Optimización de datos para TTS mediante evaluación de calidad de audio y reducción avanzada de ruido

Ignacio Correa[1]

Matias Di Bernardo[1]

Emmanuel Misley[1]

Gala Lucia González Barrios[2]

[1] Ingeniería de Sonido, Universidad Nacional de Tres de Febrero, Argentina

[2] Doctoranda en Diseño Centrado en el Humano, Virginia Tech, USA

n.correa.213@gmail.com

matias.di.bernardo@hotmail.com

emmanuel.misley@gmail.com

gala@vt.edu

Resumen – Este trabajo propone evaluar la calidad de audios para entrenar modelos de TTS (Text-to-Speech) mediante un análisis no intrusivo con NISQA y explora la mejora de dicha calidad con algoritmos de denoising como DeepFilterNet. En la primera fase se cuantifica la calidad sin referencia limpia, obteniendo parámetros de MOS (Mean Opinion Score), ruido y discontinuidad que facilitan la selección de segmentos adecuados para el entrenamiento. A continuación, se aplica DeepFilterNet para reducir el ruido lo más posible mientras se busca al mismo tiempo mantener las características acústicas fundamentales del habla. Las evaluaciones objetivas PESQ (Perceptual Evaluation of Speech Quality), MCD (Mean Cepstrum Distortion) y subjetivas CMOS (Comparative Mean Opinion Score) muestran mejoras significativas en la inteligibilidad y fidelidad vocal al entrenar modelos de TTS entrenados bajos estas condiciones, lo que indica el potencial de este procesamiento en las señales para optimizar datos de entrenamiento para TTS.

Palabras claves Speech Quality Assessment, Non-intrusive Evaluation, Noise Reduction, Denoising Algorithms, Speech Data Preprocessing, Corpus Optimization, Text-to-Speech, Deep Learning.

1. INTRODUCCIÓN

El campo del text-to-speech (TTS) ha avanzado fuertemente en los últimos años gracias a arquitecturas neuronales profundas que alcanzan niveles altos de naturalidad y calidad [1]. No obstante, estos modelos requieren grandes cantidades de audio de alta calidad para su entrenamiento, lo que es problemático para acentos o lenguas con pocos recursos [2], como es el caso del español castellano en sus diferentes variantes provinciales. Una vía práctica es emplear grabaciones "de calle" (ITW por sus siglas en inglés) [3] o de calidad no profesional y compensar con técnicas de *denoising* y *cleaning* antes del entrenamiento.

Para hacer una selección del material ITW, es fundamental, evaluar a gran escala la calidad de audio (y su idoneidad para TTS), y hacerlo de forma eficiente, ya que con pruebas subjetivas es muy costoso. Por ello existen métricas objetivas intrusivas (PESQ/POLQA) y no intrusivas (p. ej. NISQA) y también métricas específicas para TTS como el MCD. Este trabajo combina (i) un experimento que compara varios algoritmos de denoising y su efecto en la evaluación subjetiva del TTS y (ii) un estudio del modelo NISQA como herramienta automática para clasificar y diagnosticar la calidad de audios sin referencia y cuantificar las mejoras al utilizar algoritmos de denoising. El objetivo del artículo combinado es proponer y validar un flujo de trabajo donde el denoising y NISQA actúen conjuntamente para producir conjuntos de entrenamiento más aptos para TTS.

2. MARCO TEORICO

2.1 Desarrollo reciente en TTS

Los modelos TTS basados en arquitecturas con redes profundas (p. ej. FastPitch) [4] han mostrado gran capacidad para modelar prosodia y timbre, pero su entrenamiento óptimo depende de datos limpios y homogéneos; la falta de datos profesionales incentiva el uso de técnicas de preprocesado sobre grabaciones "in the wild".

2.2 Ventajas del denoising + cleaning

El proceso de denoising ideal preserva la identidad del hablante mientras reduce ruido de fondo, discontinuidades y artefactos [5]; el *cleaning* (segmentación, eliminación de segmentos no útiles, normalización) evita introducir ejemplos degradados al entrenamiento. Ambas operaciones reducen la varianza indeseada del dataset y, en muchos casos, mejoran la calidad percibida del

TTS final. Sin embargo, es importante analizar el funcionamiento de la etapa de denoising y de clasificación de audio por separado para después poder cuantificar su interacción.

2.3 Métricas de evaluación

- **Mean Opinion Score (MOS):** Es una métrica subjetiva ampliamente reconocida que cuantifica la calidad percibida de un audio a partir de evaluaciones realizadas por oyentes humanos. Los participantes asignan una puntuación en una escala de 1 (mala calidad) a 5 (excelente calidad), y el MOS final se obtiene promediando aritméticamente estos puntajes. Esta métrica es la propuesta por las normas ITU-T P.800 [6] y P.808 [7] para evaluar la calidad de los sistemas de comunicación.
- **Comparison Mean Opinion Score (CMOS):** A diferencia del MOS absoluto, el CMOS es una prueba subjetiva comparativa. En este test, los participantes evalúan la calidad de un clip de audio en relación con otro, utilizando una escala que va desde -3 ("Mucho peor") hasta +3 ("Mucho mejor"), con 0 indicando "Más o menos igual". El CMOS se emplea para cuantificar la mejora en el rendimiento de los sistemas TTS cuando se aplica *denoising* a los datos de entrenamiento.
- **Perceptual Evaluation of Speech Quality (PESQ):** PESQ [8] es una métrica objetiva y ampliamente aceptada para evaluar la calidad del habla. Su escala es análoga a la del MOS (de 1 a 5). Este método es intrusivo, lo que significa que requiere un audio de referencia limpio para su cálculo. PESQ considera diversos factores como la nitidez del audio, el ruido de fondo y el recorte de la señal. Fue el estándar ITU-T P.862, aunque posteriormente fue reemplazado por POLQA [9].
- **Perceptual Objective Listening Quality Assessment (POLQA):** Es la recomendación actual de ITU-T P.863 y, al igual que PESQ, es un método intrusivo de evaluación objetiva de la calidad de audio.
- **Non Intrusive Speech Quality Assessment (NISQA):** NISQA [10] es un modelo de aprendizaje profundo diseñado para predecir el MOS de audios de voz de forma no intrusiva. Además de la predicción del MOS, NISQA proporciona otros cuatro parámetros que caracterizan la calidad del audio: Ruido, Discontinuidad, Coloración y Volumen [11]. Los tres primeros se consideran dimensiones ortogonales del espacio multidimensional de la calidad del habla, mientras que el volumen, aunque no ortogonal, es relevante para la evaluación de la calidad. La inclusión de estos parámetros busca

ofrecer una caracterización más detallada de la calidad del audio y ayudar a identificar las causas de su degradación.

- **Short-Time Objective Intelligibility (STOI):** STOI [12] es una métrica objetiva utilizada para medir la inteligibilidad del habla en señales de audio, especialmente en entornos ruidosos. Compara una versión *denoised* de la señal con una versión sin procesar, y su puntuación oscila entre 0 y 1, donde 1 indica inteligibilidad perfecta y 0 indica ausencia de inteligibilidad.
- **Mean Mel-Cepstral Distortion (MCD):** MCD [13] es una métrica diseñada específicamente para la evaluación de sistemas TTS. Cuantifica las diferencias en las características del habla entre las señales de audio originales y procesadas. Se calcula como la distancia euclidiana entre los vectores de coeficientes Mel-frecuenciales cepstrales (MFCC) de las dos señales. Los MFCC encapsulan información acústica relevante y son ampliamente utilizados en el procesamiento del habla. Valores más bajos de MCD generalmente indican una mejor preservación de las características acústicas del audio original.

La diversidad de métricas disponibles (subjetivas, objetivas, intrusivas, no intrusivas, generales y específicas de TTS) refleja la complejidad inherente a la evaluación de la "calidad de audio". Cada métrica posee sus propias fortalezas y limitaciones, lo que sugiere que no existe una única métrica universalmente aplicable. Esta fragmentación subraya la dificultad de establecer correlaciones directas y consistentes entre diferentes tipos de métricas, un punto crucial que se explorará en la discusión de los resultados. La emergencia de modelos como NISQA, que aprenden a predecir puntuaciones subjetivas de manera objetiva, representa un intento sofisticado de cerrar esta brecha, buscando una evaluación más integral de la calidad percibida.

3. METODOLOGIA

3.1 Estudio Denoising

En esta sección se propone evaluar la correlación entre la aplicación de algoritmos de *denoising* a los datos de entrenamiento y la mejora subjetiva de los sistemas TTS.

Selección de Algoritmos: Se seleccionaron tres algoritmos de *denoising* basados en redes neuronales: Wave U-Net [14], HiFi-GAN (referido como Demucs en las tablas y figuras del estudio) [15] y DeepFilterNet (DFN) [16]. Para la tarea de TTS, se eligió el modelo FastPitch,

una arquitectura basada en *transformers*, debido a su capacidad de ajuste fino, su carácter de código abierto y su rápida velocidad de entrenamiento.

Generación de Estímulos de Prueba: Los estímulos fueron audios sintetizados por el modelo FastPitch. Se utilizó un subconjunto de 15 minutos de la colección de *datasets* ArchiVoz, que contenía un nivel nominal de ruido representativo de condiciones reales. Este *dataset* fue procesado por los algoritmos de *denoising* seleccionados, y posteriormente, el modelo FastPitch fue entrenado con cada *dataset* resultante.

Parámetros de Entrenamiento: Para asegurar la consistencia, el proceso de entrenamiento de FastPitch se mantuvo constante en todos los experimentos. El mejor *checkpoint*, basado en la pérdida de validación, se registró después de una hora de entrenamiento, y todos los hiperparámetros (tasa de aprendizaje, tamaño de *batch*, tasa de *dropout*) permanecieron fijos.

Texto de Entrada: El texto utilizado como entrada para el modelo TTS fue un párrafo diseñado por Gurlekian et al., [17] para maximizar la cobertura de fonemas, lo que lo hace efectivo para la evaluación de TTS. Las salidas de TTS generadas tuvieron una duración de 3 a 8 segundos para simplificar la evaluación subjetiva. Todas las muestras se normalizaron a la misma frecuencia de muestreo y nivel nominal de salida.

Diseño de la Prueba Subjetiva (CMOS): Se utilizó la prueba Comparison Mean Opinion Score (CMOS) para evaluar la mejora del rendimiento de TTS. Se compararon cuatro condiciones (sin *denoising* y con los tres algoritmos de *denoising*), resultando en seis pares de comparación, más un par adicional de línea base ("Normal vs. Normal-V2") para evaluar la variabilidad estocástica del modelo. Los participantes fueron instruidos para usar auriculares en un ambiente silencioso y podían escuchar las muestras cuantas veces fuera necesario.

Parámetros Objetivos Medidos: El estudio empleó métricas objetivas de vanguardia para la evaluación de *denoising* y TTS: PESQ, STOI y MCD (Mean Mel-Cepstral Distortion). Todas estas métricas fueron calculadas utilizando el paquete TorchMetrics en Python.

3.2 Estudio NISQA

El estudio sobre el modelo NISQA se centró en evaluar su utilidad para clasificar audios de voz según su calidad. Para las pruebas iniciales, se emplearon cinco pares de *datasets*, cada uno conteniendo la grabación de una persona leyendo un libro. Estos audios fueron seleccionados de

diversos locutores y condiciones de grabación para asegurar una amplia representatividad. Cada grabación fue segmentada en clips de no más de *15* segundos mediante un algoritmo de detección de habla, eliminando segmentos no útiles. Posteriormente, cada segmento fue procesado con el algoritmo de *denoising* DeepFilterNet, elegido por su rendimiento superior y su capacidad de no empeorar la calidad del audio incluso en los peores casos.

Cada par de *datasets* consistía en los segmentos originales y sus contrapartes *denoised*. Para el análisis inicial, se graficaron los resultados de NISQA, enfocándose en el MOS. Se realiza una comparación directa del MOS entre audios originales y *denoised*, y también se grafica un histograma de la distribución de los puntajes de cada *dataset*. Este mismo análisis gráfico se replicó para los otros parámetros de salida de NISQA: Ruido, Discontinuidad, Coloración y Volumen.

Para profundizar en los parámetros de Ruido y Discontinuidad, se realizó un análisis adicional degradando audios "limpios" de forma artificial. Se consideraron limpios a aquellos audios con MOS superior a 4.5 y Ruido y Discontinuidad superiores a 4. Se aplicaron dos tipos de degradación: ruido blanco con diferentes magnitudes y *dropouts* (pérdida de muestras) de varias longitudes aleatorias. Luego, se generaron dos copias degradadas de los audios limpios, una para cada tipo de degradación. Los valores utilizados fueron aleatorizados, para el ruido blanco fue de 0 a 1% del volumen máximo y para los *dropouts* de 0 a 500 muestras por pérdida, la cantidad de pérdidas también fue aleatoria. El objetivo de este último estudio fue determinar la correlación de estos valores aleatorios con los parámetros entregados por el modelo. Esto permitiría validar si los valores proporcionados por NISQA eran coherentes con la realidad de la degradación, para poder confiar en los resultados sin necesidad de escuchar el audio e incluirlo en la cadena.

4. ANÁLISIS y RESULTADOS

4.1 Resultados Denoising

Resultados Metricas Objetivas:

Los resultados de las métricas objetivas evaluadas en el dataset de entrenamiento son:

Tabla 1: Resultado evaluación objetiva por parámetro y algoritmo.

Parámetro	Wave U-NET	Demucs	DeepFilterNet
PESQ	3.08 ± 0.12	4.43 ± 0.13	4.44 ± 0.09
STOI	0.98 ± 0.004	1.00 ± 0.001	1.00 ± 0.002
MCD (dB)	23.05 ± 1.61	6.86 ± 3.36	15.06 ± 5.36

Resultados Métricas Subjetivas:

El estudio subjetivo recopiló un total de 26 respuestas, de las cuales 2 fueron descartadas por no seguir las instrucciones, resultando en 24 respuestas válidas. De estas, 8 participantes declararon tener experiencia previa con sistemas TTS o modelos de IA generativa, siendo clasificados como "expertos", mientras que el resto conformó el grupo del "público general" o "no expertos".

La distribución de los resultados del análisis subjetivo se presenta en la (Figura 1). Se observa una preferencia por el resultado del TTS entrenado con el dataset procesado por DeepFilterNet. Por el contrario, el TTS entrenado con el dataset pasado por Wave U-Net es el que registra una peor preferencia subjetiva.

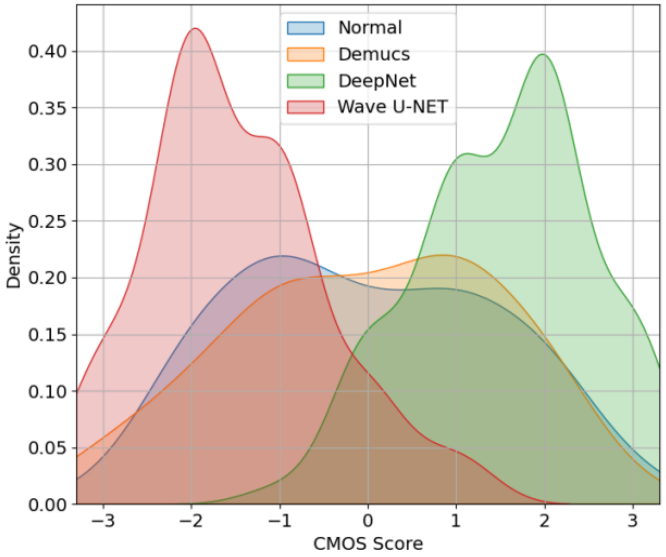

Figura 1: Estimación de la densidad de CMOS para cada algoritmo.

La comparación general por algoritmo es presentada en (Figura 2). Los *datasets* sin procesar (Normal) y los procesados con Demucs arrojaron evaluaciones subjetivas similares. Se incluyó

un caso de prueba de normalización ("Normal vs. Normal-V2") para establecer una línea base de la variabilidad intrínseca del modelo; el resultado de este caso base mostró una ligera preferencia por una de las versiones "Normal", lo que sugiere que la variabilidad del modelo aún podría influir en el resultado final de los modelos TTS, a pesar de los esfuerzos por reducirla.

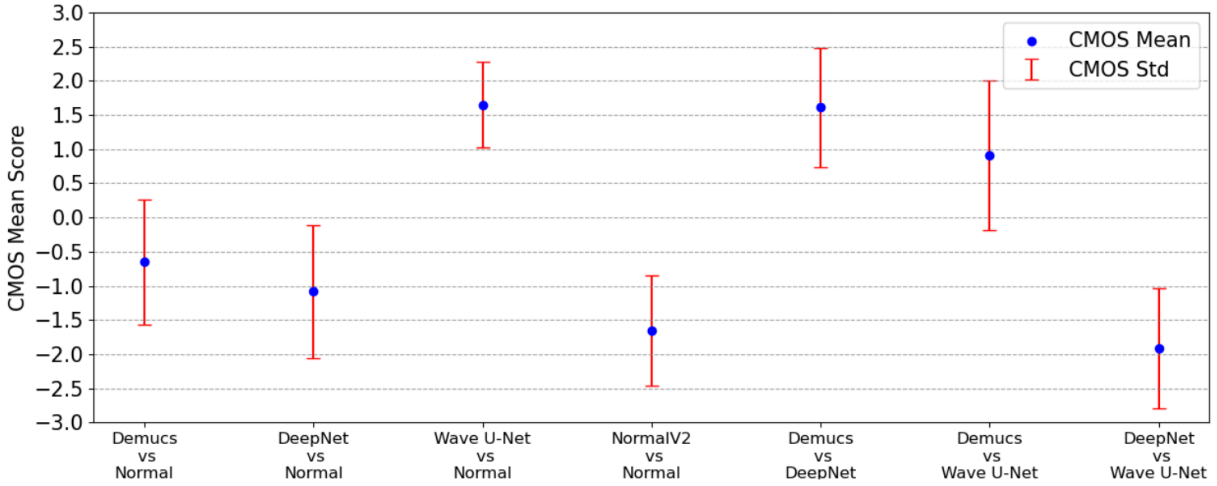

Figura 2: Resultados de CMOS para cada par de algoritmos comparados.

Análisis estadístico:

Para validar la significancia de los resultados, se realizó un análisis estadístico exhaustivo. La prueba de Shapiro-Wilk reveló que la mayoría de las distribuciones de las respuestas CMOS, agrupadas por algoritmo, no seguían una distribución normal (p-value < 0.05), con la excepción de las preguntas 4 y 6. Dada la no normalidad predominante, se aplicó la prueba de Kruskal-Wallis a los cuatro grupos de algoritmos, concluyendo que sus distribuciones no compartían varianzas similares (p-value < 0.05).

Para evaluar la significancia estadística de las medias en datos no normales, se realizó una prueba de permutación por pares. Los resultados, detallados en la Tabla 2, indicaron significancia estadística en las medias para las comparaciones entre "Normal vs. DeepNet" (p=0.012) y "Normal vs. Wave U-Net" (p=0.001). Sin embargo, no se encontró significancia estadística entre las medias de los grupos "Normal" y "Demucs" (p=0.229).

Tabla 2: Valores del test de permutación para los diferentes algoritmos.

	r	p-value
Normal vs Demucs	0.88	0.229
Normal vs DeepNet	-1.16	0.012
Normal vs Wave U-Net	1.83	0.001

La relación entre PESQ y CMOS se ilustra en la Figura 3. Demucs y DeepNet exhibieron valores PESQ similares, lo que sugiere un rendimiento de *denoising* comparable. Sin embargo, sus valores CMOS difirieron significativamente, lo que indica que mejoras objetivas similares en la calidad de audio no se traducen necesariamente en una calidad subjetiva equivalente en las salidas del modelo TTS. Además, Wave U-Net mostró una mejora en la calidad de audio a través del *denoising* (medida por PESQ) pero un rendimiento subjetivo inferior en comparación con la señal "Normal" sin procesar. Esto confirmó que PESQ no se correlaciona con las preferencias CMOS para el modelo TTS específico analizado.

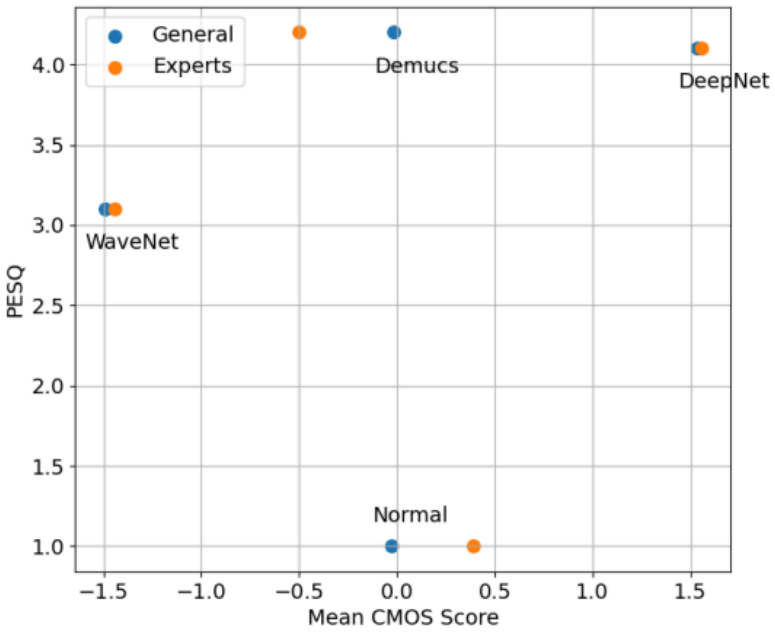

Figura 3: PESQ vs CMOS para cada algoritmo

En cuanto a la relación entre MCD y CMOS, los resultados se presentan en la Figura 4. Se observó una tendencia general donde valores MCD más bajos correspondían a un mejor rendimiento subjetivo para Wave U-Net y Demucs. Sin embargo, DeepNet logró un rendimiento subjetivo superior a pesar de tener un valor MCD más alto que Demucs. Para explorar esta

relación con mayor profundidad, se realizó un análisis de correlación de Spearman, obteniendo un coeficiente *r* de -0.64 y un p-value de 0.11. Estos resultados indicaron una evidencia estadística insuficiente para confirmar una correlación directa entre MCD y CMOS.

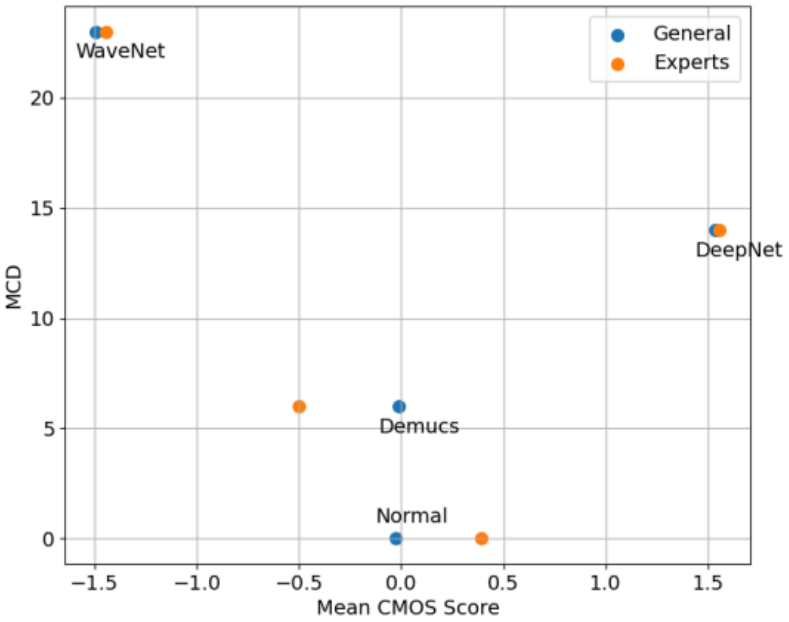

Figura 4: MCD vs CMOS para cada algoritmo

4.2 Resultados NISQA

El análisis de los resultados surge a partir de la observación de los gráficos obtenidos para cada conjunto de datos, acompañado de la escucha de algunos audios llamativos. En particular, se eligieron los audios que representaban los diferentes comportamientos del algoritmo para estudiarlos con mayor profundidad.

En la Figura 5 se presentan los resultados generales de la primera prueba. Se muestran el promedio y el intervalo de confianza de cada conjunto antes y después del *denoising*. De dicho gráfico, en conjunto con la escucha de los audios, se observó que para todos los conjuntos hubo mejoras en la calidad. Para los conjuntos de Ana y Sebastián fue leve pero en el resto es notoria a lo largo del conjunto. La principal diferencia percibida es la eliminación de ruidos estacionarios.

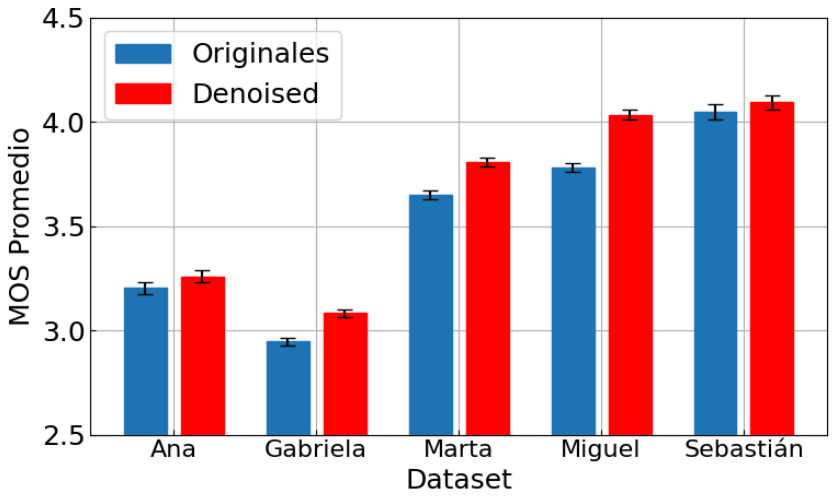

Figura 5: Promedio e intervalo de confianza del MOS de los conjuntos antes y después del denoising.

Luego, la Figura 6 presenta el mismo análisis para cada parámetro de NISQA. Dichos resultados concuerdan con el análisis percibido previamente, indicando que las mejoras en la calidad se derivan de una reducción general del ruido en los audios. Esto se ve en los puntajes mayores del parámetro Noisiness para los conjuntos *denoised.* Para algunos conjuntos también se ven mejoras en el Loudness o Volumen, lo que se atribuye a una mejora en la relación señal-ruido.

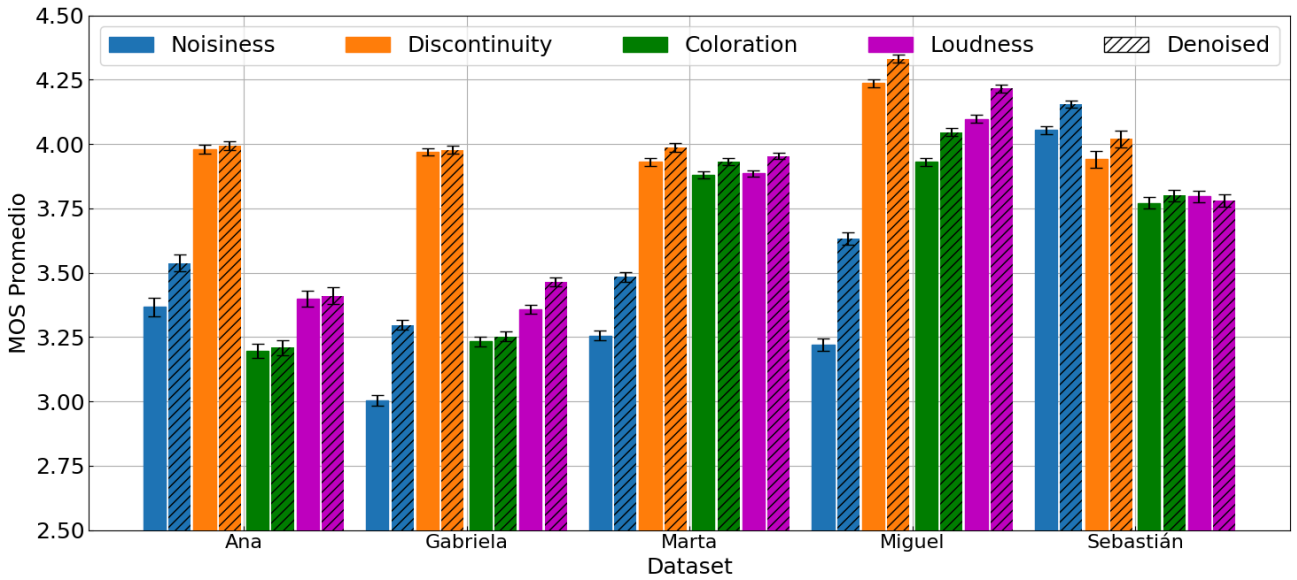

Figura 6: Promedio e intervalo de confianza del cada parámetro de NISQA para los conjuntos antes y después del denoising.

Por último se presenta la Figura 7 con los gráficos de dispersión de cada degradación, donde se puede observar la correlación entre los factores de la degradación aplicada y los resultados predichos por NISQA. Para ambos parámetros, se calculó la correlación de Spearman y se obtuvieron resultados significativos. En el caso de las discontinuidades, la correlación fue fuerte

(ρ = -0.651, $p < 0.01$) mientras que para el ruido, fue muy fuerte (ρ = -0.864, $p < 0.01$). Estos resultados indican que NISQA responde adecuadamente a la influencia de ambas degradaciones en la calidad percibida.

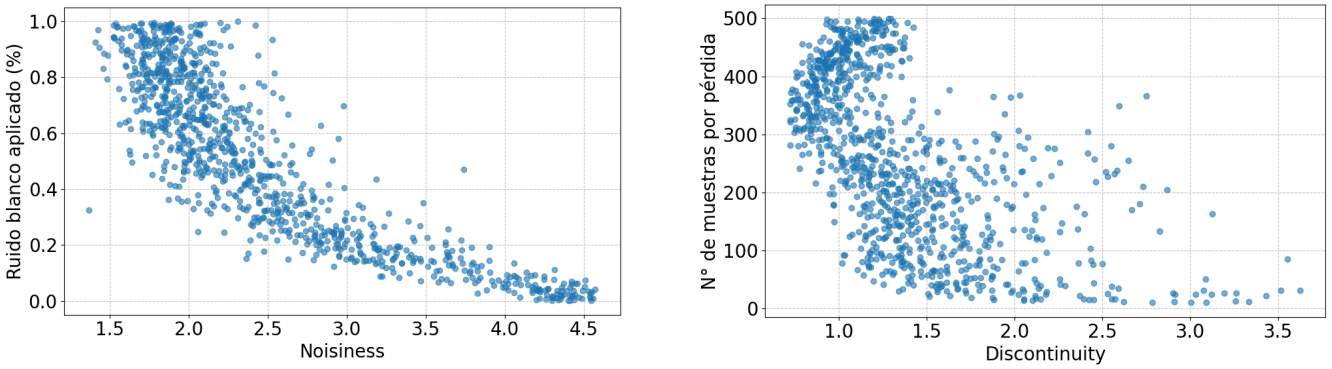

Figura 7: Correlación entre los factores de degradación artificial y las predicciones de los parámetros de NISQA.

5. DISCUSIONES

5.1 Impacto del denoising y evaluación de calidad en TTS

La evaluación conjunta de sistemas de *denoising* y TTS es intrínsecamente compleja, y la interacción entre ambos añade capas adicionales de dificultad. Los hallazgos de los estudios de DeepFilterNet y NISQA revelan que, a pesar de la eficacia individual de métricas objetivas como PESQ y MCD para evaluar el *denoising* o los modelos TTS por separado, no se encontró una relación directa y consistente entre estas métricas objetivas y las puntuaciones subjetivas de CMOS en la configuración específica analizada.

Un resultado notable es que DeepFilterNet superó a Demucs en la evaluación subjetiva (CMOS), a pesar de que DeepNet presentó una distancia MCD más alta y valores PESQ similares a los de Demucs. Esta observación contradice la hipótesis inicial de que un MCD más bajo, indicativo de una mejor preservación de las características acústicas, resultaría en un rendimiento TTS superior. Una posible explicación para este fenómeno radica en que DeepFilterNet podría alterar las características acústicas subyacentes del audio de voz de una manera que se alinee más estrechamente con el entorno de grabación profesional de los datos originales con los que se entrenó el modelo FastPitch. Esta hipótesis sugiere que la efectividad de un algoritmo de *denoising* en el contexto de TTS no solo depende de su capacidad para eliminar ruido, sino

también de cómo sus transformaciones se adaptan a las particularidades del modelo TTS posterior.

5.2 Observación de NISQA como herramienta de evaluación

El estudio sobre NISQA proporciona un contexto valioso para comprender la elección y el rendimiento de DeepFilterNet. NISQA se utilizó para evaluar la eficacia de DeepFilterNet en la mejora de las puntuaciones MOS de audios para la generación de *datasets* de ASR. Los resultados de este estudio demostraron que DeepFilterNet no solo mejoró las puntuaciones MOS promedio, sino que también fue seleccionado por su capacidad de no degradar la calidad del audio incluso en los peores escenarios. Esta validación inicial de la eficacia de *denoising* de DeepFilterNet, medida por un modelo robusto y no intrusivo como NISQA, establece una base sólida para su selección en el contexto de TTS.

5.3 Interacción de las 2 etapas

Aunque el estudio de DeepFilterNet no utilizó directamente NISQA para evaluar el rendimiento de TTS, los resultados del estudio de NISQA proporcionan una justificación indirecta y sólida para la elección y el rendimiento superior de DeepNet en el contexto de la síntesis de voz.

Se demostró que DeepFilterNet mejoró consistentemente las puntuaciones MOS promedio de los *datasets* de audio y, no degradó la calidad del audio incluso en los casos más desfavorables. Esta validación inicial de DeepFilterNet, realizada a través de un modelo de evaluación de calidad de voz no intrusivo y de alto rendimiento como NISQA, subraya la capacidad de DeepFilterNet para producir audio de mayor calidad.

La selección de DeepFilterNet en la comparación de algoritmos de *denoising* se basó en su robustez y su diseño fundamentado en el aprendizaje profundo. El hecho de que DeepFilterNet lograra las puntuaciones CMOS subjetivas más altas en el contexto de TTS, a pesar de las complejidades en la correlación con métricas objetivas como PESQ y MCD, refuerza la idea de que un *denoising* efectivo, como el validado por NISQA, se traduce en una mejor calidad subjetiva percibida en la voz sintetizada. Esto demuestra cómo la validación de un componente clave, como un algoritmo de *denoising*, a través de una métrica de calidad de audio fiable como NISQA, puede influir positivamente en el éxito de sistemas más complejos como los de síntesis de voz.

6. CONCLUSIONES

El trabajo presenta y valida un flujo práctico para optimizar datos de audio destinados a TTS combinando una evaluación no intrusiva (NISQA) con procesamiento de denoising (comparando Wave U-Net, Demucs y DeepFilterNet) y entrenamiento de un modelo FastPitch.

Los resultados muestran que DeepFilterNet produjo la mejor preferencia subjetiva en las salidas TTS (CMOS), con diferencias estadísticamente significativas frente a la condición sin denoising en las comparaciones relevantes, mientras que Demucs ofreció rendimiento subjetivo similar al original y Wave U-Net empeoró la preferencia. Al mismo tiempo, métricas objetivas tradicionales (PESQ, MCD) no se correlacionaron de manera consistente con las preferencias CMOS del TTS, lo que evidencia la limitación de confiar sólo en medidas intrusivas u objetivas para predecir calidad percibida en síntesis.

NISQA demostró ser una herramienta útil para evaluar calidad general a gran escala: sus parámetros (MOS, noisiness, discontinuidad, coloración, volumen) permitieron cuantificar mejoras tras el denoising y detectar comportamientos concretos (p. ej. eliminación de ruidos estacionarios) sin necesidad de escucha manual. Además, DeepFilterNet mostró robustez al no degradar archivos en los peores escenarios, lo que lo convierte en una opción segura para preprocesado de corpora encontrados.

En términos prácticos se recomienda usar NISQA para filtrar y diagnosticar segmentos antes del entrenamiento, y aplicar un denoising validado (como DeepFilterNet en este estudio) para mejorar la inteligibilidad y fidelidad subjetiva del TTS. Esto resulta especialmente valioso para construir voces regionales o entrenar con grabaciones no profesionales.

REFERENCIAS

[1] X. Tan, T. Qin, F. K. Soong, and T.-Y. Liu. A survey on neural speech synthesis, arXiv preprint arXiv:2106.15561. 2021.

[2] E. Cooper. Text-to-speech synthesis using found data for low-resource languages, Doctoral dissertation, Columbia University. 2019.

[3] H. Yu, et al. Autoprep: an automatic preprocessing framework for in-the-wild speech data, in Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2024, pp. 1136–1140.

[4] A. Łancucki. FastPitch: parallel text-to-speech with pitch prediction, in Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2021.

[5] Y. Kuo, S. Aryal, G. Degottex, S. Kang, P. Lanchantin, and I. Ouyang. Data selection for improving naturalness of TTS voices trained on small found corpuses, in Proceedings of the IEEE Spoken Language Technology Workshop (SLT). 2018, pp. 319–324.

[6] International Telecommunication Union. ITU-T Recommendation P.800: Methods for subjective determination of transmission quality. ITU. 1996. url: https://www.itu.int/rec/T-REC-P.800.

[7] International Telecommunication Union. ITU-T Recommendation P.808: Subjective evaluation of speech quality with a crowdsourcing approach. ITU. 2018. url: https://www.itu.int/rec/T-REC-P.808.

[8] International Telecommunication Union. ITU-T Recommendation P.862: Perceptual Evaluation of Speech Quality (PESQ): an objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs. ITU. 2001. url: https://www.itu.int/rec/T-REC-P.862.

[9] International Telecommunication Union. ITU-T Recommendation P.863: Perceptual Objective Listening Quality Assessment (POLQA). ITU. 2018. url: https://www.itu.int/rec/T-REC-P.863.

[10] G. Mittag, B. Naderi, A. Chehadi, and S. Möller. NISQA: a deep neural network-based end-to-end speech quality assessment model, Proceedings of Interspeech. 2021, pp. 2127–2131. url: https://www.isca-speech.org/archive/interspeech_2021/mittag21_interspeech.html.

[11] M. Wältermann. Dimension-based quality modeling of transmitted speech, T-Labs Series in Telecommunication Services. Springer, Berlin, Heidelberg. 2013. isbn: 978-3-642-35018-4. doi:10.1007/978-3-642-35019-1. url: https://link.springer.com/book/10.1007/978-3-642-35019-1.

[12] W. D. Voiers. Speech intelligibility and speaker recognition. 1977.

[13] R. Kubichek. Mel-cepstral distance measure for objective speech quality assessment, in Proceedings of the IEEE Pacific Rim Conference on Communications, Computers and Signal Processing, vol. 1. 1993, pp. 125–128.

[14] T. Walczyna and Z. Piotrowski. Wave-U-Net speech denoising, in K. S. Soliman (Ed.), Artificial Intelligence and Machine Learning. Cham: Springer Nature Switzerland. 2024, pp. 52–57.

[15] J. Su, Z. Jin, and A. Finkelstein. HiFiGAN: high-fidelity denoising and dereverberation based on speech deep features in adversarial networks, in Proceedings of Interspeech 2020. 2020, pp. 4506–4510.

[16] H. Schroter, A. N. Escalante-B, T. Rosenkranz, and A. Maier. Deepfilternet: a low complexity speech enhancement framework for full-band audio based on deep filtering, in Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2022, pp. 7407–7411.

[17] J. Gurlekian, M. M. Guemes, D. A. Evin, and M. Torres. Normalización del texto “Los sentidos” y su aplicación en la evaluación de habla continua, Onomazein. 2021.