AGRADECIMIENTOS

(a completar)

ÍNDICE DE CONTENIDOS

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

LISTA DE ACRÓNIMOS

ALG	Algoritmo de Griffin-Lim
ASR	Reconocimiento automático de voz
BERT	Codificación bidireccional basada en transformadores
BLSTM	Memoria de corto y largo plazo bidireccional
CNN	Aumento artificial de datos
DA	Aprendizaje neuronal profundo
DNN	Red neuronal profunda
FR	(Algoritmo de) Referencia completa
MOS	Promedio de la opinión subjetiva
ReLU	Unidad lineal rectificada
RMSE	Error cuadrático medio
STFT	Transformada de tiempo corto de Fourier
TTS	Sistema de texto a voz
VCC	Competencia de conversiones de voz

RESUMEN

En esta investigación se aborda el desarrollo de un procedimiento para la determinación objetiva de la calidad de la voz humana generada por sistemas de síntesis artificiales. Se presenta la metodología adoptada para la implementación de un sistema basado en redes neuronales que sea capaz de predecir una valoración subjetiva sobre la naturalidad de una voz sintetizada. El entrenamiento y evaluación de dicho modelo fue realizado sobre una base de datos creada a partir de distintas voces sintetizadas por algoritmos de texto a voz, grabaciones de discurso humano reales, y grabaciones procesadas digitalmente de ambos grupos previamente mencionados. Dichas voces fueron juzgadas subjetivamente en un test tipo-MOS realizado de forma online por [*Completar con el numero de participantes de la encuesta subjetiva*] . A partir de los resultados obtenidos se observa [*Completar con los resultados obtenidos (correlación de la metrica obtenida respecto de las evaluaciones subjetivas, y añadir conclusiones mas relevantes*]

Palabras clave: calidad del habla, texto a voz, evaluación objetiva, deep learning

ABSTRACT

In this research, the development of a procedure for objectively evaluating the quality of human-like speech generated by artificial synthesis methods is proposed. Detailed in this document is the methodology adopted in the implementation of a neural network system, capable of predicting the subjective score of a synthesized voice. Training and evaluation of said model is carried out on a custom database generated from a number of different text to speech algorithms, as well as recordings of real human speech, and digitally altered versions of both of those groups. These recordings were then assessed subjectively via an online MOS-like test performed by [*Completar con el numero de participantes de la encuesta subjetiva*] . From the obtained results we conclude that [*Completar con los resultados obtenidos (correlación de la metrica obtenida respecto de las evaluaciones subjetivas, y añadir conclusiones mas relevantes*]

Keywords: speech quality, text to speech, objective assessment, deep learning

1. INTRODUCCIÓN

1.1. FUNDAMENTACIÓN

La síntesis del habla consiste en la tarea de generar una voz humana a partir de otro tipo de entrada, ya sea texto, movimiento de labios o fonemas. En la mayoría de sus aplicaciones modernas, estos sistemas toman el texto como método de entrada. Esto se debe en parte a los avances en el campo del procesamiento del lenguaje natural. Un sistema de texto a voz (TTS por sus siglas en inglés, text to speech system) apunta a convertir el lenguaje escrito en discurso humano audible.

Históricamente esta tarea fue llevada a cabo por sistemas que concatenan fonemas pregrabados (sistemas concatenativos) o que modelan un audio a través de parámetros acústicos definidos arbitrariamente (sistemas paramétricos). A lo largo de la última década, hubo avances en el poder computacional que permitieron explorar y desarrollar diversos modelos de TTS basados en el aprendizaje automático profundo (*Deep Learning*), a partir de diversas metodologías: En 2016 el equipo de DeepMind introduciría WaveNet [**?**] revolucionando el campo del TTS con el primer modelo que sintetizaba el habla humana muestra por muestra. Este pilar fue seguido por numerosos avances y mejoras basadas en sistemas de paralelización [**?**], transformadores [**?**] y sistemas de tipo Flow [**?**].

Evaluar la calidad de estas distintas soluciones implica, entre otras cosas, juzgar la "naturalidad" de la voz humana generada. El estándar para realizar esa evaluación son las pruebas subjetivas, realizadas sobre sistemas entrenados con bases de datos estandarizadas, usualmente en idiomas inglés o chino. El Mean Opinion Score (MOS) [**?**] (puntaje promedio subjetivo) es el método más frecuentemente utilizado para llevar a cabo esa prueba. Dicha métrica tiene un rango de 1 a 5, en la que el habla humana real yace entre las puntuaciones de 4,5 a 4,8. El test MOS se conduce sobre las voces sintetizadas para dar un idea de que tan naturales son los resultados de los sistemas TTS.

Realizar un test subjetivo es costoso monetaria y temporalmente, e indefectiblemente presenta una barrera a la hora de evaluar pequeñas modificaciones o iteraciones en el desarrollo de un sistema TTS. Este documento detalla el desarrollo de un procedimiento de evaluación ob-

jetiva para sistemas de texto a voz. Dicha evaluación busca tener un alto grado de correlación con los resultados de las pruebas subjetivas. Se planea ofrecer la métrica desarrollada de forma gratuita y como código abierto.

Intercambios Transorgánicos (Dir. Gala González Barrios) es un programa de investigaciones radicado en el Muntref Centro de Arte y Ciencia, IIAC, UNTREF. Desde este programa se realizan proyectos de investigación que desarrollan interfaces interactivas desde las artes electrónicas y las ingenierías en relación con el campo de la salud. En este momento se encuentran desarrollando un sistema TTS en español argentino, orientado a funcionar como parte de una prótesis para personas que se encuentren en la situación de comprometer su voz, parcial o totalmente. La investigación planteada en esta tesis busca proveer una herramienta para evaluar y ayudar al progreso y desarrollo de dicha herramienta.

El trabajo propuesto es una investigación cuantitativa de alcance exploratorio. Su propósito es el de brindar a la comunidad de investigadores que desarrollan sistemas de texto a voz, una evaluación objetiva automatizada que presente un alto grado de correlación con las pruebas subjetivas que conforman el estándar de la industria para juzgar el habla. Se plantea extraer un descriptor de cada audio a juzgar, y entrenar una pequeña red neuronal de forma supervisada, de modo que la misma pueda predecir el valor MOS que obtendría el audio si fuese juzgado subjetivamente por un grupo de individuos.

1.2. OBJETIVOS

1.2.1. OBJETIVO GENERAL

El diseño, implementación y validación de un sistema computacional capaz de predecir la preferencia subjetiva promedio (MOS), sobre distintas voces sintetizadas por computadoras. La investigación se condujo en el idioma castellano.

1.2.2. OBJETIVOS ESPECÍFICOS

Se proponen los siguientes objetivos específicos:

- **Recolección de audios sintetizados.** Recolectar audios sintetizados por sistemas TTS de variada calidad de clonado de voz y procedencia, además de audios de hablantes humanos

reales.

- **Transformación de audios recolectados.** Alterar la calidad de una porción de los audios recolectados mediante distintas técnicas de procesado de señales y voz, obteniendo así una base de datos mas balanceada.

- **Extracción de descriptores objetivos de cada audio.** Extraer representaciones vectorizadas de cada audio mediante una red neuronal convolucional que aprenda a representar la características acústicas de cada voz a evaluar. Estos embeddings son utilizados para reconocimiento de hablantes se extraen mediante una red neuronal que deberá ser configurada y posiblemente re-entrenada para funcionar con el idioma castellano.

- **Diseño de una prueba subjetiva para etiquetar los audios.** Diseñar y llevar a cabo una prueba subjetivas para obtener una puntuación para cada audio obtenido, seguido de una validación de los datos obtenidos.

- **Diseño de una red neuronal para predecir la naturalidad de cada audio.** Entrenar una pequeña red neuronal de forma supervisada, con los audios recolectados como entrada y sus calificaciones MOS como salida deseada.La función de costo y el ajuste de la red tendrán como objetivo acercar sus predicciones a los valores correctos MOS recolectados. Para poner a prueba el modelo entrenado, se reserva una parte del conjunto de datos recolectados para llevar a cabo una evaluación del sistema.

1.3. ESTRUCTURA DE LA INVESTIGACIÓN

En capítulo 2 se detalla un marco teórico vinculado a los procesos detrás de las distintas implementaciones posibles para sintetizar voces artificialmente, la evaluación subjetiva MOS, y la predicción de parámetros subjetivos mediante métricas objetivas. También se provee una breve explicación de las distintas técnicas detrás de los métodos de alteración de hablantes que se utilizaron en el transcurso de la investigación, así como también información vinculada la la vectorización de hablantes utilizada. En el capitulo 3 se presenta el estado del arte vinculado a la evaluación objetiva de sistemas TTS. El capítulo 4 consta del desarrollo de la investigación,

en el cual se evidencian las distintas características de la base de datos obtenida, el diseño de la prueba subjetiva y el diseño e implementación de la red neuronal clasificadora de TTS. En el capitulo 5 y 6 se presentan y analizan los resultados obtenidos. Finalmente, en el capitulo 7 se informan las conclusiones de la investigación desarrollada, y el capitulo 8 ofrece posibles lineas de investigación futuras que se desprenden de los resultados obtenidos.

2. MARCO TEÓRICO

2.1. Sistemas de texto a voz

2.1.1. Síntesis concatenativa y paramétrica

La síntesis del habla consiste en la tarea de generar discurso humano, a partir de alguna entrada arbitraría. Son de particular interés para desarrollar interfaces de comunicación entre humanos y computadoras, los sistemas de texto a voz, o TTS (text to speech system). Históricamente se emplearon dos metodologías para llevar a cabo esta tarea: La síntesis concatenativa, donde distintos fonemas y palabras pre-grabadas son utilizadas para completar una frase solicitada, y la síntesis paramétrica, donde un modelo acústico es condicionado para modificar nuevamente voces pre-grabadas de acuerdo a variables arbitrarias solicitadas por un usuario. En ambos casos es necesario almacenar fonemas o palabras pre-grabadas, y la calidad de la voz resultante no es ideal, exhibiendo una característica "roboticidad". Es aquí donde entran en juego técnicas de síntesis basadas en el aprendizaje profundo neuronal, o Deep Learning

2.1.2. Aprendizaje profundo neuronal aplicado a TTS

A partir de 2016 el campo de los TTS fue revolucionado por distintas arquitecturas basadas en Deep Learning, que mejorarían considerablemente la calidad de las voces sintetizadas. Previo a adéntranos en una breve explicación detrás del funcionamiento de las distintas arquitecturas, podemos observar como un sistema TTS puede ser formulado matemáticamente a partir de la siguiente ecuación:

$$X = argmaxP(X|Y,\theta) \tag{1}$$

Dado X, el habla sintetizada objetivo, Y la secuencia de caracteres que compone el texto de entrada y θ los parámetros del modelo. Normalmente las distintas metodologías implementadas en esta tarea dividen el labor en dos partes:

- Un primer modelo que se encarga de generar las características acústicas de la voz a sintetizar. Es común que se obtenga un espectrograma como la salida de esta primer parte

del sistema.

- Un vocoder, o codificador de voz, también basado en redes neuronales es utilizado para generar en una segunda instancia la voz sintetizada. Es normal que esta parte de generación de señal audible este acompañada por distintos algoritmos de mejora del habla (speech enhancement).

Posterior a esto, es común incluir una etapa de post-procesado, implementando distintos filtros y el re-sampleo de la señal, que pueden ayudar a disminuir artefactos y otros tipos de ruidos e imperfecciones generados durante la inferencia de voz.

2.2. Evaluación de voces sintetizadas

Mean Opinion Score (MOS) [**?**], o promedio de la puntuación subjetiva, en castellano, es una métrica que proviene del campo de las telecomunicaciones utilizada para determinar la "calidad de la experiencia" de un usuario o conjunto de usuarios, sobre un estimulo o sistema en particular. Normalmente el MOS opera sobre una escala ordinal (símil escala Lickert), típicamente usando un rango discreto entre 1-5, donde las puntuaciones representan una valoración **Mala** a **Excelente** (Tabla (**??**)), aunque también es posible utilizar otras escalas.

Tabla 1. Posible escala para un MOS

Puntuación	**Calidad**
5	Excelente
4	Buena
3	Aceptable
2	Mediocre
1	Mala

Uno de los posibles problemas de este tipo de evaluación, es que los sujetos de prueba suelen percibir que los "saltos" entre cada categoría no son equidistantes (por ejemplo, puede haber una separación más grande entre las valoraciones **Buena** y **Excelente**, que entre **Aceptable** y **Buena**. Otro sesgo recurrente es el denominado "ecualización de rango", el cual lleva a sujetos a tratar de utilizar todas las puntuaciones posibles a lo largo de una prueba. En otras palabras, se tiende a querer utilizar todas las puntuaciones posibles al menos una vez. Esto hace que sea

imposible comparar la opinión entre dos sujetos de prueba distintos, a menos que el rango de calidad esperada de los ejemplos sea equivalente entre ambas pruebas.

MOS es el test subjetivo más frecuentemente utilizado para determinar la calidad de voces sintetizadas. La prueba presenta una serie de ejemplos que deben ser evaluados por distintos oyentes, de acuerdo a algún parámetro especifico. En general, las voces sintetizadas son juzgadas de acuerdo a su "naturalidad". La naturalidad de una voz es un termino difícil de definir, muchas metodologías de evaluación subjetiva incluso prefieren no aclarar el significado de dicha variable, proponiendo que sugerirle una definición a los sujetos de prueba puede sesgar la evaluación pretendida. Sin embargo, dentro del alcance de esta investigación, podemos decir que la naturalidad de una voz sintetizada representa un valor que nos informa acerca de que tanto se asemeja esa voz, a la de un humano. El test MOS emplea una escala discreta de 5 puntos (1 a 5), en la cual el discurso humano real se encuentra entre los valores de 4,5 a 4,8.

Existen distintos modelos de calidad que apuntan a predecir el resultado de un MOS (típicamente desarrollados utilizando el resultado de tests MOS realizados previamente). Algunos ejemplos de estos sistemas, orientados a la calidad de la voz, son desarrollados en la sección 3 de este documento.

2.3. Técnicas posibles de alteración del hablante

El proceso conocido como Data Augmentation (DA, o aumento artificial de datos), tiene el objetivo de aumentar la cantidad de información o muestras recolectadas en una base de datos, manteniendo ciertos rasgos elementales de los ejemplos originales, y sin modificar la distribución de la totalidad de las muestras. Para esta investigación, es necesario poder modificar y variar incluso sutilmente las muestras recolectadas con el fin de obtener un espacio de datos más variado.

2.3.1. Vocal Tract Lenght Perturbation

Se propone implementar técnicas de alteración del largo del tracto vocal [**?**] (Vocal Tract Length Perturbation, VTLP). Dicha técnica fue desarrollada para mejorar sistemas de reconocimiento del habla, e involucra la deformación en espacio y tiempo del espectro de cada audio. El

resultado es semejante a un desplazamiento frecuencial, pero el resultado es el de una voz más natural. En la Figura (**??**) puede observarse una comparación entre un desplazamiento frecuencial y una transformación VTLP.

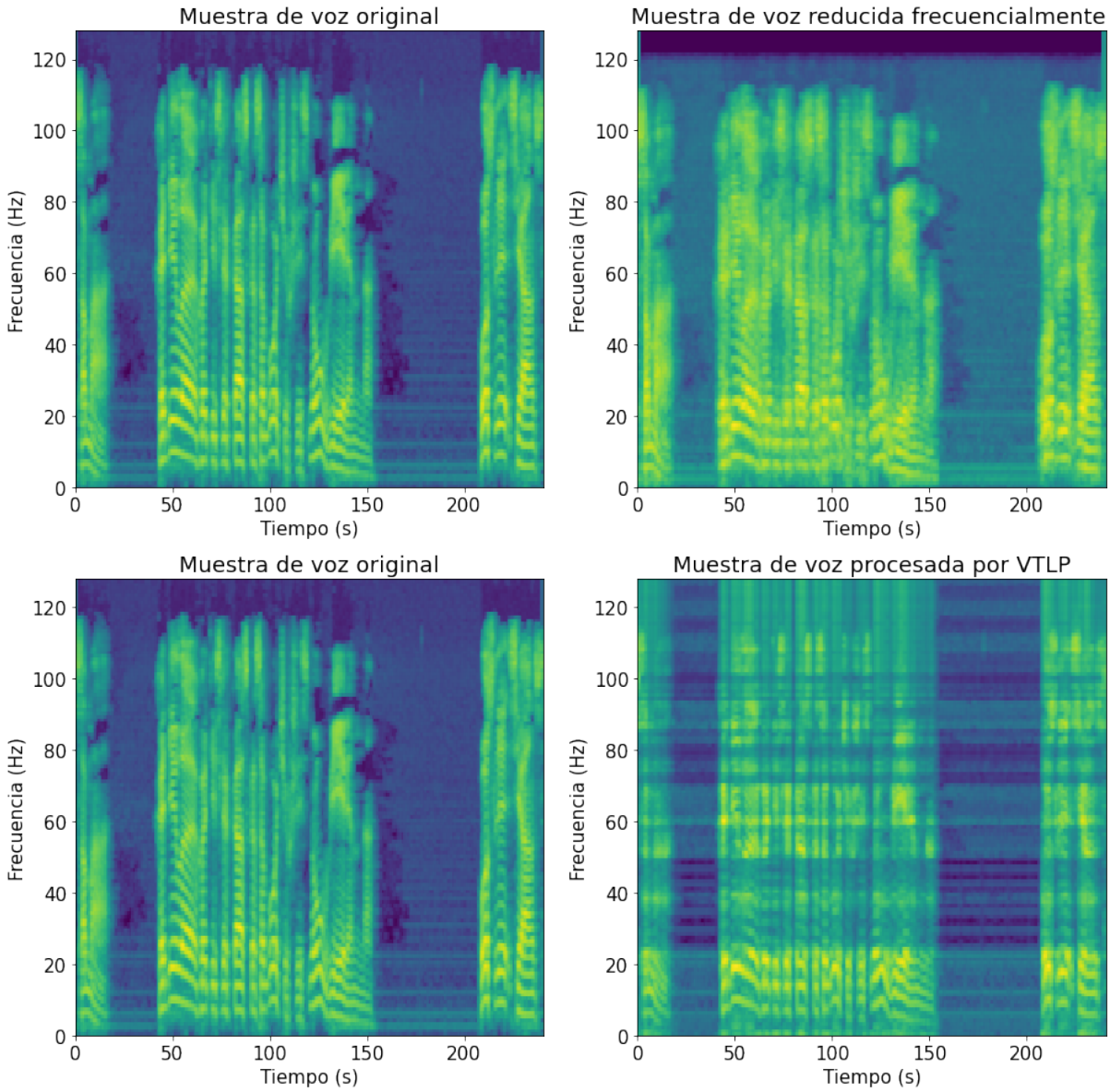

Figura 1. Espectro de una voz modificada frecuencialmente, y mediante una transformación VTLP.

2.3.2. Algoritmo de Griffin-Lim

El algoritmo de Griffin-Lim (ALG) [**?**] es un método de reconstrucción de fase para señales de las que únicamente se tiene su componente de magnitud. El método de estimación de

fase consiste de los siguientes pasos:

- 1. Inicializar la fase aleatoriamente como ruido uniforme.
- 2. Realizar la transformada inversa de Fourier (inverse short-time Fourier transform, ISTFT).
- 3. Realizar la transformada de fourier (STFT) sobre la señal temporal obtenida. Esto deriva mínima información de fase de la señal temporal.
- 4. Reemplazar la magnitud obtenida por la STFT realizada por la magnitud de la señal original. Esto mantiene intacta la información de magnitud de la señal en el espectro, y agrega la minima información de fase de la señal que se deriva de la redundancia de la STFT.
- 5. Iterar los pasos 2-5 hasta obtener un resultado satisfactorio.

Iteración a iteración la información de fase resultara más pertinente a la componente de magnitud, dato original de la señal en el espectro. Muchos modelos de vocoder ignoran o tienen problemas para modelar la fase de una voz sintetizada. Este algoritmo puede ayudar a recrear en parte los artefactos característicos de algunos sistemas TTS.

2.4. Vectorización de hablantes (speaker embeddings)

La extracción de un descriptor numérico de cada audio a evaluar es un proceso necesario para el posterior entrenamiento de la red neuronal que se encargará de calificar cada modelo TTS. Los vectores de hablantes, (Speaker Embeddings) permiten extraer información critica de cada locutor a partir de una representación sonora del mismo, obteniendo un único descriptor capaz de codificar identidad de hablante, genero, velocidad del habla y contenido semántico del texto [**?**]. El proceso de extracción y la información codificada varía de implementación a implementación.

Los X-Vectors [**?**], consisten en una representación vectorizada de cada hablante que aprovecha el uso de técnicas de DA. La representación resultante ha sido útil para mejorar la eficiencia de sistemas de reconocimiento de locutores. Una implementación de la red neuronal que extrae este tipo de descriptor se encuentra disponible en el Kaldi toolkit [**?**].

Por otro lado HuBERT [**?**] presenta una metodología auto-supervisada para extraer speaker embeddings. Existen tres problemas principales a la hora de generar este tipo de representaciones a partir de audios de una manera auto-supervisada, es decir, utilizando una base de datos no etiquetada: (1) existen más de una unidad sonora dentro de cada audio a procesar, (2) no existe un vocabulario o léxico de sonidos posibles en la etapa de pre-entrenamiento, y (3) la unidad sonora no tiene una segmentación explicita. Hubert presenta una implementación novedosa para la extracción auto-supervisada de speaker embeddings. Concretamente, el modelo aprende a agrupar (clustering) distintas unidades sonoras, enmascarando parte de la información, similar a la metodología planteada por **BERT**. La función de perdida se aplica únicamente sobre las regiones enmascaradas, forzando al modelo a aprender representaciones de alto nivel de la parte desenmascarada de la unidad sonora, para poder inferir correctamente respecto de como clasificar las regiones enmascaradas. Hubert extrae tanto información acústica, como lingüística como parte de su proceso de vectorización.

3. ESTADO DEL ARTE

Determinar la calidad del habla sintetizada es una problemática que atraviesa distintas áreas y tecnologías: sistemas de texto a voz (TTS), mejora del habla (Speech Enhancement), y conversión de la voz (Voice Conversion) entre otros. Para el desarrollo de estos tipos de sistemas, donde las características de la señal de salida deben ser evaluadas repetidamente, surge la necesidad de utilizar modelos de calidad automáticos basados en criterios matemáticos y psicoacústicos para poder aproximar la apreciación subjetiva humana que se obtendría, por ejemplo, con un test MOS. El estado del arte de estas técnicas esta conformado por los siguientes sistemas:

3.1. PESQ y POLQA

Dentro del campo del speech enhancement, el PESQ [**?**] (Perceptual Evaluation of Speech Quality o, evaluación percibida de calidad del habla), ITU T P.862, consiste en una evaluación intrusiva para cuantificar la calidad del habla. Es un algoritmo Full Reference (FR, o referencia completa), lo que quiere decir que para realizar una evaluación sobre un sistema requiere de la señal de entrada y de salida del mismo. Su funcionamiento parte de un modelo psicoacústico, refinado empíricamente, que estima un valor MOS comparando la referencia original con la salida degradada del modelo, usando distancias paramétricas entre ambas señales. Al comparar la señal original y la señal degradada, las alinea en tiempo y normaliza en amplitud, por lo que no tiene en cuenta los efectos de distorsión temporal y de atenuación de la señal. Sin embargo, en muchos casos, para sistemas de TTS, no contamos con las señales originales utilizadas para entrenar una red neuronal (voz original), por lo que no se puede depender de algoritmos de este tipo.

En 2011, POLQA (Perceptual Objective Listening Quality Analysis) [**?**], ITU-T P.863, fue desarrollado como sucesor a PESQ. Este algoritmo compara muestra a muestra una señal degradada por un sistema, contra un señal original tomada como entrada de dicho sistema. Se analizan ambas señales en es dominio frecuencial, en distintas bandas criticas. Las diferencias encontradas en cada banda son consideradas distorsiones, que luego son consideradas a la hora de asignar

una puntuación tipo-MOS en una escala de 1-5. El aporte más relevante de este algoritmo es su modelo perceptivo, que toma en consideración ciertos factores humanos (*Idealización*) de las tareas de categorización que se realizan durante tests MOS.

3.2. ViSQOL

ViSQOL (Virtual Speech Quality Objective Listener o, calidad del habla objetiva virtual) [**?**], fue desarrollado para emular la percepción humana sobre la calidad del habla. Evalúa una distancia calculada sobre un **neurograma**, análogo a un espectrograma, pero cuya intensidad (variable asignada al color del gráfico) está referida a la actividad neuronal. Nuevamente se trata de un algoritmo FR. Una comparación con las métricas desarrolladas por ITU, PESQ y POLQA, se realizó teniendo en cuenta la capacidad de cada algoritmo de detectar distintos tipos de transformaciones, incluyendo el añadido de distintos tipos de ruido de fondo, filtrado de señal, mejora del habla y variación de relación señal a ruido.

Los resultados de la investigación demostraron que ViSQOL y POLQA tienen un desempeño comparable, ambos superando el algoritmo PESQ.

3.3. MOSNet

Desarrollado para asistir en las tareas de evaluación de conversión de voz, MOSNet [**?**] es un predictor de valores MOS. El método propuesto consiste en entrenar una red neuronal sobre una base de datos construida a partir de evaluaciones de escucha realizadas durante el Voice Conversion Challenge 2018 (VCC). Para modelar la percepción humana tres diferentes arquitecturas son puestas a prueba y comparadas a lo largo de la investigación conducida por Chen-Chou, et al.

El primer modelo, basado en una red convolucional concatenada a una capa completamente conectada, fue derivado del trabajo previo desarrollado por Yoshimura et al. [**?**]. Las capas convolucionales fueron configuradas empíricamente para segmentar el discurso evaluado en secciones de 400 ms a modo de capturar información temporal más corta. El segundo modelo consiste en una red BLSTM (Bidirectional Long Short-Term Memory) previamente implementada en el paper Quality-Net [**?**], posee la habilidad de integrar la información de dependencias

en el tiempo y de características secuenciales propias de una voz humana. Finalmente el tercer modelo es diseñado como una combinación de los dos previamente mencionados. Para cada arquitectura propuesta, el entrenamiento se realiza sobre características espectrales extraídas del VCC, con los puntajes MOS de dicha competencia como la solución objetivo. Los resultados indican una correlación alta entre los puntajes MOS derivados de los modelos entrenados, y los obtenidos por medio de pruebas subjetivas.

3.4. NISQA

En 2021, Mittag y Moller [**?**] presentaron un evaluador de naturalidad del habla sintetizada, basada en una red neuronal CNN-LSTM obteniendo resultados satisfactorios para oraciones, con pequeñas limitaciones cuando el espectro de la onda resultante se ve acotado. La base de datos utilizada en el entrenamiento esta compuesta de 16 fuentes distintas extraídas de distintas competencias realizadas de forma online, divididas en 12 idiomas distintos, para desarrollar una red neuronal capaz de procesar distintos lenguajes. Una implementación abierta del código desarrollado por esta investigación se encuentra disponible. La misma permite ser re-entrenada con una nueva base de datos.

3.5. Sinopsis de las distintas metricas de calidad objetiva presentadas

En la Tabla (**??**), se presenta una comparación entre las distintas arquitecturas discutidas en esta sección.

Tabla 2. Sinopsis de las distintos modelos de calidad propuestos para predecir la preferencia humana

Año	**Referenica**	**Arquitectura**	**Comentarios**
2001	PESQ	Comparación intrusiva	Primer metodología automatizada adoptada por ITU
2011	POLQA	Comparación intrusiva con modelo perceptivo	Sucesor de PESQ desarrollado por ITU
2015	VISQOL	Comparación intrusiva con modelo perceptivo	Introducción del Neurograma como modelo perceptivo
2021	MOSNet	CNN-BLSTM	Red neuronal entrenada sobre resultados de encuestas tipo MOS de naturalidad
2021	NISQA	CNN-LSTM	También calcula otros parámetros acústicos (ruido, distorsión y discontinuidad)

4. METODOLOGÍA

4.1. Obtención de datos

Para poder entrenar un red neuronal capaz de predecir el resultado de un test tipo MOS realizado sobre un sistema de texto a voz, se necesita generar una base de datos con los resultados de un gran numero de algoritmos de síntesis vocal, acompañados de una etiqueta que represente su puntuación final obtenida de una prueba MOS. También se puede incluir en esa base de datos, ejemplos de voces humanas reales, y señales de voz sintetizadas, procesadas digitalmente.

Con el objetivo de generar esta robusta base de datos, en primera instancia se recolectaron ejemplos de un gran numero de sistemas de generación de voz humana disponibles. Los ejemplos a sintetizar fueron tomados de la lista de frases que forman parte del cuerpo de la base de datos de openSLR [**?**]: una base de datos generada por un equipo de investigación de Google, con el fin de entrenar sistemas de TTS y de ASR para idiomas de bajos recursos. Las lista completa de frases utilizadas son incluidas en el Anexo I.

En la Tabla (**??**) se detallan los sistemas de texto a voz utilizados en la generación de la base de datos. Todos los ejemplos fueron sintetizados en castellano. El código de región exhibido en la tabla esta basado en el estándar ISO 639-1 para determinar la región de la voz sintetizada.

La base de datos incluye distintas voces sintetizadas con servicios profesionales de síntesis como Amazon Polly, Microsoft Azure, Speechello y Neurasound, sistemas concatenativos como Loquendo y la implementación TTS de Thomas Dewitte, servicios experimentales basados en Fastpich, y voces humanas reales pertenecientes al banco de voces OpenSRL.

Tabla 3. Composición de la base de datos generada. Código de región de acuerdo a ISO 639-1.

	Descripción	Región	Cant. de voces
Amazon Polly	Implementación privada	es-us/es-mx /es	8
Microsoft Azure	Implementación privada	es-ar/es-bo /es/es-mx	8
Speechello	Implementación privada (I.A.)	es-us/es-mx	2
	Implementación privada	es-us/es-mx /es	5
Neurasound	Implementación privada	es-ar/es-cl/es-bo/ es-pe/es-pr	14
Loquendo	Sistema concatenativo	es	1
text-to-speech	Librería de Python	es	1
Fastpitch - HiFiGan	Implementación con transformadores	es-ar	4
DC-TTS	Red convolucional	es-ar	7
TacoTron2	Modelo secuencial	es	1
OpenSRL	Grabaciones de personas	es-ar	48

Esta base de datos inicial fue limpiada, transformada y reducida durante el diseño de la prueba subjetiva, como será expuesto en la sección 4.3.

4.2. Expansión artificial de datos

Con el objetivo de variar los tipos de voces obtenidos en la sección previa, se llevo a cabo un proceso de expansión artificial de datos basada en distintas técnicas de procesamiento digital. Las mismas son detalladas a continuación:

- **Alteración de largo tracto vocal (VTLP)**: 300 ejemplos de voces sintetizadas fueron procesados por este algroritmo, la implementación utilizada y el factor de deformación de VTLP (elegido aleatoriamente entre 0,9 y 1,1 para cada ejemplo) se basaron en recomendaciones detalladas por Jaitly et al. [**?**].

- Alteración de fase (Algoritmo de Griffin-Lim): 500 ejemplos de voces reales y 100 ejemplos de voces sintetizadas fueron procesadas por este algoritmo de acuerdo al procedimiento especificado en la sección 2.3.2.

4.3. Diseño de la prueba subjetiva

El diseño de la prueba subjetiva se basa en las especificaciones provistas por las recomendaciones del estándar ITU-T Rec. P.807. Todos los sujetos encuestados cumplieron con la condición de ser normo-oyentes.

El test consiste en la evaluación subjetiva de una serie de audios cortos que contienen distintas voces (2 a 6 segundos de duración). La duración total del test es de aproximadamente 8 minutos. Los ejemplo deben ser evaluados en una escala de tipo Likert de 5 puntos. La cantidad máxima de audios que un sujeto puede evaluar es de 50. El propósito de la encuesta subjetiva es el de etiquetar los audios recolectados previamente, con una puntuación. La cantidad de etiquetas necesarias está determinadas por el entrenamiento de la red neuronal que se desarrollará a posteriori. Un precedente útil se puede tomar del trabajo de Deja et al.[13] en el cual se llevó a cabo una metodología similar. Sujeto a la cantidad de audios que evalúe cada persona, en principio son necesarios alrededor de 100 sujetos de prueba, asumiendo que cada sujeto de prueba evalúe alrededor de 50 audios.

Una síntesis de las instrucciones presentadas a los participantes de la encuesta es provista a continuación:

Instrucciones

A continuación vas a escuchar una serie de distintos tipos de voces generados por computadoras. El propósito de este test es evaluar la calidad de cada archivo, para poder subsecuentemente utilizar esa información en un sistema de evaluación automático de voces sintetizadas.

Para cada ejemplo se deberá proveer una calificación de acuerdo a la siguiente escala. (Escala MOS para la naturalidad de una voz, Tabla (**??**))

Tabla 4. Escala MOS para la naturalidad de una voz

Puntaje	Calidad del habla	Naturalidad
5	Excelente	Completamente natural
4	Buena	Bastante natural
3	Aceptable	Natural y antinatural en partes iguales
2	Mediocre	Bastante antinatural
1	Mala	Completamente antinatural

Los siguientes ejemplos ilustran el significado de cada puntaje. Sin embargo para realizar la prueba es importante tener en cuenta que se escucharan otros tipos de voces muy distintas, con distorsiones o artefactos no presentes previamente. Por lo tanto, estos ejemplos no cubren la totalidad de las posibilidades que pueden esperar escuchar.

Ejemplos

El siguiente ejemplo presenta una voz humana y tiene un puntaje de referencia de 5.0

El siguiente ejemplo presenta una voz sintetizada, puntaje de referencia de 4.0

El siguiente ejemplo presenta una voz sintetizada, puntaje de referencia de 3.0

El siguiente ejemplo presenta una voz sintetizada, puntaje de referencia de 2.0

El siguiente ejemplo presenta una voz sintetizada, puntaje de referencia de 1.0

Tener en cuenta que la calidad de los ejemplos que deberán clasificar puede ser distinta a la escuchada en estos ejemplos.

4.3.1. Selección de respuestas validas

Con el fin de descartar respuestas atípicas, se toman ciertos criterios para considerar validos los resultados de una prueba subjetiva:

- Para cada ejemplo que se debe calificar, se tiene previamente un valor estimativo de la calidad esperada, extraído de encuestas previas y análisis objetivos. Por lo tanto si un candidato responde azarosamente, o simplemente deriva un criterio de calificación incorrecto por falta de claridad en las instrucciones, es posible identificar y descartar sus respuestas.

- Se mide el tiempo de respuesta de cada estimulo. Se descartan los resultados que hayan sido entregados en un tiempo menor a un cierto umbral.

- Se toma los ejemplos a calificar de voces humanas reales como un "ancla", símil al concepto presente en otros tests subjetivos como el MuSHRA.

4.3.2. Redistribución de audios a clasificar

Para prevenir el sesgo de "ecualización de rango", los ejemplos provistos a los sujetos de prueba deben seguir una distribución uniforme respecto del rango de calidad total MOS definido para la naturalidad de la voz. La distribución de la base de datos debe ser perpetuamente balanceada en la medida que se obtienen resultados de distintas encuestas para asegurar la minimización de este sesgo.

Se llevó a cabo una prueba piloto de la evaluación propuesta que involucro a 5 participantes que calificaron 50 audios cada uno. Con los resultados obtenidos, se determinó un leve desbalanceo en en la calidad de los ejemplos sintetizados. En base a esto, se redefinió la base de datos obtenida previamente, para incluir más ejemplos de calidad estimada en el rango de 4,0 a 5,0.

4.4. Sistema de predicción de MOS

4.4.1. Funcionamiento general

El funcionamiento del modelo de predicción cuenta de 3 etapas fundamentales: 1) Extractor de espectrograma de Mel y segmentación, 2) Extracción de características acústicas por franja, y 3) Extracción de características temporales. En la Figura (**??**) se presenta un diagrama de la arquitectura implementada, que consiste de una modificación del modelo propuesto en [**?**].

En primer lugar se extraen espectrogramas de Mel del audio a evaluar, separados en distin-

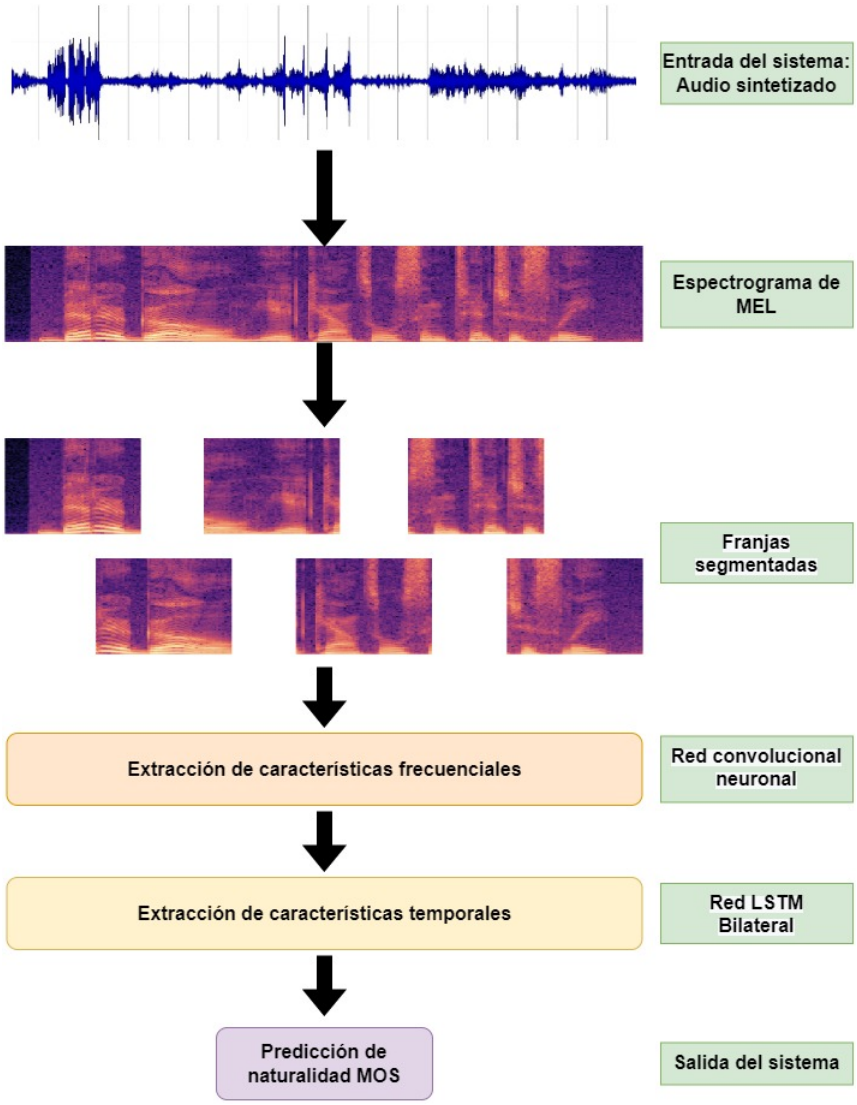

Figura 2. Funcionamiento general del modelo de predicción de naturalidad MOS.

tos segmentos con un cierto grado de solapamiento. Luego dichos segmentos son tomados como entrada de una red neuronal convolucional que extrae características útiles para predecir calidad sonora. Estas características se entregan a una red bilateral de larga-corta duración (BLSTM) que modela las dependencias temporales propias del discurso humano. La ultima capa completamente conectada de este segmento del modelo tiene como salida la predicción de naturalidad MOS.

4.4.2. Segmentación en espectrogramas de MEL

La entrada de la red neuronal convolucional toma espectrogramas de Mel generados a partir de una FFT con un tamaño de ventana de 20 ms y salto de 10 ms. El ancho de cada segmento es de 15, lo que equivale a 150 ms, y la altura es de 48 (48 x 15). La frecuencia máxima de análisis es de 20 kHz (muchos de los audios en el dataset tienen frecuencias de muestreo distintas). El tamaño de salto entre cada segmento es de 4 (40 ms), dando un cierto solapamiento de segmento a segmento.

4.4.3. Red CNN-BiLSTM

Esta sección del modelo esta basado en la red siamés propuesta en [**?**]. En la Figura (**??**) se ofrece un diagrama para representar las distintas capas de este segmento del algoritmo de predicción.

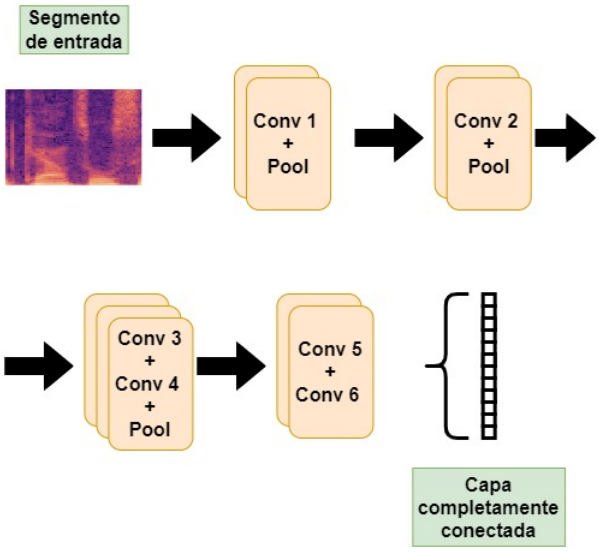

Figura 3. Arquitectura de la red neuronal convolucional. Cada capa esta acompañada por una activación de tipo ReLU.

La red esta tiene 6 capas convolucionales concatenadas, formadas por filtros de distintos tamaños como se puede observar en la Tabla **??**, donde N se corresponde con el numero de segmentos extraídos en la etapa previa, que depende de la duración del audio analizado. La salida de cada capa atraviesa una activación de tipo ReLU.

La salida de esta etapa es utilizada como la entrada de una red de larga-corta duración

Tabla 5. Arquitectura de la red neuronal convolucional

Capa	**Dimensión**
Entrada	Nx1x48x15
Conv1	Nx16x48x15
Pool	Nx16x24x8
Conv2	Nx32x24x8
Pool-Perdida(20 %)	Nx32x12x4
Conv3	Nx64x12x4
Conv4	Nx64x12x4
Pool-Perdida(20 %)	Nx64x6x2
Conv5-Perdida(20 %)	Nx64x6x2
Conv6	Nx64x6x2
FC	Nx20

bidireccional, de 1 sola capa con 128 unidades escondidas que se encarga de modelar las dependencias temporales de la señal y estimar finalmente la naturalidad del habla.

4.4.4. Entrenamiento y detalle de la implementación

El entrenamiento partió de pesos pre-entrenados con el NISQA Corpus [**?**] una base de datos de audios en ingles evaluados subjetivamente por un gran numero de participantes. El ajuste fino de la red final se realizo sobre ese modelo, dividendo los resultados obtenidos de la encuesta subjetiva realizada en 80 % para el entrenamiento, y 20 % para la validación.

El código necesario se implemento en el lenguaje de programación Python, sobre la biblioteca de DNN PyTorch y se encuentra disponible de en linea, en un repositorio [**?**]. La velocidad de aprendizaje fue fijada en 0,001, se utilizo el optimizador Adam y se eligió la función de perdida consiente de sesgos, basándose en la metodología propuesta por [**?**].

4.5. Evaluación objetiva del sistema propuesto

Para refinar el funcionamiento del modelo propuesto, (configuración de hiperparametros y características de la red propuesta), se analizan 2 métricas objetivas: correlación cruzada de Pearson (PCC) y el error cuadrático medio (RMSE). *En base a estas métricas se definirá el modelo final.*

Para comparar el modelo propuesto con otras redes de predicción de calidad, ambas métri-

cas propuestas son contrastadas con distintas soluciones del estado del arte: NISQA V1, NISQA V2 y ANIQUE+. La comparación se realiza sobre un subconjunto de ejemplos tomados de la base de datos desarrollada y evaluada subjetivamente a lo largo de este trabajo. Se evalúa el RSME y el PCC promedio sobre la totalidad del conjunto de audios, y también en distintas categorías como se presenta en la sección 5.

En total, para cada modelo predictivo se evaluaron 200 audios, comparando el pronostico obtenido con el resultado subjetivo obtenido previamente. Se calcula RSME y PCC por estimulo, y seguidamente se obtiene un promedio de ambas métricas a nivel de cada sistema evaluado.

5. RESULTADOS Y ANÁLISIS

5.1. Resultados iniciales

Se presentan los resultados iniciales en la Tabla **??**. El RMSE y a nivel de sistema se calcularon para el modelo predictivo propuesto, y para la segunda iteración del software NISQA. En ambos casos, se procesaron y evaluaron audios juzgados bajo 2 condiciones:

- RMSE y coeficiente de correlación calculado sobre las encuestas realizadas previas a balancear la base de datos (250 ejemplos utilizados en el entrenamiento del modelo propuesto)
- RMSE y coeficiente de correlación calculado sobre las encuestas realizadas posterior a balancear la base de datos (500 ejemplos utilizados en el entrenamiento del modelo propuesto)
- RMSE y coeficiente de correlación calculado sobre las encuestas realizadas posterior a balancear la base de datos (750 ejemplos utilizados en el entrenamiento del modelo propuesto)

Tabla 6. Resultados iniciales obtenidos con 15 % de los audios de la base de datos evaluados

	Modelo propuesto		**NISQA V.2**	
		RSME		RSME
Previo a balancear el dataset	0,32	2,63	0,17	3,19
Posterior a balancear el dataset	0,35	1,63	0,31	2,07
Posterior a balancear el dataset +250 ejemplos	0,49	1,21	0,24	2,04

5.2. Tiempo de procesamiento

5.3. Modelo final

6. DISCUSIÓN DE LOS RESULTADOS

(*Discusión de los resultados iniciales de la entrega **Avance 4** del Taller de Tesis.*)

Los resultados iniciales obtenidos nos permiten inferir ciertas hipótesis que pueden resultar útiles a la hora de continuar el desarrollo del trabajo:

- En primer lugar, como es de esperarse los resultados mejoran al incorporar más ejemplos de entrenamiento. Por ahora, la cantidad de datos recolectados son 10 % de la totalidad que se espera obtener para el entrenamiento del modelo final, por lo que se espera una mejora significativa de las métricas evaluadas en el futuro, simplemente por tener una base de datos más robusta.

- La segunda iteración de los resultados fueron calculados luego de balancear los estímulos de prueba. Es posible que parte de la mejora en el RMSE se deba a este ajuste de la encuesta realizada, por lo que se propone realizar un balanceo adicional antes de liberar la evaluación subjetiva al publico general para obtener el resto de las respuestas necesarias

- Si bien el modelo propuesto supero el desempeño de la red NISQA V.2, esto se debe en parte a que la red NISQA nunca había sido expuesta a audios en el idioma castellano en su entrenamiento. Al mismo tiempo, el modelo propuesto fue entrenado con audios que guardan cierta similitud a los que fueron usados en la evaluación (ambos subconjuntos de la misma base de datos), lo que puede aportar otro factor que beneficie a un modelo predictivo sobre el otro.

- Para futuras comparaciones es importante incluir otros modelos como NISQA V1 y ANIQUE+

- Es posible adoptar variaciones sobre la arquitectura propuesta, para poder refinar el sistema aún más. La mayoría de los cambios significativos son propios de la red neuronal convolucional.

7. CONCLUSIONES

8. TRABAJOS FUTUROS

Al finalizar el desarrollo, de acuerdo a los resultados alcanzados, se podría evaluar el modelo realizado en otros idiomas para verificar su funcionamiento multilingüe. Otro posible desarrollo será obtener una base de datos más robusta, con un número de encuestados mayor, para re-entrenar y refinar el funcionamiento de la red neuronal. También se plantea la posibilidad de empaquetar el modelo para poder ser utilizado como una librería de Python disponible como código abierto, para facilitar su uso en producción de sistemas TTS.

9. Anexo I: Listado de frases utilizadas para sintetizar audios

Para la caída del cabello, tengo un nuevo jabón

¿Qué color favorito es el más popular?

El pijama de rayas es azul

Las máquinas de escribir antiguas pueden ser muy caras.

Al circular menos automóviles en la ciudad se logra una mejor calidad del aire

¿A qué hora querés ir al cine?

Estoy destinada a triunfar

Tengo una junta de negocios con la firma de abogados, ¿se acuerda?

Para abrir los poros de la piel lo mejor es ir a un sauna, o tomar un baño de vapor

Para hacer esa receta necesitás fruta y alcohol

Quiero que me ayudes a elegir un regalo.

La obscuridad del pozo era obscena por naturaleza

Aquí doblamos a la derecha

Las tejas rojas del tejado no son de México, son de Texas

Estaba pensando en ir a acampar.

Lo podés pagar con la tarjeta verde de Visa

¿Cuáles son las principales diferencias?

¿Cuál es la estación de metro más cercana?

Si el cuerpo pide descanso hay que dárselo dicen los médicos

Mañana vaya abrigada, un gorro de lana y guantes es lo más recomendable

El viaje en avión es más rápido que en barco.

¿Quién subió el mejor video?

La pronunciación del francés es difícil

Me ayudas a plantar los naranjos

El Llano en Llamas es un clásico de Juan Rulfo

Quiero averiguar como se llama un director de cine

¿Cuándo sucedió la Primera Guerra Mundial?

Este personaje malverso una cantidad extraordinaria de fondos públicos

En las noticias de anoche, no vi que al presidente de Estados Unidos.

Las bocinas estas que te dije son enormes, ¿sabés que se puede hacer con ellas?

Es muy fácil usar las bicis de la ciudad, con tu tarjeta de crédito alquilas tu bici.

Mañana va a hacer mucho viento y habrá tormentas

El dictador Franco estuvo cuarenta años en el poder.

¿Quién subió el mejor video?

Te voy a mandar por whatsapp las especificaciones de los invitados

La calle está llena de hoyos

Tenés que denunciar inmediatamente el accidente para que lo cubra el seguro

Quiero un jabón para relajarme.

Necesito tu ayuda urgente.

¿Qué tan destruida quedó la ciudad de Nueva Orleans después del huracán Katrina?

Los hamsters comen zanahorias

En el salón de fiestas entran alrededor de mil quinientas personas

Hoy a la tarde saqué mi tarjeta para la bicicleta urbana

Ayudáme a encontrar un buen gimnasio pero que esté cerca de mi casa

No te entiendo nada

La granizada destruyó toda la plantación de lechuga

Quiero irme de vacaciones a Hawái pero no sé si el volcán esté activo.

Se hicieron versiones en todos los idiomas de esa canción

Voy a necesitar que lo pagués con la tarjeta roja

Te recomiendo que te llevés un buen libro o bajés una buena película

¿Queres que te recomiende algo en particular?

Además, no necesitás comprar una bicicleta, podés usar las de la ciudad.

No sé hablar malayo

¿Deseas avisar a tu familia que estás bien?

Te quiero pedir cincuenta bolsas de basura extra grandes y veinticinco carpas

¿Podés verificar si hay alguna estación de tren cerca del aeropuerto?

La piel es muy porosa

Del subte al museo son quince minutos

¿Hay algún video viral esta mañana?

Quiero ver una película pero que sea en una pantalla grande.

¿Querés aprender el idioma a nivel de negocios?

Te quiero pedir cincuenta bolsas de basura extra grandes y veinticinco carpas

¿Me ayudas a hacer los deberes?

Creo que se le disloco el hombro

El siguiente semestre quiero empezar a estudiar licenciatura en cine

Los atajos están más libres de circulación que nunca en los últimos tiempos

No entiendo nada

Me gustaría organizar unas charlas sobre mecanica industrial

Él sabe que con eso va a pasar a la historia.

La parte de las cataratas del Niágara que está del lado estadounidense es aburrida

Les recomiendo que escalen siete montañas de más de mil metros sobre el nivel del mar

Los champiñones silvestres son los más sabrosos

Corea del Norte logró lanzar un misil a una distancia de quinientos km por primera vez.

Últimamente he visto muchas bicicletas en la calle.

Últimamente he andado muy acelerada.

El caballo está amarrado

Tocar el xilófono es mi hobby favorito

¿Es un vuelo directo o hace escalas?

Quiero que me des tu opinión sobre la nueva obra de teatro.

¿Sabes de electricidad?, necesito cambiar un enchufe y no se como

Te voy a mandar una lista de los mejores exitos en español directo a tu celular.

Estoy acá con una amiga cocinando comida Tailandesa

Estoy aburrido con la música de mi teléfono, quiero que me recomiendes algo nuevo.

La película me llegó al corazón

Le recomiendo los Alpes Suizos

El Llano en Llamas es un clásico de Juan Rulfo

¿Podrías comprar por favor el disco y enviarselo a mi mamá?

Este libro nos muestra una teoría sobre la importancia del arte en el ámbito científico

La obra de teatro El Flautista fue un exito rotundo

Mas o menos ¿qué presupuesto está dispuesto a pagar?

Estoy buscando un restaurante de cocina tradicional

¿Me podés ayudar a coser un parche?

Los hamsters comen zanahorias

Tiene el tiempo contado

No entiendo nada

Quiero ver una película de comedia

Primero que nada vamos a hacer un ejercicio de respiración

Los hamsters comen zanahorias

La ciudad es bellísima y tan alegre

10. Anexo II: Código Implementado

BIBLIOGRAFÍA

[1] Oord, Aaron van den and Dieleman, Sander and Zen, Heiga and Simonyan, Karen and Vinyals, Oriol and Graves, Alex and Kalchbrenner, Nal and Senior, Andrew and Kavukcuoglu, Koray. WaveNet: A Generative Model for Raw Audio, arXiv (2016).

[2] Oord, Aaron van den and Li, Yazhe and Babuschkin, Igor and Simonyan, Karen and Vinyals, Oriol and Kavukcuoglu, Koray and Driessche, George van den and Lockhart, Edward and Cobo, Luis C. and Stimberg, Florian and Casagrande, Norman and Grewe, Dominik and Noury, Seb and Dieleman, Sander and Elsen, Erich and Kalchbrenner, Nal and Zen, Heiga and Graves, Alex and King, Helen and Walters, Tom and Belov, Dan and Hassabis, Demis. Parallel WaveNet: Fast High-Fidelity Speech Synthesis, arXiv (2017).

[3] Ren, Yi and Ruan, Yangjun and Tan, Xu and Qin, Tao and Zhao, Sheng and Zhao, Zhou and Liu, Tie-Yan. FastSpeech: Fast, Robust and Controllable Text to Speech, arXiv (2019).

[4] Prenger, Ryan and Valle, Rafael and Catanzaro, Bryan. WaveGlow: A Flow-based Generative Network for Speech Synthesis, arXiv (2018).

[5] ITU-T Rec. P.800. Methods for subjective determination of transmission quality (1996). (*p. 18-21*)

[6] Navdeep Jaitly and E. Hinton. Vocal Tract Length Perturbation (VTLP) improves speech recognition, Proc.of the 2013 IEEE International Conference on Acoustics, Speech, and Signal Processing, (2013).

[7] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey and S. Khudanpur. X-Vectors: Robust DNN Embeddings for Speaker Recognition, IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (2018).

[8] Daniel Povey. Kaldi Speech Recognition Toolkit. Extraído el 12 de septiembre de 2022, https://github.com/kaldi-asr/kaldi.

[9] R. Kubichek. Mel-cepstral distance measure for objective speech quality assessment, Proc. IEEE Pacific Rim Conference on Communications Computers and Signal Processing (1993).

[10] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra. Perceptual evaluation of speech quality (PESQ) a new method for speech quality assessment of telephone networks and codecs, Proc. ICASSP (2001).

[11] Guevara-Rukoz, Adriana and Demirsahin, Isin and He, Fei and Chu, Shan-Hui Cathy and Sarin, Supheakmungkol and Pipatsrisawat, Knot and Gutkin, Alexander and Butryna, Alena and Kjartansson, Oddu. Crowdsourcing Latin American Spanish for Low-Resource Text-to-Speech. Proceedings of The 12th Language Resources and Evaluation Conference (LREC), mayo , Marseille, France (2020) (*p. 6504-6513*)

[12] Kamil, Deja and Ariadna, Sanchez and Julian, Roth and Marius, Cotescu. Automatic Evaluation of Speaker Similarity, arXiv (2022).

[13] Benjamin van Niekerk and Marc-Andre Carbonneau and Julian Zaidi and Matthew Baas and Hugo Seute and Herman Kamper. A Comparison of Discrete and Soft Speech Units for Improved Voice Conversion. International Conference on Acoustics, Speech and Signal Processing (ICASSP) (2022).

[14] Shuai Wang and Yanmin Qian and Kai Yu. What Does the Speaker Embedding Encode? Proc. Interspeech 2017. Stockholm, Sweden (2017). (*p. 1497-1501*)

[15] Hsu, Wei-Ning and Bolte, Benjamin and Tsai, Yao-Hung Hubert and Lakhotia, Kushal and Salakhutdinov, Ruslan and Mohamed, Abdelrahman. HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units. arXiv (2021)

[16] https://paperswithcode.com/method/griffin-lim-algorithm

[17] https://www.itu.int/rec/T-REC-P.863/en

[18] A. Hines, J. Skoglund, A. C. Kokaram, and N. Harte, “Visqol: an objective speech quality model,” EURASIP Journal on Audio, Speech, and Music Processing, vol. 2015, no. 1, pp. 1–18, 2015.

[19] Chen-Chou Lo and Szu-Wei Fu and Wen-Chin Huang and Xin Wang and Junichi Yamagishi and Yu Tsao and Hsin-Min Wang MOSNet: Deep Learning-Based Objective Assessment for Voice Conversion. Interspeech 2019

[20] Manocha, Pranay and Jin, Zeyu and Zhang, Richard and Finkelstein, Adam CDPAM: Contrastive learning for perceptual audio similarity. arXiv, 2021.

[21] T. Yoshimura, G. Eje Henter, O. Watts, M. Wester, J. Yamagishi, and K. Tokuda. A hierarchical predictor of synthetic speech naturalness using neural networks, Proc. Interspeech (2016).

[22] S.-W. Fu, Y. Tsao, H.-T. Hwang, and H.-M. Wang, “Quality-Net: An end-to-end non-intrusive speech quality assessment model based on BLSTM,” in Proc. Interspeech, 2018.

[23] Gabriel Mittag and Sebastian Möller. Deep Learning Based Assessment of Synthetic Speech Naturalness, Interspeech 2020 ISCA (2020).

[24] G. Mittag and S. Moller, “Full-reference speech quality estimation with attentional siamese neural networks,” in ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2020, pp. 346–350

[25] https://github.com/gabrielmittag/NISQA/wiki/

[26] G. Mittag, S. Zadtootaghaj, T. Michael, B. Naderi, and S. Moller, “Bias-aware loss for training image and speech quality prediction models from multiple datasets,” in Accepted at QoMEX 2021.

[27] https://github.com/asosawelford/Tesis (en progreso)