PROTOCOLOS Y HERRAMIENTAS PARA LA CONSTRUCCIÓN DE UNA SÍNTESIS DE VOZ.

Luis Miguel Marulanda Torres

UNIVERSIDAD TECNOLÓGICA DE PEREIRA

FACULTAD DE INGENEIRÍAS

INGENIERÍA DE SISTEMAS Y COMPUTACIÓN

2020

PROTOCOLOS Y HERRAMIENTAS PARA LA CONSTRUCCIÓN DE UNA SÍNTESIS DE VOZ.

Luis Miguel Marulanda Torres

Director de Proyecto

Saulo De Jesús Torres Rengifo, Dr. En Informática.

Trabajo para obtener el título de Ingeniero en Sistemas y Computación.

UNIVERSIDAD TECNOLÓGICA DE PEREIRA

FACULTAD DE INGENEIRÍAS

INGENIERÍA DE SISTEMAS Y COMPUTACIÓN

PEREIRA 2020

Este trabajo es una demostración de todo lo que puedo lograr impulsado por el amor de mis padres, hermanos, familiares y amigos, que siempre me han acompañado de manera incondicional durante mi trayectoria académica.

Agradecimientos

Agradezco primeramente a Dios por permitirme conocer la informática y despertar en mí el deseo por obtener nuevos conocimientos.

Agradezco a mis padres, hermanos y familiares, quienes han hecho posible culminar de manera exitosa otra etapa de mi vida llena de experiencias inolvidables.

También, doy las gracias al Doctor Saulo De Jesús Torres Rengifo, quien ha estado acompañando todo este proceso, haciendo posible la realización de este proyecto gracias a su experiencia en el tema.

Finalmente, agradezco a la Universidad por brindarme un lugar para forjar amistades con personas que llegaron a mi vida como compañeros de clase o docentes.

TABLA DE CONTENIDO

Índice de Ilustraciones

Índice de Tablas

Resumen

Este proyecto estudia los diferentes métodos para la construcción de una síntesis de voz, los cuales utilizan herramientas diferentes entre sí y permiten obtener resultados con distintos niveles de calidad.

La primera parte muestra de manera general cómo se ha formulado el proyecto, de dónde ha surgido la idea y qué hipótesis o preguntas pretenden demostrarse a través de unos objetivos y metodología establecidos.

La segunda parte estudia el estado del arte, es decir, cómo se encuentran los sistemas de síntesis de voz, cuál ha sido su trayectoria y cuáles son los conceptos, teorías y antecedentes que complementan este tema, además, de los requisitos preexistentes para crear una síntesis de voz.

La tercera parte describe cómo se construye una síntesis de voz utilizando tres herramientas diferentes, las cuales se han escogido aleatoriamente para extraer de estas sus generalidades, ventajas y desventajas asociadas.

La cuarta parte del documento expone a través de las conclusiones los hallazgos frente a la pregunta de investigación e hipótesis planteadas.

Finalmente, se expone la bibliografía que ha complementado este proyecto y los anexos que complementan la temática.

Abstract

This project studies the different methods for the construction of a voice synthesis, which use different tools from each other and allow obtaining results with different levels of quality.

The first part shows in a general way how the project has been formulated, where the idea has come from and what hypotheses or questions are intended to be demonstrated through established objectives and methodology.

The second part studies the state of the art, that is, how the voice synthesis systems are, what their trajectory has been and what are the concepts, theories and antecedents that complement this topic, in addition to the pre-existing requirements to create a speech synthesis.

The third part describes how a speech synthesis is constructed using three different tools, which have been chosen randomly to extract from them their generalities, advantages, and associated disadvantages.

The fourth part of the document exposes through the conclusions the findings against the research question and hypotheses.

Finally, the bibliography that has complemented this project and the annexes that complement the subject are exposed.

PARTE I INTRODUCCIÓN A LA INVESTIGACIÓN

1. Capítulo 1 Introducción

1.1 Introducción

En la actualidad es posible identificar algunas herramientas de síntesis de voz que han permitido establecer la interacción entre usuarios con alguna limitación visual y las computadoras, ya que estas síntesis a través de su capacidad de convertir el texto en voz son aquellas mediadoras entre programas como lectores de pantalla y los usuarios. Sin embargo, crear una síntesis de voz sigue siendo un proceso complejo desde diferentes puntos de vista.

Mediante este proyecto investigativo, se pretende estudiar de manera avanzada los elementos fundamentales y los saberes necesarios en materia de física y computación para la construcción de una síntesis de voz, entendiendo esta construcción como una secuencia de pasos enmarcada en el conocimiento y dominio básico de herramientas para Text To Speech y lenguajes de programación ya existentes.

Es importante mencionar, que mediante la comprensión de esta investigación, deberá resultar posible la construcción de una síntesis de voz utilizando recursos básicos para la grabación, como un micrófono doméstico y la voz propia, con el fin de ahorrar en licencias de alto costo para la adquisición de voces que puedan resultar más agradables para determinado usuario, teniendo en cuenta que hoy en día los sintetizadores de voz le resultan interesantes no solo a los limitados visuales sino también a otros usuarios como aquellos que prefieren llevar un proceso de lectura a través de audiolibros.

1.2 Antecedentes de la idea

A través de la historia, se ha observado cómo la humanidad ha tratado de generar soluciones para los diferentes problemas que han ido surgiendo en los diferentes campos, como por ejemplo en el caso de las comunicaciones interpersonales, en el cual cabe mencionar el caso del científico

Stephen Hawking quien ha sido una de las personas más famosas en emplear sintetizadores de voz para comunicarse como solución a su incapacidad para hacerlo por sí mismo.

En materia de accesibilidad, las diferentes síntesis de voz ya construidas y que se encuentran en el mercado, han representado una solución para las comunidades de limitados visuales de todo el mundo, ya que mediante programas computacionales y estas síntesis de distintos protocolos e idiomas, han podido escuchar el contenido en sus pantallas mediante los lectores de pantalla con los cuales se puede escoger entre diferentes voces.

Sin embargo, a pesar de que los sintetizadores de voz han representado una solución en distintas necesidades de las personas, e incluso de las empresas como las de telecomunicaciones (en el caso de las contestadoras automáticas o conmutadores), los sintetizadores de voz también pueden representar una herramienta cuya construcción es difícil y requiere de conocimiento avanzado desde los diferentes puntos de vista, resultaría entonces grato contar con la posibilidad de construir una síntesis de voz desde un punto de vista sencillo y sin tanto tecnicismo para los diferentes lectores en las diversas áreas del conocimiento.

1.3 Planteamiento del problema

La síntesis de voz es un proceso a través del cual se pretende que una máquina genere habla o que simule el habla del humano partiendo desde un texto que se quiera escuchar y de una voz previamente grabada por alguien o creada a partir de un programa computacional. Este término no tiene una definición formal, pero a nivel de informática se ha ido popularizando gracias a diferentes trabajos de investigación y adelantos en el tema, sin embargo, también han surgido problemas como la dificultad que conlleva el proceso de creación de estas.

En el tiempo que ha transcurrido de este siglo, algunos proyectos como Festival (TTS) en conjunto con las Speech Tools del centro de Investigación de Tecnologías del Lenguaje de la universidad de Edimburgo han avanzado en el tema de la síntesis de voz, sin embargo, de manera general, se pueden identificar causas en la dificultad del proceso de creación de síntesis de voz, tales como el nivel de tecnicismo requerido para dominar de manera básica el tema, los problemas de adaptación que tiene el software preexistente (como en el caso de festival) o la alta demanda de

recursos económicos, ya que es costoso contar con un estudio de grabación, o difícil encontrar voces con las que sea posible la creación de toda una síntesis completa, sin olvidar, el tiempo requerido para todo este proceso, si como resultado final se quiere una síntesis de un lenguaje natural completo. (Soledad, s.f.)

Por otra parte, es de reconocer, que la accesibilidad a nivel de informática se ha convertido en una parte fundamental para el desarrollo de los nuevos programas computacionales (Astelehena, 2009) y sin duda alguna, la síntesis de voz ha sido parte indispensable en el proceso de adaptar el contenido para el caso de los limitados visuales, permitiendo reemplazar de manera casi completa un periférico tan importante como lo es el monitor. Habiendo entonces obviado el aporte que realiza la síntesis de voz en el apartado de la accesibilidad, existen consecuencias debido a la complejidad de la construcción de estas, como, por ejemplo, que las personas no estén dispuestas a crear una síntesis con su propia voz, o a motivar a otros para la creación de estas, además de los altos costos en las voces ya existentes para los diferentes programas como los lectores de voz.

Ya que la accesibilidad supone un derecho que otorga a un individuo la posibilidad concreta y real de entrar, permanecer y recorrer un lugar (en el caso de este trabajo un aplicativo o sitio de internet) de manera concreta, real y con la mayor autonomía posible (Definición.de, s.f.). Resulta entonces de gran importancia tratar de demostrar que es posible romper la creencia de que es muy difícil crear una síntesis de voz con las herramientas existentes hoy en día.

1.4 Pregunta de Investigación

¿Cómo es el proceso de construcción de un conversor de texto a voz (CTV, Text To Speech o TTS) utilizando la propia voz o una personalización de esta?

1.5 Hipótesis

¿Es posible que, llevando a cabo un estudio del estado del arte de las herramientas y protocolos existentes para la construcción de un sintetizador de texto a voz, se pueda establecer un modelo a seguir que integre estas herramientas para la construcción de una síntesis de voz personalizada?

1.5 Objetivo General

Analizar desde el punto de vista informático el proceso de construcción de una síntesis de voz mediante la utilización de software libre, lenguajes de programación y programas de Texto a Voz preexistentes.

1.6 Objetivos específicos

1. Definición de los conceptos fundamentales para la construcción de un sistema de síntesis de voz.
2. Estudiar los diferentes sintetizadores y protocolos desde el punto de vista teórico y de antecedentes para la implementación o creación de una síntesis de voz.
3. Identificar los requerimientos de hardware o software en el proceso de construcción de una síntesis de voz.
4. Analizar procesos de construcción del sintetizador de voz a través de herramientas que faciliten este proceso.
5. Determinar de manera concisa la dificultad en todo este proceso y las posibilidades de reducción de esta.

1.7 Metodología

El proyecto contará con varias etapas, las cuales se complementarán entre sí y harán posible una culminación exitosa y con resultados contundentes:

1. Inicialmente se recopilarán los artículos y material de interés para la profundización en el tema y la construcción del estado del arte.
2. Posteriormente, será estudiada de manera crítica toda la información que previamente se haya recopilado, con el fin de seleccionar los contenidos más pertinentes a la temática desarrollada.
3. Se procederá con la construcción del estado del arte en manera detallada, basándose en la información que haya sido seleccionada en la etapa anterior.
4. Una vez finalizado el estado del arte, será posible comenzar a identificar los protocolos adecuados de la síntesis de voz e instalar las herramientas necesarias para la realización de diferentes pruebas a través de diferentes programas.
5. Será concluida la investigación a través de la elaboración de las conclusiones analizando el cumplimiento de los objetivos previamente establecidos.

PARTE II ESTADO DEL ARTE

2. Capítulo 2 Los Sistemas Text To Speech

2.1 Componente Histórico

"Podría considerarse que Las voces electrónicas "sintéticas", son comunes ahora, pero el camino a la síntesis del habla está lleno de restos de dispositivos que prometieron traernos la voz del futuro, pero no duraron más allá de su valor de novedad". (GrundHauser, 2017)

A pesar de ser un tema con bastante utilidad, es difícil identificar los orígenes de la síntesis de voz. Algunos autores, como Alan Black y Kevin Lenzo del Lenguaje Technologies Institute, Carnagie Mellon University, son enfáticos en que la síntesis de voz es un tema con el que la humanidad ha tratado hace menos de un siglo, al cual en ocasiones se le desconocen la cantidad de procesos que abarca.

"El primer intento de crear una máquina que podría generar sonidos de voz similares a los humanos se atribuye típicamente a Wolfgang Von Kempelen en la segunda mitad del siglo XVIII" (College of Health & Rehabilitation Sciences: Sargent College). Von Kempelen logró fabricar un dispositivo mecánico que se trataba de un fuelle con el que forzaba el aire a través de una caña en una cámara, cuyas propiedades resonantes servirían para ser manipuladas utilizando las manos para producir sonidos situados en la gama de frecuencias propias de la voz.

En su obra, Black y Lenzo afirman que una de las primeras aplicaciones de la síntesis de voz, se dio en la década de los años 30 del siglo XX, donde la compañía de teléfonos del Reino Unido logró introducir una línea con reloj de oratoria, utilizando dispositivos rudimentarios, almacenaban las palabras, frases y sílabas previamente grabadas, las cuales concatenaban para obtener de manera adecuada oraciones completas.

Por otra parte, en varios artículos se habla acerca del "Voder", desarrollado por Homer Dudley en Bell Laboratories. Este dispositivo de antaño funcionaba de manera mecánica mediante pedales y llaves como si se tratara de un órgano y a través de un correcto manejo de este, se podrían generar sonidos que con una técnica adecuada terminaban sonando casi como el habla.

El Voder representó un gran cambio para la ciencia del habla y los conocimientos sobre esta, ya que, hasta antes de la aparición de este dispositivo, era posible realizar una síntesis de voz, pero a través de grabaciones previamente realizadas, como en el caso del sistema de reloj introducido por la compañía de teléfonos del Reino unido en los años 30 del siglo XX. Por su parte, el Voder permitió crear el habla apartir de un único instrumento y su operación humana, partiendo de no tener prácticamente nada, como está descrito en las noticias científicas del Smithsonian de enero de 1939. A pesar de los alcances que tenía el Voder, existió un periodo de tiempo luego de su aparición en el que la síntesis de voz no era concatenaria.

En la obra *Building Synthetic Voices* se confirma que fue en la década de los 70 cuando inició el auge de la síntesis de voz tal y como la conocemos hoy en día, sin embargo, en aquel entonces limitada por las restricciones de memoria, algo que con el pasar del tiempo se ha ido solventando gracias a que el almacenamiento electrónico se ha vuelto "barato" y robusto. Es destacable que para 1972 Unix en su tercera generación ya incluía en su manual estándar algunos comandos para procesar de texto a voz, hacer análisis de texto, predicción prósodica, generación de fonemas y síntesis de formas de onda a través de hardware especializado (Black & Lenzo, 2014), lo cual a pesar de representar los avances más significativos para la síntesis de voz en aquella época, estaba limitado por el hardware especializado requerido y por los altos costes de almacenamiento. Desde aquel entonces ha existido la tendencia de comprimir (codificar) el habla de tal manera que se pueda utilizar en diversas applicaciones.

Gracias a la incorporación de motores simples de TTS en sistemas operativos como los proporcionados por Apple, éste tema se ha popularizado y ha generado interés para los investigadores, ya que con una producción en masa de computadoras hogareñas que incluyan motores de texto a voz, ha sido posible terminar con las restricciones que antes de los 80 existían, tales como necesitar de un laboratorio para hacer grabaciones y de un almacenamiento muy grande para las primeras generaciones de computadoras.

Finalmente, hay que resaltar que en lo que va de este siglo han aparecido nuevos sistemas de síntesis de voz, que utilizan una lógica diferente y revolucionaria en comparación a las producidas a través de TTS. Un importante investigador en el tema es Keiichi Tokuda del instituto de tecnología de Negoya, quien situó a Japón como pionero en el desarrollo de un método estadístico para la síntesis de voz, llamado HTS con el cual demostró que construir modelos generativos de

habla, en lugar de seleccionar instancias aunitarias puede generar un habla de confiable y de alta calidad.

2.2 Marco Referencial

A través de los conceptos y teorías que se exponen a continuación se pretende tener un punto de partida para este proyecto, su realización y los resultados que busca.

2.2.1 Marco Conceptual

Para tener entendimiento sobre la temática abordada en esta investigación, es adecuado comprender previamente algunos conceptos relevantes.

Síntesis de voz (TTS en algunas partes de este documento)

Entendiéndose síntesis[1] como la Composición de un todo por la reunión de sus partes y a Voz[2] como el Sonido Producido por la vibración de las cuerdas vocales, podría asociarse la síntesis de voz a un conjunto de sonidos producido por la vibración de las cuerdas vocales, sin embargo, a nivel de informática y computación este término se asocia a la producción artificial del habla, realizada a través de hardware o software simulando las pistas de audio propias de una voz humana.

Sistema para síntesis de voz

Un sistema para síntesis de voz puede ser también llamado un sintetizador de voz. Comúnmente el término sintetizador se asocia a la música y las definiciones oficiales lo describen como un "Instrumento musical electrónico capaz de producir sonidos de cualquier frecuencia e intensidad y combinarlos con armónicos, proporcionando así sonidos de cualquier instrumento conocido, o

[1] Diccionario de la lengua Española Real Academia Española https://dle.rae.es/s%C3%ADntesis
[2] Diccionario de la lengua Española Real Academia Española https://dle.rae.es/voz?m=form

efectos sonoros que no corresponden a ningún instrumento convencional"[3]. Sin embargo, un sistema para síntesis de voz es un conjunto de elementos que se reconoce como sistema al tener entradas, salidas y un proceso interno asociado a etapas.

En cuanto a entradas cuenta con: Texto.

Salidas: Voz artificial producida por la concatenación de fonemas y la formación de difonemas.

Etapas: Analizador de texto, Analizador Lingüístico y generador de formas de onda.

Fonemas

Los fonemas son la "Unidad fonológica que no puede descomponerse en unidades sucesivas menores y que es capaz de distinguir significados."[4]. De manera general, los fonemas son la articulación mínima de un sonido vocálico y consonántico (Arroyo Cantón & Berlato Rodríguez, 2012) Para esta investigación los fonemas son relevantes ya que representan el elemento atómico para todo proceso de estudio o construcción de una síntesis de voz. Los fonemas son el elemento fundamental en las bases de datos utilizadas por los sistemas de síntesis de voz.

Difonemas

Los difonemas son los segmentos acústicos que incluyen la transición entre dos fonos consecutivos, formado por la parte estacionaria del primero, la transición del primero al segundo y la parte estacionaria del segundo. (Llisterri, 2020)

Este concepto es quizás uno de los más importantes en esta investigación, ya que a través de su utilización se hace posible la construcción completa de los sintetizadores de voz desde hace muchos años atrás.

Selección de unidades

[3] Diccionario de la lengua Española Real Academia Española https://dle.rae.es/?w=sintetizador+
[4] Diccionario de la lengua Española Real Academia Española https://dle.rae.es/fonema

La selección de unidades es una técnica que emplea bases de datos de voces que se han grabado previamente. Durante las grabaciones que alimentan las bases de datos ya mencionadas, cada enunciado que es segmentado en los elementos comunes de una síntesis de voz, tales como los fonemas y difonemas.

"La premisa básica de la selección de unidades es que podemos sintetizar nuevos enunciados que suenen naturalmente seleccionando unidades de subpalabras apropiadas de una base de datos de habla natural" (Black A. W., 2002)

Analizador de texto

El analizador de texto es un elemento fundamental en el protocolo TTS, el cual toma como entrada cualquier texto y le da el formato adecuado para que sea entendible al módulo que convertirá de texto a fonemas (o difonemas) dependiendo la metodología utilizada para la construcción de la síntesis de voz. El objetivo que pretende este importante elemento es obtener un formato en el que la secuencia de palabras esté libre de ruido, además de asignar los elementos como pausas entre las diferentes frases.

Sintetizador de Dominio Limitado

Un sintetizador de dominio limitado es aquel en el que solo se tiene en cuenta una parte fundamental del lenguaje o conjunto de expresiones en el que se generará la síntesis.

"El objetivo de construir un sintetizador de dominio limitado no es solo construir una buena síntesis. Se sigue esta ruta ya que es un método muy práctico y rápido para construir sistemas de texto a voz." (Black & Lenzo, 2014)

A pesar de que se trata el tema de la síntesis de voz partiendo de querer generar una síntesis completa, es importante resaltar este término, ya que en la mayoría de los casos solo es posible construir síntesis de dominio limitado, por motivos como la falta de instrumentos que permitan grabaciones de calidad avanzada, o herramientas para alinear los elementos como fonemas o difonemas.

2.2.2 Marco de Antecedentes

En el pasado diversos proyectos y herramientas han hecho grandes aportes en el tema de síntesis de voz. Es importante resaltar algunos a través de los cuales se hace posible profundizar en el tema y hacer descubrimientos significativos:

- Festival: Ofrece un marco general para construir sistemas de síntesis de voz, así como también incluye ejemplos de varios módulos. En general, ofrece la posibilidad de texto completo a voz a través de un número de API: desde el nivel de Shell, a través de un intérprete de comandos de Scheme, como una biblioteca C++, desde Java y una interfaz Emacs.[5]

  Festival fue el principal punto de partida para este documento, ya que apartir de aquí empezó todo el proceso de comprensión de lo que es una síntesis de voz y se despertó la curiosidad por saber la manera en cómo se construye.

- FestVox: Este proyecto es parte del trabajo en el grupo de oratoria de la Universidad Carnegie Mellon destinado a avanzar en el estado de la Síntesis del Habla.[6]

  A través de este proyecto se hace posible la construcción de las síntesis de voz que se utilizan en el programa Festival. El proyecto incluye todas las herramientas necesarias para la construcción de un sintetizador de voz, y a pesar de que su documentación está toda en inglés pretende que se construya una síntesis de voz natural, flexible y eficiente como la voz humana sin importar el idioma y a partir de pocos recursos computacionales.

  Un importante participante del proyecto festvox es el profesor Alan W Black, quien aparece como autor y coautor de gran parte del material utilizado para la construcción de este documento.

- SpeechTools: Es el conjunto de herramientas complementarias al software Festival a través de las cuales se hace posible compilar el léxico para construir una voz. Todos los elementos en speech tools están comúnmente compilados y optimizados para ejecutarse correctamente en sistemas UNIX y Windows (Microsoft).

[5] The Festival Speech Synthesis System: http://www.cstr.ed.ac.uk/projects/festival/
[6] FestVox: http://festvox.org/index.html

Gran parte del código fuente de las speech tools está construido en lenguaje C y C++, y tienen alta dependencia de librerías que podrían aparecer a lo largo de esta investigación.

- HTK: HTK consta de un conjunto de módulos de biblioteca y herramientas disponibles en forma de fuente C. Las herramientas proporcionan instalaciones sofisticadas para el análisis del habla, capacitación HMM, pruebas y análisis de resultados. El software es compatible con HMM que utilizan gaussianos de mezcla de densidad continua y distribuciones discretas y se puede utilizar para construir sistemas HMM complejos.[7]

- Prosodylab Aligner: A través de ProsodyLab aligner se hace la alineación forzada en el proceso de construcción de una síntesis de voz.

  La alineación forzada puede considerarse como el proceso de encontrar los momentos en que los sonidos y las palabras individuales aparecen en una grabación de audio con la restricción de que las palabras en la grabación siguen el mismo orden que aparecen en la transcripción.

  Resulta fundamental resaltar que HTK y ProsodyLab Aligner son ejemplos prácticos de herramientas que se utilizarán mientras se desarrolla este documento debido a su simplicidad y a las posibilidades que ofrecen cuando se tienen recursos limitados, sin embargo, existen otras herramientas similares con procesos más refinados.

  Los elementos presentados anteriormente cuentan con licencias, con documentación asociada y con dependencia a librerías o sistemas operativos.

2.2.3 Marco Teórico

Tras el estudio de la historia y los antecedentes de la síntesis de voz, resulta evidente que la evolución de la computadora, su producción en masa, las nuevas tecnologías y el interés general de muchos autores han hecho posible la mejoría de las técnicas y protocolos que se utilizan para

[7] HTK: http://htk.eng.cam.ac.uk/

la síntesis de voz. A continuación, serán expuestas algunas de las teorías más significativas que se construyen desde los autores que han incursionado recientemente en la temática:

Sintetizadores de voz

En los últimos 20 años se han podido ver ejemplos prácticos de sistemas de texto a voz que pueden decir cualquier texto que reciben, aunque sea de manera incorrecta (Black & Lenzo, 2014). Estos sintetizadores se han ido volviendo parte del día a día después de la invención de máquinas que pudieran generar habla, pudieran realizar lecturas o incluso entablar conversaciones mediadas inteligencia artificial, dejando en claro que "la inteligencia artificial habla, pero no entiende lo que dice". (Maroti, 2019)

Para este estudio, será tenida en cuenta la implementación de la síntesis de voz desde el Software, donde "las aplicaciones no interactúan directamente con los componentes de audio de una computadora, sino que son abstraídas del Hardware subyacente a través del motor de habla" (Morales, 2013) ya que esta es la manera en cómo se hace posible que los protocolos y herramientas utilizados sean funcionales en diferentes plataformas o sistemas.

Gran parte de la teoría en esta temática apunta a que el hombre persigue el interés de comunicarse con las computadoras y la síntesis de voz en todo su proceso evolutivo ha abierto bastantes posibilidades para la humanidad.

3. Capítulo 3 Los requisitos para la construcción de una síntesis de voz

La síntesis de voz en la actualidad se puede construir de diferentes maneras, entre las cuales se destacan las siguientes para esta investigación:

1. Mediante la utilización del conjunto de herramientas a las que se les denomina framework.
2. Mediante la utilización de determinados lenguajes de programación y sus librerías asociadas.
3. Mediante la utilización de la inteligencia artificial ofrecida por aplicaciones en la web.

Como ya se ha especificado en capítulos anteriores, es posible identificar que hay dos partes fundamentales para la construcción de una síntesis de voz, una de ellas es el hardware y la otra es el software. En cuanto al hardware, existen dispositivos como el Voder con funcionamiento similar al de un instrumento musical como el órgano o algunos elementos más modernos como Arduino y algunos de sus módulos. Aunque estos dispositivos son importantes ya que representan las bases del conocimiento en este tema, es necesario dejarlos en segundo plano ya que el enfoque de este trabajo es el informático.

Dejando claro que sin el hardware adecuado no es posible trabajar esta temática, hay que recalcar que el software cobra protagonismo y es la principal herramienta ya que es el recurso al que se puede acceder de manera más sencilla y favorable en vista de la ausencia de recursos materiales o humanos para las grabaciones de voz necesarias para construir una síntesis.

Los requisitos que se exponen a continuación se han seleccionado en base a los que exponen los diferentes autores que han trabajado en la construcción de sintetizadores de voz en la historia más reciente, por lo anterior, para metodologías de trabajo diferentes, podrían aparecer otros requisitos o incluso menos.

La importancia de cumplir con estos requisitos radica en que el proceso de construcción de un sintetizador de voz es susceptible a muchas fallas y contar con medios de mala calidad puede aumentar las probabilidades de error o la obtención de resultados negativos y alejados del producto final deseado o las metas por cumplir.

3.1 Los Requisitos de Hardware

Entendiéndose el hardware como el conjunto de elementos físicos que componen un sistema informático, existen algunos elementos específicos que son necesarios para realizar las grabaciones de pistas de audio y extraer la información que es necesaria para el funcionamiento adecuado de un sintetizador de voz TTS.

Sin importar la metodología a utilizar para construir un sintetizador de voz utilizando el software como herramienta principal, son necesarios algunos recursos. A continuación, son mencionados algunos de estos recursos:

- Procesador de velocidad razonable: Es difícil establecer una velocidad base a la que debería funcionar un procesador para elaborar un sintetizador de voz, sin embargo, dependiendo la profundización que se desee hacer en el proceso de construcción de este, debería ser posible acceder a uno con buena velocidad. En el proceso de elaboración mediada por el framework Festival y las herramientas FestVox y Speech Tools, los autores Black y Lenzo afirman que "Muchas de las técnicas descritas requieren una buena cantidad de tiempo de procesamiento para lograrlo, aunque las máquinas se están volviendo cada vez más rápidas y esto se está convirtiendo en un problema menor" (Black & Lenzo, 2014). Afirmaciones así, demuestran que estas herramientas se han elaborado para algunos procesadores más limitados que los existentes en la actualidad y que las diferentes tareas que implica la construcción de un sintetizador de voz como el proceso de alineación para etiquetar de manera difusa los textos y el audio, serán tareas por realizar en menos tiempo.

- Equipo de grabación: A pesar de que el proceso de construcción de una síntesis de voz difiere según los autores, existe un interés particular por generar unos tonos de voz similares a los propios, y esto se evidencia en que autores como Pilar Soledad elaboren síntesis de voz utilizando grabaciones de su propia voz.

  En muchos casos, es posible optar por conseguir una base de datos libre que se pueda utilizar en un sintetizador de voz, pero esto no permitiría tener una síntesis personalizada y con las tonalidades deseadas. Para construir una síntesis con voz propia, lo más adecuado

es realizar las grabaciones personalmente, y para esto lo mínimamente necesario es un micrófono que se pueda conectar a una computadora. En la obra de Black y Lenzo se considera que “Un micrófono barato pegado en la parte posterior de la PC estándar no es ideal” (Black & Lenzo, 2014) pero también se reconoce que esta es la solución más común empleada por quienes desean hacer un sintetizador de voz utilizando sus propias voces, ya que existen dificultades para acceder a recursos como un estudio de grabación insonorizado, un micrófono de alta calidad o una placa de sonido con características especializadas.

Normalmente, es posible encontrar computadoras con micrófonos integrados, o con micrófonos domésticos que se han conectado ante la necesidad de entrada de grabaciones o comandos que luego serán procesados, sin embargo, este tipo de equipos van a aumentar la complejidad para la construcción de una síntesis de voz ya que no cuentan con características como la cancelación de ruidos.

- Placa de sonido: Conocida también como la tarjeta de sonido, es un dispositivo que no es totalmente necesario para el funcionamiento de una computadora, pero sí para la construcción y utilización de un sintetizador de voz. Este dispositivo puede conectarse a la tarjeta madre o estar integrado en esta.

  La placa de sonido permite reproducir contenidos sonoros como la música o la voz, pero también permite la conexión de otros periféricos de entrada y salida de audio, tales como los parlantes, auriculares, micrófonos e incluso instrumentos musicales.

  Para la construcción de una síntesis de voz, lo más adecuado es tener una tarjeta de sonido que no sea integrada. Muchas computadoras, especialmente las portátiles, cuentan con placas de sonido integradas que se conectan directamente al micrófono propio de la computadora, pero estas no proporcionan un elemento aceptable para la construcción de un sintetizador de voz, ya que, en conjunto con los micrófonos de baja calidad, terminan capturando bastante ruido exterior o generando perdida de información en las pistas que se graben y procesen.

  Por otro lado, existen algunas dificultades en materia de adaptabilidad ya que en algunos casos a pesar de contar con una placa de sonido de calidad puede existir dificultad para la conexión y aprovechamiento de esta, por la ausencia de controladores para su gestión, sin

embargo, todos estos son aspectos de software, el otro componente fundamental para los procesos de creación y gestión de un sintetizador de voz.

- Periféricos de Salida de Audio para diferentes formatos: Para la salida de audio de una computadora no es completamente necesario contar con una placa de sonido, algunas computadoras ofrecen puertos a los que podemos conectar auriculares, bocinas u otros elementos para reproducir el audio, sin embargo, para la construcción de una síntesis de voz es importante considerar que se van a tener pistas en varios formatos, los cuales solo pueden ser reproducidos en determinados dispositivos, por dar un ejemplo, en algunos trabajos donde se han construido sintetizadores de voz, se trabaja con archivos tipo wav, mientras que en otros procesos similares trabajan con pistas de tipo mp3.

  En las computadoras y sistemas modernos es posible que no exista mayor dificultad para adaptar o reproducir cualquier tipo de contenido, ya que el contenido se puede transformar para ser reproducido en otro formato, pero esto comprometería seriamente la calidad e información que puede proporcionar una pista de audio. La opción más adecuada para escuchar y determinar la calidad de las grabaciones que se realicen para la construcción del sintetizador de voz es usar auriculares, capaz de reproducir partes del sonido como los bajos, sin distorsionar las grabaciones o darles tintes agudos o graves que puedan afectar una síntesis como producto final.

  Los periféricos de salida de audio pueden ser costosos, pero es necesario encontrar alguno que se adapte bien a los formatos en los que se desee grabar y que sean compatibles con el hardware y software escogido para este proceso.

La literatura en cuanto a síntesis de voz apunta a que a pesar de que existan muchas posibilidades a la hora de grabar contenidos, la técnica de construir una síntesis de voz se perfecciona en la medida que se tengan elementos más sofisticados que permitan trabajar de manera avanzada esta temática, es decir, con cualquier micrófono doméstico es posible producir contenido y grabaciones, con cualquier placa de sonido existe la posibilidad de procesar audio y extraer la información necesaria, pero esto no implica que se obtengan grabaciones de calidad y bases de datos adecuadas para un sintetizador de voz.

En términos generales, debería existir un equilibrio entre la calidad de los implementos de hardware, ya que sería inútil contar con una placa de sonido con muchas prestaciones si se cuenta con un procesador que no pueda hacer uso completo de los recursos que ofrece la misma, o tener un micrófono “barato” que no permita producir pistas de calidad para ser procesadas, y lo mismo pasa de manera contraria, si se tiene un micrófono de alta calidad, pero no se tienen los medios para la utilización de las grabaciones realizadas, entonces el proceso de construcción de la síntesis de voz, arrojará un resultado de poca calidad.

3.2 Los Requisitos de Software

El software es el conjunto de programas que hace posible la interacción entre el humano y la computadora. Como se evidencia en el apartado histórico, a lo largo de la evolución de la síntesis de voz el software ha jugado un papel fundamental ya que es una herramienta que aporta gran valor al momento de construirlas.

Se ha establecido que el software será la herramienta principal para el componente investigativo en esta temática. Lo anterior por la facilidad que existe para acceder a este recurso, facilidad que se debe a que las diferentes herramientas a utilizar son software libre, sin embargo, los requisitos de software son diferentes según la forma por la cual se decida trabajar el proceso de construcción:

1. Utilizando el conjunto de herramientas a las que se les denomina framework.

    - Herramientas de FestVox: FestVox es el proyecto que para esta forma de construcción se reconoce como herramienta principal al ofrecer una estructura de framework conformada por el conjunto de herramientas Festival y Speech Tools de la universidad de Carnegie Mellon. Es decir, es necesario contar con la herramienta Festival en su versión 2.5 o anterior y Speech Tools en su versión 2.5 o anterior, además de FestLex 2.5 y en caso de que se necesite hacer pruebas previas al proceso de construcción, alguna voz del conjunto que ofrece la página del proyecto FestVox en el apartado Download. Estas herramientas son gratuitas y están disponible para

la descarga en sus múltiples versiones para ser adaptadas fácilmente a diferentes plataformas. La obra de los principales autores en esta temática Black y Lenzo se ha construido utilizando Festival 2.4 y Edimburgo Speech Tools 2.4.

- Sistema Operativo: Alan W. Black y Kevin A. Lenzo en su obra documentan la utilización de las herramientas Speech Tools para la construcción del sintetizador TTS y afirman que "Debido a que estamos más familiarizados con un entorno Unix, los scripts, herramientas, etc. asumen un entorno tan básico." (Black & Lenzo, 2014), especificando que los scripts necesarios se pueden ejecutar en múltiples plataformas, pero lo más recomendable es utilizar un sistema Unix ya que la mayoría de las pruebas se han realizado en Linux.

- Compilador: De manera específica es necesario contar con un compilador de C y C++.

  Una de las tareas para el proceso de instalación y utilización de las Speech Tools es la compilación de estas herramientas y de la herramienta Festival, lo cual se hace a través de algún compilador de C++. Esto no implica necesariamente saber manejar el lenguaje C o C++, o contar con un compilador específico, sin embargo, implica conocer el proceso de instalación en las plataformas que así lo requieran al no traerlos instalados o como parte del intérprete de comandos o Shell.

- Alineador: Cuando se requiere la modificación de las pistas grabadas, es necesario contar con un alineador, por ejemplo, en el caso que queden mal etiquetadas algunas pistas de audio con su respectivo texto. En materia de alineadores no es necesario alguno en específico, con alguno que sea software libre basta, con el fin de evitar problemas de licencias.

2. Utilizando determinados lenguajes de programación y sus librerías asociadas.

Para este proceso de construcción se utiliza el proceso elaborado por Pilar soledad, quien ha construido el artículo "Construyendo un sintetizador de texto-a-voz usando Python y

selección de unidades" (Soledad, s.f.) basada en la obra de Black y Lenzo, pero simplificando algunos procesos y adaptándolos a la actualidad en un lenguaje tan versátil y con una curva de aprendizaje sencilla como Python. La autora Soledad, expone los siguientes requisitos:

- Python 2 y 3: Python es un lenguaje de programación interpretado enfocado en que los códigos sean más legibles para el usuario, cuenta con librerías para la construcción de software de ámbitos diferentes. A través de Python y algunas de sus librerías, puede ser facilitada la construcción de un sintetizador de voz en comparación con las ofrecidas por otras técnicas o lenguajes como C o C++.

- Pip 2 y 3: "PIP es un acrónimo que significa "Paquetes de instalación PIP" o "Programa de instalación preferida". Es una utilidad de línea de comandos que le permite instalar, reinstalar o desinstalar paquetes PyPI con un comando simple y directo: "pip"." [8] A través de Pip la autora descarga las herramientas adicionales que se utilizan en la construcción de un sintetizador de voz.

- Sistema operativo: En cuanto al sistema operativo, este método de construcción requiere de preferencia un sistema operativo Unix, de manera específica Mac o Linux, ya que algunas de las herramientas libres están diseñadas de manera específica para estos sistemas y pueden no funcionar en los sistemas Windows de Microsoft.

- Otros elementos: Al hacerse grabaciones con micrófonos de poca calidad o con hardware integrado de las computadoras, es necesaria la utilización de un alineador de pistas para etiquetar el audio según los textos; y algún programa que permita ver de manera gráfica las ondas que produce algún sonido que se haya grabado.

3. Utilizando la inteligencia artificial ofrecida por aplicaciones en la web.

[8] Daniel Pérez Fernández en: https://tecnonucleous.com/2018/01/28/como-instalar-pip-para-python-en-windows-mac-y-linux/

Otro proceso de construcción es el mediado por herramientas en la web, y a pesar de que no haya una metodología para construir una síntesis de voz de esta forma, existen tutoriales que guían este proceso de manera superficial, estos tutoriales se encuentran en redes de vídeo como YouTube y despiertan gran interés entre las comunidades por la facilidad que ofrecen. Es posible encontrar títulos como "¡PRUÉBALO! Crear un CLON DIGITAL DE TU VOZ con Inteligencia Artificial" de Juan Merodio, donde se utiliza la herramienta de la organización canadiense llamada "Lyrebird" a través de la cual se construye una síntesis con voz personalizada utilizando únicamente una computadora y un micrófono doméstico. En materia de software para hacer esto es necesario:

- Navegador web actualizado: Así, sin más, basta con tener un navegador en su versión más reciente, a través del cual la página proporcionada por Lyrebird solicitará la grabación de pistas, y en base a estas grabaciones irá generando una síntesis de la voz que se haya escogido grabar.

3.3 Otros requerimientos

No importa el modo que se utilice para crear un sintetizador de voz, es de gran importancia tener en cuenta que en el proceso de construcción de una síntesis de voz se requieren algunos elementos computacionales, sin ignorar que más allá de eso, existen otros componentes fundamentales que trascienden y no están relacionados con el área de sistemas. Estos componentes se pueden clasificar en dos grupos, los ideales y los necesarios.

El conjunto de elementos ideales es posible definirlo como aquellos recursos con los que se podría construir una síntesis de voz robusta, con una tonalidad vocal agradable, un margen de error mínimo en las expresiones y una cantidad significativa de posibilidades para generar las frases y oraciones para el habla. Es importante hacer un énfasis en que a los recursos ideales se puede acceder en medida de que se tengan los recursos económicos y que se desee hacer una síntesis con tendencias a la perfección. Algunos de los elementos ideales podrían ser:

- Un estudio de grabación: Sin lugar a duda, esta es la herramienta de más difícil acceso, principalmente por los altos costos asociados a cada sesión de grabación.

  La idoneidad de un estudio de grabación radica en la cantidad de recursos de hardware y software especializados, a través de los cuales se refina a niveles avanzados las diferentes grabaciones que se utilizarían para elaborar una síntesis propia.

- Un orador: Entendiéndose el orador como la "persona que habla en público, pronuncia discursos o imparte conferencias de forma elocuente y con estilo elevado"[9] se puede considerar como una parte ideal para la construcción de un sintetizador de voz ya que con las capacidades vocales que tiene y las habilidades para una pronunciación clara y contundente, se podría extraer un conjunto de fonemas y difonemas tan refinados que facilitarían el proceso de alimentación de las bases de datos que hacen posible el funcionamiento del sintetizador de voz.

En cuanto a los recursos necesarios, es posible definirlos como aquellos indispensables para el trabajo con grabaciones como las que se requieren para la construcción de un sintetizador de voz. Construir una síntesis de voz podría considerarse un proceso crítico desde algunos puntos de vista como los expuestos en el marco referencial, y esa criticidad se ve incrementada por algunos elementos como los que se exponen a continuación:

- Un entorno con poco ruido ambiental: Al momento de construir una síntesis de voz, se deben hacer grabaciones desde las cuales se extrae una gran cantidad de información que es separada en elementos que utiliza el sintetizador de voz al momento de generar los sonidos, por lo anterior, es evidente que estas grabaciones deben estar lo más limpias posibles, sin ruidos de fondo ni interferencias. Algunas personas son amantes de ruidos como el producido por la lluvia, o el canto de los pájaros, pero es necesario encerrarse en un espacio donde estos sonidos no puedan penetrar, ya que es la única manera en que se garantizan unas grabaciones de audio con la calidad aceptable para alimentar las bases de datos de un sintetizador de voz.

[9] Orador: https://dle.rae.es/?w=orador

- Bastante paciencia: A pesar de intentar superar los obstáculos impuestos por el ruido exterior, en los procesos de grabación para construir un sintetizador de voz no es posible garantizar que una grabación salga perfecta en el primer intento. Algunos investigadores que han trabajado esta temática han abordado la parte de grabación de diferentes maneras con el fin de mitigar el hecho de tener que repetir grabaciones muy largas, por ejemplo, en algunos casos las grabaciones se hacen por oraciones o frases cortas y específicas las cuales duran pocos segundos y permiten volverlas a grabar de forma sencilla ante los posibles errores. Sin embargo, entre más robusta sea la síntesis de voz que se pretende construir, existe la posibilidad de que los textos que hay que grabar sean más largas, incrementando así la probabilidad de fallo. Es por lo anterior, que resulta necesario armarse de paciencia, para un proceso que, a pesar de parecer sencillo, puede volverse repetitivo y engorroso.

- El texto que se utilizará para hacer las grabaciones: A lo largo de este capítulo se tratan aspectos importantes para grabar como audio aquel conjunto de elementos que son de gran utilidad para la construcción de un sintetizador de voz, sin embargo, construir el texto que será grabado "es un paso fundamental al que no siempre se le da suficiente importancia" (Soledad, s.f.). La importancia de este texto radica en que las expresiones utilizadas en este van a ser leídas y guardadas en las grabaciones que alimentan la base de datos del sintetizador y servirán para extraer los fonemas y difonemas a utilizar.

  La construcción de un texto de calidad es bastante importante, ya que en pocas palabras, frases u oraciones se pretende extraer el mayor número de fonemas y difonemas que sea posible, haciendo énfasis en que este texto debe ser lo más optimo posible, ya que no se dispone de tanto tiempo para crear un audio para todas las palabras y oraciones que puedan existir en un lenguaje, y si se contara con tanto tiempo, entonces no se justificaría la existencia de las síntesis concatenarias como lo son las TTS. Es necesario refinar este texto de tal manera que minimice el tiempo que se tiene que invertir para hacer las grabaciones.

- Dominio básico de herramientas informáticas: Es importante aclarar que este dominio no se encuentra relacionado con conocer y manejar algún lenguaje de programación, sin embargo, es fundamental que al trabajar en la construcción una síntesis de voz, sepa usar

una terminal de sistema operativo y conozca de manera superficial el funcionamiento del lenguaje de programación en el que se hayan construido las herramientas con las que vaya a trabajar.

- Dominio básico de conceptos: Bastantes autores que hacen grandes aportes en el tema del habla producida por computadoras utilizan conceptos que para ellos pueden ser elementales, pero para los lectores no, por lo que es una ventaja conocer de qué se tratan muchos de estos conceptos, tales como conjunto prosódico, prosodia, fonemas o difonemas.

- Algunos conocimientos previos: En el caso de la construcción mediada por algún Framework, algunos autores esperan que los lectores de sus obras estén "familiarizados con términos básicos como F0, fonema y cepstrum, pero no con ningún detalle real. Se dan referencias a textos generales (cuando sabemos que existen). Un conocimiento básico de programación en Scheme (y / o Lisp) también facilitará las cosas. Una capacidad básica en la programación en general facilitará la definición de reglas, etc." (Black & Lenzo, 2014), dejando el mensaje de que no es necesario un nivel avanzado de programación, pero sí es de gran utilidad conocer el funcionamiento de algunos lenguajes y manejar algunos términos importantes a la temática.

PARTE III DESARROLLO DE LA INVESTIGACIÓN

4. Capítulo 4 Construcción de un sistema TTS.

De manera general, es importante seguir una secuencia de pasos para la construcción del sintetizador de voz. Habiendo identificado los requisitos necesarios en materia de hardware y software, es posible pasar al proceso de construcción, y para esto, es fundamental diseñar la síntesis y posteriormente escoger la alternativa de construcción más adecuada para cada caso.

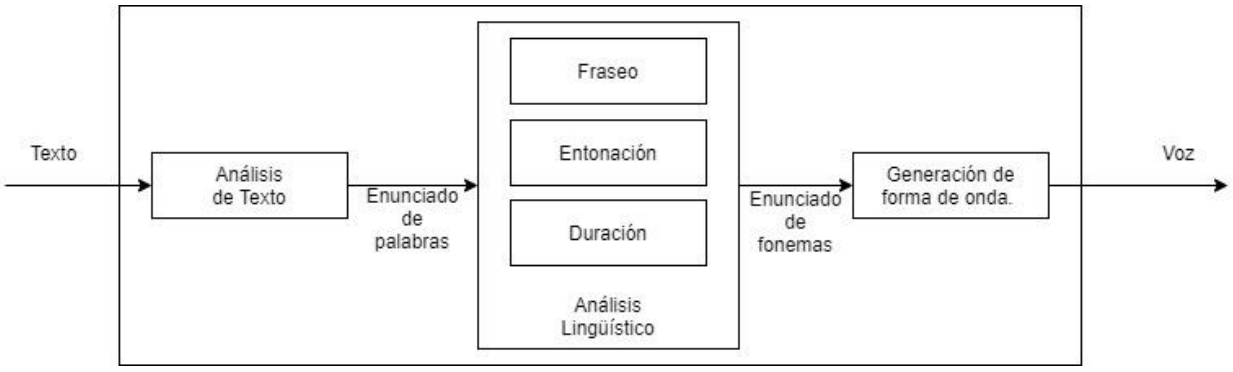
4.1 Diseño de la síntesis de voz concatenaria.

4.1.1 Estructura general de un sintetizador de voz.

El diseño es una de las partes fundamentales a la hora de construir cualquier producto informático, por su parte, la síntesis de voz como herramienta informática, requiere ser diseñada independientemente de la metodología que se vaya a utilizar para su construcción o de las herramientas mediadoras que estas metodologías puedan llegar a requerir.

A pesar de que existan diferentes alternativas para la construcción de un sintetizador de voz, una estructura destacada para esta construcción es la que propone el protocolo Text To Speech de manera general:

Ilustración 1 Estructura General de un sintetizador de voz



Fuente: Elaboración Propia.

A través de esta ilustración es posible identificar que como en todo sistema, en este existen entradas y salidas que se deben contemplar desde la fase de diseño, así como también una serie de subsistemas internos capaces de elaborar las salidas deseadas en base a sus entradas y funcionamiento.

Como entrada se tiene el texto, el cual puede llegar como una palabra, una oración, una frase, un párrafo y en algunos casos como un documento completo. Los sintetizadores más básicos pueden admitir como entrada una sola palabra, o un conjunto pequeño de palabras, mientras que los sintetizadores más robustos pueden recibir documentos de cientos de páginas con grandes cantidades de texto. En el anexo 01 es posible visualizar algunos ejemplos de entrada para dos sintetizadores de texto a voz, en los cuales se envía una palabra en el más básico y un documento de texto en el más robusto.

En cuanto a las salidas, se tiene la "voz" o "habla" dependiendo la región para la que esté diseñado el sintetizador. La voz saliente de un sintetizador de voz es la equivalente a la lectura del texto que ha llegado como entrada. Por su parte, esta voz debe ser el sonido producido por la interpretación de las formas de onda que genera el sintetizador de voz durante su funcionamiento.

Sin embargo, más allá de entradas y salidas, es necesario considerar también las diferentes partes que componen el sintetizador de voz en su estructura interna. Es importante entonces, resaltar que es parte fundamental estudiar el funcionamiento interno del sintetizador de voz, ya que de allí salen las diferentes partes a ser construidas y el sentido de la elaboración de estas:

1. Análisis de Texto: En el análisis de texto es necesario dividir el texto entre palabras, en la forma más trivial, esto significa identificar los espacios en blanco para determinar el fin de una palabra y el comienzo de una nueva. De esta manera, se hace posible que el sintetizador pueda trabajar de una manera más modular.

   Múltiples programas son capaces de realizar este tipo de análisis con diferentes métodos, en muchos casos, es posible obtener programas que a través de la inteligencia artificial puedan llegar a analizar los textos de manera avanzada.

2. Análisis Lingüístico: El análisis lingüístico es aquel en el cual se toma un conjunto de palabras y se les asocia un fraseo, una entonación y una duración, ésta es una de las partes más complejas del sintetizador de voz, ya que requiere de técnicas avanzadas y poco convencionales para el etiquetado de palabras con sus respectivas características. El análisis lingüístico debe estar en la capacidad de recibir enunciados de palabras y de entregar estos como fonemas, es decir, las partes atómicas de las palabras para el funcionamiento de un sistema de síntesis de voz. Es importante resaltar, que de manera convencional las palabras se dividen en sílabas, sin embargo, son los fonemas los que hacen posible que a cada sílaba se le puedan asociar determinadas características.

   Por otra parte, de una forma más sencilla, podría considerarse el análisis lingüístico como un sistema inteligente capaz de tomar de una base de datos la información solicitada por las diferentes entradas de texto, para entregarlas como fonemas.

3. Generación de formas de onda: En esta fase son recibidos los fonemas como la entrada principal, los fonemas son importantes ya que son aquella herramienta que concatenar para hacer posible la producción de un sonido que se escuche como una voz. Como es sabido, todos los sonidos poseen una onda asociada que tiene unas características definidas como lo son la frecuencia o la amplitud, con la voz pasa un proceso similar, las voces graves tienen una tonalidad y frecuencia específica mientras que las voces agudas tienen otra, es por esto, que los fonemas son construidos en base a un fraseo, una tonalidad y una duración. Es necesario poder determinar de manera exacta la onda a generar, ya que son bastante específicas las ondas en cada sintetizador.

   En la actualidad, existen programas de gran utilidad a través de los cuales es posible identificar las ondas asociadas a cada fonema, como por ejemplo el que se observa en el anexo 02.

   Por otra parte, el generador de formas de onda es entonces el encargado en establecer esa relación abstracta con el hardware, en la cual se envían las ondas que toma un dispositivo de salida de audio para reproducir los sonidos solicitados.

A pesar de que la estructura general del protocolo TTS parece trivial, no hay que ignorar que tiene una alta dependencia de otros sistemas informáticos, lo que puede aumentar la probabilidad

de que existan errores en el proceso de construcción o dificultades de adaptabilidad para determinados sistemas.

De manera general, al construir un sintetizador de voz realmente lo que se hace es dotar un sistema con la información mínimamente necesaria para obtener los resultados deseados.

En cuanto se conoce la estructura del sintetizador de voz, es posible comenzar a pensar en el proceso de construcción que se seguirá para una culminación exitosa de la síntesis de voz. Como es sabido, los procesos que destacan a lo largo de este documento son: El proceso mediado por el conjunto de herramientas de festival, a las que se les denomina framework; el proceso mediado por el lenguaje de programación Python, sus librerías y otras herramientas para el tratamiento de pistas de audio; y el proceso a través de tecnologías de inteligencia artificial disponibles en la web. Sin embargo, a pesar de que todos estos métodos son útiles y tienen ventajas y desventajas asociadas, es de aclarar que se hace posible la utilización de alguno de estos de acuerdo con el recurso computacional (hardware y software) y humano con el que se cuente además de la robustez que se desee para la síntesis de voz a construir. Con el fin de aclarar cuál de los métodos puede ser más útil, es necesario verificar el cumplimiento de los requisitos para poder estimar las capacidades de la síntesis de voz a construir y escoger alguno de los procesos de construcción.

4.1.2 Verificación de cumplimiento de requisitos.

En cuanto se desee iniciar el proceso de construcción de la síntesis de voz, es necesario verificar el cumplimiento de algunos requisitos con el fin de establecer cuáles son las herramientas están asociadas al método de construcción que se haya escogido.

Es de destacar que los requisitos expuestos en el tercer capítulo de este documento están enfocados a una construcción mediada por el software libre y pueden cambiar según el entorno o sistema operativo sobre el que estén corriendo estos programas.

Para identificar el cumplimiento de los requisitos se puede proceder de la manera expuesta a continuación, sin embargo, no hay una metodología enfocada en verificar que se cumplan los requisitos mínimos o recomendados en el proceso de construcción de una síntesis de voz:

1. Establecer idoneidad del entorno y la disposición de personal para el proceso de construcción:

   Es importante verificar las características del entorno en el que se han de realizar las grabaciones, ya que en algunas ocasiones se pasan por alto algunos ruidos asociados al mismo, además, es necesario verificar las capacidades del personal con el que se cuenta, o las capacidades propias si se está realizando un proceso de construcción desde la autonomía. La siguiente tabla relaciona las herramientas para la construcción de una síntesis de voz con las características idóneas del entorno

Tabla 1: Para verificación de entorno y personal en el proceso de construcción.

Herramienta para escoger	Características idóneas del entorno	Personal idóneo para la construcción.
Framework Festival y sus Herramientas asociadas FestVox.	- Estudio de grabación. - Espacio insonorizado (Incluyendo cabinas). - Espacio doméstico libre de ruido. - Laboratorios de sonido.	- Especialista en sonido y tratamiento de pistas de audio. - Persona cuya voz cuente con las características adecuadas para las grabaciones (Un orador).
Lenguaje de programación Python, sus librerías y otras herramientas.	- Espacios domésticos sin ruido exterior o interferencias. - Habitaciones sin eco.	- Cualquier persona que domine de manera básica la temática de la síntesis de voz y esté en la capacidad de ejecutar comandos en una computadora.
Tecnologías de inteligencia artificial disponibles en la web. (Lyrebird)	- Espacios domésticos sin ruido exterior o interferencias. - Habitaciones sin eco.	- Cualquier persona sin conocimientos previos, pero con capacidad de navegar en la web.

Fuente: Elaboración Propia.

Posterior al análisis de estas tablas, es menester reconocer que aquel que esté en la capacidad de cumplir los requisitos de entorno y personal para escoger como herramienta de construcción el Framework Festival y sus herramientas asociadas FestVox, podría también tener éxito si escoge cualquiera de las otras dos herramientas, ya que de manera jerárquica la herramienta con más requisitos asociados es festival y la que tiene menos requisitos asociados es la que utiliza tecnologías de inteligencia artificial en la web como Lyrebird.

2. Verificar el hardware disponible:

Para crear algún producto informático, en muchas ocasiones solo es necesario el hardware más básico con el que opera un sistema operativo, es decir, los dispositivos físicos de entrada y salida como mouse, pantalla y teclado. Sin embargo, en el caso de la creación de una síntesis de voz, se requieren elementos dedicados a la creación y procesamiento de audio. La tabla que se expone a continuación permite establecer una relación entre el hardware necesario y la herramienta a utilizar.

Tabla 2: Verificación de requisitos mínimos y recomendados de Hardware.

Herramienta	Hardware Mínimo	Hardware Recomendado.
Framework Festival y sus Herramientas asociadas FestVox.	- Auriculares domésticos. - Micrófono integrado en la computadora o en dispositivos como auriculares. - Computador con dispositivos de entrada/salida y procesamiento de audio.	- Micrófono dedicado con cancelación de ruidos. - Tarjeta de sonido dedicada. (Externa a la integrada en la computadora) - Periféricos de calidad para la Salida de audio.
Lenguaje de programación Python, sus librerías y otras herramientas.	- Auriculares domésticos. - Micrófono integrado en la computadora o en dispositivos como auriculares.	- Micrófono dedicado (no necesariamente con cancelación de ruidos) - Dispositivo con salida de audio de calidad.
Tecnologías de inteligencia artificial disponibles en la web. (Lyrebird)	Para este método de construcción de la síntesis de voz, basta con tener los mismos elementos que para la construcción de una síntesis utilizando el lenguaje de programación Python, sus librerías y otras herramientas.	

Fuente: Elaboración Propia.

"Un elemento para juzgar la calidad de la síntesis de voz es su parecido con la voz humana" (Azcona, 2008) sin embargo, en el caso de las grabaciones de voz que alimentan las bases de datos que hacen posible el funcionamiento del sintetizador de voz, el hardware en conjunto con las condiciones del medio en el que se esté trabajando determina la calidad

del producto a obtener, es decir, en la medida de que el hardware sea más avanzado y dedicado a la grabación y tratamiento de pistas de audio, mayor será la calidad y capacidad de la síntesis de voz obtenida.

Para cualquiera de los procesos de construcción es adecuado y recomendable contar con hardware de calidad que esté disponible gran parte del tiempo, ya que en casos específicos podría optarse por rentar un estudio de grabación por horas para hacer las grabaciones, lo cual generaría resultados bastante positivos, pero crearía una limitación temporal en los recursos físicos, comprometiendo así la posibilidad de repetir en varias ocasiones las grabaciones que cuenten con imperfectos.

3. Verificar la disponibilidad de software y su adaptabilidad:

Sin lugar a dudas hasta aquí el hardware ha representado un elemento muy importante para el funcionamiento y construcción de un sintetizador de voz y una síntesis de voz respectivamente, sin embargo, el software es la herramienta insignia que tomará mucho más valor al hacer posible que el hardware actúe en la manera como es necesario que lo haga, ya que, de nada sirve contar con las mejores herramientas físicas si no se dispone de software que pueda darles la utilidad que corresponde al caso.

A lo largo de este documento se hace énfasis en construir la síntesis de voz utilizando herramientas de software libre, lo que facilita el proceso de acceso a estas, sin embargo, es importante tener en cuenta que pueden existir dificultades de adaptabilidad dependiendo el sistema operativo que se vaya a utilizar.

Independientemente al conjunto de herramientas o metodología que se decida utilizar para construir la síntesis de voz, los resultados y posibilidades de instalación de las herramientas dependen del sistema operativo.

Los espacios de producción de contenido sonoro como, por ejemplo, un estudio de grabación, podrían contar con software especializado que pueda hacer un manejo adecuado de las pistas de audio que se vayan a grabar, sin embargo, cuando se hace el proceso de construcción de la síntesis de voz en espacios domésticos o poco adecuados, es necesario cumplir con unos requisitos mínimos de software, y es adecuado intentar cumplir con unos requisitos recomendados de software.

Los requisitos que se exponen a continuación aplican para cualquiera de los métodos de construcción:

Tabla 3: Verificación de requisitos mínimos y recomendados de software.

	Requisito mínimo	Requisito recomendado.
Sistema Operativo	Cualquier sistema operativo: GNU/Linux, MacOS, Solaris, Unix, Windows.	Sistemas operativos Unix, GNU o MacOS.
Controladores:	Controladores propios del sistema operativo en cualquiera de sus versiones.	Controladores privativos de hardware dedicados (para tarjetas de sonido de referencias específicas).
Compiladores:	Propios de los sistemas operativos.	Dedicados a lenguajes de programación específicos (C, C++, Python, Java, entre otros)

Fuente: Elaboración propia

A pesar de considerar el sistema operativo de manera general, hay que tener en cuenta todo el software que este incluye, es decir, la consola de comandos, sus compiladores, las capas que hacen posible su funcionamiento y los protocolos para salida de audio.

Para este tipo de proyecto, existe versatilidad a la hora de escoger un sistema operativo, y por el comportamiento del mercado podría existir una creencia de que lo más recomendable es utilizar Windows, sin embargo, no hay que ignorar todas las posibilidades que ofrecen los sistemas GNU/Linux y la cantidad de distribuciones de este que existen, ya que los autores destacados en esta temática han desarrollado gran parte de sus trabajos en sistemas Linux, como se menciona en los primeros capítulos de esta investigación.

El hecho de escoger Linux para desarrollar el proceso de construcción de una síntesis de voz significa contar con la posibilidad de solucionar bastantes problemas de una forma más sencilla en dicho desarrollo cuando se trata de un proceso de construcción hecho desde casa. Es importante tener en cuenta, además, que en principio Linux puede representar un

reto para todo aquel usuario acostumbrado a Windows, sin embargo, con un poco de estudio de la documentación y de manejo básico de comandos, puede ser una alternativa a la que se le pueda sacar mayor ventaja e incluso gusto.

Tras verificar los requisitos, es posible optar por alguno de los métodos de construcción de una síntesis de voz. Para tomar esta decisión, en un proceso de construcción autónomo primero debe analizarse que se cumplan los requisitos de software y hardware para posteriormente disponerse a instalar las herramientas que corresponden a cada método, sin ignorar que los métodos que destacan para una construcción autónoma son: La construcción mediante el lenguaje de programación Python y La construcción mediada por herramientas de inteligencia artificial en la web.

Es importante tener en cuenta que cada sistema operativo tiene diferentes versiones, pero la recomendación a seguir es que el sistema a utilizar aún cuente con soporte técnico y con actualizaciones, ya que de estas depende en gran medida un correcto funcionamiento de los controladores de sonido, permitiendo así la explotación completa de las diferentes herramientas físicas u ofrecidas por programas de cómputo.

Probablemente en las fases más tempranas del proceso, la instalación de determinadas herramientas es compleja, pero, no hay que ignorar la existencia de los recursos en la web orientados a la solución de estos inconvenientes, como por ejemplo los sistemas operativos y su respectiva documentación disponible en la web oficial de cada organización que los fabrica, e incluso su soporte técnico para fallas específicas que pueden presentarse en algunos componentes.

Entre otras cosas, es importante que se considere también que cada una de las herramientas a utilizar cuenta con su respectiva documentación y especificación técnica realizada por los propios desarrolladores, ya que, muchas de las herramientas son proyectos de software libre que han sido creados por expertos en sistemas o apasionados de la temática, sin respaldo de alguna gran organización que pueda brindar un servicio dedicado de soporte.

4.2 Construcción mediada por las herramientas Festival, FestVox y SpeechTools reconocidas como Framework.

La información consignada a continuación se encuentra concentrada con mayor profundidad en el título oficial “Building Synthetic Voices” (Black & Lenzo, 2014) el cual se encuentra en la página oficial del proyecto FestVox, sitio en la web en el que se recogen las herramientas necesarias para la construcción de un sintetizador de voz y se ponen a disposición para su descarga de manera libre.

En algunos casos, de entrada, puede parecer complejo el proceso de comenzar a manejar un framework, sin embargo, con un poco de disciplina y tiempo, este proceso se puede ir normalizando. De manera trivial, el primer paso para embarcarse en el proceso de construcción a través de este conjunto de herramientas es verificar que los periféricos físicos como auriculares o micrófonos estén correctamente conectados a la computadora y posteriormente instalar los programas necesarios, los cuales se pueden instalar de la siguiente manera en un entorno GNU/Linux no virtualizado (Esta instalación puede consultarse de manera gráfica en el Anexo 03):

a. Instalación de Festival:

1. Para esta tarea, es necesario utilizar el comando “sudo apt-get install festival festvox-ellpc11k” en la terminal de cualquier distribución de GNU/Linux. Además, de ingresar la letra “S” cuando se solicite la confirmación de nuevo programa instalado.
2. Introducir el comando “festival” en la terminal para verificar que se haya instalado correctamente la herramienta.
3. Introducir algún comando dedicado para festival, podría probarse con help, para obtener una lista con los comandos oficiales con los que dispone el programa, entre los cuales es posible encontrar el de sacar como voz algún texto introducido por teclado.
4. Verificar el funcionamiento de los diferentes comandos, con el fin de establecer que no haya problemas de controladores o adaptabilidad para la salida de audio que requieren estos.

b. Instalación de FestVox: Es importante que antes de proceder a instalar esta herramienta se verifique que un compilador del lenguaje C y C++ esté instalado en el sistema, ya que festvox se reconoce como una librería, la cual se ha escrito en C++ y requiere ser compilada para su utilización.

1. Dirigirse al sitio web oficial http://festvox.org/, en el cual podrá encontrarse el enlace de descargas a través del botón "FestVox Download" y proceder a descargar la última versión, en este caso, la 2.7.0.
2. Ubicar el directorio en el que se almacena el fichero descargado desde el sitio oficial, el cual debe estar comprimido.
3. Descomprimir el fichero descargado a través de los comandos de GNU/Linux disponibles para este proceso.
4. Verificar el nuevo directorio con los archivos necesarios para una instalación.
5. Establecer los criterios de ubicación en los archivos de configuración para almacenar la librería una vez se compile en el sistema que se esté utilizando.
6. Ejecutar el comando "make" para que el compilador de C y C++ inicie el proceso de compilar la librería y la ponga a disposición del usuario.

c. Instalación de las Edinburgh Speech Tools: Speech Tools es el conjunto de herramientas que permitirá ejecutar comandos especiales para el manejo de algunas pistas relevantes para un correcto desarrollo del proceso de construcción del sintetizador de voz. Para su instalación es necesario contar con el compilador de C y C++, de preferencia, el utilizado para la compilación de la librería FestVox.

1. Dirigirse al sitio web oficial http://festvox.org/, en el cual podrá encontrarse el enlace de descargas a través del botón "Festival -> Download" y proceder a descargar la última versión de speech tools, en este caso, la 2.4.0.
2. Repetir los pasos 2, 3, 4, 5 y 6 del proceso de instalación de la librería FestVox, sin ignorar que, en este caso, en lugar de compilar una librería, se están compilando un conjunto de herramientas que serán de utilidad en el proceso de construcción.

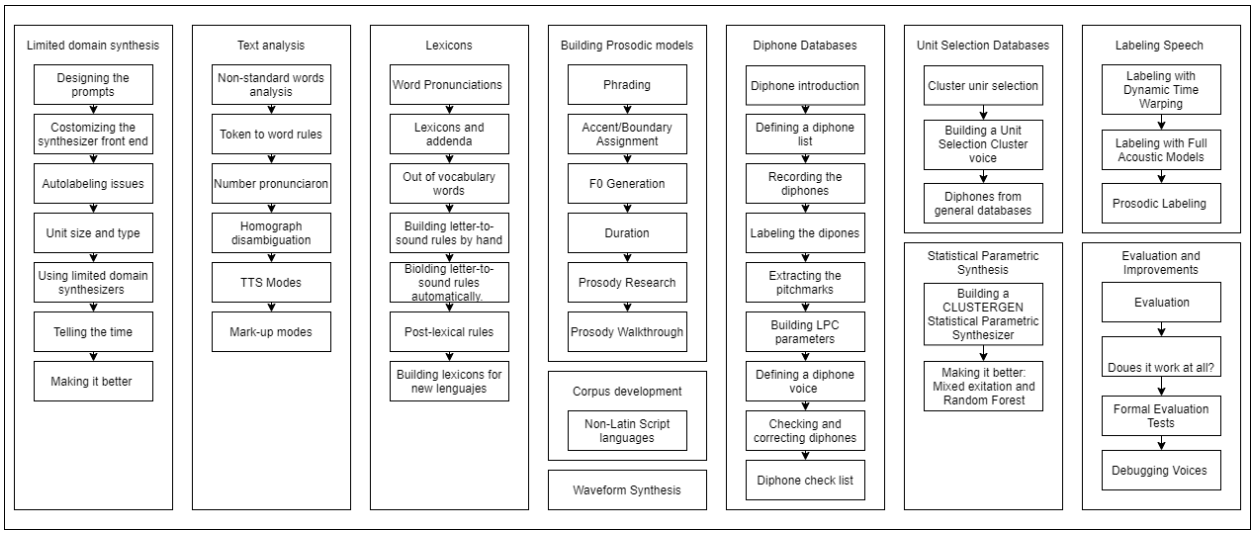
d. Instalar las herramientas adicionales: A pesar de que festival, speech tools y la librería festvox pueden ofrecer bastantes utilidades en conjunto, para la realización del sintetizador de voz es importante contar con otros elementos, como, por ejemplo, un programa de etiquetado/visualización de forma de onda y una herramienta de visualización para depurar y diagnosticar el etiquetado automático que hace FestVox. Existen varios programas de etiquetado o visualización de forma de onda que son libres y gratuitos, la documentación oficial del proyecto festvox recomienda la utilización de la herramienta EMU Speech database system "emulabel" la cual es distribuida por la universidad de Macquarie. En cuanto al software de Visualización para el etiquetado automático que hace festival, existen varias alternativas para los diferentes sistemas operativos, sin embargo, con el ánimo de que la totalidad del proyecto se construya mediado por software libre, la recomendación es a instalar herramientas gratuitas y libres que no incurran en gastos para el proyecto de construcción de la síntesis de voz.

En cuanto se instalen las herramientas, ya es posible continuar con el proceso. Es importante verificar que las herramientas funcionan de la manera necesaria, sin embargo, suponiendo que en esta parte ya se domina el proceso de verificación de requerimientos, ya no debería haber problema para continuar la construcción de la síntesis.

La documentación oficial de FestVox es propiedad del Lenguage Technologies Institute - Carnegie Mellon University, y los derechos de autor son oficialmente de Alan W. Black y Kevin A. Lenzo, siendo el material más importante en cuanto al acompañamiento requerido para este proceso de construcción. Dicha documentación al estar en inglés puede representar un obstáculo para algunos investigadores de la temática, sin embargo, sirve para construir síntesis de voz en bastantes lenguas y ofrece soporte para todos.

Una vez se han superado los requisitos, la construcción está separada en varias partes, donde una a una debe estudiarse en detalle a fin de comprender de manera completa la temática y obtener resultados exitosos tras el proceso de construcción: Síntesis de dominio limitado, Análisis de textos, léxico, construcción de los modelos prosódicos, desarrollo del Corpus, La síntesis de formas de onda, las bases de datos de difonemas, las bases de datos de selección de unidades, la síntesis estadística paramétrica, el etiquetado de la voz y la evaluación y mejoras.

Ilustración 2 Partes para la construcción de una síntesis de voz



Fuente: Elaboración Propia.

A pesar de que la ilustración 2 se ha construido para este documento, es importante tener en cuenta que está elaborada en inglés porque no es pertinente ni recomendado traducir algunas partes en el proceso de construcción, en vista de que podrían perder su sentido o propósito según como las exponen los autores en sus obras.

Es evidente que, tras analizar estas partes, se pueda identificar una secuencia típica de proyecto de software, por ejemplo, en la primera fase se contextualiza completamente la síntesis de dominio limitado y se ejemplifica el proceso a través de la construcción de un sintetizador básico capaz de pronunciar la hora.

Entrando en materia, en la parte de análisis de texto, son analizados algunos posibles problemas en el análisis de texto para poder convertirse en voz, ya que, aunque la pronunciación de palabras parezca trivial, algunos elementos que componen las lenguas como por ejemplo los números y símbolos específicos, pueden tener pronunciaciones muy específicas y poco convencionales. Es decir, si se quisiera escuchar como salida un valor numérico, el analizador de texto debería estar capacitado para recibir una entrada del número como símbolo una entrada textual de la escritura del número.

La parte de léxico es aquella en la que se específica el método para encentrar la pronunciación de determinadas palabras. Los autores Black y Lenzo destacan como métodos el léxico, en el que se utiliza una gran lista de palabras y sus pronunciaciones o el de reglas de sonido específico por cada letra de entrante.

La parte de construcción de los modelos prosódicos es bastante amplia, y se compone de etapas de gran importancia para la naturalidad de un sintetizador de voz, como por ejemplo el fraseo, etapa en la que se marcan los discursos en grupos más pequeños, tal y como lo hacen las personas a través del uso de puntuaciones. Por otro lado, se asigna también el acento y los límites para la voz saliente, y así, se procede con la generación de la F0, la cual se construye de los acentos y límites para establecer como van a ser las duraciones para la voz construida. Posteriormente, gracias al F0 es posible predecir la duración para las pistas de audio a producir según las entradas, su longitud y palabras. Así, la documentación oficial hace la invitación a investigar a profundidad los modelos prosódicos y expone el tutorial para realizar la llamada prosodia.

En la siguiente parte, la de desarrollo del Corpus, los autores Black y Lenzo exponen las técnicas para desarrollar un buen corpus, el cual cumpla con requisitos como que esté prosódicamente equilibrado, dirigido al dominio que se haya estimado, fácil de decir por la voz que realice las grabaciones sin incurrir en errores y que sea lo bastante corta para una pronunciación que no canse. Como complemento al capítulo dedicado al corpus, los autores también realizan la guía para su construcción en lenguas que no sean latinas.

Por otro lado, y como una parte importante de la documentación, también se explica la manera en cómo se generan las formas de ondas para las descripciones fonéticas y prosodias completas. Para esto, se exponen cuatro métodos diferentes, cuyo resultado deberá ser el mismo: La forma de onda para el habla correspondiente a determinado texto.

Para la siguiente parte, se construye la base de datos de “diphone”, es decir, aquella que almacena los elementos que aparecen como difonemas a lo largo de este documento. Como se trata de construir una base de datos, es necesario tener un diseño y un conjunto de grabaciones para alimentarla, para posteriormente extraer de allí información y darle la utilidad deseada. Esta fase en conjunto con la de selección de unidades puede considerarse como la más importante, ya que aquí sucede el etiquetado de las pistas grabadas, es decir, la identificación de las formas de onda según el texto que le corresponde. Además, antes de pasar a la siguiente fase, es pertinente verificar

la integridad de los difonemas almacenados, ya que son las unidades que compondrán la voz formada.

En las partes siguientes de la obra de Black y Lenzo, se expone como es el proceso de construcción de la síntesis de voz a través de formas de ondas utilizando las técnicas de selección de unidades que se encuentren en bases de datos, teniendo en cuenta que las unidades pueden ser una sílaba, una palabra o incluso una frase entera, según como se haya escogido la unidad fundamental para el sintetizador de voz. Diversos autores han elaborado obras completas sobre la selección de unidades, y es posible ver en la práctica aplicaciones de la selección de unidades, por ejemplo, los servicios telefónicos en capacidad de responder con voz de manera automática algunas interacciones.

Ya finalizando, este método de construcción expone también las diferentes posibilidades de etiquetado que existen, ya que hay que tener en cuenta todas las posibilidades. A pesar de que es importante el etiquetado de los difonemas planteado en la fase de construcción de la base de datos de difonemas, hay que considerar las otras alternativas compatibles con la herramienta Festival.

Y, por último, pero no menos importante, está la parte de evaluación y mejoras, en la cual, es posible analizar los resultados obtenidos y su funcionalidad a través de un conjunto de pruebas diagnósticas que suministra la misma documentación del proyecto festival.

Es de gran importancia, tener en cuenta que el proceso de construcción de una síntesis de voz con la metodología planteada por festival incluye otro conjunto de conocimientos, como, por ejemplo, un dominio básico y manejo de los lenguajes con estructura scheme, con el fin de ejecutar algunos comandos de prueba o realizar avances en el proceso a través de códigos de programación. Además, de otro conjunto de herramientas orientadas a la elaboración de discursos, ya que existe bastante documentación frente al tema, suministrada por diferentes universidades que investigan la temática.

Hay algunos aspectos que destacan en este método de construcción, y otros, que dificultan en gran medida el proceso.

4.2.1 Ventajas

Es evidente que la cantidad de trabajo que supone la construcción utilizando este método es bastante, sin embargo, es proporcional a los resultados a obtener, los cuales son robustos, completos y de calidad en la medida de que se siga a cabalidad el proceso planteado por los principales autores de la temática.

Adicionalmente, aparte de aprender lo suficiente en materia de síntesis de voz, este proceso supone aprendizajes en otros saberes diferentes al de la computación, es decir, no será un proceso en el que únicamente se escriba código o se construyan diagramas, sino que también servirá para aprender bastantes cosas en materias de discursos.

Incluso, con una síntesis creada con este método, podrían adaptarse documentos de texto completos a través de la utilización de festival en la línea de comandos.

También, se puede considerar una ventaja que las herramientas a través de las cuales se construye con este método son software libre, por lo que no habría que adquirir licencias de software para la mayoría de los programas.

4.2.2 Desventajas

Este método a pesar de sus ventajas puede representar problemas para aquellos investigadores que no tengan el suficiente interés en la temática, ya que requiere un entendimiento un poco más avanzado que el de otras alternativas.

También, es posible considerar como desventaja que este método tiene una cantidad elevada de requisitos, por lo que puede ser difícil acceder a estos.

4.2.3 Problemas encontrados

Festival y las demás herramientas que hacen posible este proceso de construcción a lo largo del tiempo han tenido actualizaciones, sin embargo, existen problemas de adaptabilidad para los

sistemas operativos con núcleo más moderno. La documentación oficial para este método expone que los sistemas operativos GNU/Linux son los más recomendados, sin embargo, por ejemplo, para el caso de la distribución Ubuntu 20.04, existe un problema bastante significativo:

Este problema es que cuando se intentan hacer grabaciones, incluso de prueba para la etapa más temprana del proceso, la consola de comandos arroja un error, producido porque las herramientas de speech tools se han programado para trabajar con los controladores Open Sound System, los cuales han sido reemplazados por Alsa en las versiones más recientes de Ubuntu. Si bien, este problema puede solucionarse, puede incluso dejar inútil el parlante de la computadora.

Esta serie de errores, suceden en gran medida, porque la versión oficial del conjunto de herramientas festival, festVox y Speech Tools se actualizaron por última vez en 2014, cuando los sistemas operativos eran un tanto diferentes, los elementos de hardware tenían arquitecturas diferentes y su correcto funcionamiento dependía de controladores que hoy en día se han descontinuado o no se adaptan a las nuevas tecnologías.

4.3 Construcción mediada por Python y la selección de unidades

Esta metodología de construcción supone la posibilidad de hacer una síntesis de voz con la que se pueda sintetizar el texto de manera sencilla, únicamente con invocar un script de python acompañado del texto que compone la frase o el conjunto de palabras que se desee escuchar.

Este método es propio de la autora Pilar Soledad, quien hace un importante aporte introductorio a la temática de la síntesis de voz y su construcción. La autora, a través de un artículo[10] expone en detalle este método e incluso menciona los motivos por los cuales se ha decidido a mediar esta construcción por un lenguaje como Python y sus librerías.

La autora hace énfasis en que ha trabajado con festival en creaciones académicas, incluso en su proceso de formación como el MSc en Speech and Language Processing, sin embargo, menciona también algunos de los problemas que supone trabajar con dicha herramienta, como por ejemplo el desconocimiento de los lenguajes tipo scheme y la dificultad que esto le suma al entender por completo el proceso de construcción.

Con el ánimo de hacer un código más comprensible y que facilite la construcción de una síntesis de voz incluso sin conocer mucho de lenguajes de programación, la autora expone en su obra el proceso de implementación para obtener un programa que permita crear una síntesis de voz propia, desde la comodidad de casa y sin poco más que un micrófono y una computadora doméstica. Para esto, se basa en el principio de "selección de unidades" el cual se encuentra explicado en profundidad en la documentación oficial de festival.

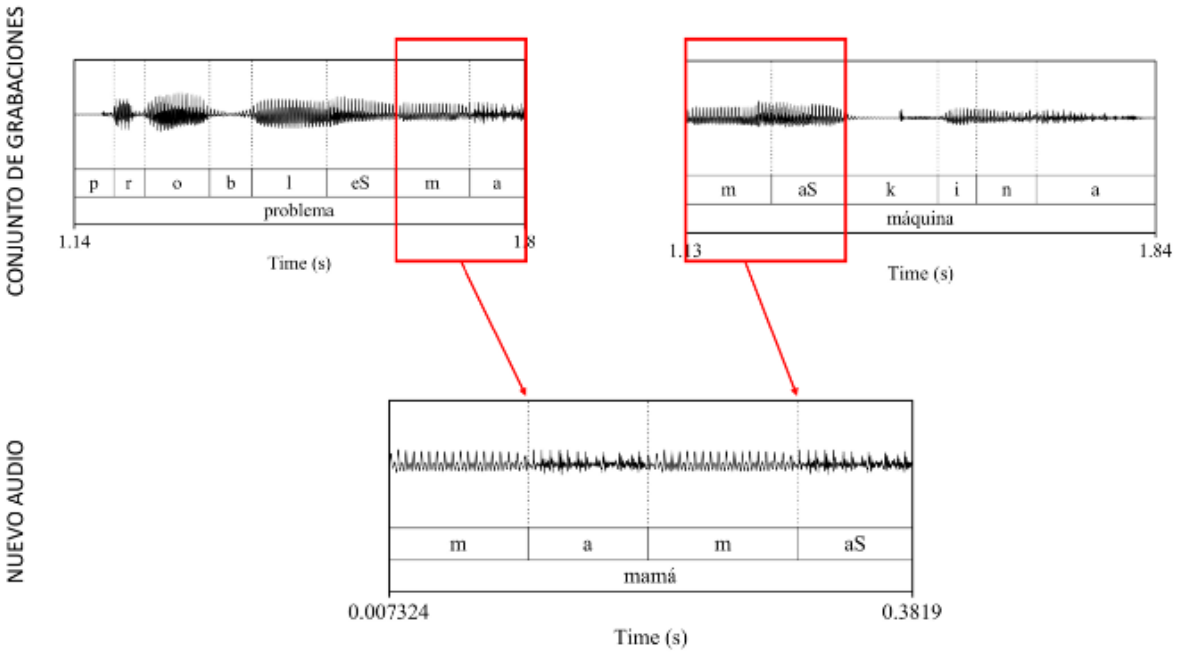
El proceso de selección de unidades consiste básicamente "en tener como base un conjunto de grabaciones anotadas, donde al momento de sintetizar una nueva palabra o frase, el sistema va a buscar las partes que necesita para crearla y las concatena: una especie de collage." (Hunt & Black, 1996)

Dentro del repositorio donde se encuentra almacenado el proyecto con los códigos que permiten construir la síntesis de voz, es posible encontrar una base de datos previamente grabada con la propia voz de la autora, con la que se podría comenzar a sintetizar algunas palabras, frases o textos

[10] Publicación en la web: https://medium.com/@pilarsoledad/construyendo-un-sintetizador-de-texto-a-voz-usando-python-y-selecci%C3%B3n-de-unidades-a5dc2e11a091

y comprender en más detalle de que se trata la construcción de una síntesis de voz, aun cuando no se tiene conocimientos avanzados en la temática.

Ilustración 3 Ejemplo del concepto que sustenta la selección de unidades.



Fuente: Construyendo un sintetizador de texto-a-voz usando Python y selección de unidades, por Pilar Soledad.

Para iniciar el proceso de construcción, lo más recomendado es instalar las herramientas que se van a utilizar, las cuales sugiere la autora que ha creado este método. Para la instalación es importante contar con la herramienta pip y el compilador de Python, versiones 2.x y 3.x en ambos casos. Pip es el sistema que gestiona la instalación y administración de paquetes de software escritos en Python. Pip puede instalarse de manera sencilla en la mayoría de los sistemas operativos y a pesar de que su instalación implica compilar un script, este, se puede obtener de manera sencilla en diferentes sitios en la web. Es importante mencionar que la autora sugiere utilizar los sistemas GNU/Linux, por lo que las herramientas a instalar son compatibles, y el proceso descrito a continuación se puede hacer sin problema en distribuciones como Ubuntu.

Para utilizar el conjunto de grabaciones que se han almacenado en el proyecto por la autora, a fin de sintetizar algunos textos con la voz de ella, y comprender la funcionalidad del sintetizador con un ejemplo funcional, es necesario contar mínimamente con dos herramientas:

1. Pydub: Es el módulo a través del cual el lenguaje de programación Python puede realizar operaciones para reproducir y convertir de formato los ficheros que son de audio. Pydub puede obtenerse a través de la utilización del comando "pip install pydub" en una terminal de comandos de sistema operativo.

2. Scipy: Como lo indica el sitio oficial donde reposa su documentación "SciPy (pronounced "Sigh Pie") is a Python-based ecosystem of open-source software for mathematics, science, and engineering. In particular, these are some of the core packages" (SciPy developers, s.f.). A través de Scipy se podrá compilar el Proyecto y obtener la utilidad deseada. La instalación de Scipy puede variar según el sistema operativo que se esté utilizando, sin embargo, el sitio oficial cuenta con la guía completa para cada sistema operativo con el que es compatible el entorno.

Una vez instaladas estas herramientas y habiendo descargado el código fuente proporcionado por la autora desde su repositorio oficial[11] o a través del comando "git clone https://github.com/pilarOG/unit_selection_tts.git" en computadoras que cuenten con git; ya es posible dirigirse al directorio donde se hayan almacenado estos archivos, descomprimirlos de ser necesario y proceder a compilar el script synthesize.py a través del comando "python synthesize.py "texto de entrada"" donde en texto de entrada se pone algún texto que se quiera escuchar como audio. Tras la ejecución de este script se mostrarán por consola las partes utilizadas en la selección de unidades para producir la pista de audio que se creará cuyo nombre de archivo será "generated" con tipo WAV y se almacenará en el mismo fichero donde se haya guardado la carpeta obtenida desde el repositorio en git.

[11] Repositorio: https://github.com/pilarOG/unit_selection_tts

Sin embargo, a pesar de que se pueden sintetizar muchos textos utilizando las grabaciones suministradas por la autora, el propósito de su proyecto es que se construya una síntesis de voz propia, concatenaria y capaz de producir las pistas de audio según las entradas de texto, por lo que se necesitan dos herramientas más para dicha construcción:

- HTK: "The Hidden Markov Model Toolkit (HTK) is a portable toolkit for building and manipulating hidden Markov models" (HTK Team , s.f.). Esta herramienta se puede instalar a través de su sitio web oficial en el apartado de Descargas[12]. A pesar de que están disponibles las versiones Beta, es recomendado instalar la versión Stable según el sistema operativo que se esté utilizando. "HTK es una de las piedras fundadoras de las tecnologías del habla." (Soledad, s.f.) y es necesario que sea compilada, para lo que se utiliza el comando sudo make en los entornos GNU/Linux.

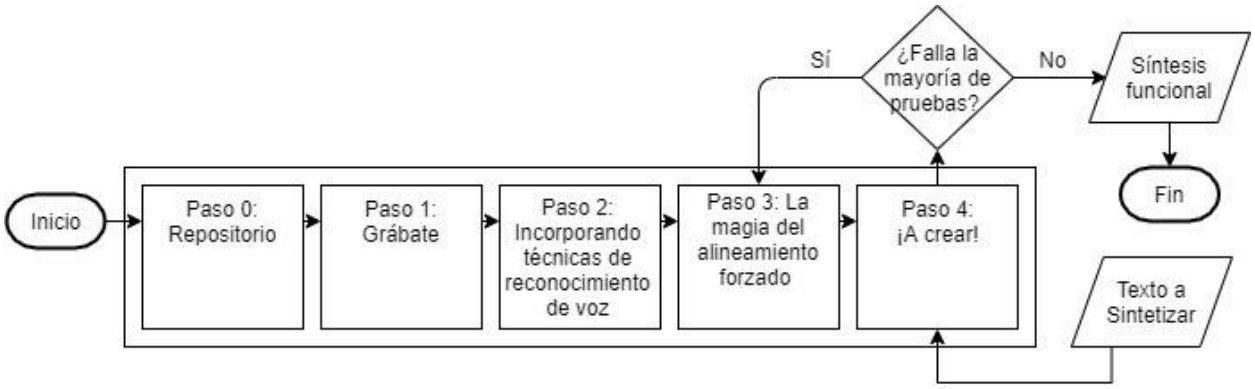
- ProsodyLab aligner: La instalación de esta herramienta es sencilla, ya que en el repositorio oficial[13] se encuentra el código fuente con los respectivos pasos de instalación y los comandos disponibles en esta. A través de ProsodyLab aligner es posible realizar la técnica de "alineamiento forzado", la cual pertenece al área de reconocimiento de voz y permite ubicar cada difono en las grabaciones que se hagan.

- Praat: Esta es una herramienta para el análisis del habla, a través de la cual es posible hacer un análisis espectral, es decir, ver el espectrograma, además de poder analizar muchas pistas de audio. Praat está disponible para Windows, Linux, Chromebook, Macintosh, Raspberry Pi, FreeBSD, SGI, Solaris, HPUX, e incluso es posible acceder a su código fuente desde el sitio oficial "Praat: Doing phonetics by computer" disponible en la web.

En cuanto se instalen estas herramientas, ya es posible proceder a construir la síntesis de voz propia. Para esto, la autora deja una serie de pasos, los cuales están representados a continuación.

[12] http://htk.eng.cam.ac.uk/download.shtml
[13] https://github.com/prosodylab/Prosodylab-Aligner

Ilustración 4 Diagrama de flujo del proceso de construcción de la síntesis de voz.



Fuente: Elaboración Propia.

A pesar de que los pasos y sus nombres dentro del proceso de construcción en el diagrama de flujo son propios de la creadora del proyecto para crear la síntesis de voz con python y la selección de unidades, este diagrama se ha construido para este documento con la finalidad de sumar un poco más de entendimiento a la temática.

En el paso 0, la autora expone el proceso para obtener desde el repositorio todo el código del proyecto, el cual se encuentra en Python junto con una base de datos de ejemplo que contiene los archivos con los que puede funcionar el sintetizador de voz con la voz de su creadora.

Para el paso 1 se utiliza el directorio de archivos llamado recording_scripts, el cual forma parte del conjunto de elementos descargados del repositorio en el paso 0. Dentro de recording_scripts es suministrado un archivo de texto que contiene el conjunto de frases que deben grabarse, así como también, un script de Python en la capacidad de hacer grabaciones, el cual deberá correrse en la terminal de comandos del sistema operativo utilizado, a través del comando "python record_samples.py". Adicionalmente, al tener acceso al código en python, existe la posibilidad de entrar en él y hacer algunos ajustes, como, por ejemplo, modificar el tiempo que se dispone para la grabación de cada una de las pistas de audio en las que se lee una a una las frases u oraciones que se encuentran en el archivo de texto. A manera de plus, este script tan importante, genera un archivo para cada grabación en el que se le asocia el texto que le corresponde, es decir, el texto

que se escucha al abrir la grabación con el mismo nombre, pero extensión .wav. Es importante aclarar que la creadora de este método no descarta la posibilidad de que las grabaciones se realicen con otro programa o en otra modalidad, siempre y cuando el resultado en cuanto a archivos sea el mismo, además, resalta que el conjunto de frases fue creado por ella misma, por lo que sugiere aumentar este banco de frases, sin olvidar que una síntesis de voz se crea precisamente para no tener que grabar todas las frases u oraciones posibles en un lenguaje.

En el siguiente paso, el 2, son incorporadas las técnicas de reconocimiento de voz, ya que disponer del texto y su respectiva grabación en formato de audio no es suficiente. Para esto, al igual que en la documentación de Festival, la autora menciona que podrían partirse las frases grabadas en diferentes unidades, pero para este caso ha escogido el difono (diphone) y fundamenta la toma de esta decisión en que "esta unidad tiene sus raíces en los conocimientos de fonética que tenemos del habla humana." (Soledad, s.f.). Una vez se define la unidad a utilizar, para identificar dónde está cada difono en la grabación, se utiliza la técnica de alineamiento forzado, a través de ProsodyLab Aligner, cuya instalación se menciona previamente, así como también se explica.

En el paso 3, se explican y elaboran dos partes, específicamente "la magia del alineamiento forzado" y la "librería de difonos". En la primera parte, se explica cómo se ejecuta el comando para construir el modelo de lenguaje para el conjunto de frases grabadas. Y en la segunda parte, se explica cómo hacer la búsqueda de los audios alineados con el texto.

Para crear el "mini" modelo de lenguaje, como lo denomina la autora, es necesario ejecutar el script "build_dict.py" que se encuentra en el fichero que se ha menciona previamente "recording_scripts". Posteriormente, tras ejecutar este script debería aparecer un nuevo elemento de extensión .dict, el cual, deberá copiarse dentro de "Prosodylab Aligner", otra de las carpetas en la raíz del proyecto y posteriormente ejecutar con python 3 el comando para alinear. A pesar de esta alineación, es importante también recortar los "silencios" al principio y final de las grabaciones, tarea para la que se recomienda en la guía oficial utilizar Praat. Posteriormente, debe reemplazarse los audios por los nuevos ya recortados, y así, ya se podría etiquetar la información de data con el "mini" modelo que se ha entrenado, esto, también a través de la utilización de python 3 y el comando correspondiente. Tras esta etapa debería quedar un archivo .TextGrid para cada archivo de audio. Es importante que la utilidad de HTK en este paso, es para entrenar modelos

ocultos de Markov, “una de las técnicas más exitosas para hacer reconocimiento de voz antes de las redes neuronales” (Soledad, s.f.)

Posteriormente, se le da un formato de diccionario a los TextGrid, primero copiando la carpeta data al directorio donde se esté almacenando el proyecto creado propiamente. Y posteriormente, ejecutando el script de python 2 “set_diphone_library.py” también suministrado por la creadora del proyecto.

Finalmente, en el paso 4 “a crear” ya es posible sintetizar textos, bastante pequeños en vista del tamaño de la data con la que se ha creado el modelo, pero, funcionales. Sin embargo, la recomendación es hacer las pruebas correspondientes para determinar la calidad del resultado obtenido, que, de no ser satisfactorio, podría mejorarse repitiendo el paso 3, es decir, volviendo a alinear y recortando los audios como corresponda para tener resultados de mayor calidad. La ejecución del paso 4 genera el archivo “generated.wav”, en donde se encontrará el audio que corresponde al texto ingresado junto con el script synthesize.py. Abriendo el archivo “generated.wav” con cualquier reproductor multimedia que soporte la extensión del archivo, se podrá escuchar el texto de entrada como audio.

4.3.1 Ventajas

Este método de construcción tiene como principal ventaja el acompañamiento proporcionado por la creadora del mismo, y el tutorial de acompañamiento disponible en la web, donde se detalla todo el proceso, resaltando en la autora un dominio avanzado del tema con el que le es útil a sus lectores para comprender de una manera sencilla todos los conceptos.

Adicionalmente, hay que agregar la calidad del código socializado a través del repositorio, el cual tiene toda la versatilidad propia de python facilitando así el entendimiento de este.

Entre otras ventajas, es posible considerar la disponibilidad de las herramientas, que siendo varias, incluso avanzadas, están disponibles para varios sistemas operativos y no tienen problemas de adaptabilidad.

Este método, además, la autora lo ha hecho público en internet en 2019, por lo que cuenta con la ventaja de ser compatible con las tecnologías más modernas, dejando de lado inconvenientes por controladoras y demás.

Finalmente, es posible destacar la posibilidad de grabación con un micrófono doméstico, y una computadora con características normales, sin tener que hacer inversiones costosas en equipo de trabajo.

4.3.2 Desventajas

Este método tiene como inconveniente que depende en gran medida de un modelo de lenguaje, el cual tiene que entrenarse como cualquier modelo computacional, por lo que es dependiente de la data introducida en el modelo, existiendo la posibilidad de sobre entrenarse o, por el contrario, carecer de información para la toma de decisiones.

Una síntesis de voz creada a través de este método sería para un uso algo medido, ya que su funcionamiento a través de selección de unidades difiere bastante de la estructura en la que se basan las grandes organizaciones para sus sintetizadores de voz.

4.3.3 Importante tener en cuenta

La probabilidad de falla aumenta en la medida que no se domine el trabajo con diferentes versiones de python, ya que los scripts deben compilarse con una versión 2.x, como por ejemplo la 2.7.15, mientras que, los comandos para entrenar el modelo o hacer la alineación automática, deben ejecutarse con las versiones 3.x, las cuales aún cuentan con actualizaciones, estando disponible actualmente la 3.8.

Por otro lado, es importante considerar que la instalación de las herramientas ProsodyLab Aligner y HTK, debe realizarse utilizando pip3, ya que los scripts funcionarán en la versión 3 de python, mientras que Scipy y Pydub deben instalarse con pip2, porque serán propios de los scripts que se correrán en la versión 2 de python.

4.4 Construcción mediada por Lyrebird.

Construir una síntesis de voz a través de la herramienta en la web Lyrebird, es básicamente utilizar la inteligencia artificial a través de un recurso web para facilitar la construcción. Lyrebird es una herramienta novedosa, que se auto reconoce como "una división de investigación de Inteligencia artificial dentro de Descript, que construye una nueva generación de herramientas para la edición y síntesis de medios que hacen que la creación de contenido sea más accesible y expresiva". (Team, s.f.)

Lyrebird está en su fase "private Beta" por lo que requiere permisos especiales para su utilización, sin embargo, existen influenciadores en la red que han promocionado el trabajo de esta herramienta.

Es importante tener en cuenta, que esta herramienta está enfocada en crear contenido en inglés, por lo que aún no hay una metodología formal que describa paso a paso todo el proceso de construcción del sintetizador de voz, sin embargo, según como se muestra el proceso en diferentes sitios de internet, vídeos tutoriales y la demostración proporcionada por la página oficial de la herramienta, es un proceso bastante sencillo, el cual consiste en grabar varias pistas de audio según lo vaya solicitando el sitio web y aceptar los permisos solicitados por el navegador utilizado para poder que Lyrebird acceda al micrófono y a las grabaciones que se realicen con este. Posteriormente, con este conjunto de grabaciones conforman una especie de base de datos con la cual se construirán las pistas del texto que se desee sintetizar. Sin embargo, es un proceso que no se conoce en detalle, por lo que no se conoce la unidad utilizada ni tampoco la unión de estas unidades para conformar el audio saliente.

4.4.1 Ventajas

Una solución como Lyrebird permite que incluso aquellos que no dominan el tema de la síntesis del habla puedan construir un sintetizador de voz sin necesidad de conocer qué está sucediendo internamente y cómo funciona.

Lyrebird no requiere hardware dedicado, ni tampoco costoso y su naturaleza web supone la posibilidad de conectarse desde cualquier sitio en el mundo.

4.4.2 Desventajas

Este proyecto se encuentra en una etapa aun temprana, por lo que acceder a él y utilizarlo podría ser complicado.

La disponibilidad de este proyecto depende en gran medida de la conexión a internet, es decir, sin conectividad no se podrá acceder a él.

Aún se desconocen las licencias o implicaciones asociadas a la utilización de una síntesis que se ha creado con este método.

4.5 Resultados obtenidos

Es complejo caracterizar los resultados obtenidos cuando se han evaluado proyectos tan diferentes como los que están consignados en este capítulo, sin embargo, la mecánica para construir una síntesis de voz en todos los casos consiste en realizar grabaciones, por lo que se podría analizar bajo criterio propio la calidad de los resultados obtenidos en síntesis construidas a través de estos métodos. Para esto, se han realizado las pruebas consignadas a continuación:

Para el caso de la síntesis propia de festival, se ha utilizado la proporcionada al instalar el programa, para la síntesis de voz creada en python a través de la selección de unidades, se ha utilizado la proporcionada por la creadora de este método, y para el caso de Lyrebird, se ha utilizado el ejemplo suministrado por la página web de Lyrebird. Lo anterior, con el objetivo de tener resultados coherentes, con los que se pueda establecer una relación entre la dificultad del proceso de creación con la calidad del resultado, ya que, si las pruebas se hicieran con síntesis creadas propiamente para este proyecto, con seguridad la calidad no sería la misma para todos los casos. Las tablas a continuación relacionan: Calidad de la síntesis de voz creada y la Dificultad en el proceso para construir una síntesis como la utilizada.

Para evaluar la calidad se tiene en cuenta la siguiente definición de acuerdo con el contexto de la síntesis de voz:

- Buena: El texto ingresado se logra escuchar de manera natural y se entiende por completo.
- Normal: El texto ingresado se entiende, pero es evidente lo artificial de la voz.
- Deficiente: El texto ingresado no se entiende.
- Sin Resultados: El sintetizador de voz no es capaz de generar el audio correspondiente al texto ingresado.

Por otro lado, la dificultad de Construcción, se refiere a lo complejo del proceso para lograr construir el sintetizador de voz, donde la dificultad Alta se refiere a una alta dependencia de requisitos y manejo avanzado de las herramientas, la dificultad normal se refiere a un proceso que puede realizar cualquiera con un poco de conocimiento informático y la dificultad baja se refiere a un proceso que puede realizar cualquiera con acceso a un computador y capaz de navegar en la web.

Tabla 4 Pruebas realizadas con la síntesis de voz en español

	Síntesis de Voz utilizada			
Prueba en español	Síntesis propia de Festival		Síntesis creada con Python	
	Calidad	Dificultad de construcción	Calidad	Dificultad de construcción
"Así es la vida"	Buena	Alta	Normal	Normal
"No te llamas"	Buena		Normal	
"No soy Santiago"	Buena		Normal	
"Tres tristes tigres"	Normal		Sin resultados	

Fuente: Elaboración Propia.

Dentro de las opciones con las que cuenta Festival, la herramienta se puede configurar también para sintetizar textos en inglés, por lo que se ha alterado esta configuración para hacer algunas pruebas en inglés y establecer una analogía con la salida proporcionada por Lyrebird.

Tabla 5 Pruebas realizadas con la síntesis de voz en inglés

	Síntesis de Voz utilizada			
Prueba en inglés	Síntesis propia de Festival		Síntesis creada con Lyrebird	
	Calidad	Dificultad de construcción	Calidad	Dificultad de construcción
"Get my engineering"	Buena	Alta	Buena	Baja
"How old was the dinosaur"	Buena		Buena	
"Completely"	Buena		Buena	
"Relatively"	Buena		Buena	

Fuente: Elaboración Propia

PARTE IV CONCLUSIONES

5. Capítulo 5 Conclusiones

El proceso de construcción de una síntesis de voz requiere de un dominio e interés por el tema para poder basarse en varios principios informáticos e incluso físicos que hacen posible el funcionamiento de un sintetizador de voz y que permiten concluir que:

- Sí bien es posible estudiar los conceptos de la síntesis de voz, como se ha hecho en el estado del arte, es complejo hacerlo para los hispanohablantes que no dominan el inglés ni manejan las herramientas básicas que ofrece la informática, ya que las definiciones formales, las investigaciones y aportes hechos por las autoridades en el tema, se encuentran en inglés y se complementan con muchos conceptos computacionales.
- Hoy en día existen diferentes sintetizadores de voz con una estructura similar entre ellos, cuyas características son propias del enfoque para el que se utilicen y su funcionamiento puede variar según el protocolo en el que se basan y la manera en que se construyen, la cual ha sido cambiante a través de la historia, partiendo desde los dispositivos netamente físicos o mecánicos, hasta llegar a las épocas más modernas de la computación donde el software se ha cobrado la mayor parte del protagonismo.
- En la medida de que se quiera obtener una síntesis de voz robusta, de calidad y capaz de generar las pistas de audio para un dominio bastante amplio de alguna lengua, los requisitos de hardware y software aumentan considerablemente el costo asociado a un proyecto de construcción de una síntesis de voz.
- Los procesos de construcción de una síntesis de voz son similares incluso cuando se hacen con diferentes herramientas, y consisten trivialmente en crear un conjunto de grabaciones a través del cual se puedan generar nuevas salidas del sistema que son producto de concatenar elementos en la base de datos donde se hayan almacenado las grabaciones, según las solicitudes entrantes desde el analizador de texto implementado o utilizado en un sintetizador de voz cuya estructura sea la del protocolo text to speech.
- La complejidad de los procesos de construcción de una síntesis de voz es proporcional a la calidad de los resultados obtenidos tras un proceso de construcción, sin embargo, esta

complejidad se puede reducir en la medida de que se adquiera experticia en el tema partiendo desde los conceptos más fundamentales.

Siendo evidente que sí es posible construir una síntesis de voz propia utilizando los protocolos y herramientas existentes, es complejo integrarlas, ya que algunos métodos por sí solos, son una construcción basada en partes específicas de los métodos más robustos.

Bibliografía

Arroyo Cantón, C., & Berlato Rodríguez, P. (2012). *La comunicación.* Lengua castellana y Literatura. España: Oxford University Press.

Astelehena. (2009). *Gobierno de España, Ministerio de educación, cultura y Deporte*. Obtenido de http://recursostic.educacion.es/observatorio/web/eu/software/software-general/689-reconocimiento-y-sintesis-de-voz

Azcona, G. A. (21 de Abril de 2008). *Bibliotecas UDLAP.* Obtenido de Nueva Voz Concatenativa de Difonemas para el Español Mexicano en Festival: http://catarina.udlap.mx/u_dl_a/tales/documentos/lis/moreno_a_ga/

Black, A. W. (2002). *Perfect Synthesis for all of the people all of the time.* Obtenido de http://www.cs.cmu.edu/~awb/papers/IEEE2002/allthetime/allthetime.html

Black, A., & Lenzo, K. (2014). *Building Synthetic Voices.* Lenguage Technologies Institute, Carnegie Mellon University.

College of Health & Rehabilitation Sciences: Sargent College . (s.f.). Von Kempelen's Speaking Machine. *Guenther Lab* .

Correa, C., Rueda, H., & Arguello , H. (2015). Síntesis de voz por concatenaci´´on de Difonemas para el Español de Colombia. Santander, Colombia.

Definición.de. (s.f.). *Definición De*. Obtenido de Accesibilidad: https://definicion.de/accesibilidad/

GrundHauser, E. (16 de Enero de 2017). *Atlas Obscura*. Obtenido de The Voder, The First Machine to Create Human Speech: https://www.atlasobscura.com/articles/the-voder-the-first-machine-to-create-human-speech

HTK Team . (s.f.). *HTK*. Obtenido de What is HTK? : http://htk.eng.cam.ac.uk/

Hunt, A. J., & Black, A. W. (1996). *Unit selection in a concatenative speech synthesis system using a large speech database. In Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings., 1996 IEEE International Conference on (Vol. 1, pp. 373–376). IEEE.*

Llisterri, J. (Marzo de 2020). *Las unidades de síntesis.* Obtenido de http://liceu.uab.es/~joaquim/speech_technology/tecnol_parla/synthesis/units/unitats_sintesi.html

Maroti, C. (26 de Abril de 2019). *understanding*. Obtenido de https://unbabel.com/blog/es/la-inteligencia-artificial-habla-pero-entiende-realmente-lo-que-dice/

Mata, C. L. (2011). Revisión de la tecnología de síntesis de voz y recursos lingüisticos existentes para el idioma Español. *Instituto Tecnológico de Chihuahua*.

Morales, Á. M. (2013). Diseño de un software para el reconocimiento de símbolos matemáticos en latex mediante síntesis de voz para personas con discapacidad visual. Pereira , Risaralda, Colombia.

Research, T. C. (s.f.). *The Festival Speech Synthesis System*. Obtenido de http://www.cstr.ed.ac.uk/projects/festival/

SciPy developers. (s.f.). *SciPy.org*. Obtenido de https://www.scipy.org/

Soledad, P. (s.f.). *Medium*. Obtenido de https://medium.com/@pilarsoledad/construyendo-un-sintetizador-de-texto-a-voz-usando-python-y-selecci%C3%B3n-de-unidades-a5dc2e11a091

Team, L. (s.f.). *Lyrebird AI*. Obtenido de https://www.descript.com/lyrebird-ai

PARTE V ANEXOS

Anexos

Anexo 01: Alternativas para ingresar texto a un sintetizador de voz.

Suponiendo que los comandos son escritos y ejecutados en una terminal de comandos de alguna distribución de Linux, basta con llamar el comando “Festival” para comenzar la ejecución del programa (siempre y cuando se tenga instalado). Una vez se esté ejecutando, las entradas se pueden ingresar así:

```
festival>  (SayText "Hello world")
```

Donde claramente se puede ver que la entrada es el texto compuesto por dos palabras, “hello” y “world”, esperando escuchar por la salida de audio de la computadora el texto “Hello World” pronunciado por una voz en inglés.

Sin embargo, aprovechando las posibilidades ofrecidas por la herramienta festival, podría entrarse un archivo de texto completo, para ello, primero se puede visualizar el texto del archivo con el comando:

```
$ cat texto_prueba.txt
  Este es un texto plano, para demostrar que festival admite textos como entrada.
```

Y posteriormente, mandarle este texto a Festival con el siguiente comando, teniendo en cuenta que es necesario tener la dirección de la terminal situada en el fichero donde se ha guardado el archivo:

```
$ festival --tts --language spanish texto_prueba.txt
```

Donde se puede esperar escuchar por la salida de audio de la computadora el texto “Este es un texto plano, para demostrar que festival admite textos como entrada.” Con una voz sintetizada en español, gracias al comando “–languaje spanish” propio de Festival.

Por otro lado, en el caso de sintetizadores como el construido en Python, la entrada de textos se puede hacer de la siguiente manera:

Situarse en una terminal de sistema operativo en el directorio donde se han almacenado los Scripts que hacen posible el funcionamiento del sintetizador y ejecutar el siguiente comando:

```
$ python synthezise.py “Hola mundo”
```

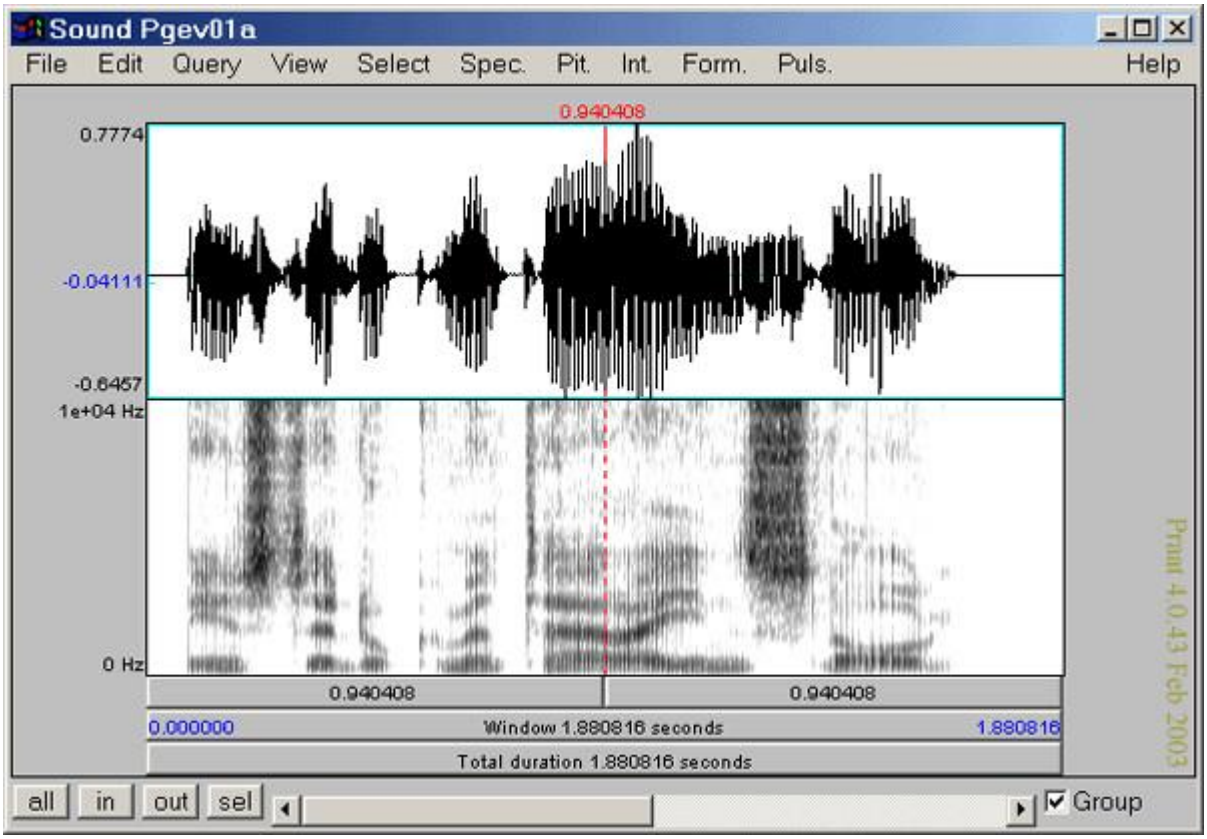
Donde el contenido de “Hola mundo” puede ser reemplazado por el texto que se quiera escuchar por la salida de voz de la computadora.

Si bien las entradas admitidas por el sintetizador construido en Python no son tantas como para Festival, podrían alterarse los scripts para que la fase de análisis de texto esté en la capacidad de recibir un archivo de texto completo en modo lectura como entrada.

Anexo 02: Ejemplo de Praat.

Praat es un programa para visualizar las ondas asociadas a cada fonema, basándose en la utilización de tiempos e imágenes.

Ilustración 5: Demostración de resultados obtenidos utilizando el programa Praat



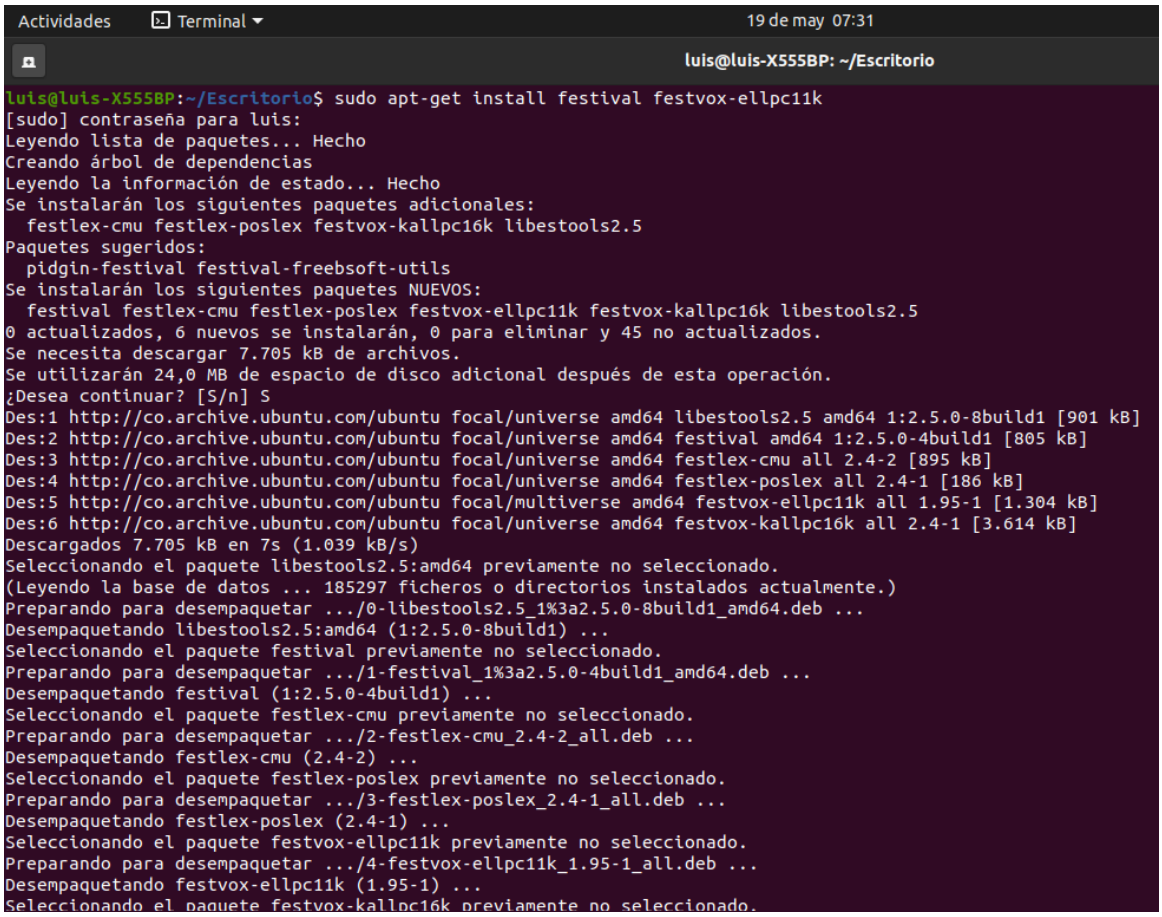
Fuente: Laboratorio de voz con Praat.[14]

[14] Blogspot: http://laboratoriodevoz.blogspot.com/p/praat.html

Anexo 03: Instalación del programa Festival y otras herramientas.

En la siguiente imagen se visualiza el comando utilizado para obtener Festival de la manera más fácil, donde basta con ingresar el comando, posteriormente ingresar “S” cuando se solicite la confirmación y esperar a que culmine la instalación.

Ilustración 6 Instalar festival en Ubuntu utilizando sudo



Fuente: Elaboración Propia.

Posteriormente cuando se haya instalado festival, es necesario probar que funcione correctamente, para ello se puede ingresar el comando mostrado en la siguiente imagen, con el cual se confirma que el programa se ha instalado correctamente, está listo para funcionar y cuenta con los comandos que se exponen.

Ilustración 7 Ingresando el comando help después de escribir festival en el terminal de Ubuntu

```
festival> help
"The Festival Speech Synthesizer System: Help

Getting Help
  (doc '<SYMBOL>)   displays help on <SYMBOL>
  (manual nil)      displays manual in local netscape
  C-c               return to top level
  C-d or (quit)     Exit Festival
(If compiled with editline)
  M-h               displays help on current symbol
  M-s               speaks help on current symbol
  M-m               displays relevant manual page in local netscape
  TAB               Command, symbol and filename completion
  C-p or up-arrow   Previous command
  C-b or left-arrow Move back one character
  C-f or right-arrow
                    Move forward one character
  Normal Emacs commands work for editing command line

Doing stuff
  (SayText TEXT)      Synthesize text, text should be surrounded by
                      double quotes
  (tts FILENAME nil)  Say contexts of file, FILENAME should be
                      surrounded by double quotes
  (voice_rab_diphone) Select voice (Britsh Male)
  (voice_kal_diphone) Select voice (American Male)
"
```

Fuente: Elaboración Propia.

Finalmente, solo hace falta instalar FestVox y Speech Tools, los cuales se encuentran disponibles junto con toda su documentación en el siguiente recurso:

http://festvox.org/index.html