UNTREF UNIVERSIDAD NACIONAL DE TRES DE FEBRERO

INGENIERÍA DE SONIDO

Desarrollo de un procedimiento de evaluación objetiva para sistemas de texto a voz

Tesis final presentada para obtener el título de Ingeniero de Sonido de la Universidad Nacional de Tres de Febrero (UNTREF)

TESISTA: Alejandro Sosa Welford (39912286)

TUTOR/A: Ing. Leonardo Pepino

Fecha de defensa: mes y año | Locación (ej. Sáenz Peña), Argentina

AGRADECIMIENTOS

(a completar)

ÍNDICE DE CONTENIDOS

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

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 la 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 [1] 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 [2], transformadores [3] y sistemas de tipo flow [4].

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) [5] (puntaje promedio subjetivo) es el método más frecuentemente utilizado para llevar a cabo esa prueba. Dicha métrica tiene un rango de 0 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

objetiva 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 (open source).

Intercambios Transorgánicos (Dir. Gala Gonzalez 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 de dicha investigación.

1.2. OBJETIVOS

1.2.1. OBJETIVO GENERAL

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.

1.2.2. OBJETIVOS ESPECÍFICOS

Se proponen los siguientes objetivos específicos:

- **Recolección de audios sintetizados.** La primera instancia de la investigación consiste en recolectar audios sintetizados por sistemas TTS de variada calidad de clonado de voz y procedencia, además de audios de hablantes humanos reales. Se decide que la investigación se conducirá en el idioma castellano.

- **Transformación de audios recolectados.** Con el objetivo de variar la calidad de los audios recolectados, se utilizaran herramientas como Vocal Tract Length Perturbation (perurbación del tracto vocal, VTLP). También los audios pueden ser comprimidos, atravesar distintos codecs o ser reverberados con este mismo objetivo.

- **Extracción de descriptores objetivos de cada audio.** Se propone extraer X-VECTORS de cada audio. Estos embeddings son utilizados para reconocimiento de hablantes. Estos descriptores 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.** Con el propósito de obtener una puntuación MOS para cada audio recolectado, se llevarán a cabo pruebas subjetivas para obtener estas etiquetas. Seguido a esto, se realizara una validación de los datos obtenidos.

- **Diseño de una red neuronal para predecir la naturalidad de cada audio.** Se entrenará una pequeña red neuronal de forma supervisada, con los audios recolectados como entrada y sus calificaciones MOS como salida deseada. La misma contará probablemente con 2 a 3 capas completamente conectadas, seguidas de una activación que aún no ha sido determinada, que indicará la predicción de cada inferencia. 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.

1.3. ESTRUCTURA DE LA INVESTIGACIÓN

En capítulo 2 se detalla un marco teórico útil para entender en más detalle los procesos detrás de las distintas implementaciones posibles para sintetizar voces artificialmente. 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

Ciertas partes de esta sección serán agregadas en la medida que se desarrolle esta tesis, y sea más evidente que información le resultara fundamental al lector para poder entender el desarrollo metodológico propuesto. Todavía no se ha decido si incluir una sección tipo-review de machine learning y deep learning, supervisados, aplicados al audio.

2.1. Sistemas de texto a voz

Como fue descripto previamente, la síntesis del habla consiste en la tarea de producir generar discurso humano, a partir de alguna entrada arbitraría. Es 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 perimétrica, donde un modelo acústico es condicionado para modificar nuevamente voces pre-grabadas de acuerdo a variables arbitrarias solicitadas por un usuario.

2.2. Evaluación de voces sintetizadas

Mean Opinion Score (MOS), o promedio de la puntuación subjetiva en castellano, es el test subjetivo más frecuentemente utilizado para determinar la calidad de voces sintetizadas. En general este test presenta una serie de ejemplos que deben ser por sujetos de prueba, 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 6 puntos (0 a 5), en la cual el discurso humano real se encuentra entre los valores de 4,5 a 4,8. Esta metrica proviene del

campo de las telecomunicaciones y se puede definir como la media aritmética calculada a partir de distintas puntuaciones realizadas por distintos sujetos de prueba sobre un mismo estimulo.

2.3. Síntesis del habla basada en técnicas de deep learning

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 eso 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 speech enhancement (mejora del habla) que mayormente involucrando distintos tipos de filtrados.

Para cada avance significativo o arquitectura nueva propuesta, detallare una breve subsubsección dentro de este título.

2.3.1. WaveNet

2.3.2. DeepVoice

2.3.3. Tacotron

2.3.4. TTS con Transformadores

2.3.5. TTS Flow

2.3.6. Conclusión

2.4. 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.

2.4.1. Vocal Tract Lenght Perturbation

Se propone implementar técnicas de alteración del largo del tracto vocal [6] (Vocal Tract Length Perturbation, VTLP). Dicha técnica fue desarrollada para mejorar sistemas de reconocimiento del habla exitosamente, e involucra la deformación en espacio y tiempo del espectro de cada audio. Esta sección será expandida con una explicación más pertinente.

2.4.2. Voice Conversion, alteración de hablante

Modelos de Voice Conversión (VC) o alteración de hablante permiten modificar la identidad de un locutor, manteniendo el contenido semántico de un mensaje. Esta tecnología permite agregar nuevas voces sintéticas a la base de datos aumentada. En esta sección se detallara el modelo especifico (A COMPARISON OF DISCRETE AND SOFT SPEECH UNITS FOR IMPROVED VOICE CONVERSION, B. van Niekerk et al. [15]) de voice conversion que será utilizada para el proceso de DA de esta tesis.

2.5. 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 [16]. El proceso de extracción y la información codificada varía de implementación a implementación.

Los X-Vectors [7], 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 [8].

Por otro lado HuBERT [17] 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 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. ayudaría mucho a la comprensión incluir graficos que ofrezcan otra explicación de este proceso de entrenamiento de Hubert

2.6. 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. https://paperswithcode.com/method/griffin-lim-algorithm

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 (la tarea de mantener el sentido de una oración, pero cambiar el locutor (Voice Conversion)). Estos sistemas pueden ser evaluados de forma tanto subjetiva como objetiva. La Distancia Cepstral de Mel (MCD por sus siglas en inglés) [9] es comúnmente utilizada para medir la calidad del habla en la tarea de conversión de voz. MCD mide la distorsión de distintos rasgos acústicos de una señal, sin embargo, los mismos no siempre se ven correlacionados con la percepción humana.

Dentro del campo del speech enhancement, el PESQ [10] (perceptual evaluation of speech quality o, evaluación percibida de calidad del habla) desarrollado por ITU-T, consiste en una evaluación intrusiva para cuantificar la calidad del habla. Estima un valor MOS comparando la referencia original con la salida degradada del modelo usando distancias paramétricas entre ambas señales. Sin embargo, en muchos casos, para sistemas de TTS, no contamos con las señales originales utilizadas para entrenar una red neuronal.

En 2016, Yoshimura et al. [11] propusieron un predictor de naturalidad en base a una red neuronal convolucional (CNN) para ser utilizada a nivel de enunciado (utterance o, la unidad mínima de duración de habla humana) y a nivel de sistema completo. El modelo se configuró con parámetros arbitrarios y se entrenó con una gran cantidad de resultados de evaluaciones subjetivas reales. A nivel de enunciado, esta solución presentó una gran nivel de varianza respecto a el valor MOS subjetivo real.

En 2021, Mittag y Moller [12] 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. Este ultimo parrafo será expandido con una descripción mas detallada del trabajo de Mittag, NISQA. Podría ser util incluir una tabla comparativa con los resultados relevantes de las distintas tecnicas detalladas en este estado del arte.

4. DESARROLLO

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 subjetiva tipo 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 [13]: 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 x. Incluir esto.

En la Tabla (1) 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 codígo 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 Archivoz.

Al menos un tipo de implementación adicional sera incluida basada en TacoTron2 a traves del servicio de CoquiTTS

Tabla 1. Composición de la base de datos generada. Codígo 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	Implementación con transformadores	es-ar	-
Archivoz	Audios de personas reales	es-ar	7

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)**: Aclarar cantidad de voces alteradas y su origen La implementación utilizada y el factor de deformación de VTLP (fijado entre 0,9 y 1,1) se basaron en recomendaciones de [6].
- **Voice Conversion**: Aclarar cantidad de voces alteradas y su origen .
- Trans-coding: Esta tecnica aún no se implemento. Completar más adelante .
- Alteración de fase: Aclarar cantidad de voces alteradas y su origen . Alteración basada en el algoritmo de Griffin-Lim. Descartando la parte imaginaria del audio de voces humanas y reconstruyendo su fase a partir del AGL, se pueden imitar los artefactos que introducen varios sistemas vocoder, que ignoran o tienen problemas para modelar la fase de audios sintetizados.

Desarrollar cada proceso de expansión de datos

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. La encuesta podría ser conducida en la página de neuropruebas.org desarrollada por el Laboratorio de Inteligencia Artificial Aplicada, ICC - UBA - CONICET. Esta página permite que usuarios completen test subjetivos de forma asincrónica. Es ideal para evaluaciones que requieren de mínimo entrenamiento como la que se propone para la investigación en cuestión.(a determinar si se usa esta pagina u otro metodo de evaluación)

El test consistirá de una serie de audios que deberán ser evaluados en una escala de tipo Likert de 5 puntos por cada sujeto. La cantidad de audios a evaluar no estará acotada en principio, permitiendo a cada usuario estudiar cuantos audios quiera. El propósito de la encuesta subjetiva es el de etiquetar los audios recolectados previamente, con una puntuación. La cantidad de etiquetas necesarias están determinadas por el entrenamiento de la red neuronal que se desarrollará a posteriori. Un precedente útil puede ser tomado 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, se propone que en principio serán necesarios alrededor de 100 sujetos de prueba, asumiendo que cada sujeto de prueba evalua alrededor de 40 audios.(los audios son de 2 a 5 segundos de duración) Este número es aproximado y será revisado a la hora de plantear la arquitectura de la red neuronal que evaluará objetivamente cada audio a partir de su representación vectorizada.

incluir una representación grafica de la base de datos aumentada obtenida

5. RESULTADOS Y ANÁLISIS

6. DISCUSIÓN DE LOS RESULTADOS

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.

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] 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).

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

[13] 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*)

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

[15] 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).

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

[17] 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)