UNTREF UNIVERSIDAD NACIONAL DE TRES DE FEBRERO

INGENIERÍA EN SONIDO

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

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

Universidad Nacional de Tres de Febrero (UNTREF)

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

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

AGRADECIMIENTOS

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

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

DEDICATORIA

Elige a quién o a qué quieres dedicárselo.

Elegir el motivo de la dedicatoria (orientativo).

ÍNDICE DE CONTENIDOS

Índice de figuras

Índice de tablas

RESUMEN

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

Palabras Clave:

ABSTRACT

Ídem que para castellano.

Keywords:

1. INTRODUCCIÓN

1.1. FUNDAMENTACIÓN

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

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

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

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

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

1.2. OBJETIVOS

1.2.1. Objetivo general

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

1.2.2. Objetivo especifico

Los objetivos específicos son:

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

1.3. ESTRUCTURA DE LA INVESTIGACIÓN

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

Organización del documento:

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

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

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

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

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

2. MARCO TEÓRICO

2.1. DESCRIPTORES DE CALIDAD DE AUDIO

2.1.1. Métricas de degradación de la señal

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

2.1.2. Métricas de entorno

Explicar: T30, C50, C80, D50

2.1.3. Métricas del habla

Explicar: F0, Speaker Rate, MCD

2.2. TEXT-TO-SPEECH (TTS)

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

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

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

2.3. REDES NEURONALES

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

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

2.4. INTELIGENCIA ARTIFICIAL GENERATIVA

Completar

2.5. MODELOS DE DIFUSIÓN

Completar

3. ESTADO DEL ARTE

3.1. MODELOS DE TTS

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

3.2. CADENAS DE PRE PROCESAMIENTO

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

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

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

4. DESARROLLO

4.1. RECOPILACIÓN DE BASES DE DATOS

4.1.1. Datos in-the-wild

Para poder evaluar cadenas de procesamientos a conjuntos de datos del habla, se utiliza la base de datos de habla en español de Argentina confeccionada por el grupo de investigación Intercambios Transorgánicos, donde se han recopilado un total de 24 horas de grabaciones del habla en diferentes condiciones, tanto de calidad de los audios como en la diversidad de los hablantes. Dado que las recolección del material se lleva a cabo principalmente de fuentes publicas de internet, este corpus *in-the-wild* es el conjunto ideal para evaluar el funcionamiento de pre procesamientos de audios. En las futuras secciones, se refiere a este conjunto de datos como original (al ser la versión si procesamientos).

En busca de armar un conjunto de datos que capture las diferencias lingüísticas de todo el país, se seleccionan datos con dialecto bonaerense y centro, según las regiones que establece (de Weinberg y de Mirande, 2004). En concreto, el conjunto de datos cuenta con 32 hablantes del acento bonaerense, y 27 del acento centro. El objetivo a futuro es armar un dataset con todos los dialectos representativos de la Argentina, pero primero se busca validar el funcionamiento de la cadena de pre procesamiento con 2 dialectos representativos para después ampliar la base de datos y cubrir todos los acentos del país.

4.1.2. Datos profesionales

Los conjuntos de datos profesionales se recolectan de trabajos previos. Como no es posible armar un conjunto de datos de referencia lo suficientemente extenso solo con audio en español de Argentina, se utiliza también el español con otras variantes dialécticas.

- Conjunto de datos de Google (Guevara-Rukoz et al., 2020): Cuenta con 8 horas de audio de dialecto bonaerense, con un total de 44 hablantes.
- Emilia (Torres et al., 2019): Con un total de 4 horas de audio de dialecto bonaerense.

En total, el conjunto de datos profesional cuenta con X horas de audio de X hablantes diferentes.

4.2. DESARROLLO DE LA CADENA DE PRE PROCESAMIENTO

Describir como es la cadena de pre procesamiento

4.2.1. Diferentes configuraciones

Describir las diferentes configuraciones evaluadas

4.3. EVALUACIÓN DE LOS CONJUNTOS DE DATOS

Explicar de forma general las métricas que se van a utilizar y como determinar cual es el mejor dataset.

4.3.1. Reducción del dataset

Como medir tamaño del dataset

4.3.2. Calidad de la grabación

PESQ, STOI, SI-SDR, SNR

4.3.3. Condiciones acústicas

T30, C50, D50

4.3.4. Diferencias del habla

F0-STD, MCD

4.3.5. Métrica conjunta

Explicar como combinar toda esta info

4.4. ENTRENAMIENTO DEL MODELO DE ESTIMACIÓN DE DENSIDAD

Explicar porque necesito hacer lo del modelo de estimación de densidad

4.4.1. Validación con medelo zero-shot

Definir modelo y explicar la justificación de este experimento

4.5. DESCRIPCIÓN DE PRUEBAS ESTADÍSTICAS

Detallar un poco las validaciónes estadísticas que se van a realizar

4.6. MODELO DE TTS ZERO-SHOT

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

5. RESTULADOS Y ANÁLISIS

5.1. RESULTADOS DE LAS DIFERENTES VARIANTES DE LA CADENA

Comparar el análisis objetivo y determinar el mejor subset

5.2. COMPARACIÓN ENTRE CONJUNTOS DE DATOS

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

5.3. COMPARACIÓN POR ESTIMACIÓN DE DENSIDAD

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

5.4. VALIDACIÓN MODELO DE TTS ZERO-SHOT

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

6. CONCLUSIONES

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

7. LÍNEAS FUTURAS DE INVESTIGACIÓN

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

Bibliografia

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

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

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

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

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

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

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

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

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*.

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

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

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*.

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)*.

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

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

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

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

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

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

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

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

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

ANEXO I. FORMATO INTERNO

AI 1. *Numeración*

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