Universidad de La Habana
Facultad de Matemática y Computación

Sintetizador de Voz en Español con Voces Cubanas

Autor:
Sandra Martos Llanes

Tutores:

Lic. Katy Castillo Rosado
Dr.Cs. Flavio Reyes Díaz

Trabajo de Diploma
presentado en opción al título de
Licenciado en Ciencia de la Computación

Fecha

github.com/username/repo

Dedicación

Agradecimientos

Agradecimientos

Opinión del tutor

Opiniones de los tutores

Resumen

Resumen en español

Abstract

Resumen en inglés

Índice general

Índice de figuras

Ejemplos de código

Introducción

La era de la informática conversacional está cambiando la forma en la que los usuarios interactuan con sus dispositivos: el Asistente de Google busca en Internet y lee las instrucciones sobre cómo preparar un pastel, Siri guía en la búsqueda de un lugar desconocido, las líneas automatizadas de servicio al cliente operan sin necesidad de esperas o botones. Uno de los primeros problemas que alguien se plantearía es la comunicación con los dispositivos, y es que no es posible determinar de qué forma va a interactuar una persona con estos. Se pueden hacer aproximaciones y suposiciones, pero cada persona puede decidir operar de una forma diferente, y sería imposible que un actor de voz grabase infinitas respuestas a las ramas de conversación que pueden surgir de una simple pregunta. En este punto entra la síntesis de voz, si en vez de unas grabaciones, el dispositivo pudiese sintetizar una voz humana, podría responder y hablar con la persona, entablando una especie de conversación.

La síntesis de voz es la producción artificial del habla humana. Se han diseñado diferentes sistemas para este propósito, llamados sintetizadores de voz y pueden ser implementados tanto en hardware como en software. Un sistema de conversión texto a voz (TTS, por sus siglas en inglés, *text to speech*) o sintetizador de voz, recibe como datos de entrada una frase escrita y como resultado produce una voz audible que reproduce la frase de entrada . Este sistema de conversión, se compone básicamente por dos componentes: un modelo que predice generalmente en forma de espectograma, la mejor pronunciación posible de cualquier texto dado, y un codificador de voz que, a partir del espectograma anterior produce ondas sonoras de voz.

El proceso de convertir texto a voz requiere un conocimiento detallado en diferentes campos de la ciencia. Si se quisiera construir un sistema TTS desde cero, se tendrían que estudiar los siguientes temas:

- Lingüística, el estudio científico del lenguaje. Para sintetizar un habla coherente, los sistemas TTS necesitan reconocer cómo un hablante humano pronuncia el lenguaje escrito; esto requiere conocimientos de lingüística hasta el nivel del

fonema[1]. Para lograr un TTS verdaderamente realista, el sistema también necesita predecir la prosodia apropiada,que incluye elementos del habla más allá del fonema, como acentos, pausas y entonación.

- Procesamiento digital de señales de voz: en el contexto de los sistemas TTS se utilizan diferentes representaciones de características que describen la señal del habla, lo que hace posible entrenar diversos modelos para generar una nueva voz.

- Inteligencia artificial, especialmente el aprendizaje profundo, un tipo de aprendizaje automático que utiliza una arquitectura informática llamada red neuronal profunda(DNN, del inglés Deep Neuronal Network). Una DNN es un modelo computacional inspirado en el cerebro humano,se conforma por redes complejas de procesadores, cada uno de los cuales realiza una serie de operaciones antes de enviar su salida a otro procesador. Una DNN aprende la mejor vía de procesamiento para lograr resultados deseados. Este modelo tiene una gran potencia informática, lo que lo hace ideal para manejar la gran cantidad de variables necesarias para la síntesis de voz de alta calidad.

Tipos de tecnologías TTS

Hasta la actualidad se han desarrollado variadas tecnologías TTS, las cuales operan de maneras distintas. Entre las más dominantes se encuentran:

1. Síntesis de formantes[1][2] y síntesis articulatoria[3]:
   Los primeros sistemas TTS empleaban tecnologías basadas en reglas, como la síntesis de formantes y la síntesis articulatoria, que lograron resultados similares a través de estrategias ligeramente diferentes. A partir de una grabación realizada a un hablante, se extrajeron características acústicas de este: formantes, cualidades definitorias de los sonidos del habla, en síntesis de formantes; y forma de articulación (nasal, oclusiva,vocal,etc.) en síntesis articulatoria. Luego, se programarían reglas que recrearan estos parámetros con una señal de audio digital. Este TTS era bastante robótico; y estos enfoques necesariamente abstraen gran parte de la variación que se encontrará en el habla humana, aspectos como la variación de tono, porque solo permiten a los programadores escribir reglas para unos pocos parámetros a la vez.

2. Síntesis de difonos[4][5]:
   El próximo gran desarrollo en el campo TTS se llama síntesis de difonos, se inició en la década de 1970 y todavía era de uso popular durante los últimos

[1]fonemas: Unidades de sonido que combinadas producen el habla, como el sonido /t/ en tierra

años del siglo XX. La síntesis de difonos crea un habla de máquina mediante la combinación de difonos, combinaciones de fonemas de una sola unidad, y las transiciones de un fonema al siguiente; es decir, no solo la /t/ en la palabra tierra sino la /t/ más la mitad del siguiente sonido /i/.

Los sistemas TTS basados en síntesis de difonos incluyen también modelos que predicen la duración y el tono de cada difono para una entrada dada, primero se conectan las señales de difono y luego se procesa esta señal para corregir el tono y la duración. El resultado es un discurso sintético con un sonido más natural que el que crea la síntesis de formantes, pero aún está lejos de ser perfecto, y tiene pocas ventajas sobre cualquier otro acercamiento.

3. Síntesis de selección de unidades[4][6]:
   La síntesis de selección de unidades constituye un enfoque ideal para los motores TTS de bajo impacto en la actualidad. Cuando la síntesis de difonos añadió duración y el tono apropiado a través de un segundo sistema de procesamiento, la síntesis de selección de unidad omite ese paso: se inicia con una gran base de datos grabados del habla, alrededor de 20 horas o más, y se seleccionan los fragmentos de sonido que ya tienen la duración y el tono deseado. La síntesis de selección de unidades proporciona un habla similar a la humana sin mucha modificación de la señal, pero sigue siendo identificablemente artificial. Probablemente el audio de salida de la mejor selección de unidades, sea indistinguible de las voces humanas reales, especialmente en contextos con sistemas TTS. Sin embargo, una mayor naturalidad requiere de bases de datos de selección de unidades muy grandes, en algunos sistemas llegando a ser de gigabytes de datos grabados, representando docenas de horas de voz.

4. Síntesis neuronal[7][8][9]:
   La tecnología de las redes neuronales profundas(DNN) es la que impulsa los avances actuales en el campo TTS, y es clave para la obtención de resultados mucho más realistas. Al igual que sus predecesores, el TTS neuronal comienza con grabaciones de voz; la otra entrada es texto, el guión escrito que su locutor de origen utilizó para crear esas grabaciones. Esas entradas alimentan una red neuronal profunda y se aprende el mejor mapeo posible entre un bit de texto y las características acústicas asociadas.

   Una vez que el modelo esté entrenado, podrá predecir sonido realista para nuevos textos: con un modelo TTS neuronal entrenado, junto con un codificador de voz entrenado con los mismos datos, el sistema puede producir un habla que es notablemente similar a la del locutor de origen cuando se expone a prácticamente cualquier texto nuevo. Esa similitud entre la fuente y la salida es la razón por la que el TTS neuronal a veces se denomina “clonación de voz”.

Hay todo un grupo de métodos de procesamiento de señales que pueden ser utilizados para alterar la voz sintética resultante y no se asemeje al locutor fuente. En la actualidad, las principales investigaciones se enfocan en lograr voces sintéticas con una calidad de audio cada vez más realista.

Entre las aplicaciones que hacen uso del TTS, se encuentran:

- Sistemas de respuesta de voz interactiva conversacional (IVR), como en los centros de llamadas de servicio al cliente.
- Aplicaciones de comercio de voz, como comprar en un dispositivo Amazon Alexa.
- Herramientas de navegación y guía por voz, como aplicaciones de mapas GPS.
- Dispositivos domésticos inteligentes y otras herramientas de Internet de las cosas (IoT) habilitadas por voz.
- Asistentes virtuales independientes, como Siri de Apple.
- Soluciones de publicidad y marketing experiencial, como anuncios de voz interactivos en servicios de transmisión de música.
- Desarrollo de videojuegos.
- Videos de marketing y formación de la empresa que permiten a los creadores cambiar las voces en off sin identificar al locutor.

Problemática

Existe un conjunto de sistemas TTS que se enmarcan bajo una licencia de software libre:

- Festival y Festvox
- Plataforma MaryTTS
- Sistema TTS
- SV2TTS
- Mozilla-TTS
- COQUI -TTS

A partir del estudio parcial de las plataformas de código abierto utilizadas para el desarrollo de conversores de texto a voz, fue posible comprobar que la mayoría se encuentran basadas en la síntesis neuronal teniendo en cuenta que alcanza mejores resultados. Estas plataformas brindan modelos previamente entrenados para idiomas específicos como inglés, francés, alemán, etc. Muy pocas presentan un modelo en español, y las que lo hacen solo poseen uno, con acento de voz española o voz neutra.

DATYS, es una empresa de desarrollo de software, que como parte de sus soluciones requiere un sintetizador de voz en español con acento propio de nuestro país; esto formará parte de un proyecto que consiste en el desarrollo del primer asistente virtual cubano.

Objetivo

Objetivo General

Desarrollar un sintetizador de texto a voz con voces cubanas.

Objetivos específicos

1. Analizar en profundidad los métodos y plataformas existentes para la síntesis de voz, principalmente las que trabajan sobre la síntesis neuronal, y seleccionar el más adecuado a utilizar para las aplicaciones de DATYS.

2. Diseñar y conformar una base de datos en español, con con voces cubanas para el entrenamiento de los modelos y muestras de voz que se desea generar.

3. Reentrenar el modelo basado en síntesis neuronal seleccionado, para su ajuste al estilo de voz cubana.

4. Evaluar el modelo entrenado y analizar los resultados obtenidos.

Capítulo 1

Sistemas para la síntesis texto a voz

Las redes neuronales han sido las encargadas de los avances actuales en el campo TTS, por tanto esta será la línea seguida en la investigación.

Para lograr la transformación de texto a voz, se siguen dos enfoques distintos. El primero es un sistema de dos estapas: una combinación de dos redes neuronales, una para convertir de texto a espectograma de mel, y luego otra que transforma el espectograma en onda sonora; y el segundo es un modelo de extremo a extremo.

1.1. Sistemas de dos etapas

El paradigma predominante en la conversión de texto a voz es la síntesis en dos etapas, es decir, primero, producir espectrogramas a escala de mel a partir del texto y, luego las ondas de sonido reales con un modelo de codificador de voz(*VoCoder*. La representación acústica de bajo nivel, espectograma de mel, es la utilizada como nexo entre las dos componentes.

1.1.1. Modelos TTS

Tacotron

Tacotron[7] es un modelo TTS de tipo secuencia a secuencia con un paradigma de atención. Este modelo toma caracteres como entrada y devuelve un espectograma

sin procesar usando técnicas para mejorar un modelo seq2seq[1] vainilla[2]. Dado un par <texto,audio>, Tacotron puede ser entrenado desde cero con una inicialización aleatoria, y no requiere alineación a nivel de fonema.

La columna vertebral de Tacotron es un modelo seq2seq con atención, que toma caracteres como entrada, y devuelve el correspondiente espectograma sin procesar, para luego pasarlo al modelo o algoritmo que sinteiza la voz. En el centro de todo esto se encuentra un codificador,un decodificador basado en atención y una red de post procesamiento.

Tacotron se basa en cuadros, o *frames*, por lo que la inferencia es sustancialmente más rápida que los métodos autorregresivos a nivel de muestra. A diferencia de otras tecnologías TTS más antiguas, Tacotron no necesita características lingüísticas diseñadas a mano ni componentes complejos como un alineador de Modelo Markov oculto(HMM). Este modelo realiza una normalización de texto simple.

Tacotron 2

Tacotron 2[10] es similar al anteriormente mencionado Tacotron; es una red recurrente de predicción de características, de tipo secuencia a secuencia con atención, que mapea incrustaciones(del inglés, *embeddings*) de caracteres en espectrogramas a escala de mel.

Un espectrograma de frecuencia de mel está relacionado con el espectograma de frecuencia lineal, es decir, la magnitud de la transformada de Fourier de tiempo corto (STFT, por sus siglas del inglés *short-time Fourier transform*). Se obtiene aplicando una transformada no lineal al eje de frecuencia de la STFT, inspirado en respuestas calificadas por el sistema auditivo humano, y resume el contenido de frecuencia con menos dimensiones.

El uso de una escala de frecuencia auditiva de este tipo tiene el efecto de enfatizar detalles en frecuencias más bajas, que son fundamentales para la inteligibilidad del habla, al mismo tiempo que se resta importancia a los detalles de alta frecuencia, que están dominados por ráfagas de ruido y generalmente no necesitan ser modelados con alta fidelidad. Debido a estas propiedades, las características derivadas de la escala de mel se han utilizado como representación base para el reconocimiento de voz durante muchas décadas.

Para Tacotron2 los espectrogramas de mel se calculan a través de un transformada

[1]Seq2Seq se basa en el paradigma codificador-decodificador. El codificador codifica la secuencia de entrada, mientras que el decodificador produce la secuencia de destino. codificador

[2]En informática, vainilla es el término utilizado cuando el *software* o los algoritmos, no se emplean a partir de su forma original.

de Fourier de tiempo corto (STFT).

La red del modelo en cuestión está compuesta por un codificador y un decodificador con atención. El codificador convierte una secuencia de caracteres en una representación oculta que alimenta al decodificador para predecir un espectrograma. Los caracteres de entrada se representan utilizando una incrustación de caracteres. La salida del codificador es consumida por una red de atención que resume la secuencia codificada completa como un vector de contexto de longitud fija para cada paso de salida del decodificador. Se usa la atención sensible a la ubicación de [11], que extiende el mecanismo de atención aditiva [12] para usar pesos de atención acumulativos de anteriores pasos de tiempo del decodificador como una funcionalidad adicional. Esto anima al modelo a seguir adelante consistentemente a través de la entrada, mitigando los posibles modos de falla donde algunas subsecuencias son repetidas o ignoradas por el decodificador.

El decodificador es una red neuronal autorregresiva recurrente que predice un espectrograma de mel a partir de la secuencia de entrada codificada un fotograma a la vez.

Este sistema puede ser entrenado directamente desde un conjunto de datos sin depender de una compleja ingeniería de características, y logra calidad de sonido de última generación cercana a la del habla humana natural. Los resultados de Tacotron 2, constituyen un paso de avance sobre Tacotron y otros sistemas previos, sin embargo dejan aún espacio para mejoras.

Deep Voice 1, 2, 3

Deep Voice de Baidu[13][14] sentó las bases para los avances posteriores en la síntesis de voz de extremo a extremo. Consta de 4 redes neuronales diferentes que juntas forman un extremo de la canalización: un modelo de segmentación que localiza los límites entre fonemas,un modelo que convierte grafemas en fonemas, un modelo para predecir la duración de los fonemas y las frecuencias fundamentales, y un modelo para sintetizar el audio final.

Deep Voice 2 [14] se presentó como una mejora de la arquitectura original de Deep Voice. Si bien la canalización principal era bastante similar, cada modelo se creó desde cero para mejorar su rendimiento. Otra gran mejora fue la adición de compatibilidad con varios hablantes.

Deep Voice 3 [14][15] es un rediseño completo de las versiones anteriores. Aquí se tiene un solo modelo en lugar de cuatro diferentes. Más específicamente, los autores propusieron una arquitectura de caracter a espectrograma completamente convolucional que es ideal para el cálculo paralelo. A diferencia de los modelos basados en RNN. También se experimentó con diferentes métodos de síntesis de forma de onda

con WaveNet logrando los mejores resultados una vez más.

Modelos TTS con Transformers

Los transformadores(*transformers*), están dominando el campo del lenguaje natural desde hace un tiempo, por lo que era inevitable que ingresaran gradualmente al campo TTS. Los modelos basados en transformadores tienen como objetivo abordar dos problemas de los métodos TTS anteriores, como Tacotron2:

- Baja eficiencia durante el entrenamiento y la inferencia.
- Dificultad para modelar dependencias largas usando redes neuronales recurentes(RNN, por sus siglas en inglés).

La primera arquitectura basada en transformadores se introdujo en 2018 y reemplazó las RNN con mecanismos de atención de múltiples cabezales que se pueden entrenar en paralelo.

FastSpeech

FastSpeech, una novedosa red de avance basada en Transformer para generar espectrogramas de mel en paralelo para TTS; toma como entrada una secuencia de texto(fonema) y genera espectrogramas de mel de forma no autorregresiva. Adopta una red feed-forward basada en la autoatención[3] en Transformer y convolución de 1D.

El modelo resuelve problemas existentes en modelos TTS antiguos de la siguiente forma:

- A través de la generación de espectrogramas de mel paralelos, FastSpeech acelera enormemente el proceso de síntesis.
- El predictor de duración de fonemas asegura alineaciones estrictas entre un fonema y sus espectrogramas, lo que es muy diferente de las alineaciones de atención automáticas y suaves en los modelos autorregresivos. Por lo tanto, FastSpeech evita los problemas de propagación de errores y alineaciones de atención incorrectas, lo que reduce la proporción de palabras omitidas y palabras repetidas.

[3]La autoatención permite a una red neuronal entender una palabra en el contexto de las palabras que la rodean.

- El regulador de longitud puede ajustar fácilmente la velocidad de la voz alargando o acortando la duración del fonema para determinar la duración de los espectrogramas de mel generados, y también puede controlar parte de la prosodia añadiendo pausas entre fonemas adyacentes.

La arquitectura para Fast Speech es una estructura de avance basada en la autoatención en Transformer y la convolución de 1D; se nombra esta estructura como Feed-Forward Transformer (FFT). Feed-Forward Transformer apila múltiples bloques FFT para la transformación de fonema a espectrograma de mel, con N bloques en el lado del fonema y N bloques en el lado del espectrograma de mel, con un regulador de longitud en el medio para cerrar la brecha de longitud entre el fonema y la secuencia del espectrograma de mel.

Posee un regulador de longitud que se utiliza para resolver el problema de la discordancia de longitud entre el fonema y la secuencia del espectrograma en el transformador de avance, así como para controlar la velocidad de la voz y parte de la prosodia. Finalmente un predictor de duración que genera un escalar, que es exactamente la duración prevista del fonema.

El entrenamiento de FastSpeech y de gran mayoría de los modelos TTS se realiza sobre un conjunto de datos que contiene clips de audios con sus correspondientes transcripciones de texto. Específicamente para FastSpeech, se divide al azar el conjunto de datos en 3 conjuntos: muestras para entrenamiento, muestras para validación y muestras para las pruebas.

El modelo FastSpeech puede casi coincidir con el modelo autoregresivo Transformer TTS en términos de calidad de voz, acelera la generación de espectrograma mel por 270x y la síntesis de voz de extremo a extremo por 38x, casi se elimina el problema de saltar y repetir palabras, y puede ajustar la velocidad de voz (0.5x-1.5x) sin problemas[8].

FastPitch

FastPitch es un modelo TTS feed-forward completamente paralelo basado en FastSpeech, condicionado por contornos de frecuencia fundamentales. El modelo predice contornos de tono durante la inferencia. Al alterar estas predicciones, el discurso generado puede ser más expresivo, coincidir mejor con la semántica del enunciado y, al final, ser más atractivo para el oyente.

Los modelos paralelos pueden sintetizar órdenes de magnitud de espectrogramas de mel más rápido que los autorregresivos, ya sea basándose en alineaciones externas o alineándose ellos mismos. El condicionamiento en la frecuencia fundamental también mejora la convergencia y elimina la necesidad de destilar el conocimiento de los

objetivos del espectrograma de mel utilizados en FastSpeech.

La arquitectura del modelo, se basa en FastSpeech y se compone principalmente de dos pilas de transformadores alimentados hacia adelante, feed-forward(FFTr) . El primero opera en la resolución de los tokens de entrada, el segundo en la resolución de los cuadros de salida. [9].

Para el entrenamiento y la experimentación los parámetros del modelo siguen principalmente FastSpeech.

Glow-TTS

A pesar de la ventaja, los modelos TTS paralelos no se pueden entrenar sin la guía de modelos TTS autorregresivos como alineadores externos.

Glow-TTS[16] es un modelo generativo basado en flujo para TTS paralelo que no requiere de ningún alineador externo. Al combinar las propiedades de los flujos y la programación dinámica, el el modelo busca la alineación monótona más probable entre el texto y la representación latente del habla en sí misma. La arquitectura del modelo consiste en un codificador que sigue la estructura del codificador de *Transformer TTS* con pequeñas modificaciones, un predictor de duración, que tiene una estructura y configuración como la de FastSpeech, y finalmente un decodificador basado en flujo, que es la parte fundamental del modelo.

Se demostró que hacer cumplir alineaciones monótonas fuerte permite un texto a voz robusto, que se generaliza a largas pronunciaciones, y el empleo de flujos generativos permite síntesis de voz rápida, diversa y controlable. Glow-TTS puede generar espectrogramas mel 15,7 veces más rápido que el modelo TTS autorregresivo, Tacotron 2, mientras obtiene un rendimiento con calidad de voz comparable. Según la literatura, el modelo se puede extender fácilmente a una configuración de múltiples hablantes.

1.1.2. VoCoders

Los VoCoders neuronales basados en redes neuronales profundas pueden generar voces similares a las humanas, en lugar de utilizar las tradicionales métodos que contienen artefactos audibles[17][18][19].

La línea principal de la investigación se basa en los modelos TTS, como los antes expuestos, sin embargo como no es posible sintetizar voz sin un VoCoder, y luego de varias pruebas realizadas, se concluye que los más adecuados para el objetivo principal son los siguientes:

HIFI-GAN

Varios trabajos recientes sobre la síntesis del habla han empleado redes generativas adversariales(GAN, por sus siglas en inglés) para producir formas de onda sin procesar. Aunque estos métodos mejoran la eficiencia de muestreo y uso de memoria, su calidad de muestra aún no ha alcanzado el de los modelos generativos autorregresivos y basados en flujo. HiFi-GAN[20] es un modelo que logra una síntesis de voz eficiente y de alta fidelidad.

Como el audio del habla consta de señales sinusoidales con varios períodos, se comprobó que modelar patrones periódicos de un audio es crucial para mejorar la calidad de la muestra.

Además se muestra la adaptabilidad de HiFi-GAN a la síntesis de voz de extremo a extremo. Para terminar, una versión pequeña de HiFi-GAN genera en CPU muestras 13,4 veces más rápido en tiempo real con calidad comparable a una contraparte autorregresiva.

UnivNet

La mayoría de los codificadores de voz neuronales emplean espectrogramas de mel de banda limitada para generar formas de onda. UnivNet[21], es un codificador neural de voz que sintetiza formas de onda de alta fidelidad en tiempo real.

Usando espectrogramas de mel de banda completa como entrada, se espera generar señales de alta resolución agregando un discriminador que emplea espectrogramas de múltiples resoluciones como entrada. En una evaluación de un conjunto de datos que contiene información sobre cientos de ponentes, UnivNet obtuvo los mejores resultados positivos, objetivos y subjetivos, entre los modelos que competían. Estos resultados, incluida la mejor puntuación subjetiva en la conversión texto a voz, demuestran el potencial para una rápida adaptación a nuevos hablantes sin necesidad de entrenamiento desde cero.

WaveGrad

WaveGrad[22] es un modelo condicional para la generación de formas de onda que estima los gradientes de la densidad de datos. El modelo se basa en trabajos previos sobre emparejamiento de puntuaciones y modelos probabilísticos de difusión. Parte de una señal Gaussiana de ruido blanco e iterativamente refina la señal a través de un muestreador basado en gradientes, condicionado en el espectrograma de mel. WaveGrad ofrece una forma natural de intercambiar velocidad de referencia por calidad de la muestra ajustando el número de pasos de refinamiento, y cierra la brecha entre los modelos autorregresivos y no autorregresivos en términos de calidad de audio. El modelo puede generar muestras de audio de alta fidelidad usando como tan solo seis

iteraciones. Los experimentos revelan que WaveGrad genera señales de audio de alta fidelidad, superando las líneas de base adversariales no autorregresivas y emparejando un fuertemente la línea de base autorregresiva basada en la probabilidad, utilizando menos operaciones secuenciales.

1.2. Modelos de extremo a extremo

1.2.1. YourTTS

1.2.2. VITS

Variational Inference with adversarial learning for end-to-end Text-to-Speech (VITS)[23] es un método TTS de extremo a extremo en paralelo. Usando un Autocodificador Variacional se conectan los dos módulos de sistemas TTS: modelo acústico y VoCoder, a través de variables latentes para permitir el aprendizaje de extremo a extremo.

Para mejorar el poder expresivo del método con el fin de sintetizar formas de onda de voz de alta calidad, se aplican flujos de normalización a la distribución condicional previa y entrenamiento adversarial en el dominio de formas de ondas.

El modelo VITS se describe principalmente en tres etapas: una formulación condicional de Autocodificador variacional; estimación de alineación derivada de la inferencia variacional; y entrenamiento adversarial para mejorar la calidad de la síntesis.

La arquitectura general del modelo consiste en un codificador posterior, un codificador anterior, un decodificador, un discriminante, y predictor de duración estocástica. El codificador posterior y discriminante solo se usan para entrenamiento, no para inferencia.

Para el codificador posterior se utilizan los bloques residuales no causales de WaveNet. Un bloque residual de WaveNet consta de capas convolucionales dilatadas con una puerta unidad de activación y salto de conexión. La proyección lineal capa por encima de los bloques produce la media y la varianza de la distribución posterior normal.

El codificador anterior consiste en un codificador de texto que procesa los fonemas de entrada, y un flujo de normalización que mejora la flexibilidad de la distribución anterior. El codificador de texto es un codificador transformador que utiliza representación posicional relativa en lugar de la codificación posicional absoluta. Mientras que el flujo de normalización es una pila de capas de acoplamiento afines [24] conformada por una pila de bloques residuales de WaveNet.

El decodificador es esencialmente el generador HiFi-GAN V1[20]. Se compone de una pila de convoluciones transpuestas, cada una de las cuales va seguida de un módulo de fusión de campo multirreceptivo (MRF). El modelo continua la arquitectura del discriminante multiperíodo discriminador propuesto en HiFi-GAN. El discriminante

multiperíodo es una mezcla de subdiscriminadores Markovianos basados en ventanas, cada uno de los cuales opera en diferentes patrones periódicos de formas de onda de entrada.

El predictor de duración estocástica estima la distribución de la duración del fonema a partir de una entrada condicional. Para la parametrización eficiente del predictor de duración estocástica, son apilados bloques residuales con bloques dilatados y capas convolucionales separables en profundidad. También se aplican flujos de ranunas neuronales[25] , que toman la forma de transformaciones no lineales invertibles mediante el uso de ranuras monótonas racionales-cuadráticas, aplicadas a capas de acoplamiento. flujos de ranunas neuronales mejoran la expresividad de la transformación con un número de parámetros en comparación con los de uso común en capas de acoplamiento.

Una vez concluido el proceso de entrenamiento, y según la literatura al compararse este modelo de extremo a extremo con sistemas de dos etapas, a través de los modelos preentrenados Tacotron2 y Glow-TTS como modelos de primer escenario e HiFi-GAN como modelo de segundo escenario, se comprueba que VITS obtiene un habla que suena más natural y cercana a la realidad. Una evaluación humana subjetiva (puntuación de opinión media, o MOS) en LJ Speech[26], un conjunto de datos de un solo hablante, muestra que el modelo supera a los mejores sistemas TTS disponibles públicamente y logra un MOS comparable a la realidad del terreno.

Ha mostrado la capacidad de ampliarse a la síntesis de voz de múltiples hablantes, generando un discurso con diversos tonos y ritmos de acuerdo a diferentes identidades de hablantes. Esto demuestra que aprende y expresa varias características del habla en un contexto de extremo a extremo.

1.3. Coqui-TTS

Existen conjuntos de sistemas TTS que abarcan la gran mayoría de los modelos explicados anteriormente, entre las más populares se encuentran **Mozilla-TTS**[27] y **Coqui-TTS**[28]. Ambas se comportan de forma similar, se instalan con el mismo comando `pip install TTS`, y el comando para ejecutarse tiene la misma sintaxis. Coqui-TTS fue fundado por miembros del equipo de Mozilla-TTS, y es su sucesor, pues Mozilla dejó de actualizar su proyecto STT y TTS.

Coqui TTS es una biblioteca para la generación avanzada de texto a voz. Se basa en las últimas investigaciones y se diseñó para lograr el mejor equilibrio entre la facilidad de entrenamiento, la velocidad y la calidad. Coqui viene con modelos preentrenados, herramientas para medir la calidad del conjunto de datos y ya se utiliza en más de 20 idiomas para productos y proyectos de investigación.

Capítulo 2

Propuesta

Con el objetivo de lograr una síntesis de voz satisfactoria y después de realizar un estudio de las tecnologías que se utilizan para este fin, se elige Coqui TTS como herramienta base. Coqui cuenta con una gran variedad de modelos preentrenados en más de 20 idiomas. El primer paso en el desarrollo del sintetizador ess instalar la biblioteca Coqui TTS, de acuerdo a las indicaciones del repositorio oficial[28]. Se realizó un estudio del comportamiento de combinaciones de parejas modelo TTS y VoCoder, para evaluar cuáles ofrecían mejores resultados, y se pueden observar en la figuras 2.1, 2.2, 2.3.

Se experimenta la combinación del modelo TTS Tacotron2-DDC preentrenado en Inglés con los VoCoders WaveGrad y HiFiGan-V2 como se evidencia en la figura 2.1. Igualmente con el modelo Fast-Pitch preentrenado en inglés con el VoCoder UnivNet igualmente en en inglés(figura 2.2), para terminar con los modelos en inglés se combina Glow-TTS y el VoCoder Multiband-MelGan(figura 2.3) Los modelos preentrenados en idioma inglés al combinarse con estos VoCoders, para un texto en español arroja distintos resultados, en algunos casos solo producen señales ruidosas mientras que los más satisfactorios, producen un discurso con una pronunciación propia de una persona de habla inglesa hablando español.

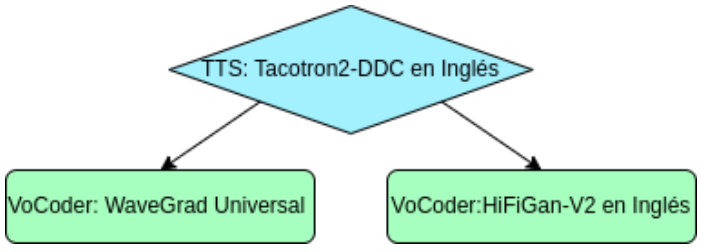

Figura 2.1: Modelo TTS Tacotron2-DDC entrenado en Inglés

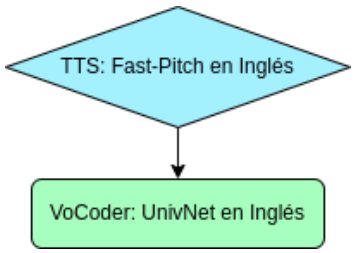

Figura 2.2: Modelo TTS Fast-Pitch entrenado en Inglés

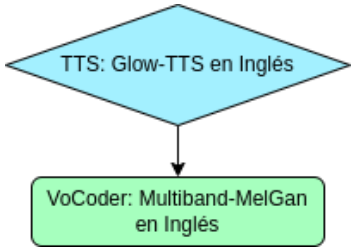

Figura 2.3: Modelo TTS Glow-TTS entrenado en Inglés

Por otro lado Coqui solo cuenta con un modelo preentrenado en español, el **Tacontron-DDC**, que está entrenado sobre la base de datos de M-AILABS[29]; este modelo fue probado junto a los VoCoders preentrenados como se muestra en la figura 2.4. Los mejores resultados fueron arrojados por las combinaciones de **Tacontron-DDC** con los modelos: Univnet entrenado sobre la base de datos Ljspeech en inglés, y Wavegrad entrenado sobre el conjunto de datos LibriTTS, aunque ambos presentan problemáticas como la mala pronunciación de la letra ñ.

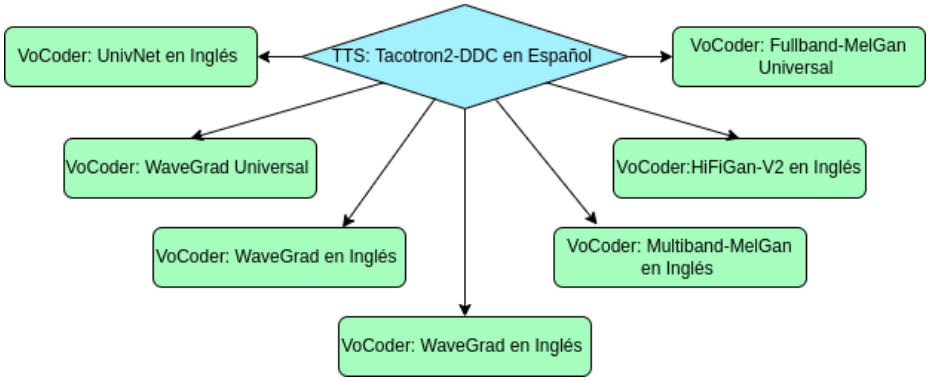

Figura 2.4: Modelo TTS Tacotron2-DDC entrenado en Español

De acuerdo con la documentación de Coqui, el objetivo de la investigación que consiste en desarrollar un sintetizador de voz en español con voz cubana, y experimentos

iniciales realizados se proponen tres enfoques:

1. Realizar un *fine-tuning* o reentrenamiento, utilizando una base de datos con voces cubanas, al modelo preentrenado de Coqui[28], **Tacontron-DDC**, que es el único disponible en español.

2. Realizar el entrenamiendo desde cero de un modelo, en este caso se podría entrenar cualquier modelo disponible. **VITS** es un buen candidato pues en general produce resultados bastante satisfactorios.

3. Realizar un *fine-tuning* a modelos VITS preentrenados en idiomas distintos al español.

4. Realizar un preentrenamiento desde cero al modelo VITS utilizando una base de datos fuente, para luego realizar un proceso de *fine-tuning* sobre este, pero utilizando la base de datos de voces cubanas contruida.

Y a partir de estos evaluar y comparar resultados.

Para el desarrollo de cualquiera de los anteriores es imprescindible la construcción de un conjunto de datos que se adapten al modelo y a las necesidades de la investigación.

2.1. Creación de la base de datos

Para entrenar un modelo TTS, se necesita un conjunto de datos con grabaciones de voz y transcripciones, en este caso fue una base de datos en español con voces cubanas. El discurso debe dividirse en clips de audio y cada clip necesita su transcripción correspondiente. La base de datos debe poseer una organización específica, de forma tal que el cargador de datos de Coqui TTS sea capaz de cargarlos para utilizar en el entrenamiento[30].

Formato wav

Los clips de audio del conjunto de datos deben poseer formato WAV. El formato WAV[31], es un formato estándar de archivo de audio, desarrollado por IBM[32] y Microsoft[33], para almacenar un flujo de bits de audio en PC. Los archivos WAV no se comprimen cuando se codifican. Eso significa que todos los elementos del audio originales permanecen en el archivo. Los editores de audio describen los archivos WAV como “sin pérdidas” porque no se pierde ninguna parte de su sonido. Como resultado, los archivos WAV tienen objetivamente una mejor calidad y proporcionan clips de audio más reales y precisos. .

Idioma Español

La base de datos conformanda debe tener una buena cobertura del idioma en el que se desea entrenar el modelo. Debe cubrir la variedad fonémica, los sonidos excepcionales y los signos especiales.

El idioma español[34] o castellano. Es una lengua romance del grupo ibérico, y tras el chino mandarín, es la lengua más hablada del mundo por el número de personas que la tienen como lengua materna. El idioma es un rasgo esencial de la nacionalidad. Por eso unos de los rasgos que define a los cubanos, es precisamente este: la comunicación mediante el idioma español.

Las variedades geográficas del español, llamadas dialectos o geolectos, difieren entre sí por multitud de razones. Entre las de tipo fonético destacan la distinción o no de los fonemas correspondientes a las grafías c/z y s (ausencia o presencia de ceceo/seseo), la distinción o no de los fonemas correspondientes a las grafías ll e y (ausencia o presencia de yeísmo) y la aspiración o no de la s ó z ante una consonante, aunque estas diferencias no suelen ocasionar problemas de inteligibilidad entre sus hablantes.

La estructura silábica más frecuente del español es consonante más vocal, de forma que tiende hacia la sílaba abierta. Caracteriza al español una tensión articulatoria alta, no tan relajada como en italiano, y estadísticamente una gran presencia de la vocal a. La acentuación es de intensidad y estadísticamente dominan las palabras llanas, o acentuadas en la penúltima sílaba, después las agudas y por último las esdrújulas.

En español hay cinco vocales fonológicas: /a/, /e/, /i/, /o/ y /u/. La /e/ y /o/ son vocales medias, ni cerradas ni abiertas, pero pueden tender a cerrarse y abrirse. Sin embargo, estos sonidos no suponen un rasgo distintivo en español general. Según la mayoría de los autores, se distinguen por lo general 24 fonemas en el español, cinco de los cuales corresponden a vocales y 19 a consonantes, además de otros fonemas dialectales y/o alofónicos[1], aunque la mayoría de los dialectos sólo cuentan con 17 consonantes, y algunos otros con 18.

El español se escribe mediante una variante del alfabeto latino con la letra adicional “ñ” y los dígrafos[2] “ch” y “ll”, consideradas letras del abecedario desde 1803, debido a que representan un solo sonido, distinto de las letras que lo componen. Así, el alfabeto español queda formado por 27 letras y 2 dígrafos: a, b, c, ch, d, e, f, g, h, i, j, k, l, ll , m, n, ñ, o, p, q, r, s, t, u, v, w, x, y, z. Además el español emplea signos gráficos de interrogación y exclamación como “¿” y “¡”, que no poseen otras lenguas. Estos signos especiales facilitan la lectura de interrogaciones y exclamaciones largas que oralmente solo se expresan por variaciones de entonación.

[1] alófono: Sonido propio de la pronunciación de un fonema, que puede variar según su posición en la palabra o en la sílaba y en relación con los sonidos vecinos, aunque sigue considerándose el mismo fonema.

[2] dígrafo: Signo ortográfico compuesto de dos letras y que representa un solo sonido.

Fine-Tuning

Entrenar correctamente un modelo de aprendizaje profundo requiere generalmente de una gran base de datos y de un extenso entrenamiento.

Si se dispone del material necesario y del tiempo para entrenar el algoritmo, estos requisitos no suponen ningún problema, pero, si la base de datos es pequeña o el modelo no se entrena lo suficiente, el aprendizaje podría no ser completo.

El *fine-tuning* consiste en aprovechar la información que se ha aprendido de la resolución de un problema y utilizarla sobre otro distinto, pero que comparte ciertas características. Se usan los conocimientos adquiridos por una red convolucional para transmitirlos a una nueva red convolucional. Esta nueva red convolucional que se crea no tiene por qué modificar la red original y puede simplemente aprender de ella, sin embargo también es válido el caso donde no solo se modifica la red original, sino que se vuelve a entrenar para aprender más conceptos.

En la presente investigación, se utiliza el reentrenamiento, para a partir modelo previamente entrenado y realizar un nuevo entrenamiento para mejorar su rendimiento en un conjunto de datos diferente.

2.2. Fine-Tuning de Tacontron-DDC

Se implementa como primera variante el reentrenamiento(*fine-tuning* en inglés) del modelo **Tacontron-DDC** en español, pues brinda ventajas tales como un aprendizaje más rápido, ya que un modelo preentrenado ya tiene aprendidas funcionalidades que son relevantes para la tarea de producir un discurso. Además convergerá más rápido en el nuevo *dataset*, lo que reducirá el costo de entrenamiento y permitirá el proceso de experimentación más rápidamente. Y de acuerdo a la teoría se pueden obtener buenos resultados con un conjunto de datos más pequeño.

Luego de tener entrenado el modelo acústico(TTS) con una base de datos construida con voces cubanas, es posible que alguno de los VoCoders preentrenados disponibles produzca una salida con las características deseadas, en caso contrario se debería entrenar el VoCoder con los datos del *dataset* construido.

El proceso de *fine-tuning* consiste en modificar la configuración original del modelo preentrenado seleccionado, es decir, especificar la base de datos a utilizar en el reentrenamiento, los detalles acústicos que reflejen las características del nuevo conjunto de datos, el nombre del nuevo modelo ajustado, la dirección donde se guardará el modelo reentrenado, entre otros aspectos. Para la mayoría de los parámetros se tomaron las características del modelo original **Tacontron-DDC** en español.

2.3. Entrenamiento de modelo VITS desde cero

Existen varios modelos de texto a voz de extremo a extremo que permiten el entrenamiento en una sola etapa y el muestreo en paralelo, sin embargo, generalmente la calidad de la muestra no coincide con la de los sistemas TTS de dos etapas.

Se selecciona VITS porque es un método TTS paralelo de extremo a extremo que genera un sonido de audio más natural que los modelos actuales de dos etapas. Y de acuerdo a una evaluación humana subjetiva[35] (puntuación de opinión media, o MOS), muestra que el modelo supera a los mejores sistemas TTS disponibles públicamente y logra un MOS comparable a la realidad.

Con este enfoque se debe entrenar la red neuronal de VITS partiendo de cero. Una desventaja es que cae una gran responsabilidad sobre el conjunto de datos de entrenamiento, pues debe tener una gran riqueza del idioma y muchos clips de audio.

2.4. Fine-tuning de modelos VITS preentrenados en otros idiomas

Se persigue experimentar el proceso de *fine-tuning* en este modelo de extremo a extremo, sin embargo no existen modelos VITS preentrenados en español disponibles. Por esto se escogen los idiomas italiano e inglés como candidatos en esta tarea.

2.4.1. Italiano

Las culturas hispánicas tienen mucho en común con la cultura italiana, sobre todo en lo que concierne al lenguaje y la comunicación. Tanto el español como el italiano son lenguas romances, derivadas del latín, siendo justamente de las más similares, incluso más que el francés o el rumano.

Es por este hecho que el italiano y el español comparten palabras muy parecidas y siguen la misma estructura gramatical. Entre ellos existe un grado de similitud léxica del 82 %, lo que además indica que son idiomas fáciles de aprender para sus respectivos hablantes. Sin embargo, las características que tienen en común van más allá de su origen y de la gramática.

Una similitud entre ambas lenguas es la cantidad de vocales en el alfabeto, aunque en italiano las vocales tienden a matizarse con sonidos más abiertos y con un acento grave. Es muy común que el parecido existente entre las palabras en italiano y español se vea afectada sólo por una o más vocales, como por ejemplo: vecino y vicino, cámara y camera, igual y uguale. Por supuesto existen un gran número de diferencias, aunque son más notables en vocabulario que en lo que respecta a la pronunciación.

Debido a todo esto un modelo entrenado en italiano sería el candidato ideal para realizar un ajuste sobre una base de datos en español.

2.4.2. Inglés

A pesar de las diferencias evidentes, en realidad el español y el inglés se parecen más de lo se cree. La primera de las semejanzas entre el inglés y el español es que usan el alfabeto romano, lo cual provoca que los sonidos de ambos idiomas sean similares, aunque no plenamente iguales. El mejor ejemplo lo constituyen las vocales: mientras en el español sólo se reconocen 5 sonidos, uno para cada vocal, en el inglés encontramos más de 14 sonidos, pues hay vocales cortas y largas, las cuales se catalogan así por la duración de su pronunciación. Estos dos idiomas son alfabéticos, de modo que a través de letras se pueden representar sus sonidos; el español comparte 2/3 de sus fonemas con el inglés.

La estructura gramatical del inglés es similar a la del español en más de un 90 %, pero la del inglés es muchísimo más simple. Además, el vocabulario se forma con prefijos y sufijos tanto en inglés como en español (que son en su mayoría de raíz latina en ambos idiomas).

Las consonantes v, ll, h, j, r, rr, z, y la x son pronunciadas muy diferentes en español y en inglés. La consonante ñ no existe en inglés; el sonido que se conoce de esta consonante en español se escribe en ingles ny. Las combinaciones de algunas consonantes con vocales cambia totalmente el sentido de la pronunciación. Por ejemplo: la combinación de la “q” y la “u” en las palabras como queen, quiet o quick la pronunciación de la “u” en inglés se percibe.

Teniendo en cuenta la similitud, y al mismo tiempo las marcadas diferencias entre ambos idiomas, se escoge un modelo VITS preentrenado en inglés para realizar un *fine-tuning* con la base de datos cubana.

Capítulo 3

Experimentación y Resultados

3.1. Instalación de la biblioteca Coqui

Coqui [28] es un repositorio de código abierto que implementa las últimas investigaciones en materia de síntesis de voz, como Tacotron 2 y VITS que son los modelos base utilizados en el presente proyecto. Este repositorio ha sido usado para generar modelos en más de 20 idiomas y cuenta además con múltiples "recetas" para el entrenamiento de modelos.

La biblioteca se instala de acuerdo a las instrucciones orientadas a desarrolladores en [28]. Con esto ya es suficiente para probar los modelos preentrenados disponibles de Coqui.

3.2. Configuración de la evaluación de los modelos

Se utiliza la medida MOS para la evaluación de los modelos obtenidos en este trabajo. Un puntaje de opinión promedio[35][36] (MOS, del inglés *Mean Opinion Score*) es una medida numérica de la calidad general de un evento o experiencia juzgada por humanos. En telecomunicaciones, MOS es terminología para clasificación de calidad de audio y video, se refiere a la calidad de escuchar, hablar o conversar, ya sea que se originen en modelos subjetivos u objetivos. Y se evalua de la siguiente forma:

Muy bueno	4.3 - 5.0
Malo	3.1 - 3.6
No recomendado	2.5 - 3.1
Muy malo	1.0 - 2.5

Debido a la tendencia humana a evitar calificaciones perfectas, entre 4.3 y 4.5

se considera un objetivo de excelente calidad. En el extremo inferior, la calidad del audio o el video se vuelve inaceptable por debajo de un MOS de aproximadamente 3.5.

Se utilizaron para la evaluación un total de 6 clips de audio obtenidos de un hablante real, y que expresan las siguientes frases:

- Mis secretos obstáculos, mi miedo inconfesado al baile de máscaras, no se habían aminorado con el cine y sus estímulos, sino que habían crecido de un modo desagradable, y yo, pensando en Armanda, hube de hacer un esfuerzo.
- Ya tengo de ti la sospecha de que tomas el amor terriblemente en serio.
- Y, al fin y al cabo, todo lo que él quería era exactamente eso: conocer mundos nuevos.
- Descubrí a un extraordinario muchachito que me observaba gravemente. Ahí tienen el mejor retrato que más tarde logré hacer de él. La culpa no es mía, las personas mayores me desanimaron cuando sólo había aprendido a dibujar boas cerradas y boas abiertas.
- Cuando yo tenía seis años vi en el libro sobre la selva virgen: Historias vividas, una grandiosa estampa.Representaba una serpiente boa comiéndose a una fiera.
- Pido perdón a los niños por haber dedicado este libro a una persona mayor. Tengo una muy seria disculpa: esta persona mayor es el mejor amigo que tengo en el mundo.

Se seleccionan estas señales por presentar rasgos característicos del idioma español y que se deben evaluar para arribar a una opinión acerca de si el discurso producido por un modelo en cuestión cumple con el objetivo de la investigación, principalmente en la pronunciación de tildes, y palabras con ñ, además de signos de puntuación.

3.3. Creación de base de datos con voces cubanas.

¿Qué hace a un buen Dataset?

- Debe cubrir una cantidad considerable de clips cortos y largos.
- Libre de errores. Se debe eliminar cualquier archivo incorrecto o corrupto.
- Para escuchar una voz con la mayor naturalidad posible, usar las diferencias de frecuencia y tono, y signos de puntuación.

- Es necesario que el *dataset* cubra una buena parte de fonemas, difonemas y, en algunos idiomas, trifonemas. Si la cobertura de fonemas es baja, el modelo puede tener dificultades para pronunciar nuevas palabras difíciles.
- Las muestras de la base de datos deben estar lo más limpio posible, es decir, se debe limpiar de ruido y cortar los espacios de tiempo entre expresiones,donde no se hable.

Cuban Voice Dataset

La base de datos está conformada por 160 clips de audio con sus respectivas transcripciones recogidas en el archivo metadata.csv. Cada clip tiene una duración de 2 a 15 segundos.

Los clips de audio poseen formato .wav y se organizan dentro de una carpeta de nombre wavs de la siguiente forma:

/wavs
| - audio1.wav
| - audio1.wav
| - audio2.wav
| - audio3.wav
...

Las transcripciones se recogen dentro del archivo metadata.csv. Donde audio1, audio2, etc se refieren a los archivos audio1.wav, audio2.wav etc.

audio1|Esta es mi transcripción 1.
audio2|Esta es mi transcripción 2.
audio3|Esta es mi transcripción 3.
audio4|Esta es mi transcripción 4.

Se adoptará en la conformación de la base de datos con voces cubanas, la misma estructura de la base de datos en Español de *The M-AiLabs Speech Dataset*, pues algunos de los modelos que se utilizaron fueron preentrenados sobre estos conjuntos de datos. Finalmente queda la siguiente organización:

/[Nombre del Dataset]/by_book/female/[creador del dataset]/[nombre del hablante]
|/wavs
|metadata.csv

3.3.1. Procesamiento de audio

RNNoise es una biblioteca basada en una red neuronal para la eliminación de ruido en grabaciones, se utliza en este proyecto para obtener clips de audio libres de ruidos y con la frecuencia de muestreo deseada.

Se realizó un procesamiento para la eliminación de ruido en el audio del *dataset* original, y se estableció una frecuencia de muestreo de acuerdo a las necesidades de cada experimento.

3.4. Herramientas

3.4.1. Google Colab

Colaboratory, o Colab[37] para abreviar, es un producto de Google Research, que permite que cualquier persona escriba y ejecute código Python arbitrario a través del navegador y es especialmente adecuado para el aprendizaje automático, el análisis de datos y la educación. Más técnicamente, Colab es un servicio de notebook Jupyter que no requiere configuración para su uso, al tiempo que brinda acceso gratuito a los recursos informáticos, incluidas las GPU. Los recursos de Colab no están garantizados ni son ilimitados, y los límites de uso a veces fluctúan. Esta herramienta fue utilizada para el entrenamiento de modelos durante la investigación, y fue imprescindible el acceso a una suscripción de pago de Colab Pro y la compra de una gran cantidad de recursos.

3.4.2. Google Drive

Google Drive[38] es un espacio de almacenamiento que permite a los usuarios con cuenta de Google mantener archivos en la nube, y poder compartirlos entre sus distintos dispositivos. La encomienda de Drive en este trabajo fue almacenar las bases de datos utilizadas para los entrenamientos, así como los modelos y archivos de configuración e información que se generan a partir del entrenamiento de una red neuronal con las características de las DNN utilizadas en esta investigación.

3.4.3. Espeak phonemizer

Espeak[39] es un *software* de texto a voz que admite muchos idiomas y es compatible con la salida IPA por sus siglas en inglés (*International Phonetic Alphabet*(alfabeto fonético internacional). El phonemizer permite la fonemización simple de palabras y textos en muchos lenguajes.

3.5. Modificación en el código fuente de Coqui TTS

El código fuente del repositorio ocasionaba problemas al conformar ruta del archivo metadata.csv, por lo que se realizó un pequeño cambio en el método `mailabs` del archivo `formatter.py`, quedando como se muestra en el siguiente código.

```
def mailabs(root_path, meta_files=None, ignored_speakers=None):
"""Normalizes M-AI-Labs meta data files to TTS format

Args:
root_path (str): root folder of the MAILAB language folder.
meta_files (str):  list of meta files to be used in the training. If
    None, finds all the csv files
recursively. Defaults to None
"""
speaker_regex = re.compile("by_book/(male|female)/(?P<speaker_name
    >[^/]+)/")
if not meta_files:
csv_files = glob(root_path + "/**/metadata.csv", recursive=True)
else:
csv_files = meta_files

items = []
print(f"{csv_files}")
for csv_file in [csv_files]:
if os.path.isfile(csv_file):
txt_file = csv_file
else:
txt_file = os.path.join(root_path, csv_file)


folder = os.path.dirname(txt_file)
# determine speaker based on folder structure...
speaker_name_match = speaker_regex.search(txt_file)
if speaker_name_match is None:
continue
speaker_name = speaker_name_match.group("speaker_name")
# ignore speakers
if isinstance(ignored_speakers, list):
if speaker_name in ignored_speakers:
continue
print(" | > {}".format(csv_file))
with open(txt_file, "r", encoding="utf-8") as ttf:
for line in ttf:
cols = line.split("|")
if not meta_files:
wav_file = os.path.join(folder, "wavs", cols[0] + ".wav")
else:
```

```
41 wav_file = os.path.join(root_path, folder.replace("metadata.csv", ""), "
       wavs", cols[0] + ".wav")
42
43 if os.path.isfile(wav_file):
44 text = cols[1].strip()
45 items.append(
46 {"text": text, "audio_file": wav_file, "speaker_name": speaker_name, "
       root_path": root_path}
47 )
48 else:
49 # M-AI-Labs have some missing samples, so just print the warning
50 print("> File %s does not exist!" % (wav_file))
51 return items
```

Ejemplo de código 3.1: Código modificado formatter mailabs

3.6. Fine-Tuning de Tacontron-DDC

Como ya se mencionó en el capítulo anterior, el *fine-tuning* resulta una idea prometedora, pues en teoría salva tiempo y recursos.

Para el proceso de ajuste de Tacotron2 a la base de datos personalizada, se utilizó la configuración del modelo preentrenado en español sobre el *dataset* de M-AILABS.

La frecuencia de muestreo(*sample rate*) que se establece en la configuración es 16000Hz, pues el conjunto de datos de M-AILABS sobre el que se preentrenó el modelo seleccionado, se encuentra en esta misma frecuencia. Finalmente se debe utilizar un cargador de datos(*formatter*) compatible con la base de datos usada, en este caso se selecciona la variante `mailabs`, que se puede apreciar en la sección anterior.

El próximo paso es descargar el modelo Tacotron2, para luego comenzar el reentrenamiento.

```
tts - -model_name tts_models/es/mai/tacotron2-DDC - -text "Hola."

>Downloading model to /home/ubuntu/.local/share/tts/tts_models-en-ljspeech
-glow-tts
```

El reentrenamiento se llevó a cabo utilizando el GPU Premium de Google Colab, y CUDA[40][41]. Requirió una gran cantidad de memoria RAM, siendo 25GB una cantidad insuficiente, se comprobó que con un procesador de 83GB de RAM, sí podía realizarse. El reentrenamiento fue extremandamente costoso, en tiempo y en *computer units*[1], consumiendo más de 1000CU

[1]Una unidad de cómputo (CU) es la unidad de medida de los recursos consumidos por ejecuciones y compilaciones de actores.

Los resultados son extremadamente lastimosos, el modelo entrenado por unas 160 *epochs* produce una señal ruidosa, sin embargo en algún momento se puede distinguir alguna que otra sílaba. A partir de las 210 *epochs* la salida del audio producida para una oración pequeña, es una señal corrupta donde no se distingue nada, este comportamiento se mantienen invariante hasta la *epoch* 580. Se tenía previsto alcanzar las 2000 *epochs*, pero debido a que el reentrenamiento consumía demasiados recursos y visto que los resultados no eran los deseados, no se continuó.

Claramente el modelo Tacotron2 no se adapta a las necesidades de la investigación, ni a las características del *dataset* conformado, y no cumplió las expectativas de ser un reentrenamiento veloz e iba a converger rápidamente en el nuevo conjunto por estar preentrenado en español. Y es por esto que se decide cambiar a otra red neuronal.

(Vamos a evaluar con MOS) (Tablitas)

3.7. Entrenamiento de modelo VITS desde cero

Como los resultados de Tacotron2 no fueron los mejores, se optó por realizar experimentos con otros modelos, se eligió VITS por ser de los que mejores resultados arroja por encima de Tacotron2 y Glow-TTS[23].

Esta vez se entrena el modelo desde cero utilizando el conjunto de datos con voces cubanas, y la receta[42] que provee Coqui[28] para entrenar VITS. Para este entrenamiento, siguiendo ejemplos de entrenamientos anteriores, se cambia la frecuencia de muestreo del *dataset* a 22050. Por otro lado el *formatter* utilizado es la variante `mailabs`, que se encuentra en `formatters.py`

El entrenamiento se produjo utilizando el GPU Premium de Google Colab, y CUDA[40][41]. Requirió mucha memoria RAM, siendo 25GB una cantidad insuficiente, se comprobó que con un procesador de 83GB de RAM, sí podía realizarse. Transcurrión completa e ininterrumpidamente por 2000 *epochs*, resultando en que el último mejor modelo se generó en la *epoch* número 967. El entrenamiento no representó un gran costo, en tiempo y en *computer units*, demorando alrededor de 7 horas y consumiendo alrededor de 200CU.

Finalmente se obtiene un modelo que permite la emisión de sonidos comprensibles, aunque no completamente inteligibles, pues produce un discurso robótico y tiene dificultad en la combinación de difonos, por lo que hay frases y palabras indescifrables para el oyente.

(Vamos a evaluar con MOS) (Tablitas)

3.8. Fine-tuning de modelos VITS preentrenados

3.8.1. Modelo preentrenado en italiano

El modelo disponible de Coqui en idioma italiano fue preentrenado sobre la base de datos en italiano de M-AILABS, cuyas grabaciones poseen una frecuencia de muestreo de 16000Hz, por tanto el *dataset* de voces cubanas que se utiliza para el *fine-tuning* fue llevado a la misma frecuencia. El *formatter* utilizado es igualmente la variante `mailabs`.

El proceso de *fine-tuning* se llevó a cabo utilizando el GPU Premium de Google Colab, y CUDA[40][41]. Requirió una gran cantidad de memoria RAM, siendo 25GB una cantidad insuficiente, no se precisa exactamente la cantidad de RAM necesaria, sin embargo, se comprobó que con un procesador de 83GB de RAM, sí podía realizarse sin problemas. Además de esto, no representó un gran costo, en tiempo y en *computer units*, demorando alrededor de 3 horas y consumiendo alrededor de 40CU.

El modelo que se genera luego de un reentrenamiento ininterrumpido durante 1000 *epochs*, arroja como resultado que, para una frase escrita dada, produce un discurso bastante comprensible, aunque un poco robótico, y entrecortado en alguna partes. Además con palabras que el oyente no puede descifrar. Es importante destacar que el modelo original de Coqui en italiano produce también una voz ruidosa, así que la cuestión del ruido es probable que venga desde el modelo inicial, agravada con el *fine-tuning* a partir del *Cuban Voice Dataset*.

(Vamos a evaluar con MOS) (Tablitas)

3.8.2. Modelo preentrenado en Inglés

La variante del modelo VITS entrenada sobre el conjunto LJ-Speech Dataset en inglés, se seleccionó por ser el inglés un idioma, más distante del español que el italiano. Se lleva a cabo el mismo proceso que en el caso anterior, con la diferencia de que la base de datos *Cuban Voice Dataset* cambia su frecuencia de muestreo a 22050Hz. El reentrenamiento se realizó siguiendo las mismas características que en el modelo en italiano, y consumió alrededor del mismo tiempo y recursos.

Por último el modelo en inglés ajustado a la base de datos cubana arroja resultados diferentes al modelo que se obtiene a partir del modelo italiano. Un aspecto a favor es que el ruido, y la pronunciación robótica no están presentes en el nuevo discurso, y la voz sintética suena bastante parecida a la del hablante, un objetivo que hasta este punto no había sido alcanzado. Sin embargo, y como era de esperar, gramatical y fonéticamente presenta más problemas, entre ellos la pronunciación de la ñ y las r unidas a vocales, entre otros bastante evidentes al escuchar la salida de audio.

(Vamos a evaluar con MOS) (Tablitas)

3.9. Entrenamiento con M-AILABS DATASET

Se comienza a sospechar, que probablemente la base de datos *Cuban Voice Dataset* no cuente con la riqueza necesaria para que un modelo realice un aprendizaje adecuado que reporte resultados aceptables. Debido a esto se considera la idea de realizar un entrenamiento desde cero sobre modelo VITS utilizando el conjunto de datos de *The M-AILABS Dataset* con su única voz femenina Angelina, este conjunto cuenta con más de 7000 clips de audio para el entrenamiento, con un *sample rate* de 16000Hz.

El entrenamiento se realiza nuevamente en Google Colab, con las mismas características de los anteriores, aunque, esta vez por la densidad de la base de datos el proceso demora más, tomando alrededor de 12 horas para llegar a 321 *epochs*, y consumiendo una cantidad proporcional de recursos. El modelo que se obtiene produce una señal inteligible y libre de ruidos, aunque mejorable en algunas expresiones, pues la meta ideal sería un entrenamiento de 1000 *epochs*.

Con un modelo que arroja resultados bastante buenos, se procede a realizar *fine-tuning* sobre la base de datos construida con un hablante cubano, se efectua con misma configuración que el entrenamiento anterior, por alrededor de 1000 *epochs*, y arroja los mejores resultados obtenidos sobre este conjunto de datos. Sin embargo, tampoco se consideran buenos, pues aunque se elimina el ruido en las grabaciones, muchas palabras resultan indescifrables para el oyente.

Para concluir se puede declarar que Tacotron2 parece ser una red demasiado grande que no se adapta a las necesidades de la investigación, produciendo finalmente señales corruptas e indescifrables.

El modelo VITS es significativamente más rápido en lo que respecta a la velocidad del entrenamiento, y además más adaptable a nuevos conjuntos de datos. Con un entrenamiento desde cero sobre una base de datos relativamente pequeña y rústica, produce un discurso bastante malo, aunque entendible en ocasiones. De forma similar sucede con el modelo VITS preentrenado en italiano y luego sometido a un proceso de *fine-tuning* con el conjunto mencionado anteriormente, a pesar de la rapidez del entrenamiento, produce una señal algo ruidosa, y solo entendible en ocasiones. El mejor resultado con diferencia es el producido por el modelo VITS entrenado desde cero con una base de datos fuerte, aunque este no cumple con el objetivo de la investigación pues la voz corresponde a un hablante mexicano. El discurso obtenido de este modelo recibe la mayor medida MOS. Al mismo tiempo al realizar *fine-tuning* sobre este modelo con la base de datos cubana, el audio empeora y resulta incomprensible en ocasiones. Debe ser aclarado que este último modelo es el mejor resultado obtenido sobre el conjunto de datos conformado con voces cubanas, con la obtención de grabaciones libres de ruido.

Conclusiones

-Ninguna de las ideas implementadas fue completamente satisfactoria
-El mejor resultado fue el de finetuning a angelina
- Es necesario un buen poder de computo

Recomendaciones

Recomendaciones

Bibliografía

[1] J. C. Tordera Yllescas, *Lingüıéstica computacional. Tecnologıéas del habla*, 2011 (vid. pág. 2).

[2] E. Rodrıéguez León, «Sintetizador de vocales sostenidas,» Tesis doct., Universidad Central"Marta Abreu"de Las Villas, 2013 (vid. pág. 2).

[3] A. Hernández Araujo y A. H. Araujo, «Sıéntesis de voz esofágica,» 2018 (vid. pág. 2).

[4] M. Beutnagel, A. Conkie y A. K. Syrdal, «Diphone synthesis using unit selection,» en *The Third ESCA/COCOSDA Workshop (ETRW) on Speech Synthesis*, 1998 (vid. págs. 2, 3).

[5] G. Al-Said y M. Abdallah, «An Arabic text-to-speech system based on artificial neural networks,» *Journal of Computer Science*, vol. 5, n.º 3, pág. 207, 2009 (vid. pág. 2).

[6] M. A. Guzmán Arreola, «Sintetizador de voz para la enseñanza de la lectura a niños mexicanos,» 2004 (vid. pág. 3).

[7] Y. Wang y col., «Tacotron: Towards end-to-end speech synthesis,» *arXiv preprint arXiv:1703.10135*, 2017 (vid. págs. 3, 6).

[8] Y. Ren y col., «Fastspeech: Fast, robust and controllable text to speech,» *Advances in Neural Information Processing Systems*, vol. 32, 2019 (vid. págs. 3, 10).

[9] A. Łańcucki, «Fastpitch: Parallel text-to-speech with pitch prediction,» en *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2021, págs. 6588-6592 (vid. págs. 3, 11).

[10] J. Shen y col., «Natural tts synthesis by conditioning wavenet on mel spectrogram predictions,» en *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, IEEE, 2018, págs. 4779-4783 (vid. pág. 7).

[11] J. K. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho e Y. Bengio, «Attention-based models for speech recognition,» *Advances in neural information processing systems*, vol. 28, 2015 (vid. pág. 8).

[12] D. Bahdanau, K. Cho e Y. Bengio, «Neural machine translation by jointly learning to align and translate,» *arXiv preprint arXiv:1409.0473*, 2014 (vid. pág. 8).

[13] Baidu. «Neural Voice Cloning with a Few Samples.» (), dirección: `http://research.baidu.com/Blog/index-view?id=81` (vid. pág. 8).

[14] S. Arik, J. Chen, K. Peng, W. Ping e Y. Zhou, «Neural voice cloning with a few samples,» *Advances in neural information processing systems*, vol. 31, 2018 (vid. pág. 8).

[15] Baidu. «Neural Voice Cloning with a Few Samples.» (), dirección: `http://research.baidu.com/Blog/index-view?id=81` (vid. pág. 8).

[16] J. Kim, S. Kim, J. Kong y S. Yoon, «Glow-tts: A generative flow for text-to-speech via monotonic alignment search,» *Advances in Neural Information Processing Systems*, vol. 33, págs. 8067-8077, 2020 (vid. pág. 11).

[17] D. Griffin y J. Lim, «Signal estimation from modified short-time Fourier transform,» *IEEE Transactions on acoustics, speech, and signal processing*, vol. 32, n.º 2, págs. 236-243, 1984 (vid. pág. 11).

[18] H. Kawahara, I. Masuda-Katsuse y A. De Cheveigne, «Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds,» *Speech communication*, vol. 27, n.º 3-4, págs. 187-207, 1999 (vid. pág. 11).

[19] M. Morise, F. Yokomori y K. Ozawa, «WORLD: a vocoder-based high-quality speech synthesis system for real-time applications,» *IEICE TRANSACTIONS on Information and Systems*, vol. 99, n.º 7, págs. 1877-1884, 2016 (vid. pág. 11).

[20] J. Kong, J. Kim y J. Bae, «Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis,» *Advances in Neural Information Processing Systems*, vol. 33, págs. 17 022-17 033, 2020 (vid. págs. 12, 13).

[21] W. Jang, D. Lim, J. Yoon, B. Kim y J. Kim, «UnivNet: A neural vocoder with multi-resolution spectrogram discriminators for high-fidelity waveform generation,» *arXiv preprint arXiv:2106.07889*, 2021 (vid. pág. 12).

[22] N. Chen, Y. Zhang, H. Zen, R. J. Weiss, M. Norouzi y W. Chan, «WaveGrad: Estimating gradients for waveform generation,» *arXiv preprint arXiv:2009.00713*, 2020 (vid. pág. 12).

[23] J. Kim, J. Kong y J. Son, «Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech,» en *International Conference on Machine Learning*, PMLR, 2021, págs. 5530-5540 (vid. págs. 13, 28).

[24] L. Dinh, J. Sohl-Dickstein y S. Bengio, «Density estimation using real nvp,» *arXiv preprint arXiv:1605.08803*, 2016 (vid. pág. 13).

[25] C. Durkan, A. Bekasov, I. Murray y G. Papamakarios, «Neural spline flows,» *Advances in neural information processing systems*, vol. 32, 2019 (vid. pág. 14).

[26] LJ-Speech. «LJ-Speech Dataset.» (), dirección: https://keithito.com/LJ-Speech-Dataset/ (vid. pág. 14).

[27] Mozilla. «Mozilla TTS.» (), dirección: https://github.com/mozilla/TTS (vid. pág. 14).

[28] Coqui. «Coqui TTS.» (), dirección: https://github.com/coqui-ai/TTS (vid. págs. 14, 15, 17, 22, 28).

[29] I. Solak. «The M-AiLabs Speech Dataset.» (), dirección: https://www.caito.de/2019/01/03/the-m-ailabs-speech-dataset/ (vid. pág. 16).

[30] C. TTS. «Formatting Your Dataset.» (), dirección: https://tts.readthedocs.io/en/latest/formatting_your_dataset.html#formatting-your-dataset (vid. pág. 17).

[31] WAV. «WAV Audio Format.» (), dirección: https://docs.fileformat.com/audio/wav/ (vid. pág. 17).

[32] IBM. «IBM.» (), dirección: https://www.ibm.com/ (vid. pág. 17).

[33] Microsoft. «Microsoft.» (), dirección: https://www.microsoft.com/en-us/ (vid. pág. 17).

[34] «Idioma Español.» (), dirección: https://www.ecured.cu/Idioma_espa%C3%B1ol (vid. pág. 18).

[35] «Mean Opinion Score.» (), dirección: https://telecom.altanai.com/2018/04/17/voip-call-metric-monitoring/ (vid. págs. 20, 22).

[36] «MOS Technology Brief - Mean Opinion Score Algorithms for Speech Quality Evaluation.» (), dirección: https://www.tek.com/en/documents/whitepaper/mos-technology-brief-mean-opinion-score-algorithms-speech-quality-evaluation (vid. pág. 22).

[37] Google. «Google Colab.» (), dirección: https://research.google.com/colaboratory/faq.html (vid. pág. 25).

[38] Google. «Google Drive.» (), dirección: https://www.google.com/drive/ (vid. pág. 25).

[39] Espeak. «Espeak phonemizer.» (), dirección: https://github.com/bootphon/phonemizer (vid. pág. 25).

[40] «Pytoch CUDA.» (), dirección: https://pytorch.org/docs/stable/notes/cuda.html (vid. págs. 27-29).

[41] «Pytoch CUDA.» (), dirección: https://pytorch.org/docs/stable/cuda.html (vid. págs. 27-29).

[42] C. TTS. «Coqui Recipe to Train VITS.» (), dirección: https://github.com/coqui-ai/TTS/blob/dev/recipes/ljspeech/vits_tts/train_vits.py (vid. pág. 28).