discrepancia entre modelos de las diferentes implementaciones es una da las motivaciones para el desarrollo de esta tesis.

3.3. CONJUNTOS DE DATOS DEL HABLA EN ESPAÑOL

Es relevante mencionar la falta de conjuntos de datos del habla en español, en principal de la variante español de Argentina. En la Tabla 2 se recopilaron diferentes conjuntos

de datos de habla en español, especificando la cantidad de horas y hablante, la región que lingüística que representa y el tipo. Los diferentes tipos hacen referencia a la fuente de donde fueron recopilados los datos. Además de los conjuntos profesionales e *in-the-wild* que ya fueron discutidos, se presentan dataset de tipo conversacional (de entrevistas o llamadas telefónicas) y los dataset *crowdsourced*, donde se recopilan audios de voces de voluntarios a través de internet.

Tabla 2. Resumen de datasets

Nombre dataset	Cantidad horas	Cantidad hablantes	Región dialecto	Tipo dataset
Google LREC - Guevara-Rukoz et al., 2020	8	44	Buenos Aires	Profesional
Emilia - Torres et al., 2019	4	1	Buenos Aires	Profesional
HaCASpa - Gabriel, 2011	10	50	Buenos Aires/Mendoza	Conversacional
CORdEBA - UNLP, 2014	6	25	Buenos Aires	Conversacional
Common Voice - Ardila et al., 2020	587	260	Diversas	Crowdsourced
YODAS - Li et al., 2023	50k	N/A	Diversas	In-the-wild
CML TTS - Oliveira et al., 2023	400	N/A	España	Profesional
MLS - Pratap et al., 2020	1.5k	120	España	Profesional
VoxPopuli - Wang et al., 2021	166	305	España	Conversacional
Tedx Spanish - Hernandez-Mena, 2019	24	N/A	México	Espontáneo

Este análisis de los dataset en el estado del arte pone en evidencia la necesidad de conformar un corpus del español de Argentina, donde es fundamental extender la representatividad cultural de nuestra región, y capturar los diferentes acentos de todas las regiones del país, ya que como se ve en la recopilación, casi todo el material disponible del dialecto de Argentina se centra en la región porteña. Los dataset calificados como dialecto diverso es porque en su confección no delimitaron entre los diferentes dialectos del español, y agruparon todo en una misma categoría, lo que incluya español de las diferentes regiones de latinoamerica y de españa.

4. DESARROLLO

4.1. RECOPIACIÓN 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 (Gabriel, 2011): aproximadamente 10 horas de audio con dialecto bonaerense, 50 hablantes.

El conjunto profesional suma un total de 22 horas.

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 evalúan 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.

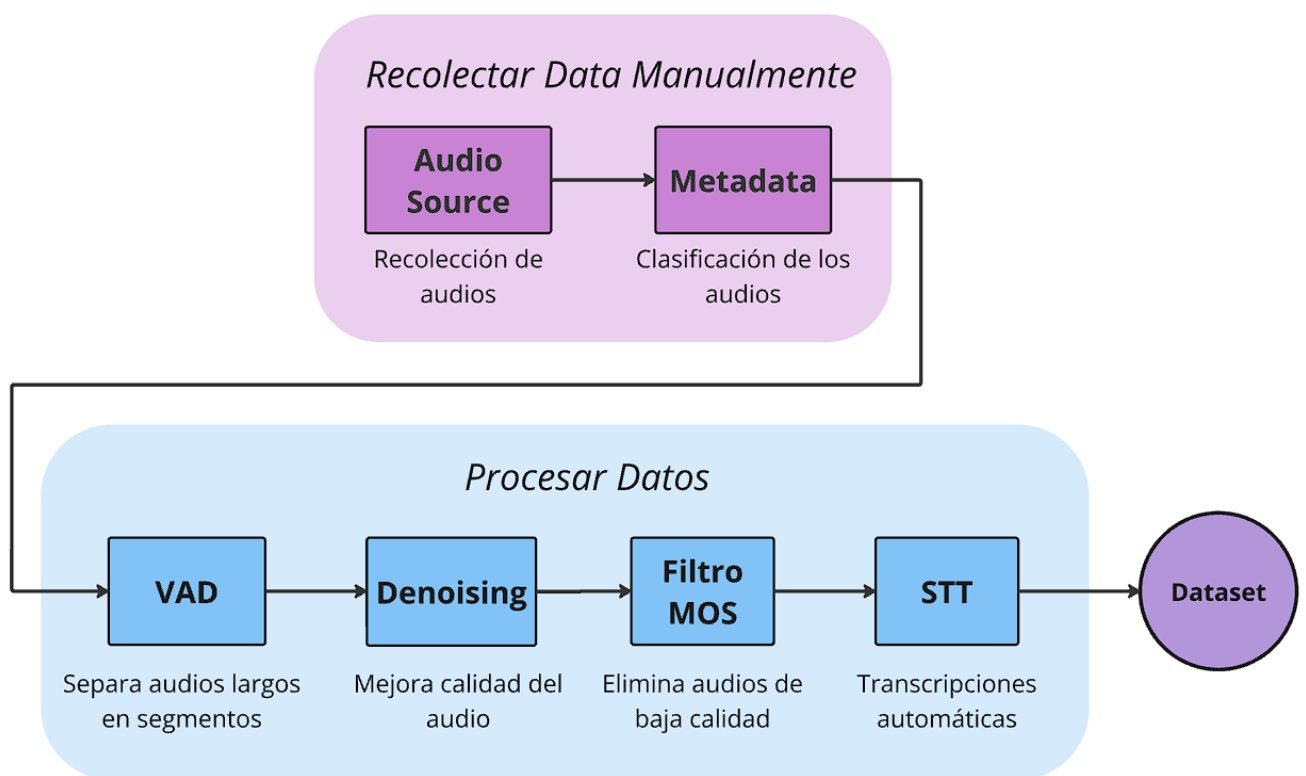


Figura 1. Diagrama de flujo del proceso completo para generar el dataset.

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 preser-

vacación 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 los más usados en la literatura, donde se utiliza uno u el otro pero no se han contrastado para determinar el modelo más óptimo.
- 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.

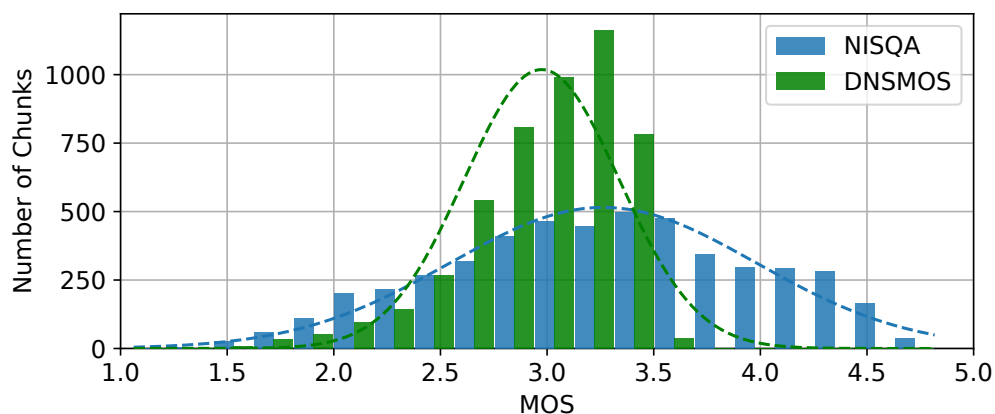


Figura 2. Resultado de MOS por cantidad de segmentos del dataset original para los 2 modelos a comparar.

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.

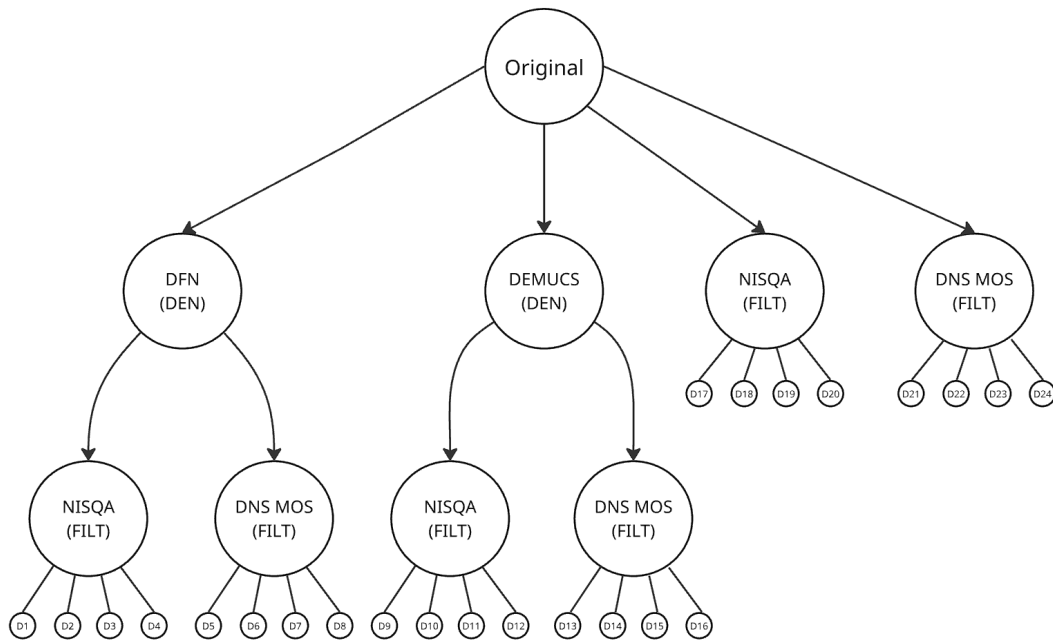


Figura 3. Variantes propuestas para evaluar diferentes configuración de la cadena de procesamientos.

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 4, 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.

$$RD_P = 1 - \frac{HORAS_P}{HORAS_O} \quad (4)$$

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 5, teniendo en cuenta que se espera que los valores de PESQ, SI-SDR y SNR suban (mejo-

ría) al aplicar la cadena de procesamientos. Los subíndices respetan la condiciones anterior donde P es por procesado y O es original, estos subíndices se mantiene consistentes para todas las métricas.

$$CS_P = \frac{PESQ_O}{PESQ_P} + \frac{SI-SDR_O}{SI-SDR_P} + \frac{SNR_O}{SNR_P} \quad (5)$$

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 reverberación 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 6, 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.

$$CA_P = \frac{T_{30,P}}{T_{30,O}} + \frac{C_{50,O}}{C_{50,P}} \quad (6)$$

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 7. 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.

$$DH_P = \left| 1 - \frac{F0std_P}{F0std_O} \right| + \frac{MCD_P}{5} \quad (7)$$

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 acústicas (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 8, donde se busca minimizar el aporte negativo de cada variable para poder identificar la mejor variante P de la cadena de procesamientos.

$$\min_{P \in \text{Conf}} \left\{ RD_P + CS_P + CA_P + DH_P \right\} \quad (8)$$

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

FALTA: Explicar porque necesito hacer lo del modelo de estimación de densidad.

4.4.1. Validación con medelo zero-shot

FALTA: Definir modelo y explicar la justificación de este experimento.

4.5. DESCRIPCIÓN DE PRUEBAS ESTADÍSTICAS

FALTA: Detallar un poco las validaciones estadísticas que se van a realizar.

4.6. MODELO DE TTS ZERO-SHOT

FALTA: 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. Ejemplo de cálculo

Se calcula el promedio de cada métrica para las diferentes configuraciones propuestas, y se calcula la mejora porcentual. Por ejemplo, para el SNR, el valor promedio calculado en el dataset original es de 19.08 dB. Al analizar el SNR para todas las variantes, se encuentra que el valor mas bajo de SNR es de 19.89 dB, mientras que el valor mas alto es de 33.78 dB. Los resultados ordenados según el porcentaje de mejora se puede ver en Tabla 3.

Tabla 3. Mejora SNR por métrica de calidad y algoritmo de denoising.

Filtro de calidad y umbral	Algoritmo de denoising	Mejora SNR (%)
NISQA = 3	Sin denoising	4.07
NISQA = 3.5	Sin denoising	4.93
DNSMOS = 2.7	Sin denoising	6.93
DNSMOS = 3.4	Sin denoising	7.56
DNSMOS = 3.0	Sin denoising	9.92
DNSMOS = 3.2	Sin denoising	10.00
NISQA = 3.8	DeepFilterNet	10.13
NISQA = 4.2	Sin denoising	11.67
DNSMOS = 3.4	DeepFilterNet	12.52
NISQA = 3.5	DeepFilterNet	14.78
NISQA = 3.8	Sin denoising	15.58
DNSMOS = 3.2	DeepFilterNet	15.65
NISQA = 3.0	DeepFilterNet	15.72
NISQA = 4.2	DeepFilterNet	17.55
DNSMOS = 3.0	DeepFilterNet	17.86
DNSMOS = 2.7	DeepFilterNet	18.84
DNSMOS = 3.4	Demucs	39.66
NISQA = 4.2	Demucs	40.34
NISQA = 3.8	Demucs	40.54
NISQA = 3.5	Demucs	41.24
DNSMOS = 3.2	Demucs	41.78
NISQA = 3.0	Demucs	42.16
DNSMOS = 3.0	Demucs	43.25
DNSMOS = 2.7	Demucs	43.52

Este análisis es el que se pondera para cada variable en las 4 categorías de métricas propuestas.

5.1.2. Métricas separadas

En una instancia exploratoria previa a la definición de la métrica conjunta, se exploran los resultados del dataset de prueba ante las diferentes métricas propuestas.

Primero se 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.

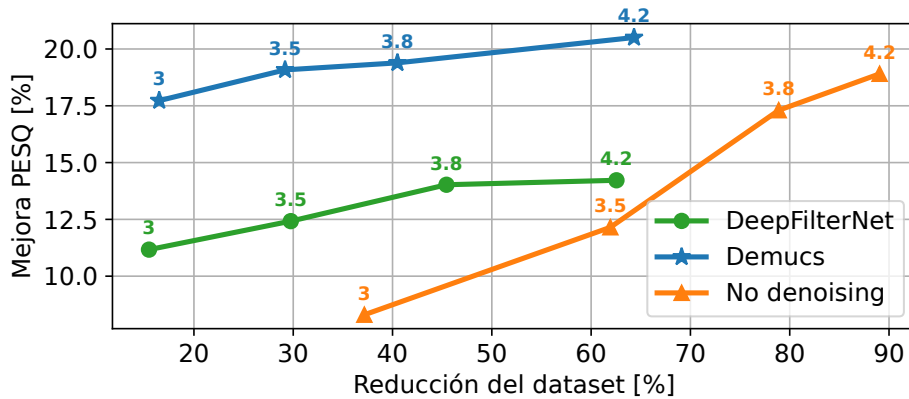


Figura 4. Comparación entre PESQ y cantidad de horas de diferentes variantes con NISQA.

En la Figura 5 se 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, cuando no se aplica realce previo el filtrado selecciona segmentos con menor reverberación y la mejora en T_{30} es sustancial. En cambio, en las condiciones ya denoised las ganancias en T_{30} son más moderadas (por debajo aproximadamente del 5 % en los umbrales considerados). Además, no se detecta una diferencia significativa en T_{30} entre DeepFilterNet y Demucs. 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.

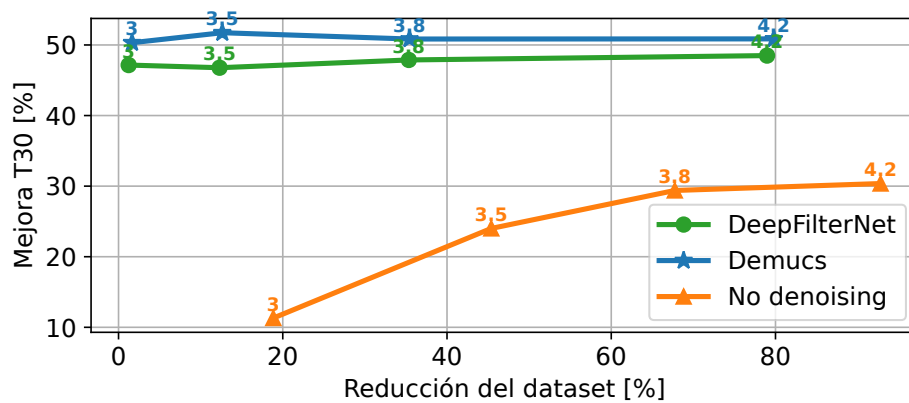


Figura 5. Comparación entre T_{30} y cantidad de horas de diferentes variantes con DNS MOS.

Además, de comparar la cantidad de horas, también se pueden relacionar otras variables, como puede ser la desviación de frecuencia fundamental $F0-STD$ y la mejoría de calidad cuantificada por PESQ Figura 6. En este análisis se ilustra la relación entre la mejora relativa de PESQ (eje horizontal) y la diferencia porcentual en la desviación estándar de $F0$ ($F0-STD$, 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 $F0-STD$. 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 $F0-STD$ 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.

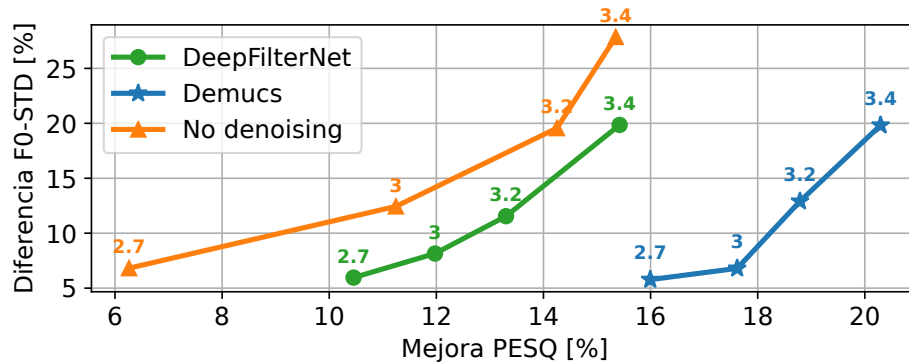


Figura 6. Comparación entre F0-std y PESQ para diferentes variantes con DNS MOS.

Este análisis pone en evidencia las relaciones de compromiso entre las métricas seleccionadas y como la mejor variante no será la que alcance un mejor resultado en particular, sino la que logre un mejor balance de todos los criterios en conjunto.

5.1.3. Métricas compuestas

Una vez validado el comportamiento de cada métricas de forma individual, se analiza las métricas compuestas definidas en la Sección 4.3. Primero, la calidad de la señal, presenta el comportamiento similar a la tendencia explicada al analizar PESQ. Esto es evidente, ya que tanto PESQ, SI-SDR y SNR sigan la misma tendencia, como se presenta en las gráficas completas en el Anexo A.

En la Figura 7, se compara la métrica compuesta de reducción de datos (RD), respecto a la calidad de la señal (CS). Como la métrica se diseñó para minimizar el objetivo total, se puede ver que el mínimo coincide con la variante que filtra menor cantidad de datos y utiliza Demucs como algoritmo de denoising, resultado que coincide con el análisis de las métricas de forma aislada.

Se analizaron y compararon todas las combinaciones de las diferentes métricas. Los resultados completos de este análisis se presentan en el Anexo B. Los resultados mas destacados son la comparación entre calidad de señal y condiciones acústicas Figura 8, donde el agrupamiento de las métricas indica que el filtrado tiene muy poco impacto.

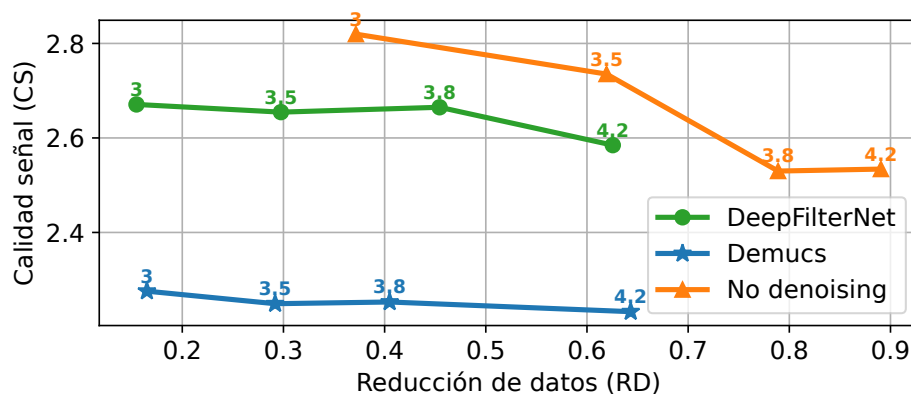


Figura 7. Comparación entre reducción de datos y calidad de señal para diferentes variantes con Nisqa.

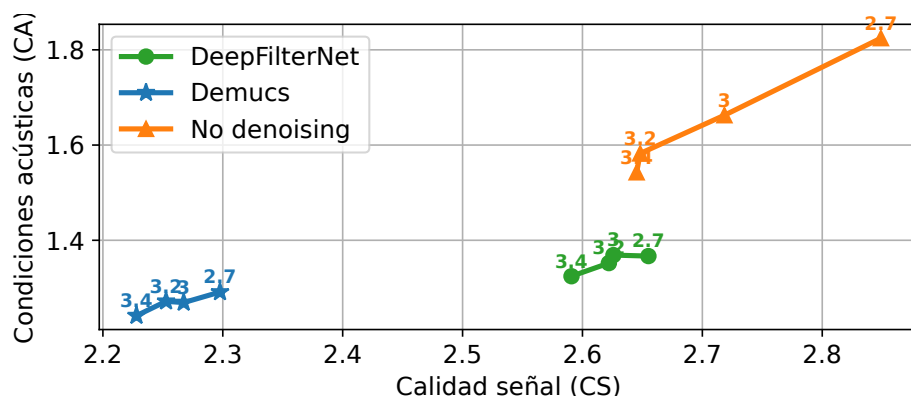


Figura 8. Comparación entre calidad de señal y condiciones acústicas para diferentes variantes con DNSMOS.

5.1.4. Optimización de la cadena

Al sumar todas las métricas compuestas, se obtiene la métrica conjunta o total, con la cual se puede ordenar las variantes de la cadena y así definir cual es la configuración optima para procesar un conjunto de datos del habla.

En la Tabla 4, se presenta el valor de cada métricas compuesta y el valor de la métrica conjunta, para todas las variantes, ordenadas de menor a mayor (siendo el mínimo la mejor variantes y máximo la pero variante).

La mejor configuración es la que utiliza Demucs como algoritmo de denoiser y DNSMOS como algoritmo de filtrado con umbral de MOS en 2.7. Esta configuración obtiene un valor total de 4.08, mientras que la peor variante alcanza un valor total de 5.53

Es interesante analizar como la mejor variante, no es la mejor opción en ninguna de las métricas individuales (marcadas en negrita), sino que ofrece la mejor relación de com-

Tabla 4. Métricas compuestas y total para todas las configuraciones.

Filtrado	Umbral	Denoiser	RD	CS	CA	DH	TOT
DNSMOS	2,7	Demucs	0,02	2,30	1,29	0,48	4.08
DNSMOS	3,0	Demucs	0,13	2,27	1,27	0,51	4,17
NISQA	3,0	Demucs	0,17	2,28	1,27	0,60	4,31
NISQA	3,5	Demucs	0,29	2,25	1,28	0,58	4,41
NISQA	3,8	Demucs	0,40	2,25	1,30	0,55	4,51
DNSMOS	3,2	Demucs	0,35	2,25	1,27	0,63	4,51
DNSMOS	2,7	DeepFilterNet	0.01	2,66	1,37	0,62	4,65
NISQA	4,2	Demucs	0,64	2.22	1,29	0,52	4,68
DNSMOS	3,0	DeepFilterNet	0,12	2,63	1,37	0,61	4,73
NISQA	3,0	DeepFilterNet	0,15	2,67	1,37	0,63	4,83
DNSMOS	3,2	DeepFilterNet	0,35	2,62	1,35	0,60	4,93
DNSMOS	2,7	No denoising	0,19	2,85	1,82	0.07	4,93
NISQA	3,5	DeepFilterNet	0,30	2,65	1,38	0,63	4,96
DNSMOS	3,0	No denoising	0,45	2,72	1,66	0,12	4,96
DNSMOS	3,4	Demucs	0,80	2,23	1.24	0,78	5,05
NISQA	3,0	No denoising	0,37	2,82	1,79	0,09	5,08
DNSMOS	3,2	No denoising	0,68	2,65	1,58	0,20	5,10
NISQA	3,8	DeepFilterNet	0,45	2,66	1,37	0,64	5,13
NISQA	3,8	No denoising	0,79	2,53	1,69	0,20	5,21
NISQA	3,5	No denoising	0,62	2,74	1,74	0,13	5,22
DNSMOS	3,4	DeepFilterNet	0,79	2,59	1,32	0,58	5,29
NISQA	4,2	DeepFilterNet	0,63	2,58	1,37	0,81	5,39
DNSMOS	3,4	No denoising	0,93	2,65	1,54	0,28	5,39
NISQA	4,2	No denoising	0,89	2,53	1,65	0,46	5,53

promiso entre las 4 condiciones evaluadas.

5.1.5. Análisis del peso de cada métrica

Se analiza la varianza de cada métrica en función de las distintas configuraciones para determinar cual métrica tiene mas peso en el ordenamiento propuesto por la metodología de evaluación.

5.2. COMPARACIÓN ENTRE CONJUNTOS DE DATOS

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

5.3. COMPARACIÓN POR ESTIMACIÓN DE DENSIDAD

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

5.4. VALIDACIÓN MODELO DE TTS ZERO-SHOT

FALTA: 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.

BIBLIOGRAFÍA

- Ardila, R., Branson, M., Davis, K., Kohler, M., Meyer, J., Henretty, M., Morais, R., Saunders, L., Tyers, F., & Weber, G. (2020). Common Voice: A Massively-Multilingual Speech Corpus. *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 4218-4222. <https://aclanthology.org/2020.lrec-1.520/>
- 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>
- Bäckström, T., Räsänen, O., Zewoudie, A., Zarazaga, P. P., Koivusalo, L., Das, S., Mellado, E. G., Mansali, M. B., Ramos, D., Kadiri, S., Alku, P., & Vali, M. H. (2022). *Introduction to Speech Processing* (2.^a ed.). <https://doi.org/10.5281/zenodo.6821775>
- 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>
- Gabriel, C. (2011). Hamburg Corpus of Argentinean Spanish (HaCASpa).
- Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press.
- Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2020). Generative adversarial networks. *Commun. ACM*, 63(11), 139-144. <https://doi.org/10.1145/3422622>
- 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>
- Hernandez-Mena, C. D. (2019). TEDx Spanish Corpus. Audio and transcripts in Spanish taken from the TEDx Talks.
- Ho, J., Jain, A., & Abbeel, P. (2020). Denoising diffusion probabilistic models. *Proceedings of the 34th International Conference on Neural Information Processing Systems*.
- 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.
- ISO 3382-2:2008 Acoustics – Measurement of room acoustic parameters – Part 2: Reverberation time in ordinary rooms. (2008). International Organization for Standardization.
- ITU-T. (2001, febrero). *Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs* (inf. téc. N.º P.862). International Telecommunication Union.
- ITU-T. (2018). *P.863: Perceptual Objective Listening Quality Prediction* (inf. téc. N.º P.863). International Telecommunication Union.
- 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.
- Kingma, D. P., & Welling, M. (2013). Auto-Encoding Variational Bayes.
- 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*.
- Kong, Z., Ping, W., Huang, J., Zhao, K., & Catanzaro, B. (2021). DiffWave: A Versatile Diffusion Model for Audio Synthesis. *International Conference on Learning Representations*. <https://openreview.net/forum?id=a-xFK8Ymz5J>

- 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>
- Le Roux, J., Wisdom, S., Erdogan, H., & Hershey, J. (2019). SDR – Half-baked or Well Done?, 626-630. <https://doi.org/10.1109/ICASSP.2019.8683855>
- 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>
- Li, X., Takamichi, S., Saeki, T., Chen, W., Shiota, S., & Watanabe, S. (2023). Yodas: Youtube-Oriented Dataset for Audio and Speech. *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 1-8. <https://doi.org/10.1109/ASRU57964.2023.10389689>
- 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>
- Oliveira, F. S., Casanova, E., Junior, A. C., Soares, A. S., & Galvão Filho, A. R. (2023). CML-TTS: Multilingual Dataset for Speech Synthesis in Low-Resource Languages. *Text, Speech, and Dialogue: 26th International Conference, TSD 2023, Pilsen, Czech Republic, September 4–6, 2023, Proceedings*, 188-199. https://doi.org/10.1007/978-3-031-40498-6_17
- 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*.
- Pratap, V., Xu, Q., Sriram, A., Synnaeve, G., & Collobert, R. (2020). MLS: A Large-Scale Multilingual Dataset for Speech Research. *Interspeech 2020*, 2757-2761. <https://doi.org/10.21437/Interspeech.2020-2826>

- 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>
- Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., & Poole, B. (2021). Score-Based Generative Modeling through Stochastic Differential Equations. *International Conference on Learning Representations*. <https://openreview.net/forum?id=PxTIG12RRHS>
- Taal, C., Hendriks, R., Heusdens, R., & Jensen, J. (2011). An Algorithm for Intelligibility Prediction of Time–Frequency Weighted Noisy Speech. *Audio, Speech, and Language*

- Processing, IEEE Transactions on*, 19, 2125-2136. <https://doi.org/10.1109/TASL.2011.2114881>
- 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.
- Van Den Oord, A., Kalchbrenner, N., & Kavukcuoglu, K. (2016). Pixel recurrent neural networks. *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, 1747-1756.
- 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*.
- Wang, C., Riviere, M., Lee, A., Wu, A., Talnikar, C., Haziza, D., Williamson, M., Pino, J., & Dupoux, E. (2021). VoxPopuli: A Large-Scale Multilingual Speech Corpus for Representation Learning, Semi-Supervised Learning and Interpretation. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 993-1003. <https://doi.org/10.18653/v1/2021.acl-long.80>
- 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.