RAE

1. **TIPO DE DOCUMENTO:** Trabajo de grado para optar por el título de INGENIERO DE SONIDO
2. **TÍTULO:** IMPLEMENTACIÓN DE SÍNTESIS CONCATENATIVA PARA CONVERSIÓN TEXTO A VOZ EN UN SISTEMA EMBEBIDO
3. **AUTOR:** Kelly Johanna Correales Ducuara
4. **LUGAR:** Bogotá, D.C.
5. **FECHA:** Octubre de 2015
6. **PALABRAS CLAVE:** Análisis y procesamiento de señales, fono, identificación de fonemas, difono, trifono, sílaba, síntesis concatenativa de voz, sistema embebido, corpus, dispositivo de conversión texto a voz, Pure Data, discapacidad para hablar.
7. **DESCRIPCIÓN DEL TRABAJO:** El objetivo principal de este proyecto es desarrollar un dispositivo de conversión texto a voz para personas con discapacidad del habla, implementando síntesis concatenativa en un sistema embebido. La implementación se realizó con el sistema embebido Raspberry Pi que procesa mediante Pure Data, los mensajes provenientes del texto ingresado en un teclado inalámbrico, para realizar la conversión texto a voz utilizando síntesis concatenativa. Para esto, el texto es analizado fonéticamente y relacionado con un conjunto de audios que al reproducirse consecutivamente en un altavoz, generan un mensaje sonoro. Estos audios corresponden a sílabas grabadas y editadas para conformar el corpus del sistema. El texto se muestra en una pantalla LCD mediante programación en Python.
8. **LÍNEA DE INVESTIGACIÓN:** Línea de Investigación de la Universidad: Tecnologías Actuales y Sociedad. Sub línea de Facultad de Ingeniería: Análisis y Procesamiento de Señales. Campo Temático del Programa: Control.
9. **METODOLOGÍA:** Es de carácter empírico – analítico debido a que los resultados son obtenidos mediante pruebas de ensayo y error, los cuales son evaluados según la teoría de la síntesis concatenativa para lograr la conversión texto a voz y que este proceso funcione adecuadamente en un sistema embebido.
10. **CONCLUSIONES:** Con la extracción de fonos y difonos se concluyó que entre más pequeña es la unidad mayor es la variación en el tono espectral. Por ende, se decidió crear un corpus compuesto por sílabas ya que proporcionan mejor inteligibilidad por ser unidades sonoras más largas en tiempo y al no ser extraídas de palabras, no tienen contexto ni acento, brindando continuidad entre fonos. Con la implementación de síntesis concatenativa en Pure Data se comprueba que el software libre es una herramienta óptima para el desarrollo de dispositivos que realicen procesamiento digital de señales. La generación de 160 palabras se debe a la limitación en la cantidad de audios a abrir en Pure Data en la arquitectura ARMHF del sistema embebido y a su procesador. El análisis de los resultados obtenidos por las encuestas, demuestra que el dispositivo tiene que mejorar aspectos de ergonomía; su funcionamiento cumple con los requisitos de un sistema de conversión texto a voz y es útil para personas con discapacidad para hablar.

IMPLEMENTACIÓN DE SÍNTESIS CONCATENATIVA PARA CONVERSIÓN TEXTO A VOZ EN UN SISTEMA EMBEBIDO

KELLY JOHANNA CORREALES DUCUARA

UNIVERSIDAD DE SAN BUENAVENTURA
FACULTAD DE CIENCIAS BÁSICAS E INGENIERÍA
INGENIERÍA DE SONIDO

BOGOTÁ, D.C - 2015

IMPLEMENTACIÓN DE SÍNTESIS CONCATENATIVA PARA CONVERSIÓN TEXTO A VOZ EN UN SISTEMA EMBEBIDO

KELLY JOHANNA CORREALES DUCUARA

Trabajo presentado como requisito parcial para optar por el título de profesional en Ingeniería de Sonido

Asesor: Ingeniero Miguel Ricardo Pérez Pereira

UNIVERSIDAD DE SAN BUENAVENTURA
FACULTAD DE CIENCIAS BÁSICAS E INGENIERÍA
INGENIERÍA DE SONIDO

BOGOTÁ, D.C - 2015

Nota de aceptación

_____
_____
_____
_____
_____
_____

_____
Presidente del jurado

_____
Jurado

_____
Jurado

Bogotá, ____ / ____ /_____

Dedicatoria

A Dios por haberme permitido llegar hasta aquí, brindándome salud, sabiduría y perseverancia.

A mis padres Orlando y Trinidad y a mis abuelos Antonio y Ana, por darme su amor, comprensión, confianza y apoyo incondicional.

A mi hermanita Ana María por ser una bendición en mi vida y por ser una de las razones para seguir adelante.

Al resto de la familia, Fernando, Francisco, Carolina, Sofía, María Fernanda y Luciana por su amor y amistad.

A mi novio Andrés porque con su amor me ha enseñado que los retos de la vida engrandecen corazón, mente y espíritu.

Agradecimientos

A Dios por haber guiado mis pasos para hacer este proyecto posible.

A mi familia y a mi novio por brindarme su amor, creer en mí y apoyarme durante la realización de este proceso.

A mi director de tesis Miguel Pérez y a los ingenieros Andrés Caballero y Lorena Aldana por su tiempo, paciencia, apoyo y orientación aún en la distancia.

A la Universidad de San Buenaventura Seccional Bogotá por abrirme sus puertas y brindarme las herramientas y conocimientos necesarios para ser una Ingeniera de Sonido.

Tabla de Contenido

Lista de Tablas

Lista de Figuras

Lista de Apéndices

Glosario

Alófono: Realizaciones sonoras de un mismo fonema dadas en contextos fonéticos diferentes.

Afasia: Incapacidad total o parcial para usar el lenguaje.

Consonante Lateral: Sonido producido mediante vibraciones en órgano articulador.

Consonante Oclusiva: Sonido consonántico obstruyente producido por una detención del flujo de aire y por su posterior liberación.

Consonante Rótica: Sonido consonántico producido por la obstrucción hecha a lo largo del eje longitudinal de la lengua.

Corpus: Base de datos o inventario de archivos de audio.

Crossfade: Transición entre dos audios mediante fade in/out.

Difono o difonema: Segmento de voz va desde el centro acústico de un fonema hasta el centro acústico del siguiente.

Fade In/Out: Aumento o disminución gradual de nivel de una señal de audio.

Fonema: Unidad segmental de una determinada lengua.no e

Fono: Sonido que representa las características acústicas particulares de los fonemas.

Frame: Segmento pequeño que contiene información.

Laringetcomía: procedimiento quirúrgico que se realiza para extirpar la laringe cuando se diagnostica cáncer

Oclusión: Cierre completo y momentáneo del canal articulatorio que se produce al pronunciar las consonantes oclusivas.

Patch: Conjunto de comandos programados para el entorno de Pure Data.

Prosodia: Se refiere al acento, ritmo y entonación de una palabra.

Pure Data: Lenguaje de programación visual de código abierto que permite crear software gráficamente sin líneas de código

Python: Lenguaje de programación.

Síntesis Concatenativa: Técnica de síntesis de voz que consiste en concatenar secciones de una señal grabada y previamente dividida en unidades, de tal forma que el sonido sea continuo e inteligible.

Sistema Embebido: Sistemas programables, diseñados para realizar tareas específicas determinadas por el usuario, con el fin de optimizar los procesos para mejorar su desempeño y eficiencia, reduciendo tamaño y costos de producción. Se caracterizan por el bajo consumo de energía, son económicos y poseen periféricos limitados.

Subpatch: Es un objeto de Pure Data que se puede comparar a un cajón o a un contenedor que tiene en su interior código de programación.

Superusuario: Cuenta que permita administrar un sistema operativo.

Tono: Frecuencia fundamental de la voz. Se conoce también como *pitch.*

TTS: Text to Speech. Texto a voz.

Trifono o trifonema: Fragmento de voz que contiene la transición acústica entre tres fonemas.

Introducción

Durante el S. XX, se realizaron investigaciones enfocadas al procesamiento digital de voz para identificar sus principales características; en consecuencia, surgieron técnicas de síntesis de voz aplicables a diferentes campos; entre ellas, la conversión de texto a voz implementada en dispositivos móviles y computadores. Actualmente, los sistemas de conversión texto a voz son de gran utilidad, no solo para facilitar algunas tareas, sino también para permitir la comunicación entre personas con discapacidad para hablar. A raíz de esto, han surgido proyectos en diferentes universidades del mundo que dan solución a esta problemática mediante sistemas implementados con diferentes técnicas de síntesis de voz. Entre esos proyectos se encuentran sistemas de concatenación basado en MIDI (Macon, Jensen-Link, Oliveiro, Clements & George, 1997), aplicación móvil TTS (Áviles, 2012), conversión texto a voz durante llamadas por celular (Rueda, Correa, Arguello, 2012), entre otros.

Los sistemas embebidos son herramientas programables y diseñadas para realizar tareas específicas determinadas por el usuario, con el fin de optimizar los procesos y así mejorar su desempeño y eficiencia, reduciendo tamaño y costos de producción. Son utilizados en el campo de la ingeniería para desarrollar sistemas específicos relacionados con procesamiento de señales, procesos de control y comunicación.

Este proyecto está dirigido a personas con discapacidad para hablar con el fin de realizar una aproximación de comunicación básica, es decir, la interacción se lleva a cabo mediante palabras y frases fundamentales de la comunicación según una publicación del gobierno de Austria (Austria, 2014). Las frases se generan a través de síntesis concatenativa por selección desarrollada en software libre e implementado en un sistema embebido.

Capítulo 1.
Planteamiento del Problema

Antecedentes

Sistema TTS. El sistema Text-To-Speech (TTS) fue inicialmente desarrollado en 1779 por el científico danés Christian Kratzenstein, quien construyó en la Academia Rusa de Ciencias, un modelo del tracto vocal humano el cual podía producir las cinco vocales (Universidad Tecnológica de Helsinki, 2006). En 1791, Wolfgang von Kempelen de Hungría creó una máquina acústico-mecánica capaz de generar sonidos del habla; se basó en el modelo de Kratzenstein, agregándole una lengua y labios con el fin de producir no solo vocales, sino también consonantes (Degen, sf). Este modelo fue retomado en 1923 (Mattingloy, 1974). En 1837, Charles Wheatstone mejoró el modelo de Von Kempelen incorporando el canto. En 1930, Bell Laboratories desarrolló el vocoder[1] el cual analizaba la frecuencia fundamental y sus armónicos. Tiempo después, Homer Dudley creó el Voder[2], un sintetizador de voz controlado por un teclado cuyo funcionamiento consistía en un radio a válvula, que se encargaba de generar los sonidos de las vocales y en un ruido de siseo producido por un tubo de gas para generar consonantes. Estos sonidos eran filtrados, amplificados, mezclados, modulados y finalmente, reproducidos. En los años 40 se llevaron a cabo investigaciones sobre la percepción acústica de los fonemas[3], a cargo de Alvin Liberman. En la década de los ochenta y noventa, MIT[4] y Bell Laboratories emplearon métodos de procesamiento del lenguaje natural para crear sistemas multilenguaje. Los sistemas Text–To–Speech,aparecieron en los años 60 gracias a la síntesis de voz. Noriko Umeda de Electrotechnical Laboratories desarrolló el primer sistema text-to-speech de inglés (Klatt, 1987).

LYRICOS. En la convención de la AES de 1997 llevada a cabo en Nueva York, se presentó un artículo de concatenación basada en MIDI, implementado en un sistema

[1] Sistema de análisis y síntesis capaz de generar voz.
[2] Vocoder Demonstration.
[3] Unidades segmentales de una determinada lengua
[4] Massachusetts Institute of Technology

sintetizador de voz en el canto. El sistema emplea TTS con síntesis concatenativa, el cual recibe un archivo MIDI, selecciona las unidades del inventario de modelos de voz sinusoidales para representar las características fonéticas[5], ejecuta un algoritmo y sintetiza. El modelo sinusoidal se usa para modificar el pitch, la duración y las características espectrales de las unidades concatenadas, tal cual se especifica en el archivo MIDI (Macon, Jensen-Link, Oliveiro, Clements & George, 1997).

Sistema TTS para idioma mandarín. En el año 2000, un grupo de investigadores en Taiwán crean un sistema TTS para el idioma mandarín cuya unidad básica de síntesis es el monosílabo. Esta unidad es elegida de una base de datos de habla continua, garantizando que la distorsión entre sílabas es mínima y que el tono, energía y duración sean constantes en diferentes combinaciones (Wu & Chen, 2000).

Sistema TTS para idioma mongol. En la década del 2000, se implementa en Mongolia el primer sistema TTS del idioma mongol basado en la síntesis concatenativa de difonos[6], aplicando la técnica TD-PSOLA[7]. La evaluación del sistema fue realizada por hombre y mujeres, quienes escribían lo que entendían de las palabras y oraciones que escuchaban. Se obtuvo un 86 % de inteligibilidad de palabras y un 60 % de inteligibilidad de oraciones. En cuanto a la naturalidad, el 37 % de las personas consideraron que la voz era natural, el 40 % afirmaron que la naturalidad de la voz era aceptable y el 23 % consideraron que la voz no era natural (Davaatsagaan & Paliwal, 2008).

Sistema TTS para idioma hindú implementado en MATLAB[8]. En el año 2012, unos investigadores de la India presentan un método para diseñar un módulo de conversión texto a voz en MATLAB mediante operación de matrices y vectores. Las palabras son grabadas en MATLAB, luego exportadas en formato .wav y finalmente se dividen en muestras para identificar sus fonemas. El diccionario de palabras es almacenado en el directorio de MATLAB y posteriormente usado para realizar la división de fonemas del texto de entrada. La síntesis

[5] Fonética: Parte de la lingüística que se encarga de estudiar los sonidos físicos del discurso humano.
[6] Unión de dos fonos. Fono: Sonido que representa las características acústicas particulares de los fonemas.
[7] Técnica de procesamiento de señales en el dominio del tiempo encargada de modificar el pitch y la duración.
[8]Lenguaje y entorno de programación de alto nivel.

consiste en comparar los fonemas del texto de entrada y de la señal, de tal forma que cuando se encuentre una correlación se concatenen los vectores de los fonemas extraídos de las grabaciones para formar palabras. Esta implementación es simple y no requiere de mucho espacio de almacenamiento (Patra, Patra, Mohapatra, 2012).

Up' N Talk. En el Instituto Politécnico Nacional (IPN) de México diseñaron un dispositivo denominado Up'N Talk, el cual funciona con una aplicación en el teléfono celular, desde donde se envía el texto mediante bluetooth hacia otro dispositivo. Este dispositivo tiene un receptor conectado a un amplificador, el cual emite el sonido a través de un altavoz pequeño. Todo el prototipo fue construido con materiales reciclables. (Áviles , 2012)

Sistema TTS bilingüe compatible con diferentes interfaces. La universidad Politécnica de Cataluña desarrolló un sistema TTS que consta de un núcleo y diferentes interfaces compatibles con aplicaciones de teléfono, de Windows y de internet; actualmente, se ejecuta en UNIX, MacOS y Windows32. Es un sistema bilingüe capaz de leer texto en español y catalán debido a que sus fonemas son similares. El sistema TTS se basa en los Modelos Ocultos de Markov (HMM) y en la concatenación de difonos (Bonafonte, Esquerra, Febrer, Fonollosa, Vallverdú).

Audiotexto. En la Universidad Autónoma de Occidente de Cali, implementaron un software de conversión de texto a voz mediante síntesis por regla y composición alofónica. El software se reconoce como Audiotexto, el cual fue realizado en Visual Basic 6.0 junto con macros para Microsoft Word. La síntesis de voz consistió en grabar fonemas[9] por separado y concatenarlos según la palabra a generar (D & Agudelo sf).

Modelo acústico fonador para vocales masculinas en español. En el 2007, una estudiante de Ingeniería de Sonido de la Universidad de San Buenaventura de Bogotá presenta su proyecto de grado cuyo nombre es Modelo Acústico Fonador para Vocales

[9] Realizaciones sonoras de un mismo fonemas dadas en contextos fonéticos diferentes

Masculinas en Español. El objetivo de la investigación es emular las vocales humanas estableciendo patrones de comportamiento espectral de la voz masculina en español, mediante mediante una teoría multi-cuadro propuesta por Yoshinori Shiga, por medio de la cual se conseguiría una función de transferencia que permitiera la construcción de un circuito eléctrico y eléctrico y acústico, con el fin de generar sonidos para personas laringectomizadas[10]. Infortunadamente no fue posible construir eléctricamente el circuito pero se pudo modelar el aparato fonador en MATLAB (Pardo, 2007).

Síntesis de voz durante llamadas de celular. En el año 2012, la Universidad Industrial de Santander desarrolló un software capaz de realizar llamadas de celular, usando síntesis de voz mediante concatenación de difonos. El objetivo es permitir a las personas comunicarse por llamadas telefónicas en situaciones donde es imposible contestar el celular. Por ende, el software consiste en recibir el texto de entrada proveniente de un computador para luego analizarlo y realizar síntesis. Esta señal se envía a través de un dispositivo de transmisión el cual se conecta al computador y al micrófono del celular. El 100 % de las personas que probaron el sistema afirmaron la señal era inteligible. Los autores de este artículo aseguran que el dispositivo de transmisión debe optimizarse usando Bluetooth (Rueda, Correa, Arguello, 2012).

Descripción y Formulación del Problema

En Colombia, la implementación de dispositivos portátiles capaces de realizar conversiones texto a voz aún no se ha posicionado; la Universidad Industrial de Santander y la Universidad Autónoma de Cali han desarrollado algún tipo de software que realice este proceso. Sin embargo, el software es ejecutado únicamente en ordenadores, perdiendo la capacidad de ser un aplicativo portátil. Debido a esto, el usuario (persona con discapacidad para hablar) dependería de computadores de escritorio o computadores portátiles.

La necesidad de comunicar mensajes básicos como un saludo, una advertencia o una solicitud de ayuda, resulta ser un factor importante para las personas con discapacidad para hablar, es decir, aquellas que fueron laringectomizadas, o que tienen algún tipo de afasia[11] o que

[10] Personas a las cuales se les extirpó la laringe
11 Incapacidad total o parcial para usar el lenguaje.

nacieron con deficiencias auditivas, impidiéndoles producir sonidos; como es el caso de las personas sordas y mudas. Las personas mudas utilizan la lengua de señas para expresarse solo con quienes tienen conocimiento sobre esta lengua, lo cual representa una minoría de la población. Las personas laringectomizadas experimentaron la capacidad de hablar en algún momento de sus vidas y por cuestiones médicas, no lo pueden volver a realizar de forma natural; por ende, se expresan a través de la escritura, gestos o señas. Las personas con afasia se comunican por medio de la escritura y/o habla dependiendo del tipo de enfermedad. Pero desafortunadamente la comunicación no se lleva a cabo con éxito en la mayoría de casos, así que el desarrollo de un sistema que permita aproximarse a una comunicación básica para personas con discapacidad para hablar, es una buena alternativa para complementar su medio de comunicación

Teniendo en cuenta lo anterior, se plantea la siguiente pregunta: ¿Cómo se puede desarrollar un sistema TTS[12] como una herramienta de comunicación básica para personas con discapacidad para hablar, implementando software libre en un sistema embebido?

Justificación

Las personas con discapacidad para hablar, emplean texto, gestos y señas (personas laringectomizadas), lengua de señas (personas mudas), escritura y/o habla incoherente (personas con afasia) para comunicarse. Sin embargo, la comunicación es limitada. En el caso de las personas mudas, solo un sector de la población entiende el lenguaje de señas, lo cual cohíbe su forma de expresión y comprensión; las personas laringectomizadas, no logran comunicarse con facilidad, mientras que las personas con algún tipo de afasia se comunican mediante la escritura o habla incoherente dependiendo de las limitaciones de su enfermedad. Por esta razón, es necesario implementar un dispositivo que facilite la comunicación de mensajes básicos a personas incapaces de entender un lenguaje no verbal. Partiendo del anterior supuesto, la conversión TTS proporciona un beneficio importante en cuanto a la generación de palabras, debido a que no es predecible y permite que la interacción se lleve a cabo con mayor libertad, con el fin de lograr que las personas con discapacidad del habla amplíen sus alcances de comunicación. Este dispositivo tendría impacto en la población descrita más adelante en las estadísticas.

[12] Text to Speech: conversión texto a voz.

Un sistema embebido, es un dispositivo óptimo y eficiente para el diseño de aplicaciones específicas, de tamaño reducido, liviano, portable y de bajo consumo de energía. En comparación con una aplicación para dispositivos móviles de conversión texto a voz, el sistema embebido procesa audio eficientemente gracias a su conversor digital-análogo; mientras que un dispositivo móvil tiene limitaciones para este tipo de procesamiento, debido a la capacidad de procesamiento y/o versión del sistema operativo. Asimismo, las aplicaciones móviles necesitan de una conexión a internet para ser descargadas. Por ende, se utilizará un sistema embebido para el desarrollo del dispositivo de conversión TTS.

La implementación de software libre permite un desarrollo óptimo, porque este es de libre distribución, modificación y copia. Una ventaja del software de código abierto es la posibilidad de compilar este software en diferentes dispositivos, por ejemplo sistemas embebidos; o diferentes distribuciones y arquitecturas de Linux. Por tal razón, el código empleado para el diseño del dispositivo de conversión TTS se va a implementar en software libre de código abierto.

Estadísticas. Según un estudio realizado por GLOBOCAN en el 2012, Colombia tiene alrededor de 720 personas con cáncer de laringe y 431 muertes por esta enfermedad. El Instituto Nacional Cancerológico (INC) reportó entre el año 2002 y 2006, 876 casos de cáncer de laringe en hombres y 183 en mujeres, para un total de 1059 casos en todo el país (INC, 2006). En el 2011, recibió 40 casos nuevos de cáncer de laringe de los cuales solo 4 fueron llevados a cirugía. (INC, 2011). Además, según el Congreso Nacional de Otorrinolaringología y Cirugía de cabeza realizado en Cartagena el 30 de mayo del 2004, las prótesis que permiten a los laringectomizados hablar de nuevo son fabricadas en Suecia y tienen una vida útil de un año, con un costo promedio de 1.5 millones de pesos. En noviembre del 2005, el DANE realizó un censo en 1098 municipios y 20 corregimientos departamentales, por medio del cual se concluyó que de cada 100 colombianos el 13.2 % tiene limitaciones permanentes para hablar. (DANE, 2005). En marzo del 2005, el Registro de Localización y Caracterización de la Población con Discapacidad, reportó 337.892 corresponden a discapacidad de la voz y el habla, de las cuales el 50% no han alcanzado ningún nivel educativo Solo el 22% de estas personas se encuentran en el mercado laboral, 6% están buscando trabajo, 46% se encuentran en condición de incapacidad permanente y sin

pensión, el 2% están incapacitados permanentemente para trabajar con pensión y el 30% se encuentran estudiando.

Objetivos

Objetivo General.

Desarrollar un dispositivo de conversión TTS para personas con discapacidad del habla, implementando síntesis concatenativa en un sistema embebido.

Objetivos Específicos.

Realizar un banco de 18 palabras en español a partir de grabaciones de voz e identificar sus fonemas.

Implementar síntesis concatenativa para la generación de 34 palabras mediante software libre.

Implementar la síntesis concatenativa y la interfaz de entrada de comandos de texto en el sistema embebido.

Evaluar el dispositivo con personas con discapacidad para hablar.

Alcances y Limitaciones

Alcances.

El dispositivo será capaz de realizar conversión texto a voz de palabras y frases ingresadas al sistema, permitiendo que las personas con discapacidad para hablar se comuniquen ampliamente con individuos sin limitaciones auditivas mediante conversaciones fluidas, rompiendo de esta forma la barrera de comunicación que esta población enfrenta.

Limitaciones.

Se grabarán 18 palabras provenientes de una voz masculina, para extraer los fonos que serán utilizados en la conversión texto a voz de 34 palabras en español, con el fin de poner a

prueba el funcionamiento del sistema TTS y de esta forma, establecer el procedimiento adecuado que permita ampliar el número de palabras en futuros proyectos. La programación del sistema TTS se realizará en Pure Data y dependiendo de la capacidad del procesamiento del sistema embebido, se puede lograr un mayor o menor rendimiento.

Capítulo 2.
Metodología

Enfoque de la Investigación

El enfoque de esta investigación es empírico - analítico, debido a que los resultados son obtenidos mediante pruebas de ensayo y error, los cuales son evaluados según la teoría de la síntesis concatenativa para lograr la conversión texto a voz y que este proceso funcione adecuadamente en un sistema embebido.

Línea de Investigación de la Universidad/ Sub-Línea de la Facultad/ Campo Temático el Programa

De acuerdo a los campos de investigación determinados por la Universidad de San Buenaventura sede Bogotá, el proyecto hace parte de la línea de investigación *Tecnologías Actuales y Sociedad*, ya que al desarrollar síntesis concatenativa para conversión texto a voz en software libre implementado en un sistema embebido, se soluciona una problemática actual a partir de la implementación de tecnología.

La sublínea de investigación de la facultad es *Análisis y Procesamiento de Señales*, debido a que se trabaja con señales de audio que serán analizadas y modificadas.

El campo temático del programa de Ingeniería de Sonido al que pertenece este proyecto, es *Control* debido a que el usuario ingresa las palabras a sintetizar mediante un teclado y son procesadas en un sistema embebido.

Hipótesis

El desarrollo de un sistema TTS que sirva como herramienta de comunicación para personas con discapacidad para hablar toma como comandos de entrada las palabras a sintetizar a través de un teclado, posteriormente estas son segmentadas en unidades básicas del lenguaje (fonemas) y son unidas entre sí para formar difonos[13], trifonos[14] y sílabas, con el fin de ser

[13] Segmento de voz va desde el centro acústico de un fonema hasta el centro acústico del siguiente.
[14] Fragmento de voz que contiene la transición acústica entre tres fonemas.

comparados con una base de datos de grabaciones de voz y de esta forma, relacionar el texto a sintetizar con los fragmentos de audio que conforman la base. El análisis de los segmentos de voz se lleva a cabo con técnicas de procesamiento digital de señales para modificar nivel, longitud y tono.

Variables

- Desarrollo de dispositivo de conversión texto a voz.
- Complementación de la comunicación básica de personas con discapacidad para hablar.
- Implementación de síntesis concatenativa para el desarrollo de sistema de conversión texto a voz.

Capítulo 3.
Marco Teórico

El Habla

Producción del Habla.

El habla generalmente se produce por la contracción de los pulmones, la expulsión de aire, el cual tiene consigo un sonido correspondiente a la distribución de frecuencias Gaussiana, y en general por nuestro aparato fonador conformado por: nariz, paladar, lengua, faringe, epiglotis, laringe, tráquea, clavícula, pulmones, cavidad torácica y diafragma. El aire viaja a través de los bronquios y luego a través de las cuerdas vocales en la cima de la tráquea, de modo que genera una vibración en ellas. En la parte posterior de la boca, se encuentran dos 'caminos' los cuales permiten que el aire salga al exterior. El primero se encuentra por encima de la lengua, pasando por los dientes y termina en la parte exterior de la boca. El segundo camino se encuentra en la cavidad nasal, como se muestra en la Figura 1.

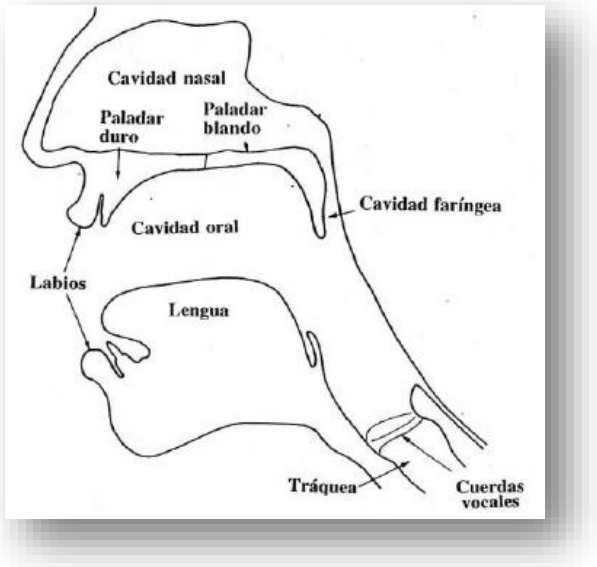

Figura 1. Tracto vocal del ser humano.

Fuente: http://liceu.uab.es/~joaquim/phonetics/fon_produccio/articulacion.html

Los sonidos generados por este proceso dependen de ciertos criterios:

1. La fuerza de los pulmones es responsable del volumen de los sonidos así como también de las pausas entre las sílabas, en especial cuando el volumen es variable.
2. El cierre de la glotis[15] genera unos sonidos explosivos en el final de una palabra.
3. La tensión de los músculos que sostienen las cuerdas vocales logran que estas vibren a diferentes frecuencias.
4. El timbre de los sonidos cambia cuando el aire viaja por la cavidad nasal y por la boca.
5. Cuando el aire viaja a través de la boca y la mandíbula se cierra, se genera una vocal, de lo contrario, una deslizada
6. Si la lengua toca el paladar o los dientes se generan gran variedad de sonidos.

Estructura Del Habla.

Los *fonemas* son las unidades estructurales más pequeñas del habla. Son sonidos que permiten distinguir palabras en una lengua. Así, los sonidos [p] y [b] son fonemas del español porque existen palabras como /pata/ y /bata/ que tienen significado distinto y su pronunciación sólo difiere en relación con esos dos sonidos.

Desde un punto de vista estructural, el fonema pertenece a la lengua, mientras que el sonido pertenece al habla. La palabra (casa), por ejemplo, consta de cuatro fonemas (/k/, /a/, /s/, /a/). A esta misma palabra también corresponden en el habla, cuatro sonidos, a los que la fonología denomina *alófonos*[16], y estos últimos pueden variar según el sujeto que lo pronuncie. La distinción fundamental de los conceptos fonema y alófono, está en que el primero es una huella psíquica de la neutralización de los segundos que se efectúan en el habla.

- Los fonemas son diferenciadores, es decir, cada fonema se define dentro del sistema por las cualidades que se distinguen de los demás.
- Los fonemas son indivisibles, es decir, no se pueden descomponer en unidades menores.

[15] Parte más estrecha de la laringe en donde se encuentran las cuerdas vocales
[16] Realizaciones sonoras de un mismo fonema dadas en contextos fonéticos diferentes.

- Los fonemas son abstractos, es decir, no son sonidos sino modelos de sonidos.

El conjunto de fonemas forman *sílabas* las cuales determinan el ritmo natural de las palabras. Constan de un inicio, un núcleo y un final; el inicio y el final generalmente se conforman de consonantes, mientras que el núcleo se conforma de una vocal. Las *consonantes* son generadas por contracciones en el tracto vocal mientras las *vocales* por una expansión de este.

Características Del Habla.

El habla sigue parámetros generales establecidos por la escucha, las limitaciones de la generación de la voz y el ambiente debido a que las características del habla dependen de la fisionomía y habilidades del cuerpo humano.

Clasificación del Habla.

Los sonidos del habla se pueden describir en términos del tono y las frecuencias resonantes (formantes). Los formantes son frecuencias de resonancia del tracto vocal los cuales se pueden identificar como picos en un espectrograma. Su ubicación en el espectrograma puede variar gracias a los cambios y movimientos que surgen en la boca a medida que transcurre el tiempo de la pronunciación. Los formantes tomados en cuenta corresponde a las tres primeras frecuencias (F1, F2, F3) que contribuyen a la inteligibilidad del habla: F1 contiene la mayor cantidad de energía y entre F2, F3 está la inteligibilidad. El tono de la voz es considerado como la frecuencia vocal fundamental (f0).

Tipos de Fonemas.

La clasificación de los fonemas se encuentra en el Apéndice A (Qullis, 1993).

Enfermedades Causantes de Discapacidad para Hablar

Laringectomía.

La laringectomía es un procedimiento quirúrgico que se realiza para extirpar la laringe cuando se diagnostica cáncer. Existen dos tipos de laringectomía: parcial y total. La laringectomía parcial consiste en la extirpación de solo la zona afectada por el tumor; la persona puede seguir hablando pero con menor intensidad después de la cirugía. En la laringectomía total se extirpa la laringe en su totalidad, desde la base de la lengua hasta llegar a la tráquea; en este caso, la persona pierde su voz como se observa en la Figura 2.

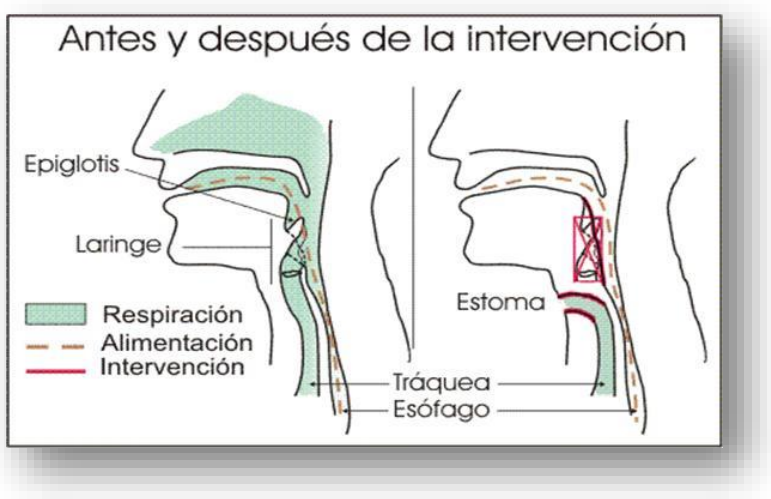

Figura 2. Antes y después de un laringectomía.

Fuente: Marín, M. (2013). Perder la voz tras un cáncer de laringe.

Sin embargo existen técnicas y prótesis especiales para las personas laringectomizadas. La técnica de la voz esofágica consiste en que el hablante introduce aire en su boca para conducirlo al esófago y posteriormente, impulsarlo hacia arriba para que el aire vibre en la neoglotis[17] y se produzca un sonido grave. (Marín, 2013). Las prótesis fonatorias son aparatos diseñados para unir la tráquea con el esófago.

[17] Estructura vibrátil que reemplaza a la glotis después de una laringectomía.

Afasia.

Es la incapacidad parcial o total para usar el lenguaje. Las personas con esta enfermedad tienen problemas para comprender lo que dicen los demás, problemas para leer, escribir u operar números. Las causas más frecuentes de la afasia son: daño cerebral por traumatismo cráneo encefálico o por apoplejía[18], incidencia insidiosa progresiva[19] e ictus[20].Las afasias que afectan el habla pero no la habilidad motriz de escribir son: Afasia de Broca y Afasia Transcortical Motora. La afasia de Broca o motora mayor, consiste en la pérdida motriz del lenguaje y la escritura; la persona dice palabras con lentitud y prosodia[21] inadecuada omitiendo artículos y preposiciones, además, pierde la motricidad en su mano derecha. La afasia transcortical motora consiste en la pérdida de iniciación y espontaneidad en el habla cuando la persona responde con frases. (Asociación Ayuda Afasia, 2014)

Sistema TTS (Text to Speech)

Un sistema Text-To-Speech consiste en la producción del habla implementada en dispositivos, mediante la fonetización[22] automática de oraciones (Dutoit, 1997).
El sistema Text-To-Speech fue inicialmente desarrollado en 1779 por el científico danés Christian Kratzenstein, quien construyó un modelo del tracto vocal humano el cual podía producir las cinco vocales. Años después se incorporó una boca, labios y nariz de goma al modelo ya existente, con el fin de producir consonantes. En 1837, Joseph Faber construyó un sistema compuesto por una faringe, la cual se utilizó principalmente para el canto y era controlado por medio de un teclado. Laboratorios Bell desarrolló un sintetizador electrónico de voz operado mediante un teclado; tiempo después Homer Dudely hizo una mejora sobre el Vocoder creandoel Voder. Haskin Laboratories logró convertir patrones acústicos del habla en sonidos a través de la generación de 50 armónicos de la frecuencia fundamental. Este proyecto estuvo a cargo del Dr. Franklin S. Cooper, John M. Burst y Caril Haskins en 1968. Desde los años 80, la mayoría de sistemas operativos han implementado sintetizadores de voz,

[18] Sangrado interno de un órgano
[19] Se refiere a enfermedades que progresan lentamente sin mostrar síntomas.
[20] Repentina sensibilidad o debilidad localizada en un lado del cuerpo.
[21] Se refiere al acento, ritmo y entonación de una palabra.
[22] Relación sonora entre el habla y la escritura.

específicamente, síntesis concatenativa. El siguiente gráfico resume el proceso que lleva a cabo un sistema Text-To-Speech.

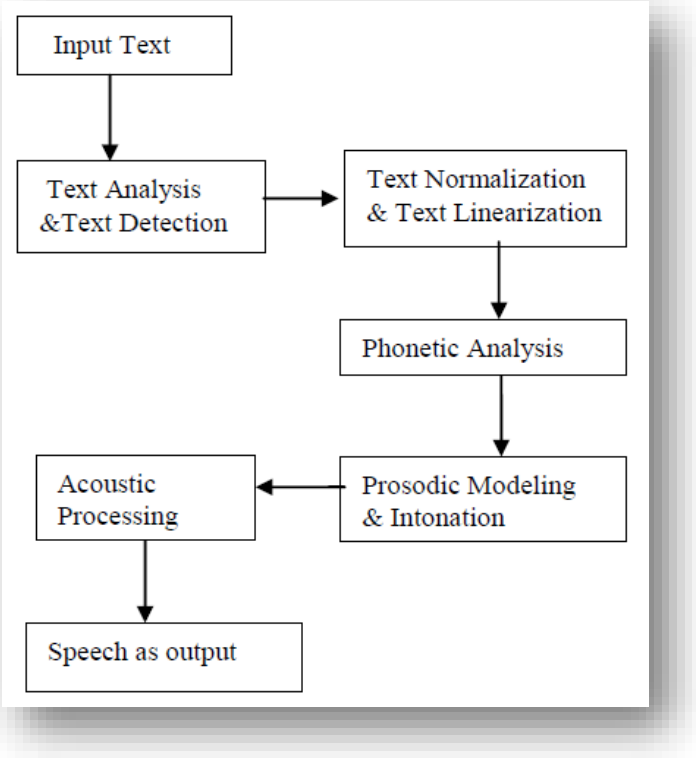

Figura 3. Sistema Text To Speech.

Fuente: Sasirekha, D., & Chandra, E. (2012). Text to speech: A simple tutorial. International Journal of Soft Computing and Engineering (IJSCE), 2(1).

Análisis y Detección De Texto.

El pre procesamiento consiste en analizar y organizar el texto de entrada en una lista de palabras, como números, abreviaciones y acrónimos. El análisis de texto se compone de tres módulos: morfológico, contextual y sintáctico. El módulo morfológico se encarga de identificar los morfemas[23] de las palabras, mientras que el módulo contextual determina el contexto de las palabras teniendo en cuenta la sintaxis. Por último se encuentra el módulo de sintaxis el cual descubre la estructura del texto. La detección permite ubicar las áreas de texto en documentos.

[23] Fragmento mínimo capaz de expresar un significado. Es también la secuencia de fonema s.

Análisis Fonético.

Los símbolos ortográficos se convierten en fonemas y alófonos mediante un alfabeto fonético, como por ejemplo, el del AFI (Alfabeto Fonético Internacional). (Ver Apéndice A)

Modelamiento de Prosodia y Entonación.

La prosodia abarca el acento, el ritmo y la entonación de una palabra, por ende, es fundamental para determinar el estado emocional del hablante El modelamiento de la entonación es importante debido a que afecta la naturalidad del habla.

Procesamiento Acústico.

Existen seis tipos de síntesis de voz:

- **Síntesis Concatenativa:** Una base de datos de palabras grabadas por voces humanas es empleada para realizar concatenación de estas, con el fin de formar oraciones. La principal ventaja es la naturalidad del habla.
- **Síntesis de Formantes:** Emplea filtros en paralelo y serie para generar voces con las frecuencias del tracto vocal, es decir, los formantes. Las voces suelen ser robóticas y artificiales.
- **Síntesis Articulatoria:** Consiste en modelar matemáticamente el tracto vocal humano y de esta forma lograr, una voz sintética.
- **Síntesis Sinusoidal:** Utiliza un modelo matemático para descomponer cada trozo de la señal, en armónicos de la frecuencia fundamental. Los parámetros de modelamiento son las amplitudes y los periodos de los armónicos.
- **Síntesis por Modelos Ocultos de Markov:** Los parámetros de la voz como el espectro de frecuencia, frecuencia fundamental y la duración son modelados estadísticamente para generar el habla.

- **Síntesis de Selección Unitaria:** La base de datos comprenden frases y palabras completas, con el fin de lograr mayor naturalidad en el habla reduciendo la variación fonética. Esta síntesis es una extensión de la síntesis concatenativa.

Técnicas para síntesis de voz

Existen tres clases de técnicas para realizar síntesis de voz, que dependen del fundamento empleado para realizar el proceso.

- **Técnica de Primera Generación:** Requiere de una descripción detallada de lo que se va a decir. En otras palabras, representa fonéticamente los componentes verbales teniendo en cuenta la duración de cada fono y su frecuencia fundamental (Fo). Para esto, se utilizan modelos que especifican las características de la fuente, con el fin de basarse en ellas y simular la voz. La síntesis de formantes y la síntesis articulatoria pertenecen a este grupo.
- **Técnicas de Segunda Generación:** Obtienen los parámetros de los componentes verbales mediante el análisis de una señal. Además, mejoran la calidad del habla gracias al uso reducido de modelamiento de pitch[24] y duración. La unidad predilecta para estas técnicas es el difono. La síntesis concatenativa y la síntesis sinusoidal corresponden a este grupo de técnicas.
- **Técnicas de Tercera Generación:** Utilizan una base de datos para analizar estadísticamente los parámetros del habla en las señales. A este grupo pertenece la síntesis por modelos ocultos de Markov y la síntesis de selección unitaria.

Difono

Es la unidad que inicia en la mitad de un fono hasta la mitad del siguiente (Taylor, 2009), generalmente tienen la misma duración de los fonemas, es decir, 100 ms. La principal ventaja en usar difonos para síntesis concatenativa de voz, es que permiten obtener el mismo estado del tracto vocal durante la transición que comprende la mitad de un fono hasta la mitad del siguiente

[24] Tono

fono. De esta forma no se presentan discontinuidades entre fonemas. Si un lenguaje tiene *N* fonemas, el número de difonos es N^2.

Para determinar el punto de inicio y final de un difono, se definen dos estados conocidos como *half-phones*, los cuales tienen su propia duración. Estos estados se denominan *estado 1* y *estado 2.* La frecuencia fundamental es independiente para cada estado y se determina justo en la mitad de cada fono (Figura 4).

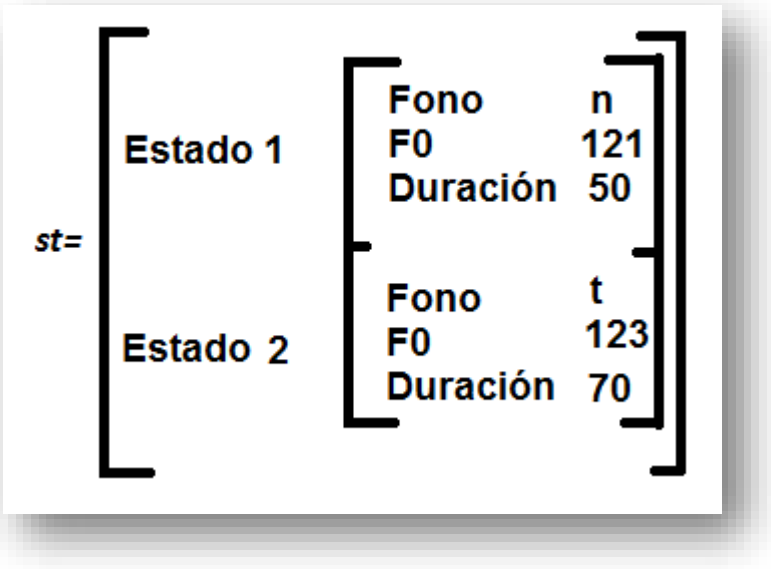

Figura 4. Estados que determinan el punto de inicio y final de un difono.
Fuente: Propia

La base de datos de grabaciones de voces debe abarcar los diferentes contextos, con el fin de asegurar todas las posibles combinaciones de difonos. La base de datos en donde se almacenan los difonos se denomina *Inventario de Difonos,* cuya creación depende de los siguientes criterios:

- La extracción de difonos se basa en la similitud acústica del último *frame*[25] del difono de la izquierda con el primer *frame* del difono de la derecha, que componen un fonema.

[25] Segmento pequeño que contiene información.

- El tono y la duración de los difonos son parámetros modificables que en algunos casos generan un efecto de distorsión. Por esta razón, es necesario elegir difonos que tengan valores promedio de tono y duración para minimizar este efecto.

Obtención de Difonos por medio del habla.

Grabar palabras por separado es una opción óptima para obtener difonos debido a que estos están incorporados en las palabras, conllevando a que se adquiera un control sobre su contexto fonético y prosódico. Las palabras pueden ser inventadas con el fin de abarcar todos los difonos necesarios. El siguiente paso, consiste en modificar el pitch y la duración de cada unidad para ajustar su prosodia.

Síntesis Concatenativa

Este tipo de síntesis concatena secciones de una señal grabada y previamente dividida en unidades, de tal forma que el sonido sea continuo e inteligible. La lista de palabras es grabada por un hablante y almacenada digitalmente, para después segmentar cada uno de sus elementos y crear el inventario de difonos. Las señales de audio traen consigo parámetros como duración, tono y frecuencia fundamental, los cuales son enlistados en el módulo NLP (Procesamiento Natural del Lenguaje) junto con sus respectivos difonos. La síntesis concatenativa tiene como ventaja la naturalidad de la señal concatenada, debido a que se conserva las características del sistema fonador en las unidades de concatenación; tiene como desventaja que en algunos casos al obtener secciones de palabras, estas no encajan en algunos casos en términos de amplitud, fase y espectro. La solución consiste en “suavizar” segmentos sucesivos. Existen tres tipos de síntesis concatenativa:

- **Por Selección:** Se concatenan unidades de voz previamente grabadas como sílabas, fonemas, palabras que se encuentran en una base de datos.
- **Por Difonos:** Se concatenan todos los difonos de una lengua contenidos en una base de datos.

- **Por Dominio:** Se utilizan palabras y frases grabadas para la concatenación y así generar mensajes completos.

Para este proyecto, se utiliza la síntesis de concatenación por selección.

Prosodia.

La prosodia se refiere a las propiedades de las señales del habla relacionadas con el tono, la longitud de las sílabas y la sonoridad. Al modificar la curva del tono, se altera la naturalidad del segmento y aún más, si las dos unidades tienen diferentes valores de tono. La longitud de los fonos con acento son más largos de lo normal El espectro de frecuencias en la parte final de los difonos es diferente para la mayoría de casos, por esta razón, es necesario ecualizar para lograr la misma amplitud espectral.

Sistemas Embebidos

Son sistemas programables, diseñados para realizar tareas específicas determinadas por el usuario, con el fin de optimizar los procesos para mejorar su desempeño y eficiencia, reduciendo tamaño y costos de producción. Se caracterizan por el bajo consumo de energía, son económicos y poseen periféricos[26] limitados (Caballero, 2012). Son implementados en diferentes sistemas operativos como Linux y Android.

Los sistemas embebidos están compuestos por un procesador, dispositivos de almacenamiento y periféricos. Algunos sistemas de uso general como la BeagleBoard, PandaBoard y Raspberry Pi utilizan un procesador de arquitectura ARM[27] , una memoria SD/Micro SD para almacenamiento y RAM.

El sistema embebido ejecuta tareas a través de un sistema operativo capaz de comunicar el hardware y software.

[26] Interfaces de conexión
[27] Procesador embebido que tiene una arquitectura especial funcional para algunos dispositivos.

- **Linux:** Sistema operativo multiplataforma, desarrollado para las arquitecturas x86, x86-64 y ARM. Por lo cual puede ser ejecutado tanto en computadores convencionales como en sistemas embebidos, aunque muchas de las aplicaciones para Ubuntu están para las tres arquitecturas (Caballero, 2012).
- **Android:** Plataforma móvil desarrollada en base Linux creada por Google, junto con aplicaciones middleware[28] está diseñada para ser utilizada en teléfonos inteligentes, tabletas, televisores, etc. basándose en la arquitectura ARM, x86 y x64. (Caballero, 2012)

Software de Código Abierto

El software de código abierto cumple con las siguientes condiciones:

- Libre redistribución.
- El código fuente debe estar disponible, como también la compilación del software.
- Modificación del software y distribución bajo la misma licencia de este.
- No discriminación para el uso del software.
- No discriminación para el campo de aplicación del software.
- Los derechos del programa no dependen de la licencia del software.
- La licencia permite que los programas realizados con el software no sean de código abierto.
- La licencia no debe limitarse aplicar a una determinada tecnología.

Dentro de las licencias de software de código abierto están la licencia GPL y MPL.

Pure Data

Es un lenguaje de programación visual de código abierto que permite crear software gráficamente sin líneas de código como se muestra en la Figura 5. Pure Data es usado generalmente para el procesamiento de señales de audio y video 2D/3D, recepción de mensajes

[28] Software que asiste a las aplicaciones para generar conexiones.

MIDI o de cualquier dispositivo de entrada. Puede trabajar con redes locales y remotas para integrar tecnología corporal, sistemas de motores, iluminación de plataforma, entre otros. Este lenguaje es multiplataforma y portable, compatible con GNU/Linux, Windows, Mac, entre otros sistemas operativos y es ejecutable en Raspberry Pi, tabletas, smartphones y computadoras. Algunas de sus librerías son desarrolladas por usuarios en el lenguaje C++. (Pure Data, s.f)[29].

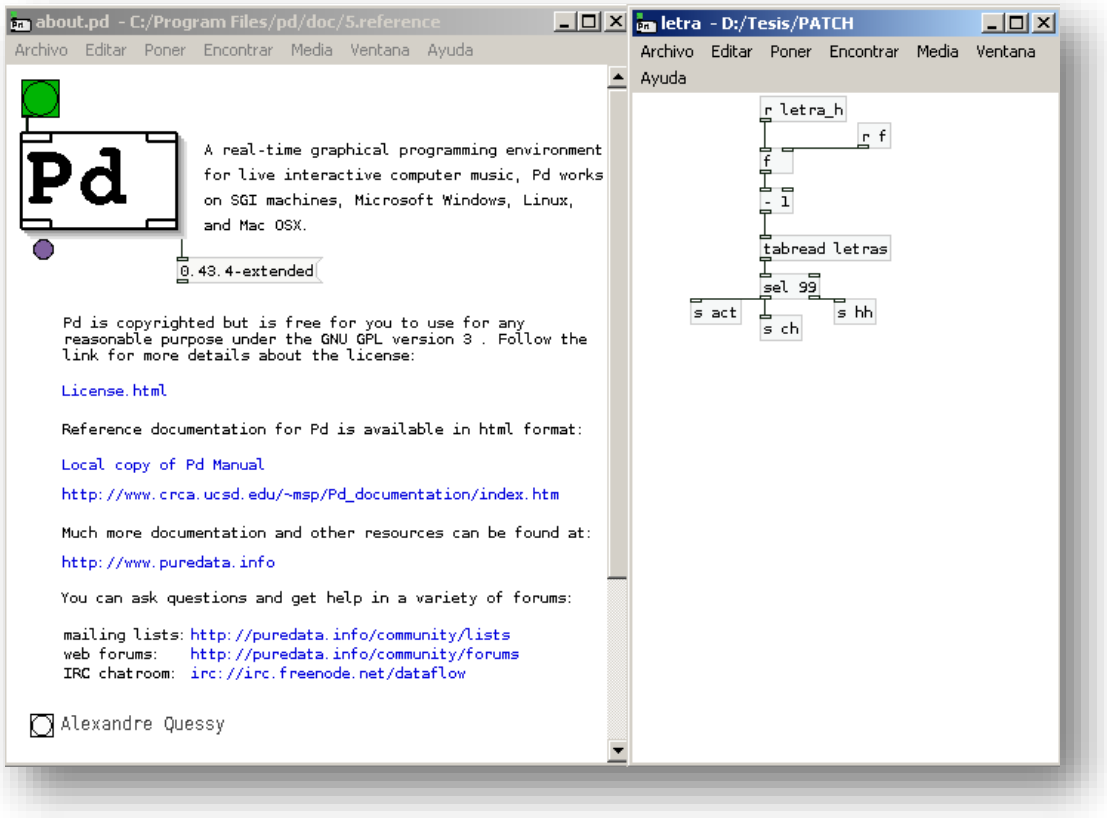

Figura 5. Interfaz Pure Data

Fuente: Propia.

Python

Es un lenguaje de programación multiparadigma[30] y mutiplataforma. Es administrado por la Python Software Foundation. Está regido bajo la licencia *Python Software Foundation License*

[29] https://puredata.info/
[30] Que soporta varios estilos de programación.

de código abierto, la cual es homologable por la Licencia pública general de GNU. Python tiene la gran ventaja de proporcionar la facilidad para leer sus códigos y programar sentencias mediante palabras en lugar de símbolos. Los códigos de este lenguaje se organizan mediante bloques, los cuales son delimitados mediante tabuladores. Es compatible con todos los sistemas operativos.

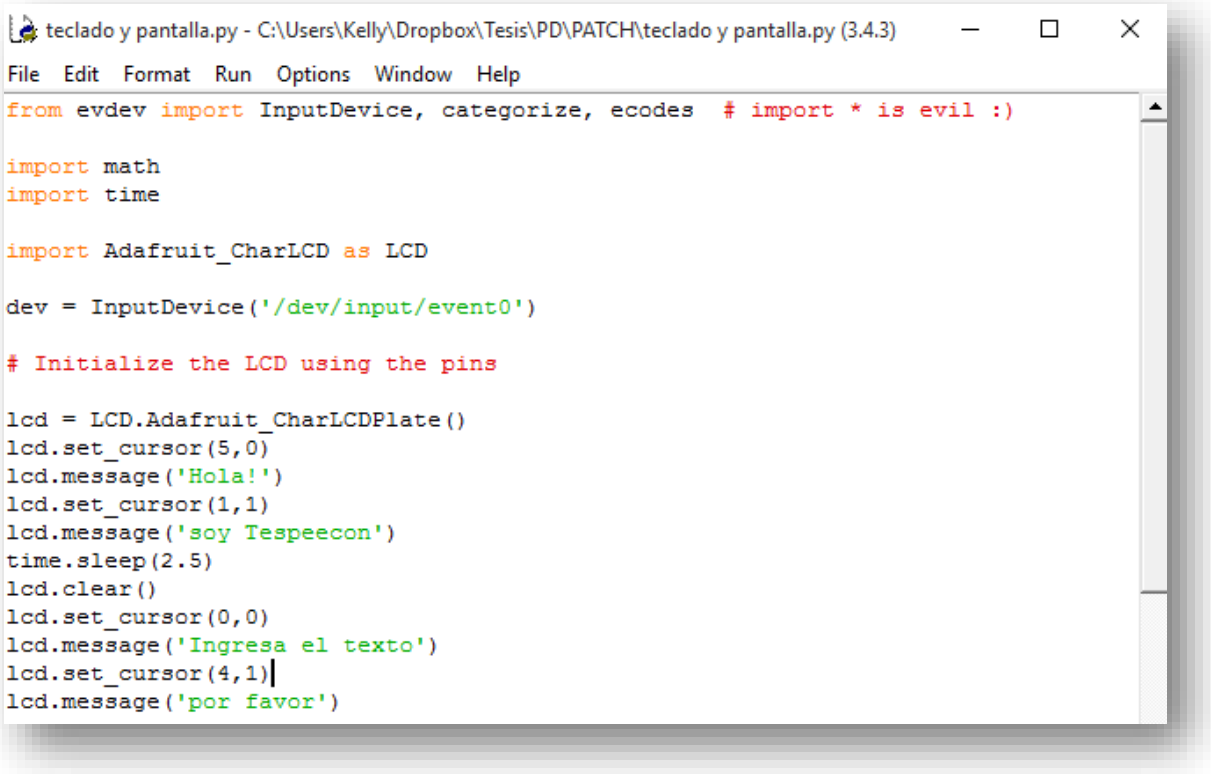

Figura 6. Entorno de programación de Python.

Fuente: Propia.

Capítulo 4.
Desarrollo Ingenieril

La implementación de síntesis concatenativa por selección para conversión texto a voz en un sistema embebido se desarrolla finalmente mediante un programa elaborado en Pure Data que recibe mensajes de un teclado para mapearlos como letras, con el fin de identificar las combinaciones de sus fonemas enlazadas a un corpus que contiene difonos, trifonos y sílabas, los cuales son utilizados para la concatenación y generación de sonido. Este programa se implementa en un sistema embebido Raspberry Pi junto con un teclado inalámbrico para ingresar el texto, una pantalla LCD programada en Python para ver el texto que se ingresa y un amplificador para el altavoz que reproduce el mensaje. La Figura 7 muestra el diagrama general del dispositivo ensamblado para ejecutar el sistema TTS. Esta sección se divide en tres partes: creación del corpus, programación conversión texto a voz utilizando síntesis concatenativa y ensamble del dispositivo (hardware y periféricos). El desarrollo del sistema TTS comienza con la creación del corpus.

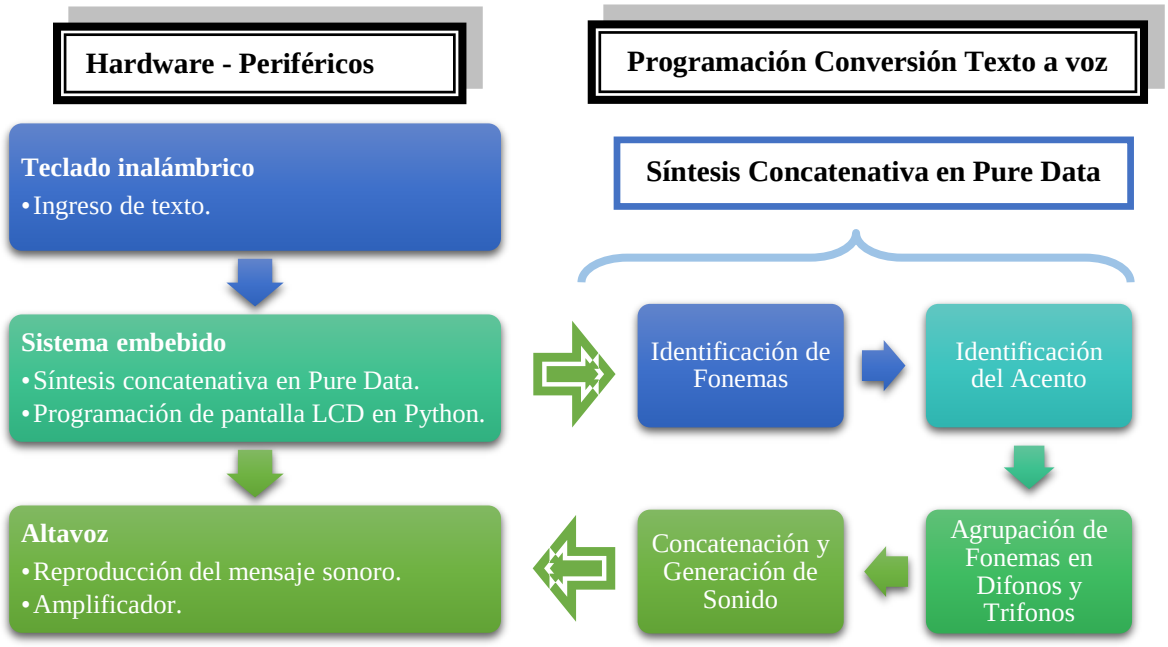

Figura 7. Diagrama general del dispositivo de conversión texto a voz.

Fuente: Propia.

Creación del corpus

La voz grabada para la creación del corpus pertenece a una persona de género masculino, nacionalidad colombiana y de 25 años de edad. El procesamiento de nivel, tono y duración de los audios del corpus se realizó en Adobe Audition CS6[31].

El primer objetivo específico señala la identificación de fonemas a partir de palabras grabadas. En total se obtuvieron 23 fonos, los cuales al ser concatenados, presentaban ininteligibilidad debido a la diferencia en la duración de estos y a la transición espectral abrupta entre ellos, enmascarándose entre sí y generando un mensaje sonoro incoherente. Por tal razón, se decidió conformar un nuevo corpus a partir de difonos extraídos de frases para garantizar una transición espectral natural entre fonos. Al realizar pruebas de concatenación, se presentaron variaciones de tono y diferencia entre la duración de los difonos, causando que el mensaje sonoro se escuchara robótico e inentendible y con falta de naturalidad. Por ende, se grabaron sílabas individuales para garantizar continuidad entre fonos y disminuir la variación espectral entre ellos. A continuación se presenta detalladamente cada proceso de la creación del corpus.

Grabación de Palabras.

En primera instancia se grabaron 24 palabras para obtener sus fonos y así crear el corpus del sistema para la concatenación. En cuanto a la elección de las palabras, se tuvo en cuenta la lista de 23 fonemas del Apéndice A, con el fin de buscar mínimo dos palabras por cada uno (cada palabra grabada contiene mínimo dos fonemas distintos); no se tuvieron en cuenta los alófonos que representan una variación acústica de más sobre el mismo fonema, por ejemplo:

Tabla 1. *Ejemplo de alófonos de un mismo fonema*

PALABRA	FONEMA	ALÓFONO
Donde	/donde/	[donde]
Dedo	/dedo/	[ðeðo]

Fuente: Propia.

[31] Software de edición de audio.

Como se puede ver en la Tabla 1, el fonema /d/ tiene dos alófonos diferentes: [d] y [ð]. Solo se tuvo en cuenta el primer alófono, debido a que se pretendía lograr un corpus reducido para modificar sus características fonéticas, mediante programación. Se eligieron los primeros alófonos de cada fonema que se muestran en el Apéndice A. A pesar de que en los objetivos específicos se planteó grabar 18 palabras, se decidió registrar más para tener mayor cantidad de material del cual escoger los fonos. La grabación se realizó en el estudio 5.1 de la Universidad de San Buenaventura Sede Bogotá con un micrófono Audiotechnica AT4050 usando el software Protools[32]. Cada canal de Protools contiene las dos grabaciones realizadas por palabra. La lista de las palabras grabadas se muestra en la Tabla 2.

Tabla 2. *Palabras Grabadas Para la Obtención de Fonos*

PALABRA	ALÓFONOS	PALABRA	ALÓFONOS	PALABRA	ALÓFONOS
Palo	[palo]	**Puerta**	[puerta]	**Hélice**	[elise]
Silla	[siʎa]	**yeso**	[ʎeso]	**Líbano**	[libano]
Mesa	[mesa]	**Está**	[esta]	**Botón**	[boton]
Nadar	[nadar]	**Cómo**	[como]	**Cámara**	[kamara]
Tarro	[taɾo]	**Días**	[días]	**Gruñir**	[gruⁿiɾ]
Fácil	[fásil]	**Qué**	[ke]	**Doble**	[doble]
Bruselas	[bruselas]	**Cuánto**	[kuanto]	**Grafiti**	[grafiti]
Tu	[tu]	**Niño**	[niⁿo]	**Caminar**	[kaminaɾ]

Fuente: Propia.

Obtención de Fonos.

Después de realizar la grabación descrita en el punto anterior, se procedió a realizar la división de fonos. Durante este proceso se determinó que la obtención de fonos no estaba dando resultados adecuados, por las siguientes razones:

- Los audios de los fonos tenían una duración entre 10 ms y 130 ms como se muestra en la Figura 8 y 9, esto implicaba que al unir los audios, se escuchaba notoriamente la

[32] Software para la grabación y edición de audio.

diferencia de duración entre los fonos, generando ininteligibilidad en el mensaje y enmascaramiento. Los fonos más cortos corresponden a las consonantes oclusivas[33], róticas[34] y laterales[35]; los fonos de mayor duración son los de las consonantes fricativas, nasales y vocales según el Apéndice A.

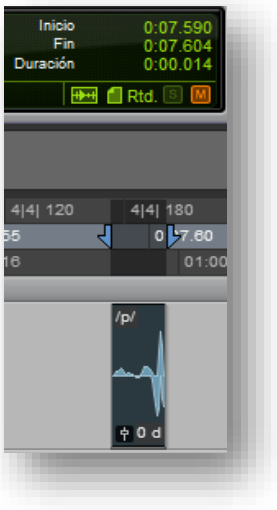

Figura 8. Duración del fono correspondiente al fonema oclusivo sordo /p/.
Fuente: Propia

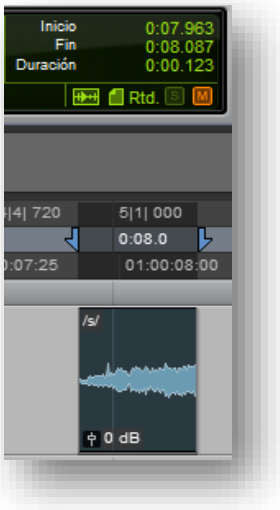

Figura 9. Duración del fono correspondiente al fonema fricativo /s/.
Fuente: Propia

[33] Sonido consonántico obstruyente producido por una detención del flujo de aire y por su posterior liberación
[34] Sonido producido mediante vibraciones en órgano articulador.
[35] Sonido consonántico producido por la obstrucción hecha a lo largo del eje longitudinal de la lengua.

- Obtener individualmente los fonos de las consonantes oclusivas no fue óptimo, debido a que no proporcionan la suficiente inteligibilidad para identificarlos, es decir, se escuchan como un impulso. Sin embargo, si están acompañados de una vocal, adquieren coherencia y contexto, debido a que la transición entre estas dos unidades tienen un suavizado espectral natural como se muestra en la Figura 10, en donde la línea blanca divide la consonante *p* de la vocal *a* de la sílaba *pa* grabada. Se puede apreciar que el espectro de la letra *p* tiene parte del espectro de la vocal *a* entre 1.5 kHz y 4 kHz, por lo tanto hay continuidad entre la transición de estas dos unidades, en contraparte a lo que se muestra en la Figura 11, donde la transición espectral es más abrupta entre 1.5 kHz y 4 kHz, ya que la sílaba *pa* fue conformada por los fonos correspondientes a los fonemas /p/ y /a/.

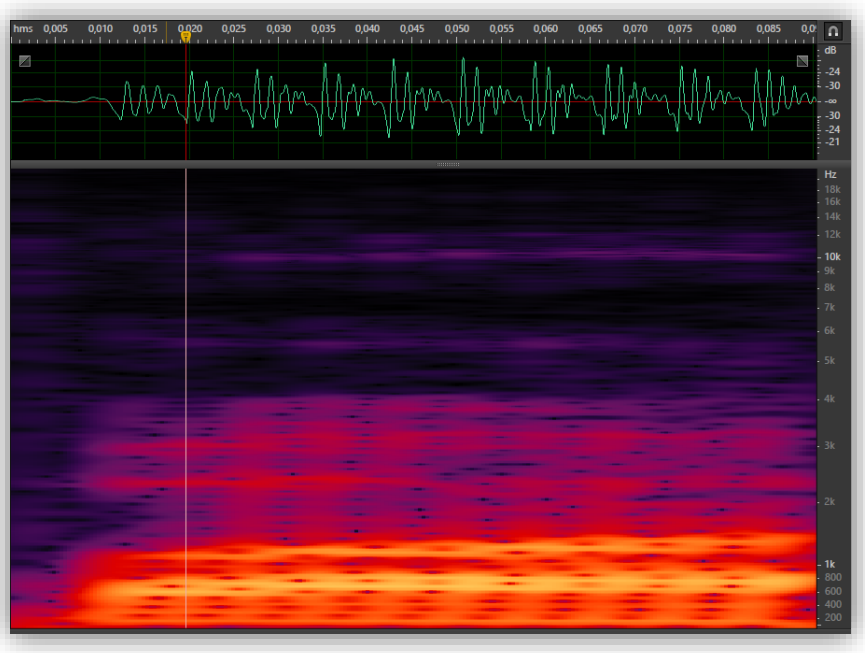

Figura 10. Espectrograma de la sílaba *pa*.

Fuente: Propia

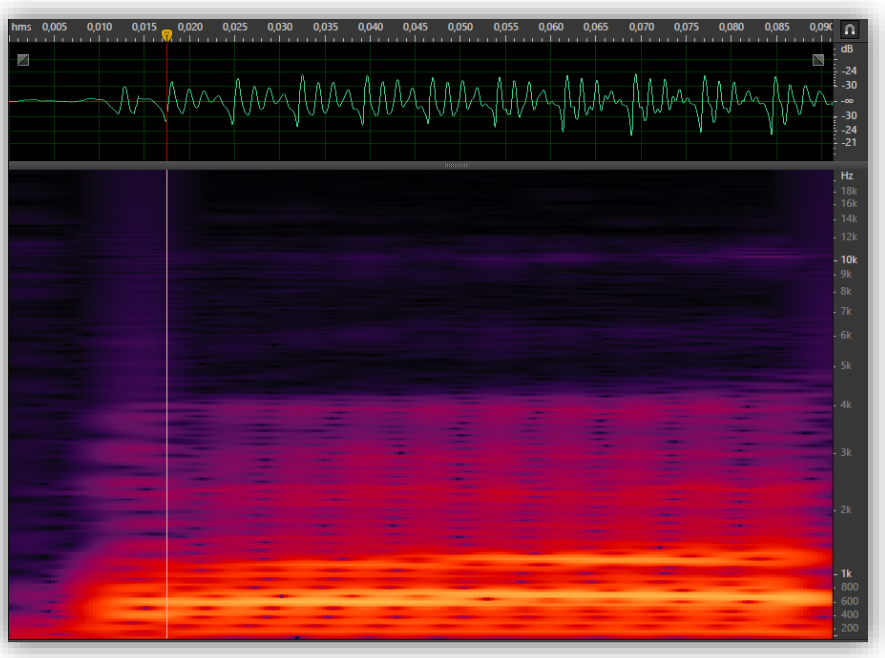

Figura 11. Espectrograma de la sílaba conformada por los fonos /p/ y /a/.
Fuente: Propia

- Los fonos de las consonantes róticas y laterales son cortos de duración pero no se escuchan como un impulso, de igual forma no tienen coherencia si no están acompañados de una vocal.

La división de fonos y su edición se realizó en el software Protools. Debido a los anteriores aspectos, se decidió trabajar con unidades más grandes de concatenación efectuando de esta forma, la obtención de difonos a partir de grabaciones de frases. Teniendo en cuenta el primer objetivo específico del proyecto, durante el proceso de la conversión texto a voz se identifican los fonemas en los difonos según la Tabla 3 para poder enlazarlos a las combinaciones de fonemas generadas para la concatenación desarrollada en Pure Data. Es decir, es necesario conocer los fonemas correspondientes a los difonos del corpus para que estos sean reproducidos correctamente según el análisis del texto ingresado y así realizar conversión TTS en el dispositivo.

Grabación de Frases.

Luego de determinar que la obtención de fonos provenientes de palabras no presentaba los resultados adecuados, se decidió grabar frases en Protools con palabras específicas compuestas por los difonos de la Tabla 3, basándose en el procedimiento descrito en el artículo "Síntesis de Voz por Concatenación de Difonemas para el Español" (Correa, Rueda, & Arguello, 2010) donde se propone realizar una tabla con las posibles combinaciones de fonemas para posteriormente formar difonos y buscar palabras que los contengan. Para lograr que los difonos estuvieran contextualizados, se incorporaron las palabras a la siguiente frase:

Me gusta comer _____ en la mañana y me gusta comer _____ en la noche.

Donde los espacios corresponden a la ubicación de las palabras, de tal forma que se obtuvieron dos grabaciones por palabra para poder elegir en cuál de las dos se encontraba el mejor difono. Los difonos obtenidos de las palabras son atónicos. La persona, cuya voz fue grabada, dijo esta frase con cada una de las palabras de la Tabla 4. La grabación de las frases se realizó en el Estudio 5.1 de la Universidad de San Buenaventura con un micrófono Audiotechnica AT4050 usando el software Protools. En total se grabaron 23 canales, cada uno tiene registradas las palabras que contienen las combinaciones de fonemas de cada columna de la Tabla 3. Se realizaron varias tomas de una misma frase debido a que algunas palabras quedaron grabadas con diferente tono, velocidad y nivel.

Tabla 3. *Combinación de posibles fonemas.*

FONEMAS	A	b	ʧ	d	e	f	g	i	x	k	l	λ	m	n	ñ	o	p	r	ṙ	s	t	U	ks
a		ab	aʧ	ad	ae	af	ag	ai	ax	ak	al	a λ	a m	an	aɲ	ao	ap	ar	aṙ	as	at	au	aks
b	ba		b ʧ	b d	be			bi	b x	b k	b l		b m	bn		bo	bp	b r		b s	bt	bu	
ʧ	ʧa				ʧe			ʧi								ʧo						ʧu	
d	da	db			de			di	d x	d k			d m			do		d r				du	
e	ea	eb	eʧ	ed		ef	eg	ei	ex	ek	el	e λ	e m	en	eɲ	eo	ep	er	eṙ	es	et	eu	eks
f	Fa				fe		fg	fi			fl					fo		fr			ft	fu	
g	ga			g d	ge			gi			g l		g m	gn		go		g r				gu	
i	Ia	ib	iʧ	id	ie	if	ig		ix	ik	il	i λ	im	in	iɲ	io	ip	ir	iṙ	is	it	iu	iks
x	xa				xe			xi								xo						xu	
k	ka				ke			ki			k l			kn		ko		k r		k s	kt	ku	
l	La	lb	lʧ	ld	le	lf	lg	li		lk			lm	ln		lo	lp		lṙ	ls	lt	lu	
λ	λa				λe			λi								λo						λu	
m	ma	mb			me			mi						m n		mo	mp					mu	
n	na	nb	nʧ	nd	ne	nf	ng	ni	n x	n k	n l	n λ	n m			no			nṙ	n s	nt	un	
ɲ	ɲa				ɲe			ɲi								ɲo						ɲu	
o	oa	ob	oʧ	od	oe	of	og	oi	o x	o k	o l	o λ	o m	on	oɲ		op	o r	oṙ	o s	ot	ou	oks
p	pa				pe			pi			p l					po		p r		p s	pt	pu	
r	Ra				re			Ri								ro						ru	
ṙ	ṙa	ṙb	ṙʧ	ṙd	ṙe		ṙg	ṙi	ṙx	ṙk	ṙl		ṙm	ṙn		ṙo	ṙp			ṙs	ṙt	ṙu	ṙks
s	sa	sb			se			Si		sk	sl		sm	sn		so	sp				st	su	
t	Ta				te			Ti			tl		tm	tn		to		tr				tu	
u	ua	ub	uʧ	ud	ue	uf	ug	ui	u x	u k	u l	u λ	u m	un	uɲ	uo	up	u r	ur	u s	ut		uks
ks	Ka				ks	ks		ks								ks	ks				ks	ks	
	sa				e	f		i								o	p				t	u	

Fuente: Propia

Los cuadros azules de la Tabla 3 indican que la respectiva combinación de fonemas no existe en el idioma español.

Tabla 4. *Difonos obtenidos de palabras grabadas.*

DIF	PALABRA	DIF	PALABRA	DIF	PALABRA	DIF	PALABRA	DIF	PALABRA
ba	lava	**ke**	palenque	**ɾx**	arjona	**aɲ**	añoranza	**gr**	gratinar
ʧa	brecha	**le**	tolemaida	**ux**	bujía	**eɲ**	empeñar	**ir**	viral
da	tandada	**ʎe**	bachillerato	**ak**	acuarela	**iɲ**	cariñoso	**kr**	croqueta
ea	fea	**me**	támesis	**bk**	subconciente	**oɲ**	retoñó	**or**	ortogonal
fa	fachada	**ne**	nebraska	**dk**	vodka	**uɲ**	gruñón	**pr**	aprobar
ga	conga	**ɲe**	bañera	**ek**	ecológico	**ao**	ahogarse	**tr**	triángulo
ia	fia	**oe**	radioescucha	**ik**	picazón	**bo**	bonita	**ur**	hurtar
xa	jamón	**pe**	amperaje	**lk**	alcachofa	**ʧo**	cachorrito	**aɾ**	arborizar
ka	acababa	**re**	acuchare	**nk**	bronca	**do**	dado	**eɾ**	ergonómico
la	bala	**ɾe**	irreprochable	**ok**	karaoke	**eo**	leoncito	**iɾ**	irradiar
ʎa	valla	**se**	sentarse	**ɾk**	arca	**fo**	foquita	**lɾ**	alrededor
ma	cámara	**te**	gente	**sk**	cascarón	**go**	algo	**nɾ**	enroscar
na	lana	**ue**	buenísimo	**uk**	cauca	**io**	ionizar	**at**	atender
ɲa	caña	**kse**	excelente	**al**	alegre	**xo**	joropo	**bt**	obtener
oa	canoa	**af**	afamado	**bl**	cable	**ko**	kokoriko	**et**	etcétera
pa	apagar	**ef**	efímero	**el**	acelerar	**lo**	golosina	**ft**	oftalmólogo
ra	vara	**if**	rechiflar	**fl**	flauta	**ʎo**	cayo	**it**	tití
ɾa	barra	**lf**	alfaro	**gl**	glosario	**mo**	amorío	**kt**	actuar
sa	casa	**nf**	chanfle	**il**	iluso	**no**	notario	**lt**	altura
ta	antagónico	**of**	cofradía	**kl**	bucle	**ɲo**	caño	**nt**	antropólogo
ua	agua	**uf**	bufón	**nl**	enlatar	**po**	popó	**ot**	otorrino
ksa	hexadecimal	**ksf**	exfutbolista	**ol**	holograma	**ro**	coro	**pt**	aptitud
ab	aburrido	**ag**	aguantar	**pl**	plosivo	**ɾo**	tarro	**ɾt**	arte
db	advierte	**eg**	egocéntrico	**ɾl**	arlequín	**so**	pozo	**st**	castidad
eb	nochebuena	**fg**	afganistán	**nb**	envolver	**to**	alto	**ut**	lutero
ib	compatibilidad	**ig**	iglesia	**tl**	atletismo	**uo**	inócuo	**kst**	éxtasis
lb	albaca	**ng**	canguro	**ul**	ultrasónico	**lg**	algarabia	**au**	audífonos
mb	membrana	**og**	holograma	**a ʎ**	ayer	**kso**	éxodo	**bu**	buscó
ob	aprovechar	**ɾg**	cargan	**e ʎ**	camellar	**ap**	apartar	**sl**	Oslo
ɾb	urbano	**ug**	uganda	**i ʎ**	anillar	**bp**	subpiso	**ʧu**	chuzar
sb	presbitero	**ai**	taichi	**n ʎ**	conllevar	**ep**	epílogo	**du**	duchar
ub	cubano	**bi**	bifurcar	**o ʎ**	coyote	**ip**	hipopótamo	**eu**	euclides
aʧ	achantar	**ʧi**	chinauta	**u ʎ**	bulloso	**lp**	golpear	**fu**	fuego
bʧ	chibcha	**di**	adiestrar	**dm**	administrar	**mp**	amperios	**gu**	agujero
eʧ	echados	**ei**	taipei	**em**	emoción	**op**	optimizar	**iu**	enviudar
iʧ	chicharra	**fi**	ficharlos	**gm**	magma	**ɾp**	torpedo	**xu**	jugar
lʧ	acolchado	**gi**	águila	**im**	cimarrón	**sp**	caspa	**ku**	acuarela

ntʃ	hinchado	**xi**	ginebra	**lm**	almíbar	**up**	agrupar	**lu**	lujoso
otʃ	sancochado	**ki**	quitarán	**nm**	conmutar	**ksp**	expiar	**ʎu**	ayuntamiento
ɾtʃ	marcha	**li**	limonazo	**om**	pomada	**ar**	carísimo	**mu**	murmurar
utʃ	duchado	**ʎi**	chantillí	**ɾm**	armada	**br**	abrazar	**un**	untar
ad	adviento	**mi**	amigable	**sm**	prismático	**dx**	adjunto	**ɲu**	buñuelo
bd	abducción	**ni**	nigeria	**tm**	atmósfera	**oɾ**	cortar	**ou**	monousuario
ed	tedioso	**ɲi**	cañihuecos	**um**	tumaco	**urr**	urrego	**pu**	publicar
gd	bagdad	**oi**	boinazo	**an**	antaño	**as**	castaño	**ru**	abrumar
id	midieron	**pi**	helicoidales	**bn**	abnegar	**bs**	abstracto	**ɾu**	arrullar
ld	caldera	**ri**	espichados	**en**	centrado	**es**	tesoro	**su**	sutana
nd	bandido	**ɾi**	arigato	**gn**	agnósis	**is**	izquierdo	**tu**	tumaco
od	poderío	**si**	charrito	**in**	invitar	**ks**	acción	**ksu**	exhuberante
ɾd	borde	**ti**	alfabético	**kn**	acné	**ls**	alzar	**aks**	acceso
ud	budismo	**ui**	huir	**ln**	malnacer	**ns**	cansón	**eks**	expiar
ae	caen	**ksi**	exiliado	**mn**	amnesia	**os**	osciloscopio	**iks**	asfixiar
be	ventajoso	**ax**	cajón	**on**	contrario	**ps**	bíceps	**oks**	ortodoxista
ʧe	caniche	**bx**	objeto	**ɾn**	carnicería	**ɾs**	arsénico	**ɾks**	marxista
de	demorar	**ex**	alejandro	**ge**	guevara	**us**	australia	**uks**	auxilio
fe	posfechar	**ix**	aguijón	**sn**	asno	**dr**	apedrear		
ie	dieciocho	**nx**	franja	**tn**	etnografía	**er**	ergonómico		
xe	genoveva	**ox**	hojear	**un**	apuntó	**fr**	fracaso		

Fuente: Propia.

En la Figura 12 se puede ver la ventana de edición de la sesión. La división de las frases en palabras se realizó en Protools como se muestra en la Figura 13.

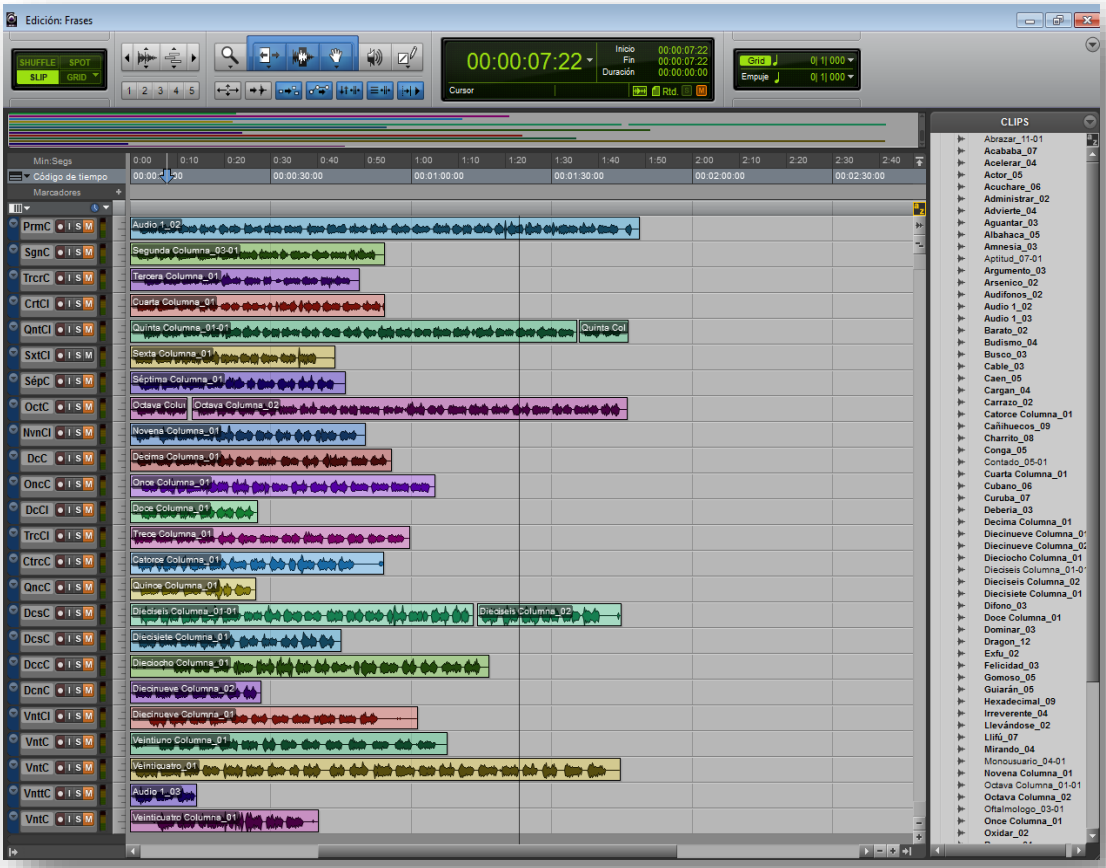

Figura 12. Ventana de edición de la sesión de grabación de frases.

Fuente: Propia.

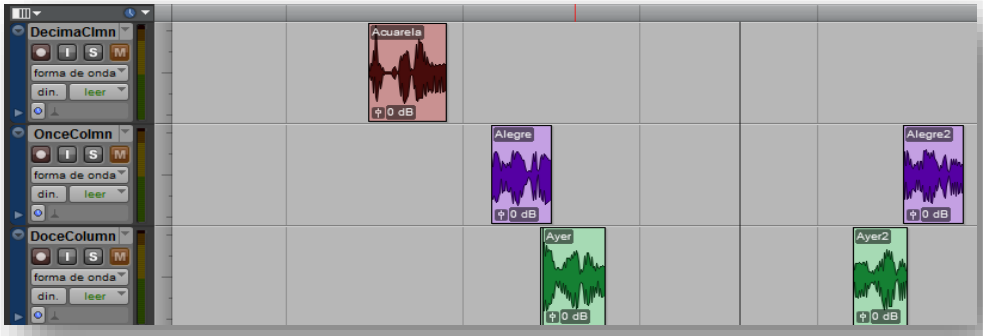

Figura 13. Ventana de edición de la sesión de grabación de frases en Protools.

Fuente: Propia

Obtención y Edición de Difonos.

Una vez divididas las frases en palabras, se procedió a obtener sus difonos teniendo en cuenta la Tabla 3. Para esto, se utilizó Adobe Audition CS6 como se muestra en la Figura 14.

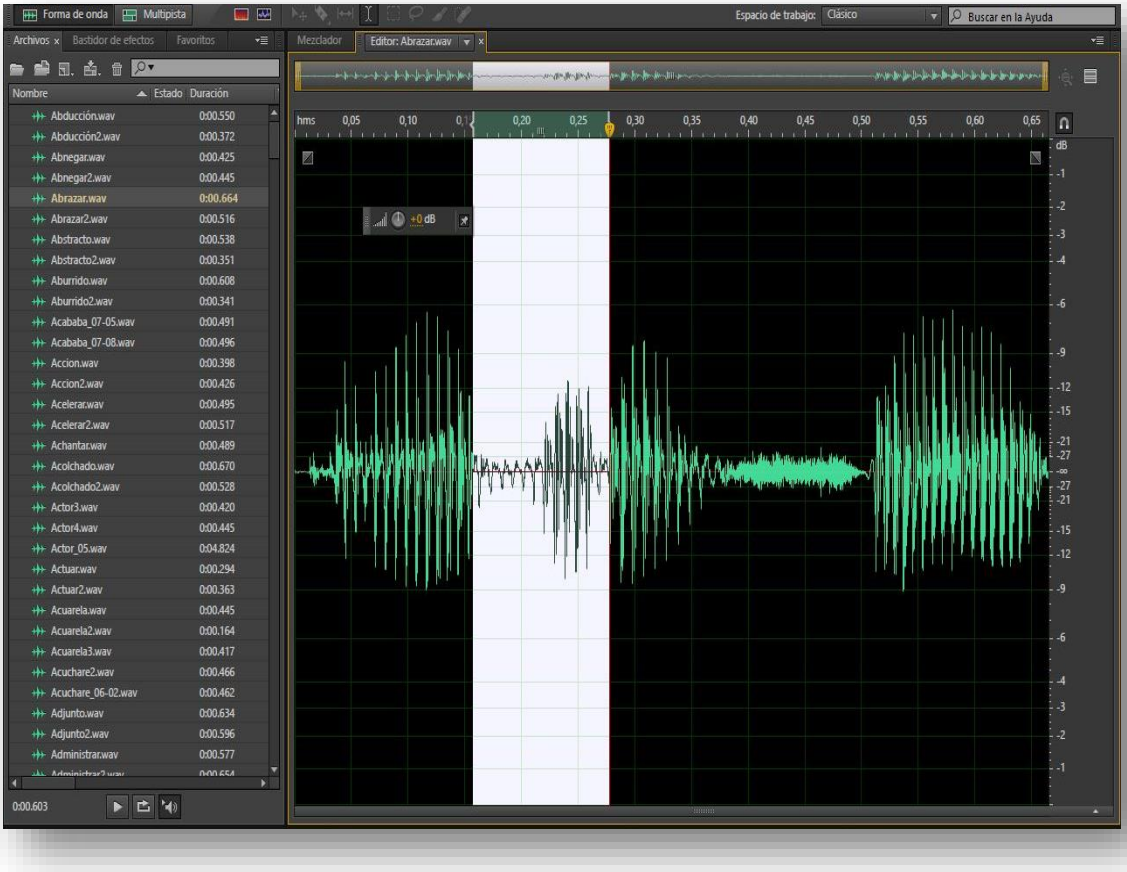

Figura 14. Obtención del difono /br/ en Adobe Audition.

Fuente: Propia

En la Figura 15 se describe el proceso de edición de los difonos en Adobe Audition. Para mayor control de edición sobre los archivos de audio, estos fueron inicialmente modificados con los mismos parámetros en cuanto a duración y nivel.

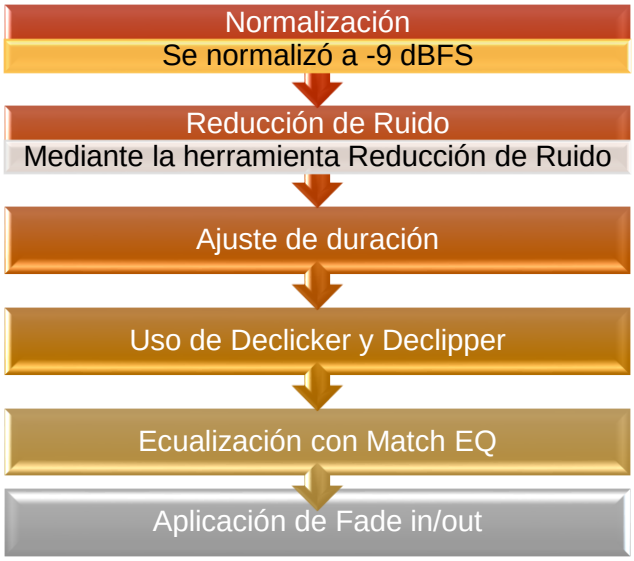

Figura 15. Proceso de edición de difonos en Adobe Audition.

Fuente: Propia.

La normalización de los audios se realizó para lograr uniformidad en el nivel del mensaje sonoro generado por la concatenación. Todos los audios se normalizaron a -9 dBFS debido a que este valor era cercano al pico máximo de la mayoría de los audios grabados. La reducción de ruido se ejecutó para disminuir y/o eliminar los sonidos no deseados en las señales de voz. En total se obtuvieron 287 difonos cuya duración oscilaba entre 70 ms y 160 ms, lo cual generaba discontinuidad al combinarse, debido a que algunos difonos se escuchaban más largos que otros creando de esta forma enmascaramiento. Para tratar de solucionar esto, se realizó un ajuste de duración mediante compresión y expansión en el tiempo con la herramienta *Ampliación y Tono* de Adobe Audition, como lo muestra la siguiente tabla:

Tabla 5. *Aproximación de duración en milisegundos de los difonos.*

Duración del difono (ms)	Aproximación (ms)
Entre 70 y 95	90
Entre 96 y 115	110
Entre 116 y 135	130
Mayor a 136	160

Fuente: Propia.

La expansión en el tiempo generó chasquidos y detonaciones en las formas de onda de los difonos, por consiguiente se usó la herramienta *DeClicker*[36] de Adobe Audition para eliminar estos problemas. También se utilizó la herramienta *DeClipper*[37] para reparar en algunos casos, formas de onda recortadas.

La discontinuidad se presentaba también por las variaciones de tono en los difonos. Durante la grabación de las frases, el tono de la voz del hablante se percibió continuo en general, pero al segmentar en unidades pequeñas como son los difonos, los cambios de tono son significativos de tal forma que la combinación de estas unidades se escucha desordenada. Por ende, se utilizó la herramienta *Match EQ* del Software *Ozone 6*[38] de *iZotope*[39]en Adobe Audition como se muestra en la Figura 17. Esta herramienta proporciona herramientas para ecualizar automáticamente una señal teniendo una referencia, que en este caso serían difonos elegidos por tener el tono de voz deseado. Se escogió un difono por cada vocal, coincidencialmente todos tenían el fono correspondiente al fonema /l/, es decir los difonos /al/ /el/ /il/ /ol/ /ul/. Se decidió realizar la ecualización con solo las vocales porque se determinó que la discontinuidad es más notoria en estas que en consonantes, teniendo en cuenta el siguiente método de concatenación de difonos:

gato: /_g/-/ga/-/at/-/to/- /o_/

El proceso de ecualización con *Match EQ* se muestra en la Figura 16.

[36] Efecto que elimina chasquidos y detonaciones.
[37] Efecto que repara formas de onda recortadas, rellenando las secciones recortadas con datos de audio nuevos.
[38] Software de masterización de audio.
[39] Compañía que desarrolla software y plug-ins para mezcla, masterización, producción musical y sonido en vivo.

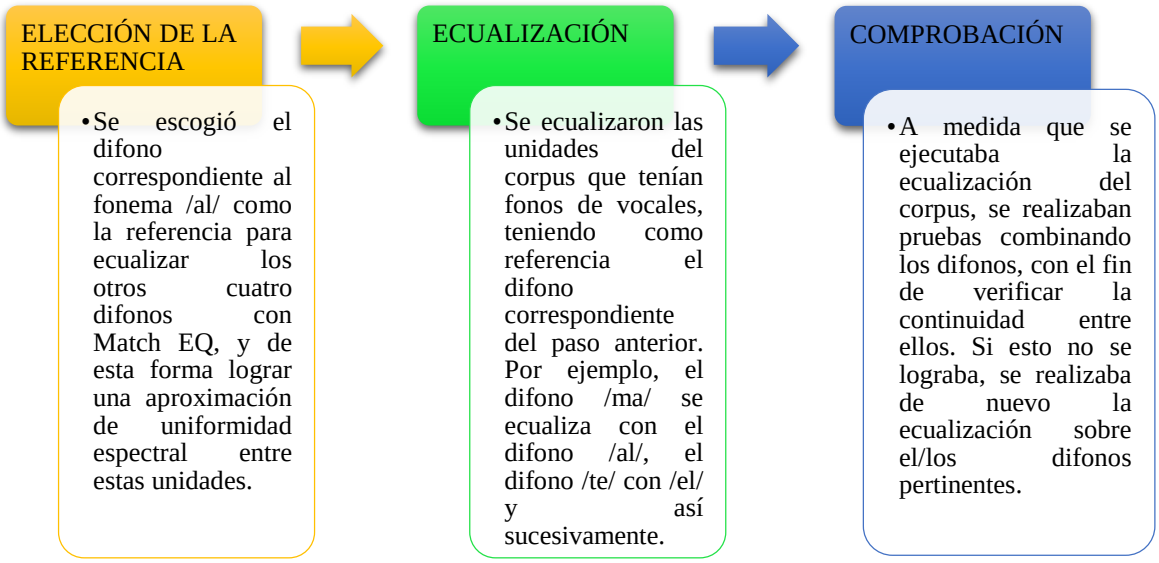

Figura 16. Proceso de ecualización con Match EQ.

Fuente: Propia.

Figura 17. Ecualización con Match EQ.

Fuente: Propia

En la Figura 17 se puede ver la forma de onda de dos difonos y el ecualizador de Ozone 6, en donde se aprecian tres curvas. La curva azul corresponde a la señal de referencia, es decir a la primera forma de onda de la parte superior de la imagen; la amarilla corresponde a la señal a la cual se va a aplicar la ecualización, es decir, la segunda forma de onda de la parte superior; la curva blanca indica la caracterización de la ecualización.

Los fade in/out[40] aplicados a los archivos de audio tienen una cobertura del 30 % al inicio y final del difono, con el fin de crear un efecto de crossfade[41] en la concatenación y suavizar la transición entre estos, es decir, los audios estarían solapados durante los milisegundos correspondientes al 30 % del final de un difono y 30% del inicio del siguiente. Aplicando este porcentaje de fade in/out a las señales se garantiza que no se pierda información necesaria para al inteligibilidad de los difonos.

Este corpus de difonos no funcionó adecuadamente debido a que la duración inicial del más del 50 % de lo difonos seguía siendo muy corta (entre 70 ms y 160 ms) lo que implicaba que la señal concatenada se escuchara robótica y en algunos casos inentendible. Se intentó expandir aún más en el tiempo a los audios pero su tono cambiaba drásticamente. Se concluyó que la extracción de difonos de palabras no es óptimo en la concatenación si no se realiza un suavizado espectral, debido a que el tono varía notablemente entre cada difono como se muestra en la Figura 18, donde se observa el cambio de tono espectral (línea azul) de la palabra *lana* conformada por los difonos /la/-/an/-/na/ entre el rango de 90 Hz y 100 Hz, a comparación de lo que se observa en la Figura 19, donde se muestra el tono espectral obtenido de la grabación de esa misma palabra, el cual es más uniforme y se encuentra entre 110 Hz a 116 Hz, presentando un rango menor al del caso anterior en cuanto a variación de la frecuencia fundamental. Las variaciones de tono causan que se pierda naturalidad en el mensaje sonoro.

[40] Aumento o disminución gradual de nivel de una señal de audio.
[41] Transición entre dos audios mediante fade in/out.

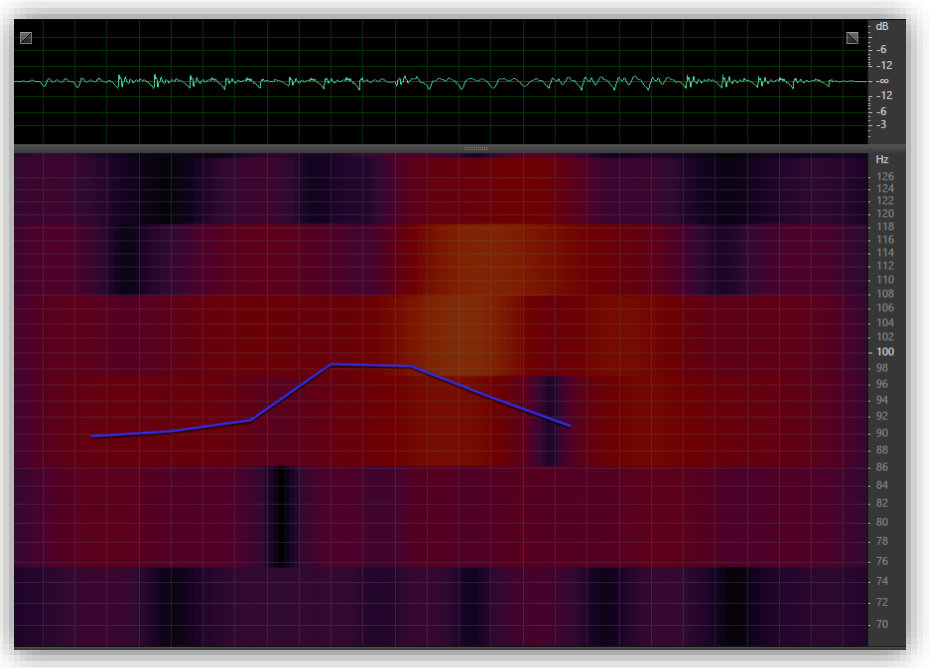

Figura 18. Tono espectral de la palabra *lana* conformada por los difonos /la/-/an/-/na/.

Fuente: Propia

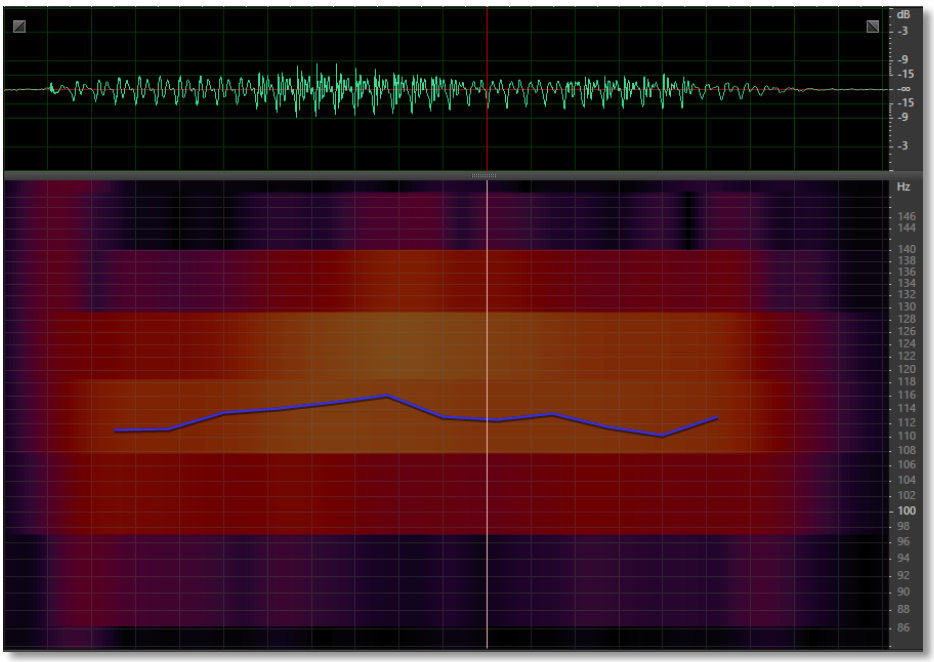

Figura 19. Tono espectral obtenido de la grabación de la palabra *lana.*

Fuente: Propia

Por consiguiente, se decidió realizar un nuevo corpus conformado por sílabas, difonos y trifonos. Teniendo en cuenta el primer objetivo específico del proyecto, durante el proceso de la conversión texto a voz se identifican los fonemas en las sílabas según la Tabla 3 para poder enlazarlas a las combinaciones de fonemas generadas para la concatenación desarrollada en Pure Data. Es decir, es necesario conocer los fonemas correspondientes a las sílabas del corpus para que estas sean reproducidas correctamente según el análisis del texto ingresado y así realizar la conversión TTS en el dispositivo.

Para este proyecto, el suavizado espectral no se desarrolló porque se habría aumentado el procesamiento de Pure Data y disminuido el rendimiento del sistema embebido Raspberry Pi. Debido a que este sistema embebido es el único que cuenta con una versión de Pure Data pre compilada y tiene la comunidad de desarrolladores más grande del mundo en este campo, no se consideró necesario adquirir en el mercado otro sistema embebido, ni cambiar el lenguaje de programación por uno compatible con un sistema embebido con mayor capacidad de procesamiento y mayor memoria, ya que Pure Data es un lenguaje optimizado para el procesamiento digital de señales y tiene una comunidad de desarrollo muy grande.

Grabación de Sílabas.

Debido a los resultados obtenidos con el corpus de difonos, se decidió grabar una nueva serie de unidades. Para este caso, se grabaron sílabas ya que al pronunciarlas individualmente, no se acentúan ni se contextualizan. Para este caso la cantidad de unidades utilizadas es menor a los procedimientos descritos anteriormente, porque al implementar el sistema TTS (Sección *Programación de Conversión Texto a Voz Usando Síntesis Concatenativa*) en el sistema embebido no se abría el patch[42] generando un error que indica que la librería *readsf*[43] de Pure Data tiene una limitación en la cantidad de audios a abrir en el patch cuando se ejecuta en la arquitectura ARMHF, por lo tanto se disminuyó la cantidad de audios en el sistema (inicialmente habían 250 sílabas); como resultado se obtuvo una ejecución adecuada con las 74 sílabas de la Tabla 6 para generar las 47 palabras de la Tabla 7, las cuales fueron obtenidas de la página oficial

[42] Conjunto de comandos programados para el entorno de Pure Data.
[43] Objeto de Pure Data que lee y reproduce archivos de audio.

de la oficina de turismo de Austria (Austria, 2014), por ser consideradas básicas e importantes para la comunicación.

Tabla 6. *Sílabas que conforman el corpus final.*

si	no	po	orr	fa	ab	bo	gra	me	ya
as	si	go	co	om	mo	en	ti	se	ed
ie	do	on	pe	err	mi	is	so	am	bre
bu	en	os	di	ia	as	as	ta	tri	te
arr	de	es	ad	io	ol	la	ke	fe	li
al|	bi	ki	ku	ua	an	to	el	ud	da
añ	ño	ti	em	ab	bla	pa	ol	ay	yu
oi	ma	ña	na						

Fuente: Propia.

Tabla 7. *Palabras a generar en el sistema.*

Sí	Per**dón**	Bi**en**	Ma**ña**na
No	Bu**en**os **dí**as	Qui**én**	Bu**en**o
Por fa**vor**	Bu**en**as tardes	Cu**ál**	A**yu**da
Gracias	Adi**ós**	¿**Dón**de es**tá** el **ba**ño?	Mal
Y	**Ho**la	¿Cu**án**do?	Fe**liz**
¿**Có**mo?	¿Qué tal?	¿Cu**án**to ti**em**po?	**Tris**te
No enti**en**do	¿**Có**mo es**tás**?	¿**Ha**bla espa**ño**l?	**Ham**bre
Con per**mi**so	¿Cu**án**to?	Hoy	Sed
Vale	Es**to**y	**Ten**go	**Di**ce

Fuente: Propia.

A diferencia de los demás procedimientos de grabación, los difonos y trifonos fueron obtenidos de sílabas, garantizando de esta forma mayor continuidad entre fonos. La grabación se realizó en el Estudio 5.1 de la Universidad de San Buenaventura con un micrófono Audiotechnica AT4050 usando el software Protools. Cada sílaba fue grabada dos veces para tener mayor material para escoger. La persona que grabó los demás corpus fue la misma para este caso.

Edición de Sílabas.

La edición de sílabas se realizó en Adobe Audition CS6 para modificar su duración, tono, nivel, como también, parámetros de restauración según el método de prueba y error, siguiendo el proceso de la Figura 20. Para mayor control de edición sobre los archivos de audio, estos fueron inicialmente modificados con los mismos parámetros en cuanto a duración y nivel.

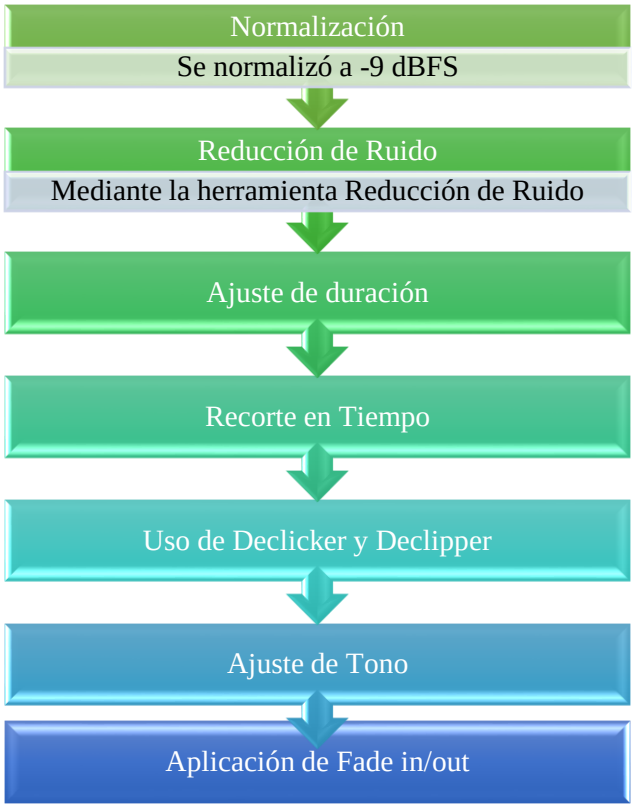

Figura 20. Proceso de edición de sílabas en Adobe Audition.

Fuente: Propia.

La normalización de los audios se realizó para lograr uniformidad en el nivel del mensaje sonoro generado por la concatenación. Todos los audios se normalizaron a -9 dBFS debido a que este valor era cercano al pico máximo de la mayoría de los audios grabados Después de normalizar y aplicar reducción de ruido a los archivos de audio, se realizó un recorte en el tiempo teniendo en cuenta la posición de la sílaba (inicio, medio, final) dentro de las palabras en el momento de identificar estas unidades en el texto ingresado al sistema.

Teniendo en cuenta la duración de todos los archivos de audio, se estableció que 348 ms es un promedio estándar para aplicarlo en la expansión y compresión de tiempo de cada uno de ellos, porque su duración inicial se aproximaba a este valor. Se realizaron pruebas de concatenación de sílabas en Adobe Audition de la siguiente forma:

gato: /ga/ - /at/ - /to/

Se concluyó que la duración de los audios debía variar según la posición (Tabla 8), es decir, la duración debía ser mayor en las sílabas del principio y final para lograr buena inteligibilidad en la primera y última letra de la palabra, ya que estas sílabas no son complementadas como las que se encuentran en medio. Las sílabas que se encuentran en el medio, debían tener menor duración que las del principio y final porque estas se complementan entre sí para conformar fonos completos. Al modificar la duración de las sílabas recortándolas en puntos específicos, se obtienen difonos y trifonos. La Tabla 8 muestra el tipo de recorte realizado a los audios.

Tabla 8. *Parámetros de recorte en el tiempo de los audios de las sílabas.*

POSICIÓN EN LA PALABRA	TIPO DE RECORTE
Inicio	Desde el inicio hasta el 60 % de la duración del audio aprox. Ejemplo: Duración: 348 ms ; Recorte: 200 ms.
Medio	Desde el 30 % hasta el 70 % de la duración del audio aprox. Ejemplo: Duración: 348 ms ; Corte en : 104 ms y 250ms; Recorte: 120 ms.
Final	Desde el 40 % de la duración del audio aprox. hasta el final. Ejemplo: Duración: 348 ms ; Corte en: 140 ms; Recorte: 200 ms.

Fuente: Propia.

Las sílabas recortadas desde el 30 % hasta el 70 % de su duración, se convierten en difonos y trifonos. Para lograr la menor variación de nivel posible entre los audios se aseguró que el punto de corte en el tiempo sobre una vocal, tuviera un nivel entre -10 dBFS y -12 dBFS como se muestra en la Figura 21.

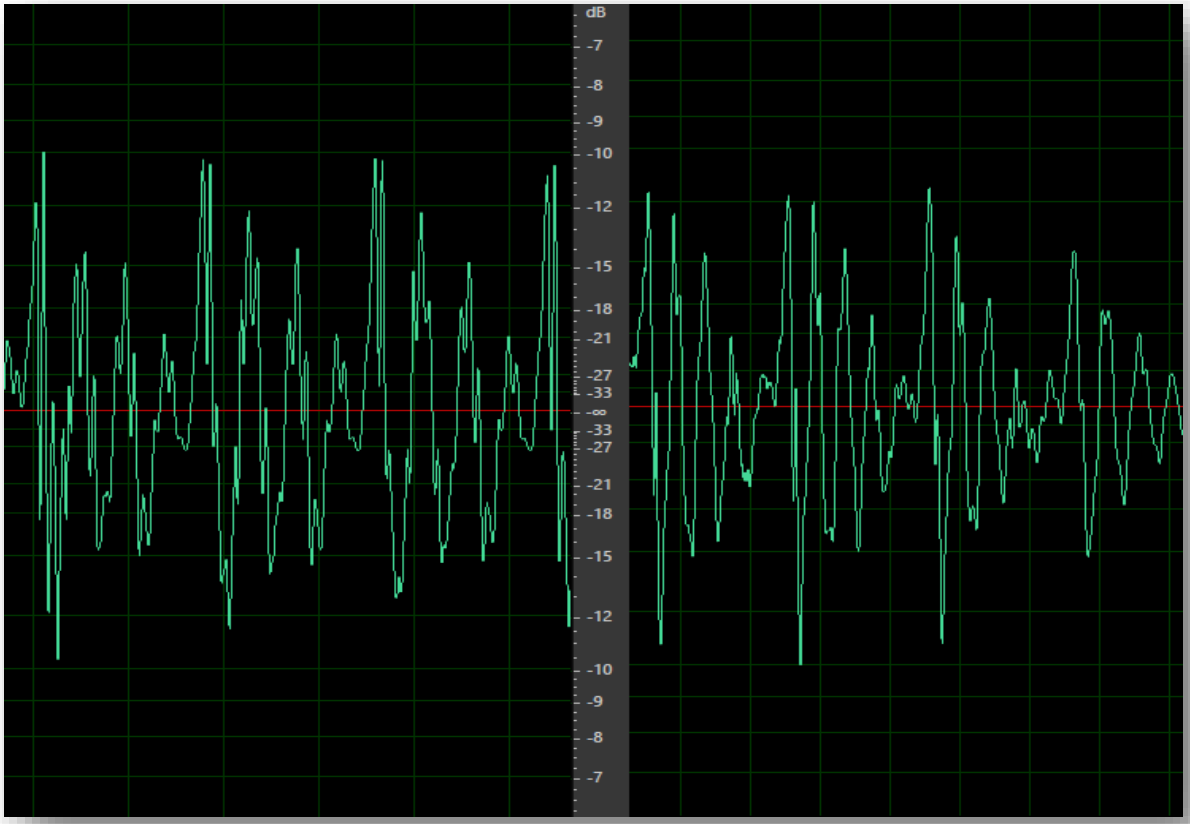

Figura 21. Nivel en el punto de corte en el tiempo de dos silabas: *pe* a la izquierda y *err* a la derecha. Fuente: Propia.

La expansión y compresión en el tiempo, como también el recorte en el tiempo generó chasquidos y detonaciones en las formas de onda de los audios, por consiguiente se usaron las herramientas *DeClicker* y *DeClipper* de Adobe Audition para eliminar estos problemas.

La acentuación se relaciona con la duración y tono (Taylor, 2009), como también la intensidad y el timbre (Cantero, 2002), por ende se realizaron pruebas en Adobe Audition para determinar estos ajustes. Las palabras tienen sílabas tónicas[44] y sílabas atónicas[45], por tal razón se modificaron los archivos de audio, que corresponden a las sílabas en negrilla (tónicas) y los demás de la Tabla 7 según los parámetros descritos en la Tabla 9. En primer lugar, se modificó la duración, luego el tono y por último el nivel. Para las silabas tónicas se normalizó a -6 dBFS, aumentando de esta forma la intensidad sonora de la señal, teniendo en cuenta que a mayor

[44] Acentuada.
[45] No acentuadas.

intensidad, mayor es la acentuación. De igual forma, se normalizó a -10 dBFS en las sílabas atónicas para disminuir este parámetro y diferenciar auditivamente estas dos tipos de sílabas.

Tabla 9. *Ajustes de tono, duración y nivel.*

TIPO DE SÍLABA	DURACIÓN	TONO	NIVEL
Tónica	5 % adicional a la duración original	+0.3 céntimos[46]	Normalización a -6 dBFS
Atónica	Sin cambios	Sin cambios	Normalización a -10 dBFS

Fuente: Propia.

El cambio de tono en los audios fue probado en Pure Data con el objeto *tabread4~*[47] el cual permite cambiar la duración; a menor longitud, mayor es el tono. Este proceso no se implementó porque no se adecuaba a la forma de concatenación secuencial en la cual está programado el sistema, es decir, retrasaba la reproducción de los audios.

Los fade in/out tienen una cobertura del 30% al inicio y final de los archivos de audio, con el fin de crear un efecto de crossfade y suavizar la transición entre estas en la concatenación programada en Pure Data, es decir, los audios estarían solapados durante los milisegundos correspondientes al 30% del final de una sílaba, difono o trifono y 30% del inicio de la siguiente unidad. Aplicando este porcentaje de fade in/out a las señales se garantiza que no se pierda información necesaria para al inteligibilidad de las sílabas, difonos y trifonos.

Programación de Conversión Texto a Voz Utilizando Síntesis Concatenativa por Selección

La programación de la conversión texto a voz se realizó en Pure Data Extended en Windows 7 utilizando un teclado de computador de escritorio. El proceso de manipulación de datos en Pure Data se simplifica si son de tipo numérico. La Figura 22 muestra el flujo de procesamiento de la información desde el ingreso del texto hasta la generación del sonido. El sistema no está programado para identificar caracteres, símbolos, números y letras en mayúscula.

[46] Fracción de un semitono. 1 semitono=100 céntimos.
[47] Objeto de Pure Data que lee una matriz que contiene señales.

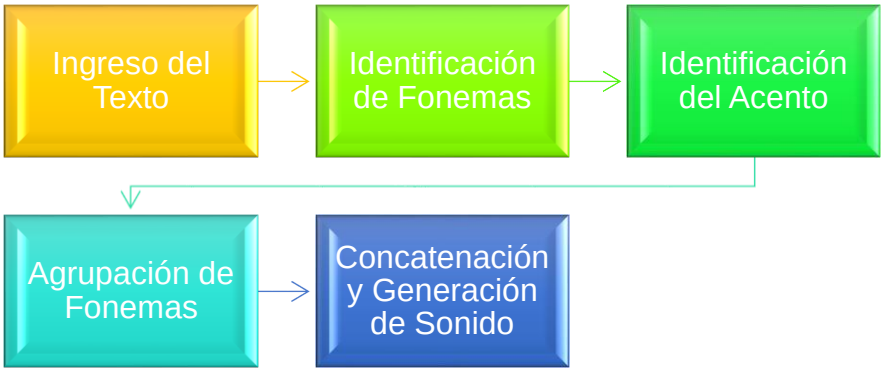

Figura 22. Flujo de procesamiento de la información en el sistema TTS.

Fuente: Propia.

Ingreso del Texto.

Después de escribir el texto, Pure Data recibe los números correspondientes a las letras del teclado como lo indica la Tabla 10. El proceso de obtención de la nomenclatura numérica del teclado se puede observar en la Figura 23, cuyo patch de Pure Data se muestra en las Figuras 24 y 25.

Tabla 10. *Nomenclatura numérica del teclado alfabético.*

Letra	Número	Letra	Número	Letra	Número	Letra	Número
a	97	j	106	r	114	á	225
b	98	k	107	s	115	é	233
c	99	l	108	t	116	í	237
d	100	M	109	u	117	ó	243
e	101	n	110	v	118	ú	250
f	102	ñ	241	w	119	Enter	10
g	103	o	111	x	120	Esc	8
h	104	p	112	y	121	Space	32
i	105	q	113	z	122	Back	8

Fuente: Propia.

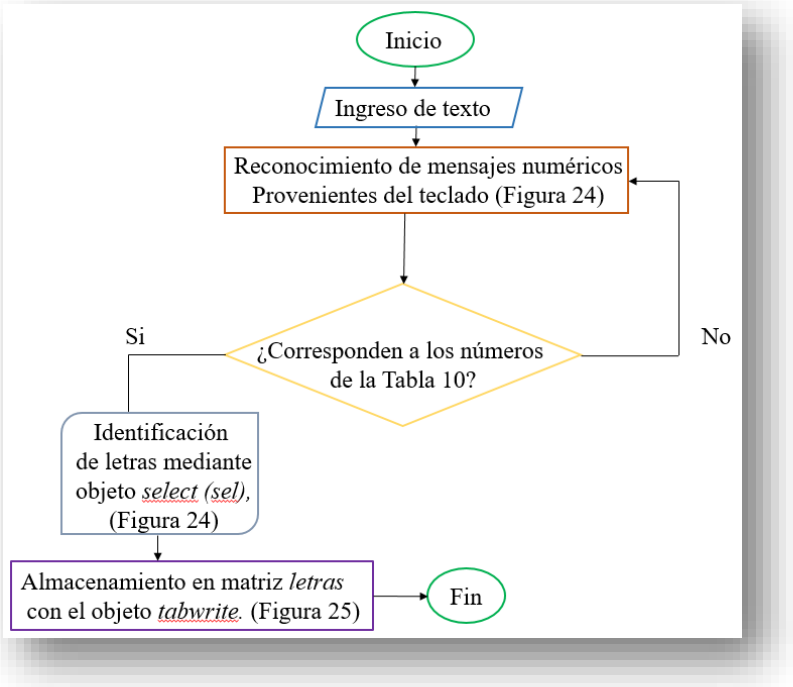

Figura 23. Diagrama de flujo de señal de la obtención de nomenclatura numérica del teclado en Pure Data.
Fuente: Propia

Los números correspondientes a cada letra de la Tabla 10 son recibidos en Pure Data por el objeto *key*[48] como se muestra en la Figura 24.

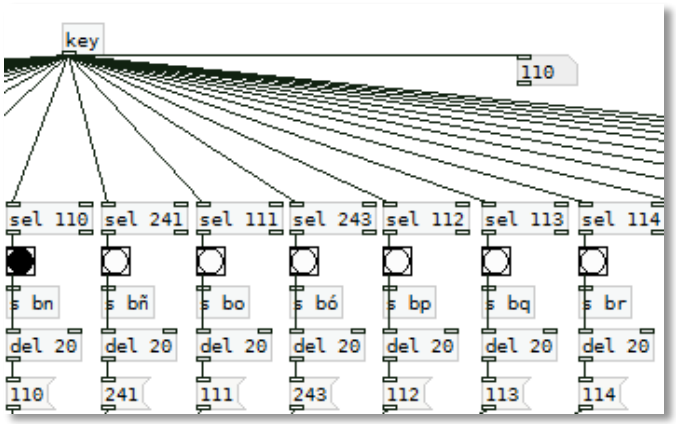

Figura 24. Patch de la obtención de nomenclatura numérica del teclado en Pure Data.
Fuente: Propia

[48] Captura los mensajes numéricos provenientes de un teclado.

Cuando se oprime una tecla, la función *key* recibe un número. En la Figura 24, este número es 110 correspondiente a la letra *n*. Inmediatamente, se realiza una comparación a través de la función *select*[49], la cual determina si el número de su argumento es igual al número que entra por su *inlet*[50], si son iguales se activa un objeto *bang*[51] . De esta forma se conoce las letras que están ingresando al sistema, incluyendo vocales con tilde.

Los números correspondientes a las letras se almacenan en una matriz llamada *letras* mediante el objeto *tabwrite*[52] como se muestra en la Figura 25. La variable *índice* contiene un contador que indica la posición de la matriz en donde se va almacenar cada dato.

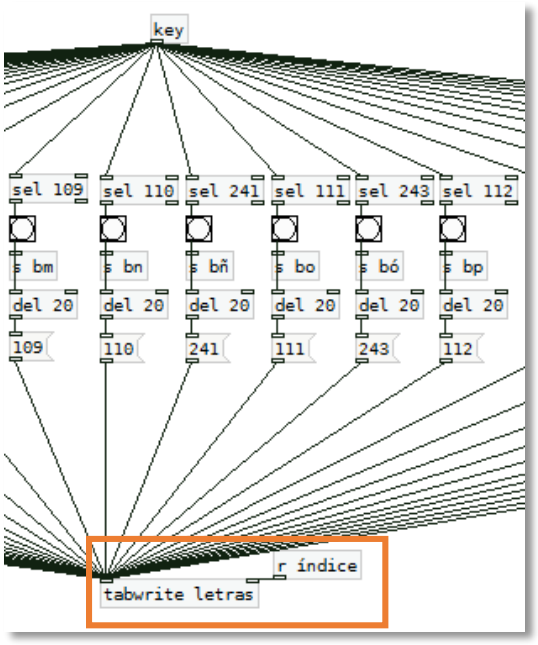

Figura 25. Almacenamiento en la matriz *letras.*

Fuente: Propia

Índice es un contador que aumenta a medida que se escribe el mensaje en el teclado. La Figura 26 muestra el proceso de borrado de letras en el sistema.

[49] Objeto Pure Data que compara si dos números son iguales.
[50] Entrada
[51] Objeto Pure Data que activa una acción.
[52] Objeto Pure Data que almacena datos en una matriz.

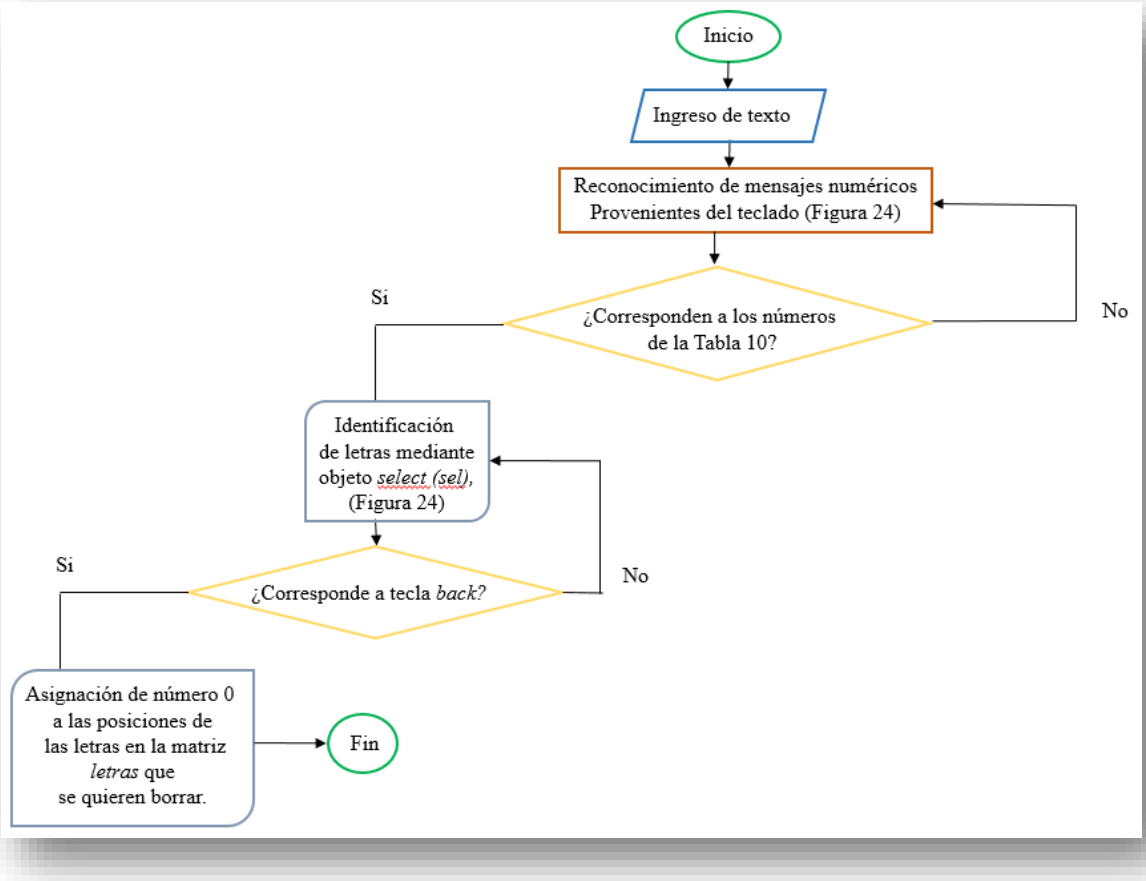

Figura 26. Diagrama de flujo de señal del proceso de borrado.

Fuente: Propia

Identificación de Fonemas.

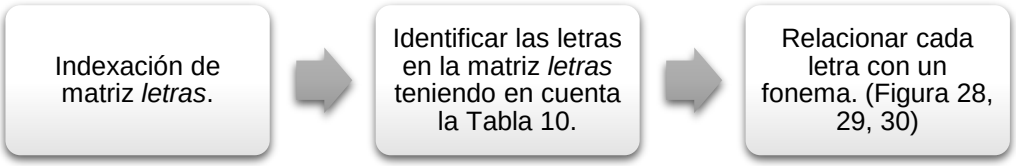

Figura 27. Proceso general para la identificación de fonemas.

Fuente: Propia

La Figura 27 presenta el proceso general para la identificación de fonemas. El primer paso consiste en indexar las posiciones de la matriz *letras,* con el objetivo de reconocer las letras almacenadas según la Tabla 10.

El segundo paso consiste en relacionar cada letra con un fonema según la Tabla 11 y 12. La Tabla 11 contiene las letras cuyos fonemas son identificados sin ningún tipo de análisis de contextualización[53].

Tabla 11. *Identificación de fonemas por letra sin análisis de contextualización.*

LETRA	FONEMA
a	/a/
b, v	/b/
e	/e/
i	/i/
o	/o/
u	/u/
d	/d/
f	/f/
k	/k/
j	/x/
m	/m/
n	/n/
ñ	/ñ/
p	/p/
q	/k/
s, z	/s/
t	/t/

Fuente: Propia.

[53] Análisis realizado para determinar los diferentes fonemas de una misma letra.

Tabla 12. *Identificación de fonemas por letra con análisis de contextualización.*

LETRA	FONEMA	EJEMPLO
c	/s/	▪ Be**c**erro
	/k/	▪ Hama**c**a
	ʧ	▪ En combinación con la letra h. Ej, Co**ch**e
g	/g/	▪ A**g**ujero, pa**g**ar, **g**orro ▪ En combinación con la letra *u* y la vocal *e* o *i* Ej, **Gu**iso
	/x/	▪ **G**énero, sa**g**itario
l	/l/	▪ E**l**efante
	/ ʎ/	▪ En combinación con la letra *l*. Ej, Be**ll**ota
r	/r/	▪ A**r**aña
	/ɾ/	▪ En combinación con la letra *r* o inicio de palabra. Ej, **R**atón, Ca**rr**o
y	/i/	▪ Re**y**
	/ ʎ/	▪ **Y**egua

Fuente: Propia

La Figura 28 presenta el diagrama de flujo de señal para el ejemplo de identificación de fonemas con análisis de contextualización realizado para la letra *c*, cuyo código se observa en las Figuras 29 y 30.

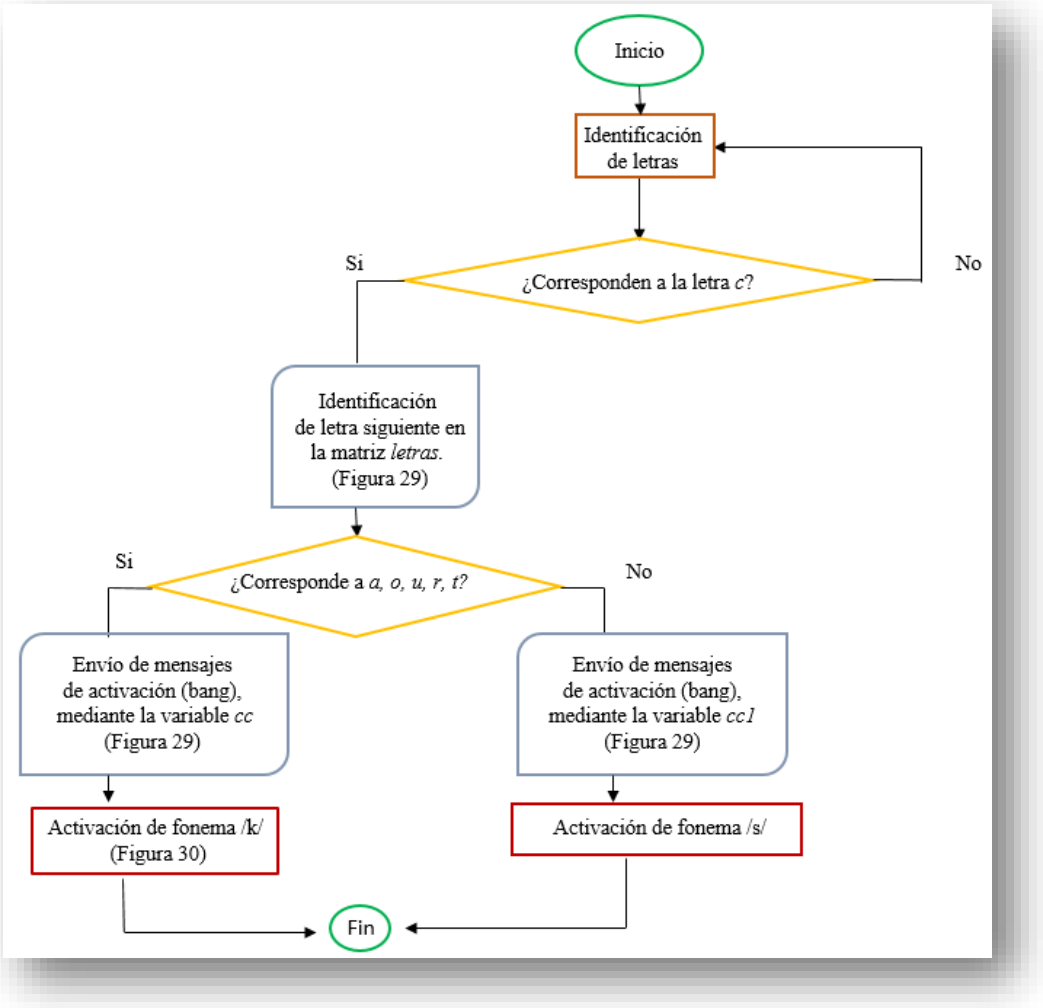

Figura 28. Diagrama de flujo de señal para la identificación de fonemas con análisis de contextualización para la letra *c*.

Fuente: Propia

En la Figura 29 se observa el código del subpatch[54] *pd letra c*, donde se analiza la letra siguiente, si es *a, o, u, r, t*, se envía la variable *cc* mediante un objeto *send (s)* y como se puede ver en la Figura 30, esta variable es recibida por un objeto *receive (r)*[55] para activar el fonema /k/. Los demás casos de la Tabla 12 tienen una programación similar para identificar sus fonemas.

[54] Es un objeto de Pure Data que se puede comparar a un cajón o a un contenedor que tiene en su interior código de programación.
[55] Objeto de Pure Data que recibe datos.

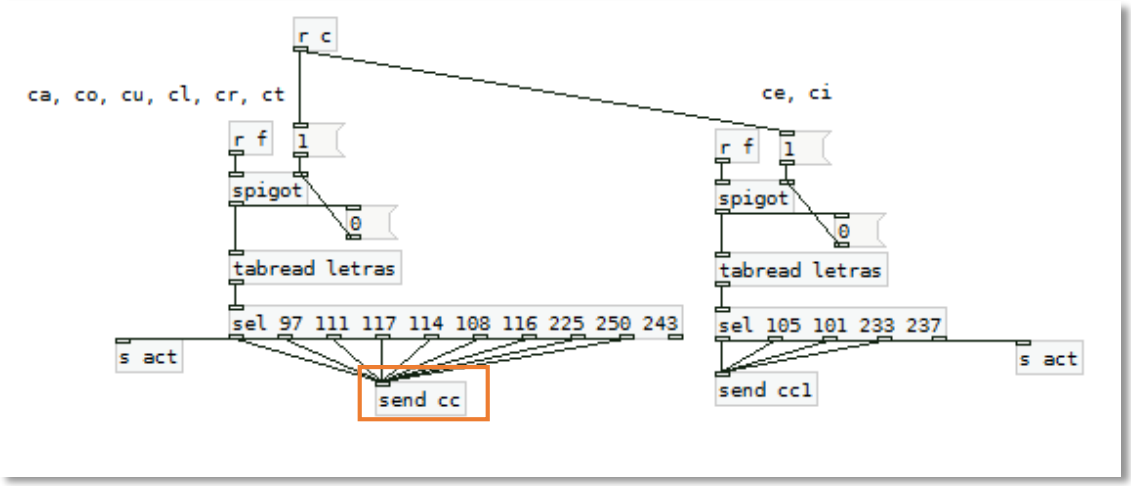

Figura 29. Identificación de fonemas con análisis de contextualización para la letra *c*.

Fuente: Propia

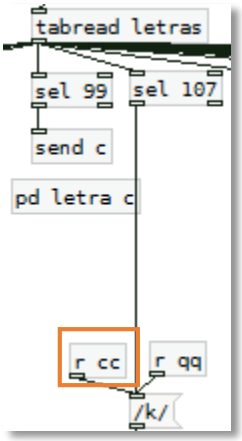

Figura 30. Identificación de fonemas con análisis de contextualización para la letra *c*.

Fuente: Propia

Identificación del Acento.

La identificación del acento depende de la presencia de tildes, es decir, si la palabra es ingresada con tilde, el sistema acentúa en la sílaba correspondiente, si no, se realiza el análisis explicado a continuación, el cual se desarrolla en dos etapas.

- La primera etapa consiste en identificar la cantidad de sílabas de cada palabra. Para esto, se realiza un conteo de las vocales; si hay un diptongo o triptongo, se cuenta como una vocal.
- La segunda etapa consiste en clasificar en palabras agudas o graves. Las palabras esdrújulas y sobreesdrújulas deben ingresarse con tilde para que el sistema acentúe correctamente, de lo contrario serán clasificadas como palabras graves. Las palabras agudas deben terminar en la letra *d, l, j, r, para* que el sistema las reconozca como tal, de lo contrario, se asume que la palabra es grave.

Agrupación de Fonemas.

La agrupación de fonemas consiste en realizar combinaciones entre sí para identificar sus correspondientes unidades de concatenación en el corpus del sistema, es decir sílabas, difonos y trifonos. Para esto, se creó una matriz llamada *concat*, en la cual se almacenan los números asignados a cada fonema siguiendo el proceso de la Figura 31, cuyo código se encuentra en la Figura 32.

Figura 31. Proceso general de almacenamiento de fonemas en la matriz *concat.*

Fuente: Propia

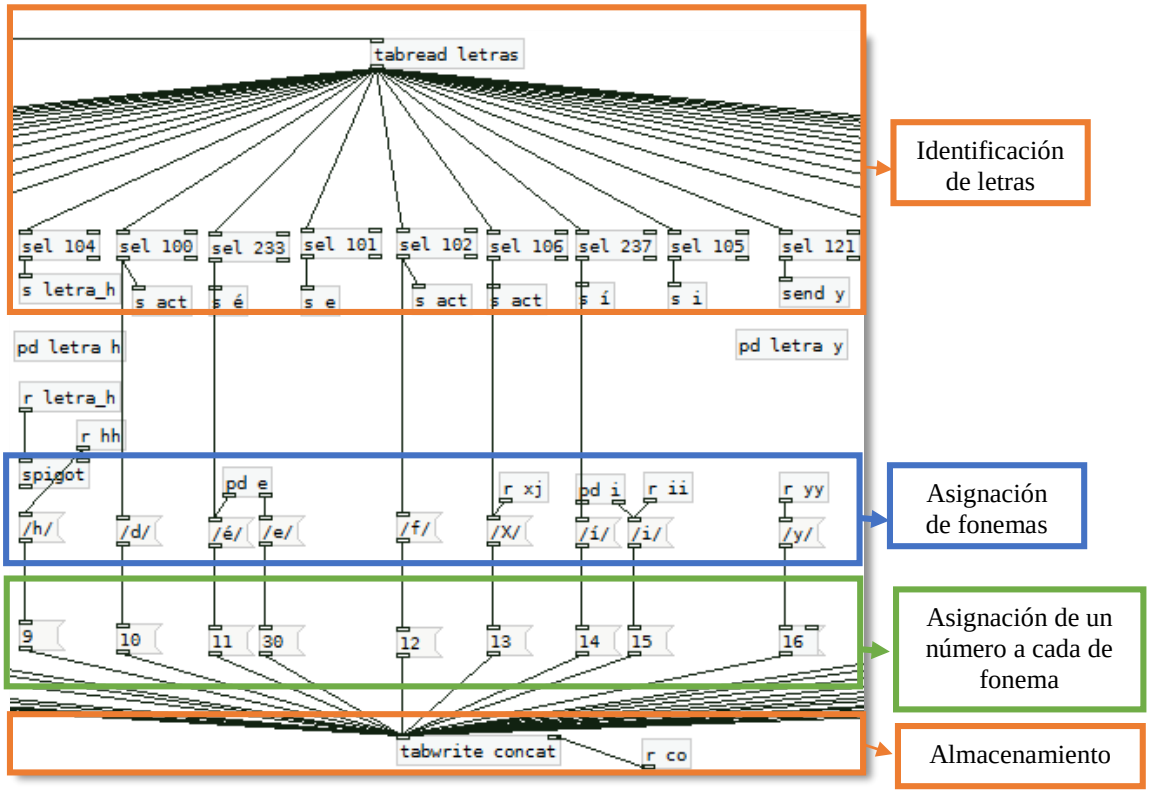

Figura 32. Almacenamiento de números asignados a los fonemas en la matriz *concat.*

Fuente: Propia

Cada sílaba tiene un subpatch con el fin de realizar individualmente la agrupación de fonemas. Todos los subpatches tienen la misma programación descrita en el diagrama de flujo de señal de la Figura 33. La Figura 34 muestra algunos de los subpatches.

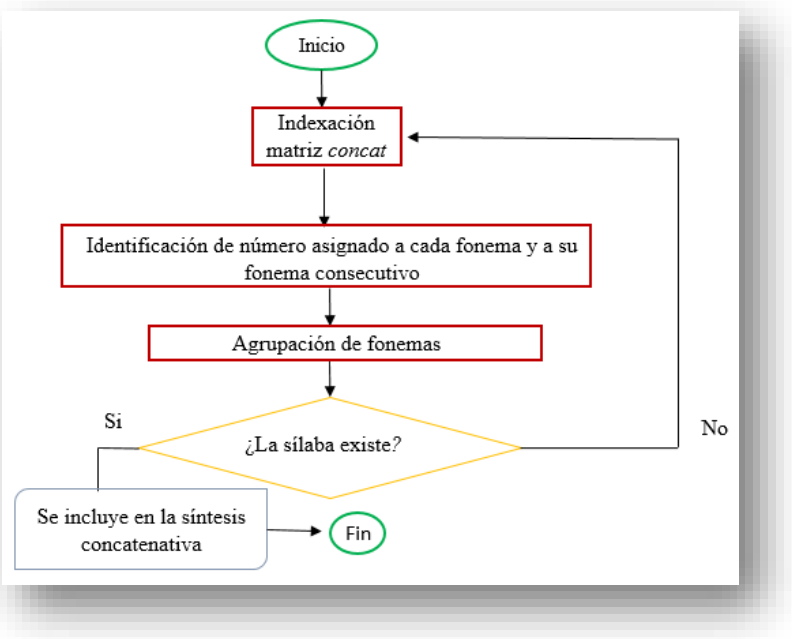

Figura 33. Diagrama de flujo de señal para la agrupación de fonemas.

Fuente: Propia

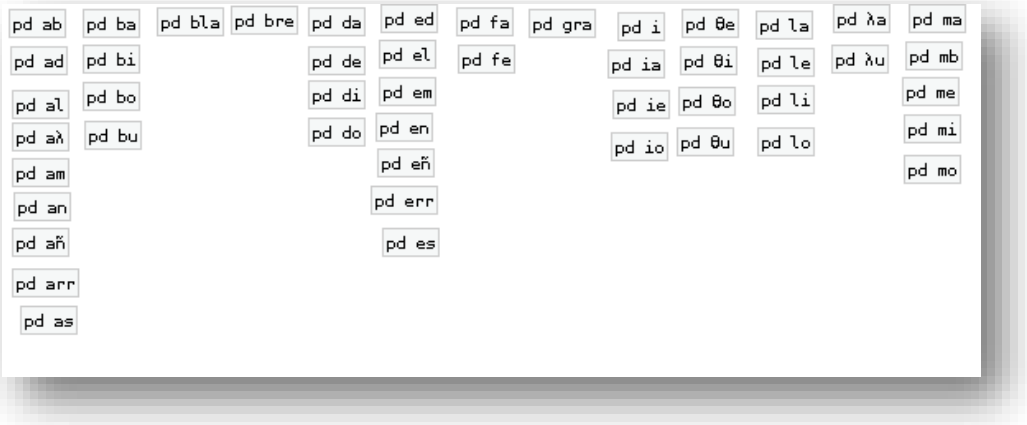

Figura 34. Algunos subpatches de sílabas.

Fuente: Propia

Para reproducir correctamente las sílabas, difonos y trifonos del corpus teniendo en cuenta la agrupación de fonemas en Pure Data, se realiza la identificación de fonemas en estas unidades. Es decir, si se analiza la palabra *araña* y la palabra *rata*, se observa que tienen en común la sílaba *ra*, la cual se pronuncia diferente en cada caso y por ende tiene diferentes fonemas. Para enlazar correctamente estos fonemas a sus correspondientes unidades en el corpus, se analiza su contexto

dentro de la palabra. En la agrupación de fonemas en Pure Data se establecen combinaciones de estas que están enlazadas a una sílaba, difono o trifono del corpus, para este ejemplo, la combinación de fonemas sería: /ra/ para la palabra *araña* y /ɾa/ para la palabra *rata*. De esta forma, la sílaba *ra* de las palabras *araña* y *rata* generadas por concatenación, sonará diferente para cada una de ellas si se ingresan al dispositivo. Por lo tanto, debe haber correlación entre los fonemas de las unidades del corpus y los fonemas identificados en el texto ingresado para realizar la síntesis concatenativa correctamente. Con este análisis se cumple la identificación de fonemas, propuesta en el primer objetivo específico, de los archivos de audio del corpus.

Concatenación y Generación de Sonido.

Para concatenar y generar sonido se sigue el procedimiento de la Figura 35. El código de este proceso se encuentra la Figura 36.

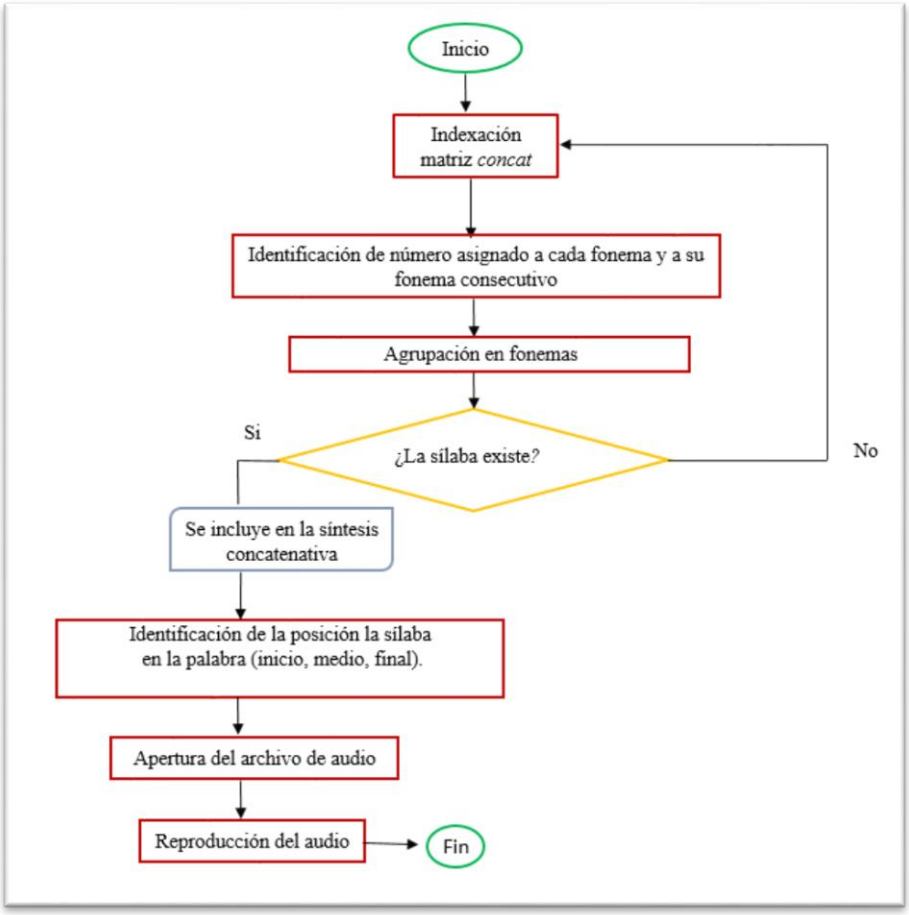

Figura 35. Diagrama de flujo de señal para la concatenación y generación de sonido.

Fuente: Propia

Los audios se reproducen secuencialmente conforme el orden de las sílabas del texto ingresado, por eso, en cada subpatch hay un objeto *delay*[56] *(del)* con la duración en milisegundos del audio en reproducción, para enviar la orden que activa el análisis para la agrupación de fonemas de la siguiente sílaba. De esta forma, se logra que la reproducción de los audios sea secuencial. Los audios son escuchados mediante el conversor digital análogo proporcionado por el objeto *dac~*[57].

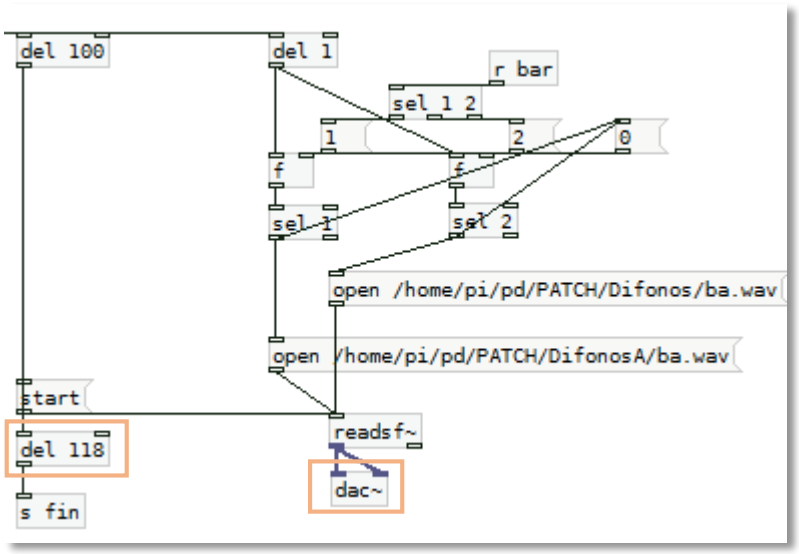

Figura 36. Generación de sonido.

Fuente: Propia

Ensamble del Dispositivo (Hardware y Periféricos)

Para presentar el sistema de conversión texto a voz como un producto o un dispositivo, es necesario realizar el ensamble correspondiente con todos los elementos necesarios para su funcionamiento. Por lo tanto, el dispositivo está conformado por el sistema embebido Raspberry Pi, pantalla LCD Adafruit i2c 16x2, teclado inalámbrico Rii Mini X1, mini amplificador digital PAM8403, interruptor redondo, conector estéreo 3.5 mm, parlante de computador, batería externa Power Bank de 1800 mAh[58]reales y un led indicador. El diagrama de la Figura 37 muestra la

[56] Objeto de Pure Data que genera un retraso en milisegundos.
[57] Objeto de Pure Data que realiza la conversión digital-análoga.
[58] Miliamperio-hora

conexión de los dispositivos y elementos previamente mencionados para realizar el ensamble, cuyas especificaciones técnicas se encuentran en el Apéndice E.

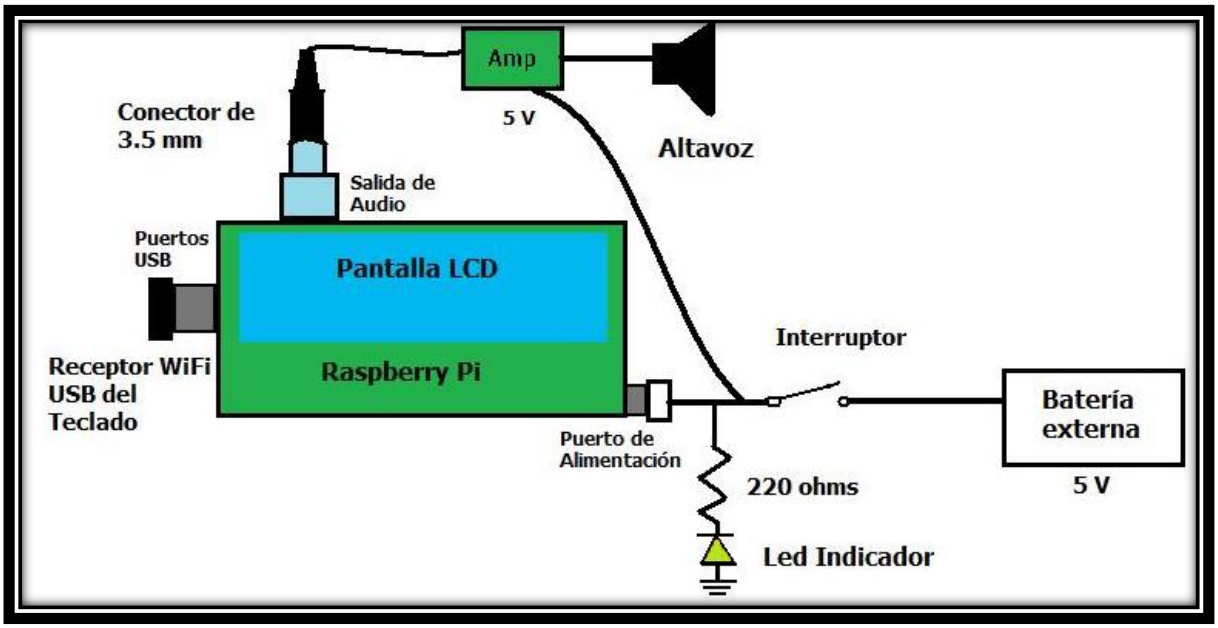

Figura 37. Diagrama de conexión del dispositivo de conversión texto a voz.
Fuente: Propia.

Sistema Embebido.

El sistema embebido utilizado para la implementación es Raspberry Pi Modelo B como se muestra en la Figura 38. El embebido tiene una salida de audio, puertos USB y micro USB, lector de tarjeta SD, salidas digitales (GPIO[59]), puerto ethernet y salida HDMI para conectar a un televisor con esta entrada y tener de esta forma, retroalimentación visual del sistema operativo del embebido. En el sistema embebido se instala el sistema operativo, el software Pure Data para ejecutar la conversión texto a voz y la librería pertinente de Python para programar la LCD.

[59] General Purpose input/output: Interfaz física entre Raspberry Pi y dispositivos externos.

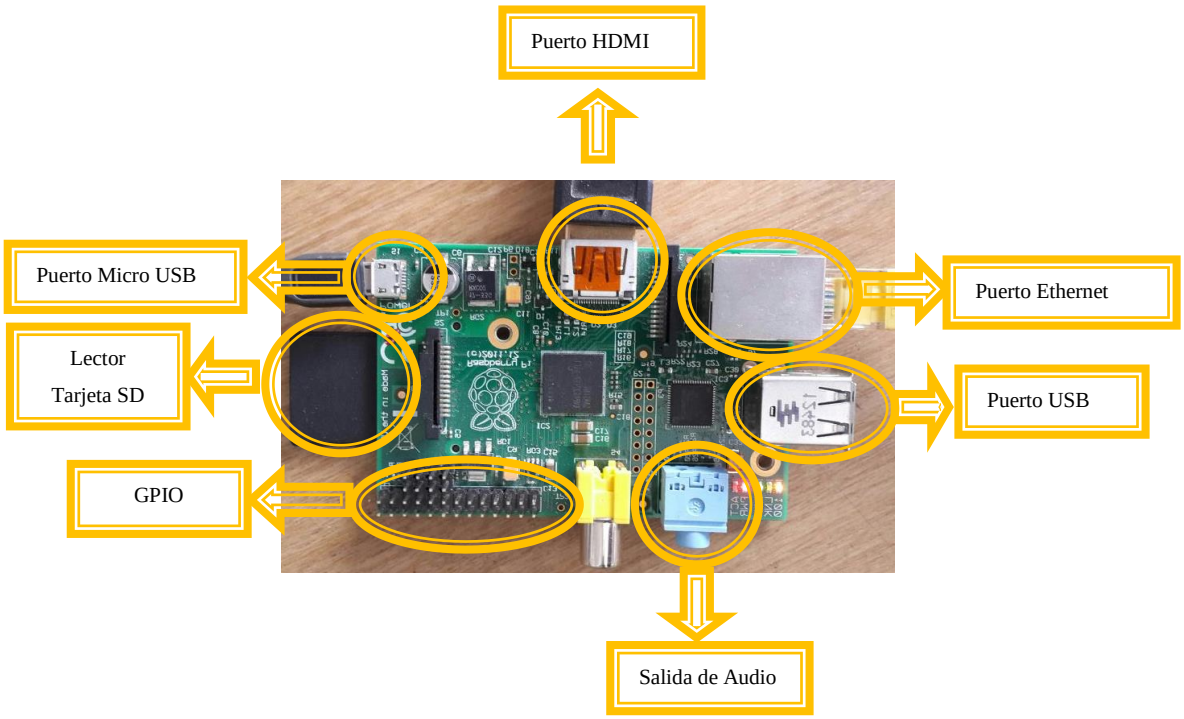

Figura 38. Raspberry Pi Modelo B.

Fuente: Propia

Instalación de Sistema Operativo y de Pure Data.

El proceso de instalación del sistema operativo Raspbian y de Pure Data en el sistema embebido Raspberry Pi se encuentra en el Apéndice B.

Implementación de Síntesis Concatenativa en el Sistema Embebido.

El patch de conversión texto a voz (TTS.pd) y los audios del sistema se guardaron en el embebido por medio de la red como se muestra en la Figura 39.

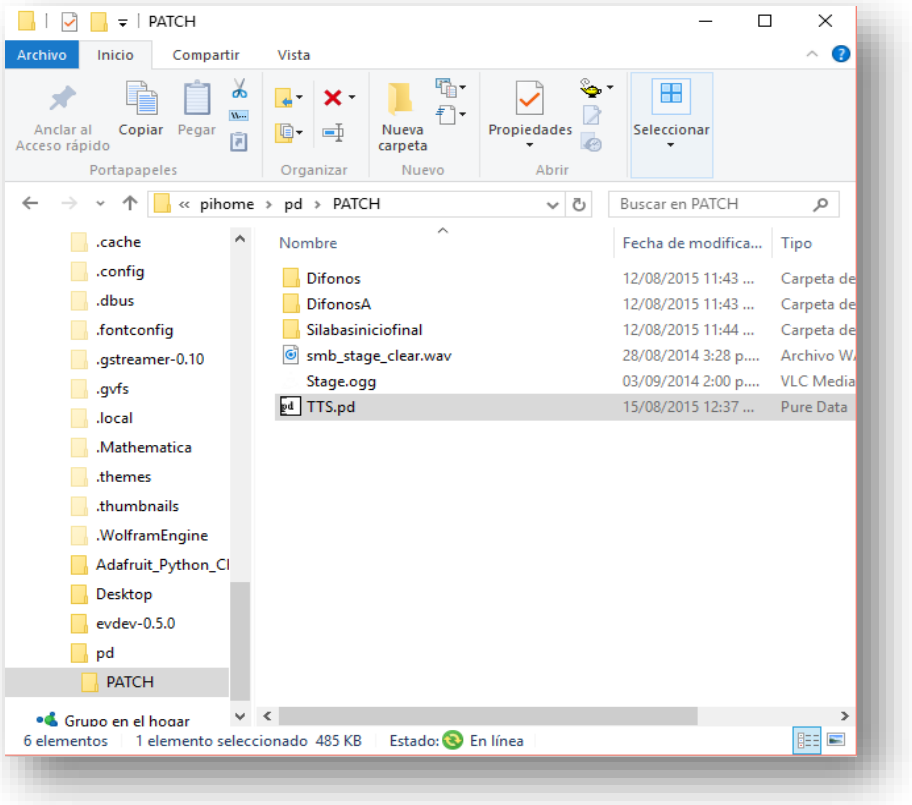

Figura 39. Acceso a los archivos del sistema embebido mediante red.

Fuente: Propia.

En un principio, estaban programados 250 sílabas en el patch, pero debido al error de incompatibilidad señalado en la Figura 40, la cantidad de sílabas disminuyó a 74 (Tabla 6). El error indica que la librería *readsf* de Pure Data tiene una limitación en la cantidad de audios a abrir en el patch cuando se ejecuta en la arquitectura ARMHF. De igual forma, el procesador y la memoria del sistema embebido también pueden influir en este inconveniente. Debido a que este sistema embebido es el único que cuenta con una versión de Pure Data pre compilada y tiene la comunidad de desarrolladores más grande del mundo en este campo, no se consideró necesario adquirir en el mercado otro sistema embebido, ni cambiar el lenguaje de programación por uno compatible con un sistema embebido con mayor capacidad de procesamiento y mayor memoria, debido a que Pure Data es un lenguaje optimizado para el procesamiento digital de señales y tiene una comunidad de desarrollo amplia y completa..

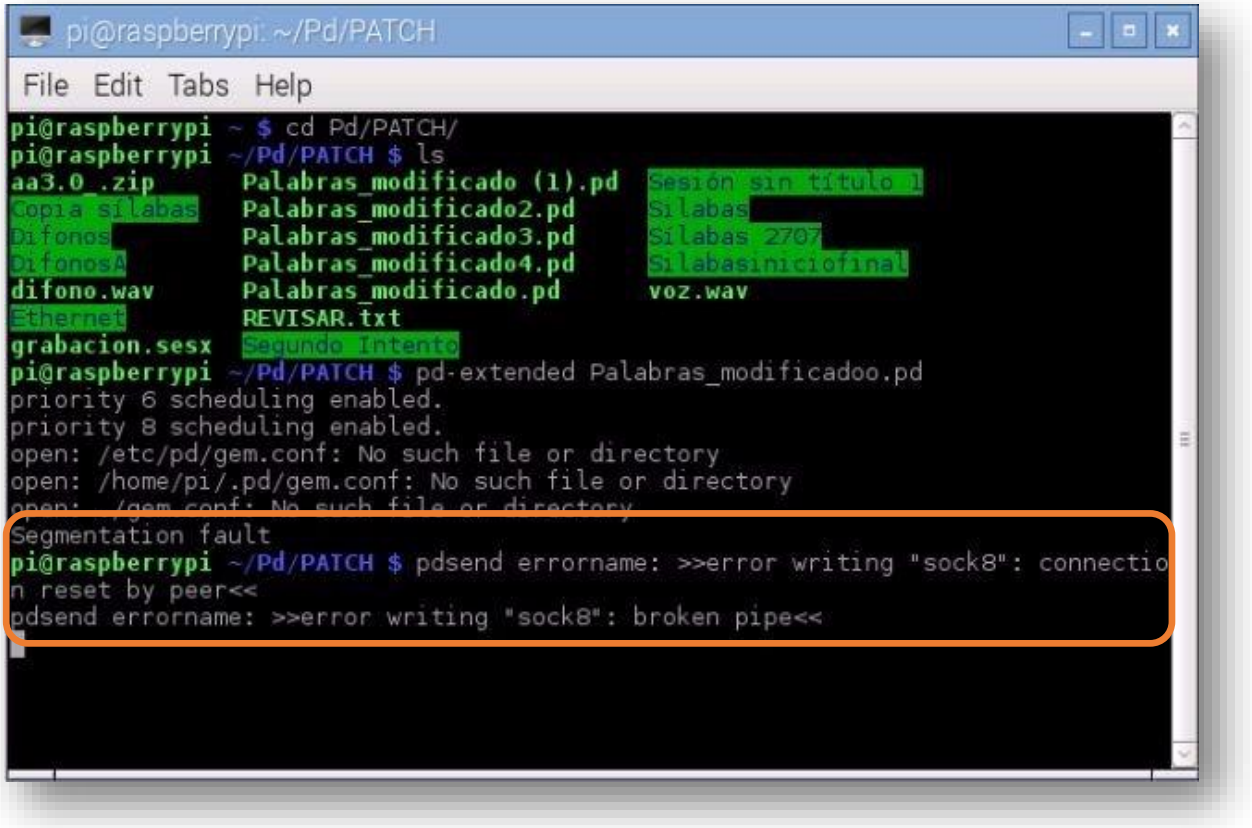

Figura 40. Error de incompatibilidad en la arquitectura ARMHF.

Fuente: Propia

Teclado Inalámbrico y Pantalla LCD.

El sistema tiene una pantalla LCD Adafruit[60] i2c 16x2 (Figura 41) y un teclado inalámbrico Rii Mini X1 (Figura 42) para ingresar el texto. Con el fin de incorporar las funciones de estos dispositivos al sistema embebido, se desarrolló un programa en Python [61] dentro del sistema operativo Raspbian. Python estaba pre instalado en el embebido (Ver Apéndice B).

[60] Compañía de aprendizaje y compra online de elementos electrónica.
[61] Lenguaje de programación administrado por la Pyhton Software Foundation

Figura 41. Pantalla LCD Adafruit i2c16x2 para Raspberry Pi

Fuente: https://learn.adafruit.com/adafruit-16x2-character-lcd-plus-keypad-for-raspberry-pi

*Figura 42.*Teclado inalámbrico Rii Mini X1.

Fuente:http://www.amazon.com/Wireless-Keyboard-Touchpad-mini-X1/dp/B00I5SW8MC/ref=pd_sim_147_4?ie=UTF8&refRID=1873R0KK7SY86Q28BRF3

La pantalla LCD se conectó en las salidas GPIO del sistema embebido. El teclado Rii Mini X1 funciona con conexión inalámbrica RF[62] de 2.4 GHz, tiene batería de litio recargable, función de suspensión y su distancia operativa es de 10 metros (Rii, 2015); fue conectado mediante un receptor WiFi USB al embebido sin necesidad de instalar drivers adicionales.

[62] Radio Frecuencia

Amplificador y Batería.

El PAM8403 de la Figura 43 es un amplificador de 3W, que funciona con un voltaje de alimentación de 5 V y con una carga de 4 ohms. El parlante Veco 35KM04-C de la Figura 44 tiene una impedancia de 4 ohm y una potencia de 1 W.

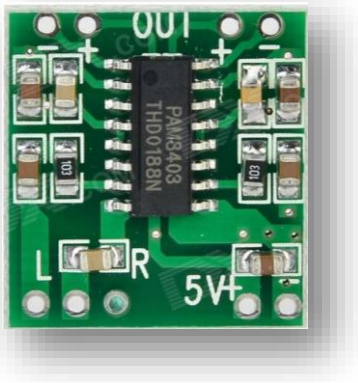

*Figura 43.*Amplificador digital PAM8403

Fuente:http://www.dx.com/p/pam8403-dc-5v-class-d-mini-digital-amplifier-board-module-green-347324#.VdDOD7J_Oko

Figura 44. Parlante de computador Veco 35KM04-C.

Fuente: Propia

Según la Figura 37, la alimentación de voltaje del sistema embebido depende de un interruptor, el cual debe ser accionado después de apagar el sistema operativo (Ver Apéndice B). El led tiene la función de indicar si el embebido esta encendido o no. Todos estos dispositivos y elementos fueron colocados dentro de una caja de plástico de 16 x 10 x 5 cm aproximadamente como se ve en la Figura 46. La entrada micro USB de la batería externa (Figura 45) tiene una extensión micro USB-USB incorporada a una de las superficies laterales de la caja. Esto se

realizó con el fin de cargar la batería desde la parte externa de la caja (Figura 47). La batería se carga en dos horas y dura dos horas de uso continuo.

Figura 45. Batería externa.
Fuente:http://www.virtualstore.com.co/accesorios/93-bateria-externa-portatil-2600mah-celular-camara-.html

Figura 46. Dispositivo final de conversión texto a voz.
Fuente: Propia.

Figura 47. Carga del Dispositivo.

Fuente: Propia.

Capítulo 5.
Análisis de Resultados

El análisis de resultados se divide en dos partes: generación de palabras y prueba del dispositivo con personas. El dispositivo se denominó *Tespeecon* (Text to Speech Converter[63]), pesa 28.2 gr. y su batería dura dos horas de uso continuo.

Generación de Palabras

La Tabla 13 tiene 160 palabras generadas por el dispositivo a partir de la combinación de sílabas de la Tabla 6, contenidos en el sistema. Para las pruebas realizadas, se seleccionaron solo las palabras que se usan comúnmente en la comunicación (Austria, 2014) (Tabla 16).

Tabla 13. *Palabras generadas por el dispositivo.*

Bueno(s)	Baño	Español	Vale	Enano(s)	Sal	Oso(s)
Día(s)	Tengo	Entiendo	Eso(s)	Enanas(s)	Sales	Mesa(s)
Soy	Hambre	Quien(es)	Dice(s)	Buen	Bola(s)	Masa(s)
Feliz	No	Es	Si	Esto	Bolos(s)	Taza(s)
Hola	Se	Con	Dolor	Son	Hada(s)	Tiza(s)
Que	En	Permiso	Muy	La(s)	Mi(s)	Queso(s)
Tal	Cuanto(s)	Ayuda	Bien	Lo(s)	Misa	Visa(s)
Cómo	Tiempo(s)	Por	Gracias	Esa(s)	Pañal(es)	Paso(s)
Está(s)	Sed	Favor	Adiós	Hermana(s)	Dado(s)	Grasa(s)
Donde	Buena(s)	Perdón	Noche(s)	Hermano(s)	Dedo(s)	Lana
Está	Tarde(s)	Estoy	Hoy	Mamá(s)	Maña(s)	Lento(s)
El	Habla	Triste	Cual(es)	Ana	Olor	Liso
Loza	Llama(r)	Pesa(s)(r)	Pesado(s)	Posada	Ser	España
Tanto	Tema(s)	Talar	Tomar	Mudo(s)	Sordo(s)	Pez
Mar	Ala(s)	Amor	Asar	Dos	Docena	Bazar
Vía(s)	Tía(s)	Bozal	Loma	Pozo(s)	Vaso(s)	

Fuente: Propia.

[63] Conversor de texto a voz.

Prueba de Funcionamiento del Dispositivo con Personas

Las encuestas elaboradas para probar el dispositivo se encuentran en el Apéndice C. Las pruebas se realizaron con:

- Personas con discapacidad para hablar: Personas de la Asociación de Sordos de Suba *Asorsub*.
- Personas oyentes, para evaluar la inteligibilidad de las palabras.

Prueba con Personas con Discapacidad para Hablar.

Las pruebas con personas con discapacidad para hablar se llevaron a cabo en la sede de la Asociación de Sordos de Suba *Asorsub*. La definición de la población depende de la cantidad de personas pertenecientes a la asociación, es decir 30 miembros. La muestra es de 14 personas a las cuales se les explicó el procedimiento mediante una persona intérprete[64]. La Tabla 14 muestra la cantidad de personas con sordera y con deficiencia auditiva moderada[65].

Tabla 14. *Nivel de discapacidad.*

NIVEL DE DISCAPACIDAD AUDITIVA	CANTIDAD DE PERSONAS
Sordera	13
Deficiencia Auditiva Moderada Con Audífono	1

Fuente: Propia.

El objetivo de la prueba era evaluar las categorías: utilidad, comodidad y funcionamiento en el uso del dispositivo. El funcionamiento se evaluó mediante la pregunta 3; la comodidad se relaciona con el peso físico, tamaño y facilidad de uso del teclado y se evaluó mediante las preguntas 1, 2 y 4; la utilidad del dispositivo para personas con discapacidad para hablar se evaluó mediante la pregunta 5. Para esto, se formularon las preguntas de la Tabla 15. El procedimiento consiste en ingresar algunas palabras de la Tabla 16 en el teclado, observarlas en

[64] Persona que comunica y traduce la lengua de señas
[65] Imposibilidad de seguir una conversación normal si existe ruido de fondo.

la pantalla LCD y responder las preguntas marcando con una **x.** La escala es de 1 a 5, donde 5 es excelente y 1 es pésimo.

Tabla 15. *Preguntas de la encuesta realizada a personas sordas.*

	PREGUNTA	1	2	3	4	5
1	¿Considera que el peso físico del dispositivo es apropiado para su función?					
2	¿Considera que el tamaño del dispositivo es apropiado para su función?					
3	¿Se entienden las palabras en la pantalla?					
4	¿Es fácil oprimir las teclas del teclado?					
5	¿Considera que el dispositivo es útil para personas con discapacidad para hablar?					

Fuente: Propia.

No todas las personas con discapacidad para hablar son sordas, como por ejemplo, las personas laringectomizadas. Infortunadamente, no fue posible realizar esta prueba con personas que aun teniendo la discapacidad para hablar, siguen siendo oyentes. Por tal razón, no se requirió elaborar una encuesta que evaluara la inteligibilidad de las palabras en este caso.

Tabla 16. *Palabras y frases utilizadas para la prueba del dispositivo con personas sordas.*

Buenos días	Soy feliz	¿Cómo dice?	Adiós
Hola	No entiendo	Tengo dolor	Mal
¿Dónde está el baño?	¿Quién es?	Si	Mañana
Tengo hambre	Con permiso	Muy bien	Hoy
¿Cuánto tiempo?	Ayuda	Gracias	¿Qué es eso?
No sé	Perdón	Bueno	¿Habla español?
Tengo sed	Estoy triste	¿Cómo estás?	Por favor
Buenas tardes	¿Cuánto vale eso?	Buenas noches	¿Qué tal?

Fuente: Propia.

Los resultados de las pruebas están dados en porcentajes que indican la cantidad de personas que respondieron las preguntas, como se muestra en la Figura 48 y 49.

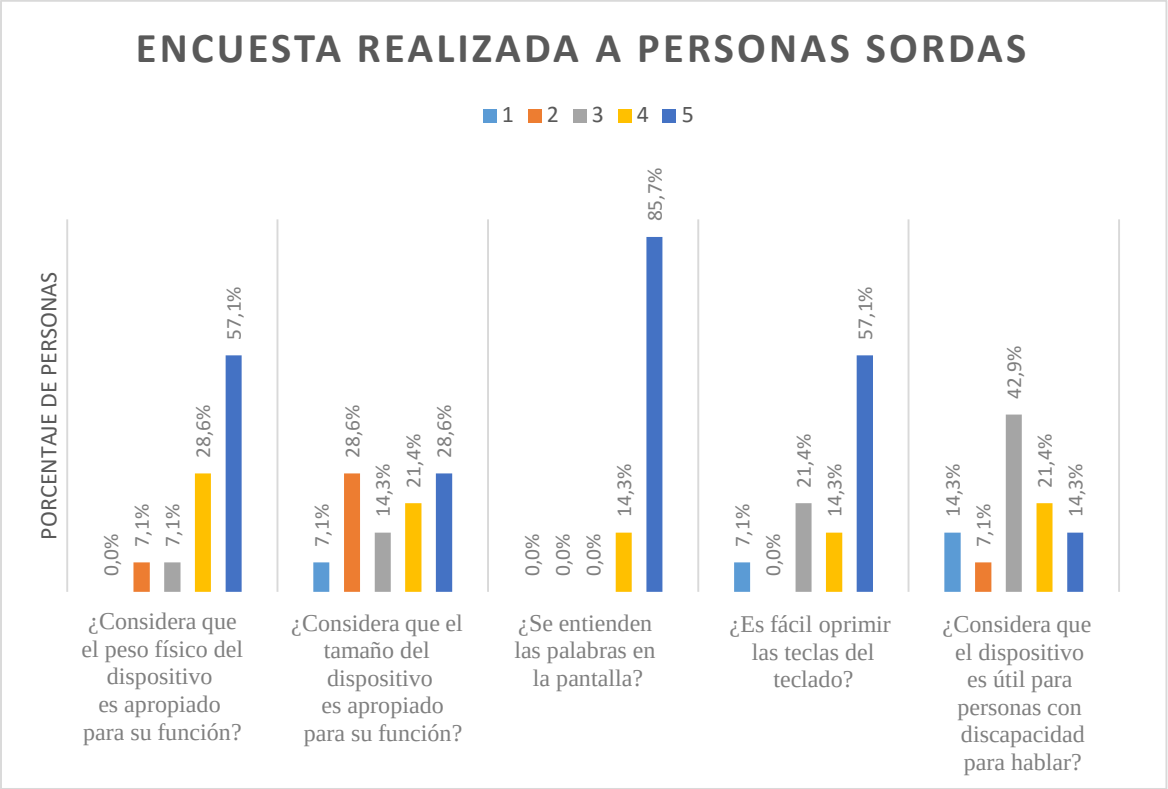

Figura 48. Resultados de la encuesta realizada a personas sordas.

Fuente: Propia.

Al final de la encuesta, se solicitó calificar la experiencia de uso del dispositivo. La escala de la pregunta tiene cinco opciones de respuesta: *Muy buena*, *buena*, *aceptable*, *mala* o *muy mala*, debido a que se requería conocer la opinión específica de las personas. La Figura 49, presenta el resultado obtenido.

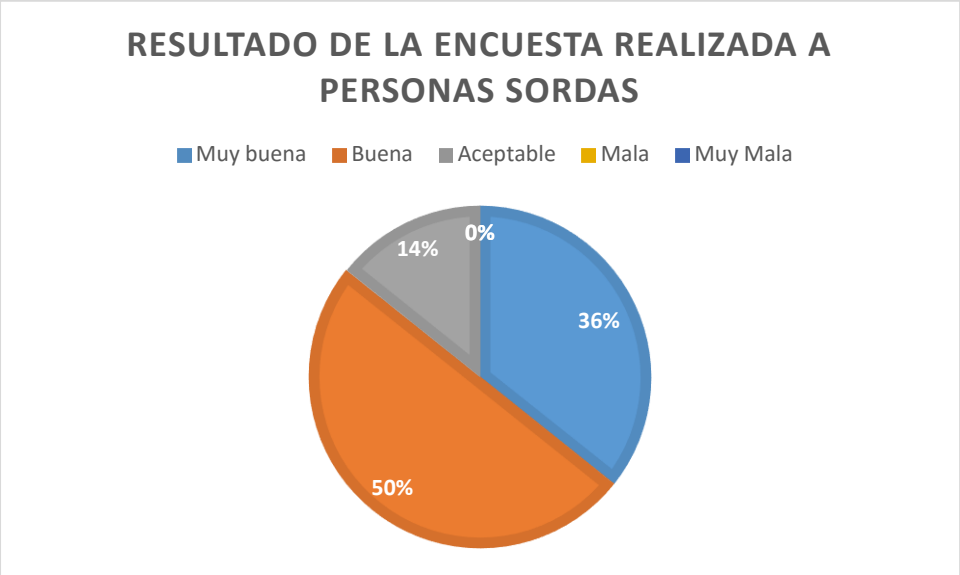

Figura 49. Resultado de la encuesta realizada a personas sordas.

Fuente: Propia.

Según los resultados de las Figuras 48 y 49, el análisis estadístico presentado en las Tablas 20, 21, 22, 23, 24 y 25 del análisis estadístico en el Apéndice D y teniendo en cuenta los comentarios de las personas, se puede interpretar que:

- Hay una tendencia general a considerar excelente el peso físico del dispositivo.
- Un tercera parte de las personas encuestadas considera que el tamaño del dispositivo es apropiado para su función y otra tercera parte considera lo contrario.
- En promedio, las personas encuestadas afirman que las palabras se entienden en la pantalla.
- Cerca de la mitad de las personas encuestadas considera que es fácil oprimir las teclas.
- Para algunas personas el dispositivo es relativamente útil, para la mayoría es útil y para otras no es útil.
- En promedio, la experiencia de utilizar el dispositivo fue buena debido a que para las personas encuestadas es agradable e innovador hacer uso de herramientas que puedan ayudar a complementar su medio de comunicación.

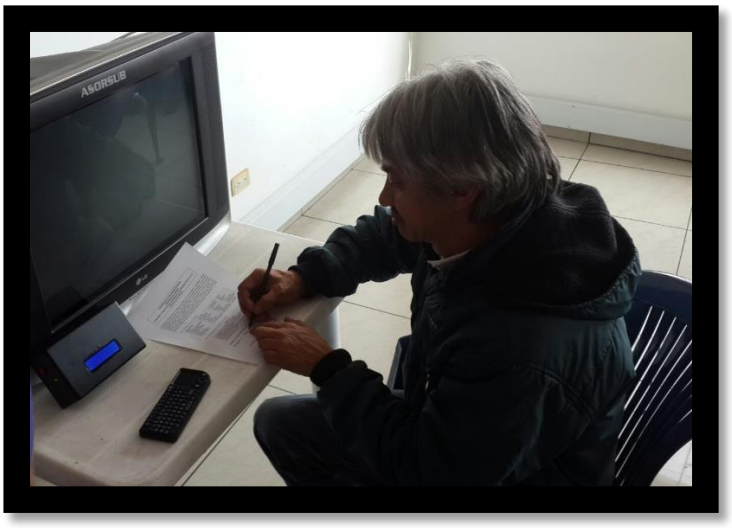

Figura 50. Persona con discapacidad para hablar realizando la prueba.

Fuente: Propia.

Prueba con Personas Oyentes.

La prueba con personas oyentes tenía el objetivo de evaluar la inteligibilidad de las palabras, al igual que las categorías: funcionamiento general, comodidad y utilidad en el uso del dispositivo. Por tal razón, el procedimiento de evaluación se dividió en dos partes:

- La primera parte consiste en que la persona encuestada escucha las frases de la Tabla 17 generadas por el sistema y escribe lo que entiende de ellas.
- La segunda parte consiste en que la persona encuestada ingresa frases y palabras de la Tabla 17 al sistema mediante el teclado y dependiendo de su percepción frente a estas, responde marcando con una **x** a las preguntas de la Tabla 19. El funcionamiento del dispositivo se evaluó mediante las preguntas 3, 6, 7, 8 y 9; la comodidad se relaciona con el peso físico, tamaño y facilidad de uso del teclado y se evaluó mediante las preguntas 1, 2 y 4; la utilidad del dispositivo para personas con discapacidad para hablar se evaluó mediante la pregunta 5. La escala es de 1 a 5, donde 5 es excelente y 1 es pésimo

En total, se encuestaron 25 personas.

Tabla 17. *Frases ingresadas al sistema para realizar prueba del dispositivo con personas oyentes.*

Buenos días, soy feliz	Ayuda por favor
Hola, ¿qué tal?, ¿cómo estás?	Perdón, estoy triste
¿Dónde está el baño?	¿Cuánto vale eso?
Tengo hambre	¿Cómo dice?
No sé en cuanto tiempo	Si tengo dolor
Tengo sed	Muy bien, gracias
Buenas tardes, ¿habla español?	Bueno, adiós
No entiendo quién es	Buenas noches
Con permiso	Es hoy

Fuente: Propia.

El resultado de inteligibilidad de las palabras para la primera parte de la prueba, está dado en porcentajes en la Tabla 18.

Tabla 18. *Porcentajes de inteligibilidad de las palabras de prueba.*

PALABRA		PALABRA		PALABRA		PALABRA	
Buenos	96%	**Español**	96%	**Baño**	96%	**Vale**	88%
Días	96%	**Entiendo**	84%	**Tengo**	96%	**Eso**	80%
Soy	96%	**Quien**	88%	**Hambre**	96%	**Dice**	92%
Feliz	88%	**Es**	96%	**No**	92%	**Si**	96%
Hola	100%	**Con**	96%	**Se**	92%	**Dolor**	96%
Que	100%	**Permiso**	96%	**En**	92%	**Muy**	96%
Tal	100%	**Ayuda**	100%	**Cuanto**	88%	**Bien**	96%
Cómo	92%	**Por**	100%	**Tiempo**	88%	**Gracias**	100%
Estás	96%	**Favor**	100%	**Sed**	96%	**Bueno**	96%
Donde	96%	**Perdón**	96%	**Buenas**	100%	**Adiós**	96%
Está	96%	**Estoy**	100%	**Tardes**	100%	**Noches**	96%
El	96%	**Triste**	100%	**Habla**	88%	**Hoy**	100%

Fuente: Propia.

Tabla 19. *Preguntas de la encuesta realizada a personas oyentes.*

	PREGUNTA	1	2	3	4	5
1	¿Considera que el peso físico del dispositivo es apropiado para su función?					
2	¿Considera que el tamaño del dispositivo es apropiado para su función?					
3	¿Se entienden las palabras en la pantalla?					
4	¿Es fácil oprimir las teclas del teclado?					
5	¿Considera que el dispositivo es útil para personas con discapacidad para hablar?					
		Si			No	
6	¿Todas las palabras son entendibles?					
7	Si respondió *No* a la pregunta anterior, ¿qué palabra(s) no entendió?					
8	¿Escuchó el acento en todas las palabras?					
9	Si respondió *No* a la pregunta anterior, ¿qué palabra(s) no escuchó con acento?					

Fuente: Propia.

Como se puede observar en la Tabla 18, los porcentajes indican que las palabras son inteligibles, esto quiere decir que los mensajes sonoros generados se entienden con claridad y con contexto. El promedio de inteligibilidad es de 95 %. Las palabras que tuvieron menos de 90% de inteligibilidad son: *entiendo, cuanto, quien, tiempo, habla, eso y vale,* esto se debe a que la variación en la distribución de energía de la transición espectral entre las sílabas y difonos de estas palabras, específicamente en las siguientes combinaciones: /ua/-/an/, /ie/-/en/ y /es/-/so/. Por ejemplo, la transición espectral entre /ua/ y /an/ de la palabra “cuanto”, no es uniforme a partir de 1.5 KHz como se ve en la Figura 51, a comparación de lo que se observa en la Figura 52 donde la transición entre los difonos /gra/ y /as/, para conformar la sílaba *gras* de la palabra “gracias", presenta mayor homogeneidad en todo el espectro y 100 % de inteligibilidad según la Tabla 18.

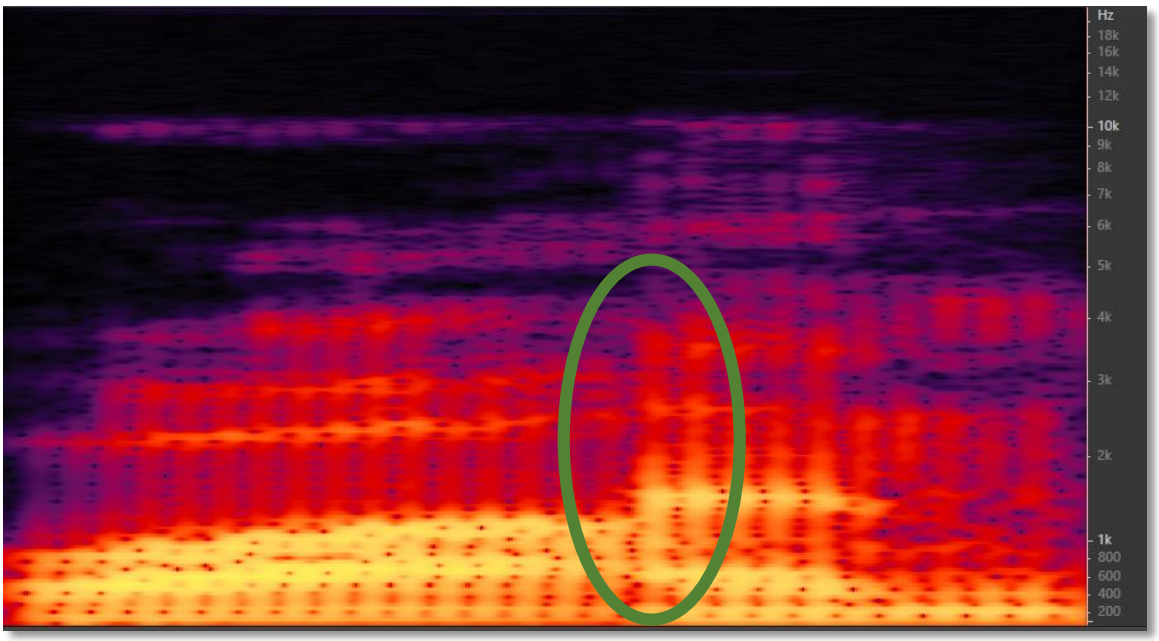

Figura 51. Variación en la transición espectral entre los difonos /ua/-/an/, de la palabra cuanto conformada por concatenación de unidades.

Fuente: Propia.

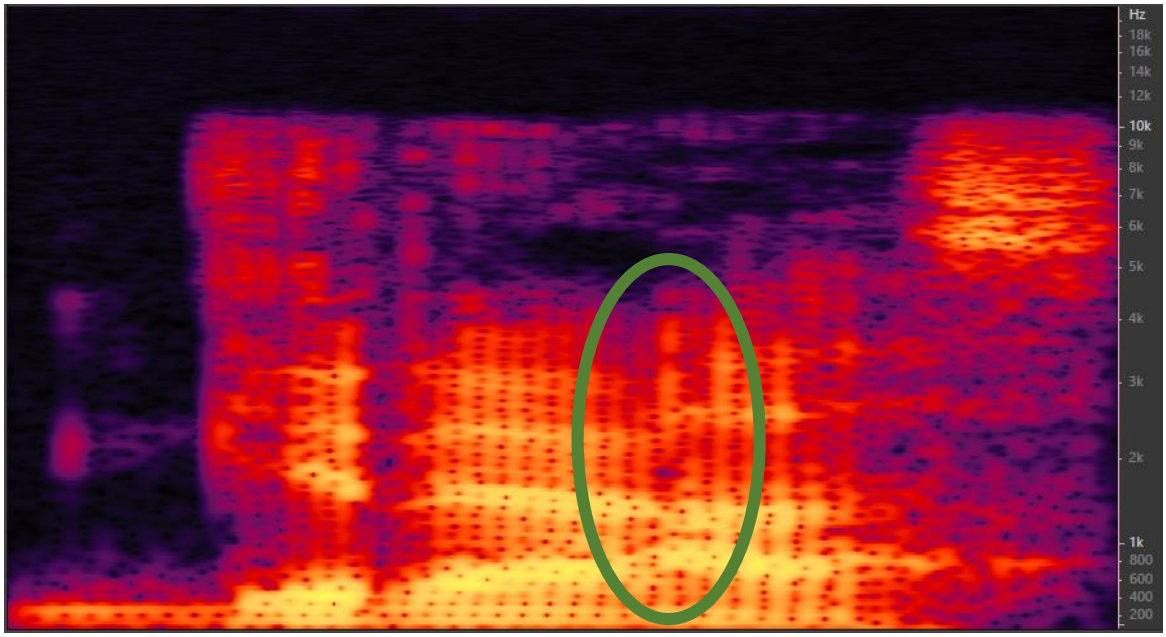

Figura 52. Homogeneidad en la transición espectral entre los difonos /gra/-/as/ de la palabra gracias conformada por concatenación de unidades.

Fuente: Propia.

Los resultados de la segunda parte de la prueba están dados en porcentajes que indican la cantidad de personas que respondieron las preguntas, como se muestra en la Figura 53 y 54.

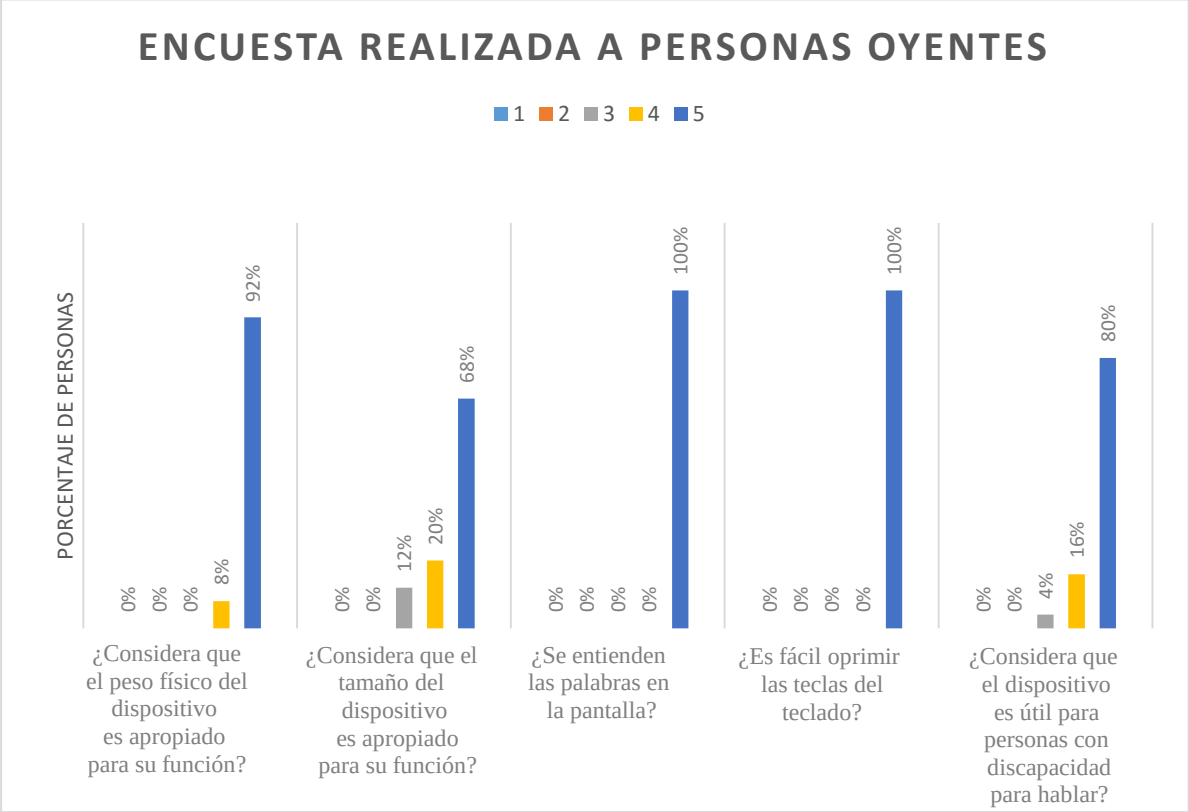

Figura 53. Resultados de la encuesta realizada a personas oyentes.

Fuente: Propia.

Al final de la encuesta, se solicitó calificar la experiencia de uso del dispositivo. La escala de las preguntas tiene cinco opciones de respuesta: *Muy buena, buena, aceptable, mala* o *muy mala*, debido a que se requería conocer la opinión específica de las personas. La Figura 54, presenta el resultado obtenido.

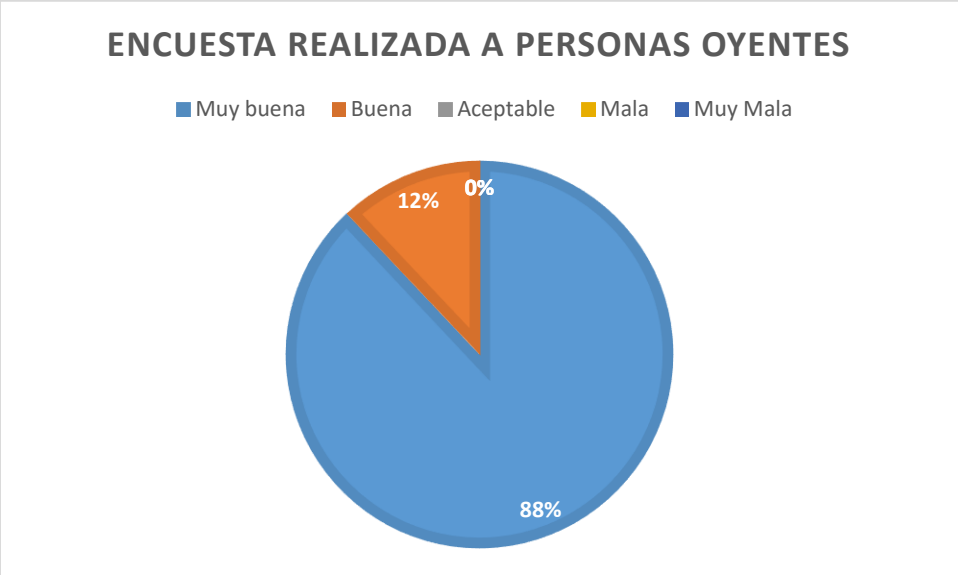

Figura 54. Resultado de la encuesta realizada a personas oyentes.

Fuente: Propia.

El 43% de las personas entendieron las frases de pregunta y las escribieron como tal. Según los resultados de la Figura 53 y 54 y las Tablas 26, 27, 28, 29,30 y 31 del análisis estadístico del Apéndice D, se puede interpretar que:

- Casi todas las personas encuestadas califican como excelente el peso físico del dispositivo.
- Hay una tendencia general a considerar que el dispositivo tiene el tamaño apropiado para su función.
- Todas las personas encuestadas consideran que se entienden las palabras en la pantalla y que es fácil oprimir las teclas.
- Hay una tendencia a la consideración de que el dispositivo es útil para personas con discapacidad para hablar.
- En promedio, la experiencia de utilizar el dispositivo fue muy buena debido a que para las personas encuestadas es agradable e innovador hacer uso de herramientas que puedan ayudar a complementar el medio de comunicación de personas con discapacidad para hablar.

Al iniciar la prueba, se les notificó a las personas que si no entendían el acento de una palabra, avisaran a la persona encargada para responder la pregunta 9 de la Tabla 19. Se obtuvo un promedio de percepción de acento en las palabras de 99 %. Tres personas no escucharon acento en las palabras en*tiendo* y *cuanto* y dos personas en la palabra *noche*, debido a la distribución de energía de las señales. Por ejemplo, en la Figura 55 se puede observar que la sílaba *che* de la palabra *noche* conformada manualmente por sílabas y difonos, tiene concentraciones de energía claramente separadas en todo el espectro, a diferencia de lo que se observa en la Figura 56, donde la energía de esta misma sílaba proveniente de la grabación de la palabra *noche*, está mejor distribuida en todo el espectro, sin cambios espectrales bruscos entre los fonos y con un decaimiento más prolongado al final, dando de esta forma "protagonismo" a la sílaba *no*, que es más contundente las concentraciones de energía que presenta.

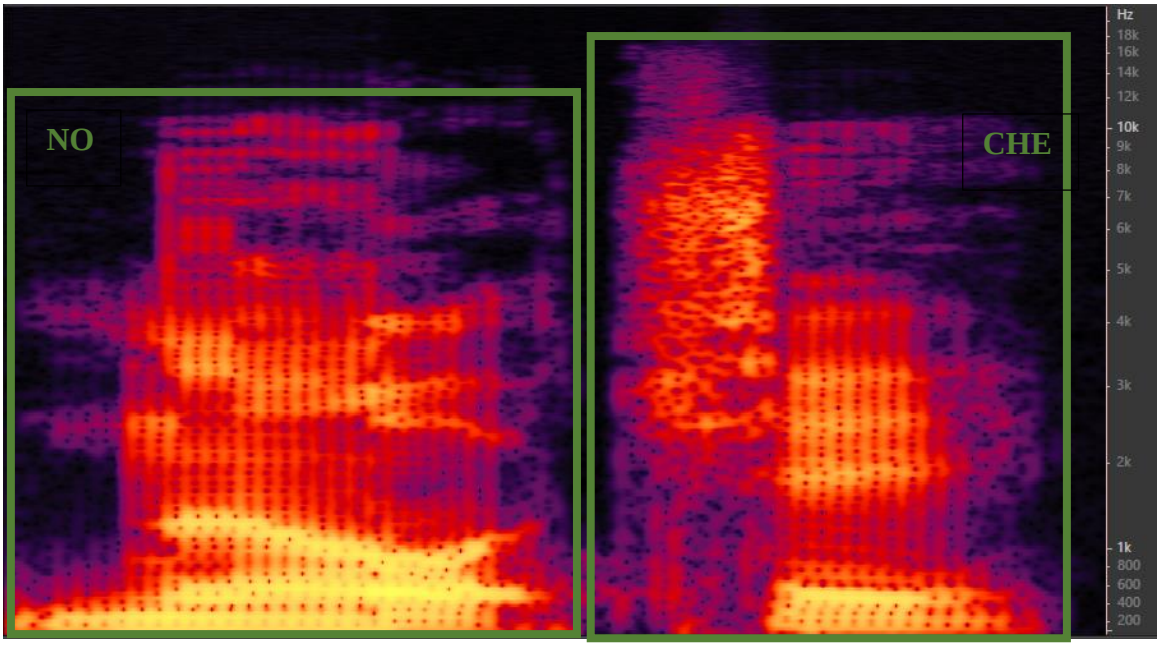

Figura 55. Espectrograma de la palabra noche conformada manualmente por difonos.

Fuente: Propia.

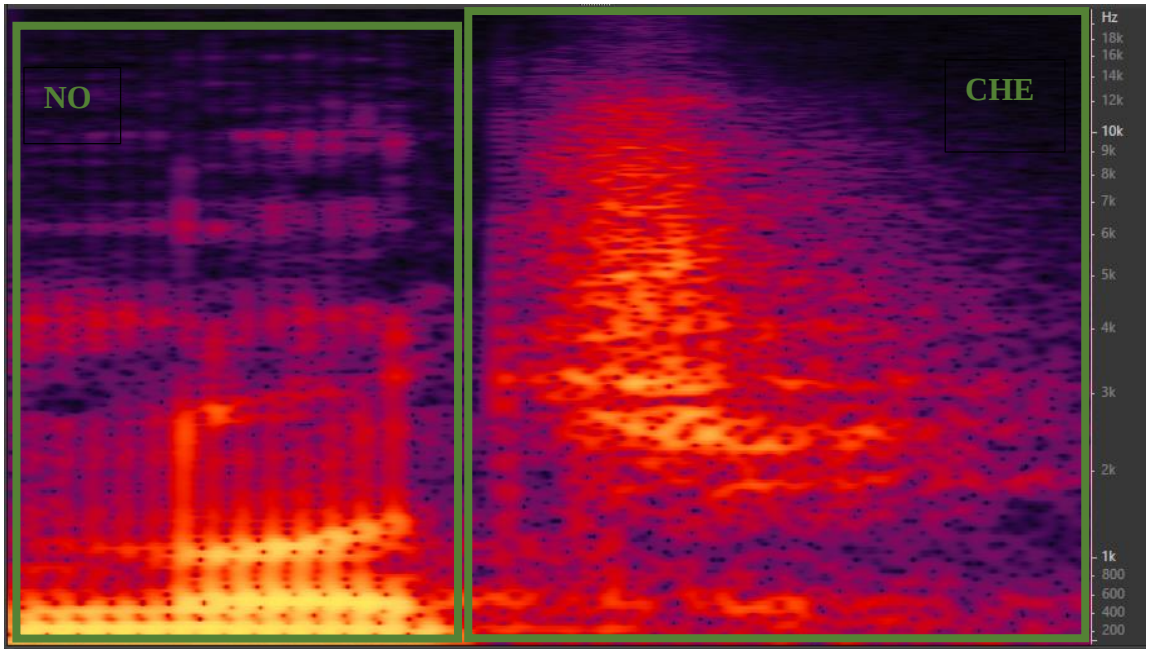

Figura 56. Espectrograma de la palabra noche grabada.

Fuente: Propia.

Los resultados de la Figuras 53 y 48 y del análisis estadístico del Apéndice D varían teniendo en cuenta:

- La pregunta N° 1 y la pregunta N° 2 indagan la expectativa de las personas respecto al peso y tamaño del dispositivo ya que las respuestas obtenidas son subjetivas. Las personas con discapacidad para hablar esperan que los dispositivos que faciliten su comunicación sean lo más livianos, pequeños y cómodos posible, por ende, el 7,1 % de las personas sordas consideraron que el tamaño y el peso no es apropiado. En promedio, las personas oyentes consideraron que estas características son apropiadas para la función del dispositivo.
- En la pregunta N°3, la mayoría de las personas encuestadas concordaron en que las palabras se ven claramente en la pantalla. Esto se debe al alto contraste de la pantalla y a las letras mayúsculas de las palabras.
- La pregunta N° 4 depende de la experiencia de las personas en utilizar un teclado. Si la persona no ha usado un teclado con frecuencia, le resulta más difícil oprimir las teclas

e identificar el lugar de estas en el teclado. Por esta razón, algunas personas sordas consideraron que es relativamente fácil oprimir las teclas.

- En la pregunta N° 5, la mitad de las personas sordas consideraron que el dispositivo es útil para la comunicación básica de palabras hacia personas que no conocen la lengua de señas y que no tienen discapacidad auditiva.

Conclusiones

- Con la extracción de fonos se concluyó que estas unidades carecen de contexto e inteligibilidad si no están acompañados de una vocal, causando cambios espectrales bruscos en la transición de unidades sonoras en la concatenación. Por tal razón se creó un corpus a partir de la extracción de difonos de palabras grabadas.

- Al extraer difonos de palabras para utilizarlos como unidades de concatenación, se concluyó que entre más pequeña es la unidad mayor es la variación en el tono espectral debido a que en este caso, los difonos tienen el contexto de las palabras de origen y al ser concatenados, el mensaje sonoro es discontinuo. Por ende, se creó el corpus con unidades de mayor longitud mediante la grabación de sílabas.

- El corpus de sílabas proporciona mejor inteligibilidad en comparación a los previamente realizados debido a que son unidades sonoras más largas en tiempo y al no ser extraídas de palabras, no tienen contexto ni acento, brindando continuidad entre fonos.

- La implementación de síntesis concatenativa para conversión texto a voz de 160 palabras se desarrolló en Pure Data, permitiendo ingresar frases al sistema, borrar letras, identificar fonemas y acento y generar el mensaje sonoro a partir de la reproducción secuencial de los audios del corpus, siendo así el software libre un herramienta para el desarrollo de dispositivos que realicen procesamiento digital de señales.

- La generación de solo 160 palabras se debe a la limitación en la cantidad de audios a abrir en Pure Data en la arquitectura ARMHF y por el procesador del sistema embebido, mas no por la cantidad de sílabas en el corpus del sistema. Si no existiera dicha limitación, en el corpus habrían 250 sílabas y se habrían podido generar mayor cantidad de palabras. No obstante, utilizando el sistema embebido Raspberry Pi se

cumple y se supera la expectativa de cantidad de palabras a generar por conversión texto a voz propuesta en los objetivos.

- El análisis de los resultados obtenidos por las encuestas, demuestra que el dispositivo tiene que mejorar aspectos de ergonomía como sus dimensiones y peso físico para sea cómodo en su uso; su funcionamiento cumple con los requisitos de un sistema de conversión texto a voz y es útil para personas con discapacidad para hablar, sin embargo es necesario desarrollar la tecnología pertinente para lograr una comunicación bilateral.

- Los resultados de las pruebas realizadas a personas oyentes demostraron que las palabras generadas tienen un promedio de 95 % de inteligibilidad y 99 % de percepción de acento en las palabras.

Recomendaciones

- Grabar voz con un micrófono cardioide, en campo directo y preferiblemente en un lugar con acondicionamiento acústico para evitar que sonidos no deseados queden registrados en la grabación.

- Utilizar unidades de voz grandes para sintetizar, debido a que el resultado sonoro de la concatenación será más continuo e inteligible, es decir, si se graba una frase para ser segmentada en unidades y así conformar un corpus, se debería dividir en palabras o sílabas, en lugar de fonos.

- Implementar el sistema TTS en el embebido desde el inicio de su desarrollo para probar su funcionamiento y solucionar cualquier inconveniente a tiempo.

- Buscar sistemas embebidos que puedan ser óptimos en cuanto a su capacidad de procesamiento y memoria RAM para sistemas de conversión texto a voz, preferiblemente de 4 núcleos de 1 GHz cada uno y 1 GB de memoria RAM.

- Se recomienda implementar un indicador de batería en el sistema.

- Desarrollar la tecnología pertinente para lograr que las personas con discapacidad para hablar, específicamente, las personas sordas y las personas sin discapacidad auditiva puedan tener una comunicación bilateral.

- Constituir el corpus del sistema con unidades que conformen la mayor cantidad de palabras en español, al concatenarse.

Referencias

Adafruit. (8 de Agosto de 2015). *Adafruit 16x2 Character LCD + Keypad for Raspberry Pi*. Obtenido de https://learn.adafruit.com/adafruit-16x2-character-lcd-plus-keypad-for-raspberry-pi/usage

Asociación Ayuda Afasia. (Abril de 2014). *Definición*. Obtenido de http://www.afasia.org/index.php/definicion

Austria, G. d. (2 de Mayo de 2014). Palabras y expresiones importantes. Obtenido de http://www.austria.info/es/consejos-practicos/palabras-y-expresiones-importantes-1412088.html

Áviles , K. (2012, Diciembre 31). Crean alumnos del IPN dispositivo electrónico para personas mudas. *La Jornada*, 35. Retrieved from http://www.jornada.unam.mx/2012/12/31/sociedad/035n1soc

Benesty, J. (2008). *Handbook of speech processing*. (pp. 438-447). Alemania: Spinger.

Bormane, D., & Shirbahadurkar, S. (2010). TTS system for devanagari script using concatenative synthesis.*International Journal of Computational Intelligence Research*, *6*(1), 1-12.

Bonafonte, A., Esquerra, A., Febrer, A., Fonollosa, J., & Vallverdú, F. (n.d.). *The upc text-to-speech system for spanish and catalan*. Informally published manuscript, Universidad Politécnica de Cataluña, Barcelona, España.

Caballero, A. (2012). *Desarrollo de un controlador MIDI no convencional, implementado en un sistema embebido, utilizando el kinect*. Manuscript submitted for publication, Facultad de Ingeniería, Universidad de San Buenaventura, Bogotá, Colombia

Cantero, F. J. (2002). *Teoría y Análisis de la Entonación.* Barcelona: Edicions de la Universitat de Barcelona.

Castañer, M., Grasso, A., López, C., Mateu, M., Motos, T., & Sanchez, R. (2006). *La inteligencia corporal en la escuela: Análisis y propuestas*. (p. 144). España: Biblioteca de Tándem.

Chabchoub, A., & Cherif, A. (2011). High quality arabic concatenative speech synthesis. *Signal & Image Processing: An International Journal (SIPIJ),2*(4).

Correa, C., Rueda, H., & Arguello, H. (2010). Síntesis de Voz por Concatenación de Difonemas para el Español. SISTEMAS, CIBERNÉTICA E INFORMÁTICA, (págs. 19-24).

D, L., & Agudelo, O. (n.d.). Implementación de un sistema de conversión de texto a voz, mediante sÍntesis por regla y composición alofónica. Informally published manuscript, Ingeniería Electrónica y Mecatrónica, Universidad Autónoma de Occidente, Cali, Colombia.

DANE. (2005). *Discapacidad personas con limitaciones permanentes*. Retrieved from http://www.dane.gov.co/censo/files/discapacidad/preva_indices.pdf

Davaatsagaan, M., & Paliwal, K. (2008). Diphone-based concatenative speech synthesis system for mongolian. *Proceedings of the International MultiConference of Engineers and Computer Scientists, 1*, 19-21.

Degen, J. B. (n.d.). Mechanismus der menschlichen sprache nebst der beschreibung seiner sprechenden

Fonseca, S. (2005). *Comunicación oral: Fundamentos y práctica estratégica*. (2nd ed., p. 51). México: Pearson Prentice Hall.

Free Software Foundation. (n.d.). Retrieved from http://www.fsf.org/

Globocan. (2012). *estimated age-standardized incidence and mortality rates: both sexes*. Retrieved from http://globocan.iarc.fr/Pages/fact_sheets_population.aspx

GNU. (n.d.). *¿qué es software libre?*. Retrieved from http://www.gnu.org/philosophy/free-sw.es.html

Hazan, V., Simpson, A., & Huckvale, M. (1998). Enhancement techniques to improve the intelligibility of consonants in noise : Speaker and listener effects. In International Conference of Speech and Language Processing. Londres: Department of Phonetics and Linguistics, UCL.

How can I get a String from HID device in Python with evdev? (8 de Agosto de 2015). Obtenido de http://stackoverflow.com/questions/19732978/how-can-i-get-a-string-from-hid-device-in-python-with-evdev

Indumathi, A., & Chandra, E. (2012). Survey on speech synthesis. *Signal Processing: An International Journal (SPIJ)*, *6*(5).

Instituto Nacional Cancerológico. (2006). *Cáncer de laringe incidencia estimada según departamentos. Hombres*. Retrieved from http://www.cancer.gov.co/documentos/Incidencia/Tabla 78.pdf

Instituto Nacional Cancerológico. (2011). *Distribución de casos nuevos de cáncer por tratamiento recibido en el inc, según localización primaria, inc, bogotá, colombia, 2011*. Retrieved from http://www.cancer.gov.co/documentos/Tablas2011/Tabla 16.pdf

Jacob, A., & Mythili, P. (2008). Developing a child friendly text-to-speech system. *Advances in Human-Computer Interaction*

Javidan, R., & Rasekh, I. (2010). Concatenative synthesis of persian language based on word, diphone and triphone databases. *Modern Applied Science*,*4*(10).

Klatt, D. (1987). Review of text-to-speech conversion for english. *Journal of the Acoustical Society of* America , 82, 739-793.

Macon, M., Jensen-Link, L., Oliveiro, J., Clements, M., & George, E. (1997, September). *Concatenation-based midi-to-singing voice synthesis*. Presented at 103rd aes convention , New York

Marín, M. (2013). Perder la voz tras un cÁncer de laringe. *REVISTA DE ESTUDIOS FILOLÓGICOS, 24*, Retrieved from https://www.um.es/tonosdigital/znum24/secciones/monotonos-Perder_la_voz_tras_un_cancer_de_laringe.htm

Mattingly, I. (1974). *Speech synthesis for phonetic and phonological models*. Informally published manuscript, Retrieved from http://www.haskins.yale.edu/Reprints/HL0173.pdf

Open Source Initiative. (n.d.). *The open source definition*. Retrieved from http://opensource.org/osd

Open Source Initiative. (n.d.). *The open source licenses*. Retrieved from http://opensource.org/licenses

Pardo, A. M. (2007). *Modelo acústico fonador para vocales masculinas en español*. Manuscript submitted for publication, Facultad de Ingeniería, Universidad de San Buenaventura, Bogotá, Colombia.

Pyhton evdev. (8 de agosto de 2015). Evdev Synopsis Obtenido de https://python-evdev.readthedocs.org/en/latest/

Rueda, H., Correa, C., & Arguello, H. (2012). Diseño y desarrollo de un software de síntesis de voz para el español de Colombia aplicado a la comunicación a través de dispositivos móviles. *Dyna*, *79*(173), 71-80

Patra, K., Patra, B., & Mohapatra, P. (2012). Text to speech conversion with phonematic concatenation. *International Journal of Electronics Communication and Computer Technology (IJECCT)*, *2*(5),

Pure Data. (s.f.). *Pure Data Info*. Obtenido de https://puredata.info/

Pure Data. (4 de Agosto de 2015). Pd-Extended 0.43.3 on Raspberry Pi (Raspbian Wheezy) - Armhf. Obtenido de https://puredata.info/downloads/pd-extended-0-43-3-on-raspberry-pi-raspbian-wheezy-armhf

Quillis, A. (1985). El comentario fonológico y fonético de textos. Teoría y práctica. Madrid: Arco/Libros.

Qullis, A. (1993). Tratado de fonología y fonética españolas. Madrid: Gredos.

Raspberry Pi Foundation. (4 de Agosto de 2015). INSTALLING OPERATING SYSTEM IMAGES. Obtenido de https://www.raspberrypi.org/documentation/installation/installing-images/README.md

Raspberry Pi Foundation. (4 de Agosto de 2015). Raspbian. Obtenido de https://www.raspbian.org/

Rii. (Agosto de 2015). *Rii mini X1*. Obtenido de http://www.riitek.com/goods/detail/58.htm

Saripella, R., C. Loizoua, P., Thibodeau, L., & A. Alford, J. (2011). The effects of selective consonant amplification on sentence recognition in noise by hearing-impaired listeners. Journal Acoustic Society Of America, 130(5), 3028-3037.

Sasirekha, D., & Chandra, E. (2012). Text to speech: A simple tutorial. *International Journal of Soft Computing and Engineering (IJSCE), 2*(1).

Sinha, P. (2010). *Speech processing in embedded systems*. (pp. 157-164). Londres

Suazo, G. (2002). Nueva Ortografía Práctica. Madrid: Edaf S.A.

Taylor, P. (2009). *Text to speech synthesis*. (pp. 422-445). Nueva York: Cambridge University Press.

Tsiros, A. A Multidimensional Sketching interface for Corpus Based Concatenative Synthesis. (2013). In: 19th International Conference on Auditory Display (ICAD2013). [online] Polonia: Georgia Institute of Technology International Community for Auditory Display. Available at: https://smartech.gatech.edu/bitstream/handle/1853/51680/36_S9-02_Tsiros.pdf?sequence=1 [Accessed 4 Jul. 2015].

Universidad Tecnológica de Helsinki. (2006, 11 4). History and development of speech synthesis. Retrieved from http://www.acoustics.hut.fi/publications/files/theses/lemmetty_mst/chap2.html

Wu, C., & Chen, J. (2000). *Automatic generation of synthesis units and prosodic information for chinese concatenative synthesis*. Informally published manuscript, Department of Computer Science and Information Engineering, National Cheng Kung University, Tainan, Taiwan.

Apéndice

Apéndice A: Lista de Fonemas y Alófonos del Español.

Las siguientes tablas contienen la lista de fonemas y alófonos según la AFI y la RFE[66]. La primera columna contiene los fonemas transcritos según la AFI, la segunda tiene los fonemas transcritos según la RFE, la tercera columna presenta los alófonos según la AFI, la cuarta columna muestra los alófonos transcritos por la RFE, la quinta columna recoge las grafías[67] y la sexta columna presenta ejemplos.

Vocales.

/i/	[i]	*i*	/pípa/ [pípa] *pipa*	
			/pisáR/ [pisár] *pisar*	
	[ĩ]		/mímo/ [mĩ́mo] *mimo*	
	[j]		/biéne/ [bjéne] *viene*	
	[i̯]		/áire/ [ái̯re] *aire*	
/e/	[e]	*e*	/pépa/ [pépa] *Pepa*	
			/pesáR/ [pesár] *pesar*	
	[ẽ]		/méma/ [mẽ́ma] *mema*	
/a/	[a]	*a*	/pápa/ [pápa] *papa*	
			/pasáR/ [pasár] *pasar*	
	[ã]		/máma/ [mã́ma] *mama*	
/o/	[o]	*o*	/pópa/ [pópa] *popa*	
			/posáR/ [posár] *posar*	
	[õ]		/móno/ [mṍno] *mono*	
/u/	[u]	*u*	/púpa/ [púpa] *pupa*	
			/puxáR/ [puxár] *pujar*	
	[ũ]		/múNdo/ [mṹn̩do] *mundo*	
	[w]		/buéno/ [bwéno] *bueno*	
	[u̯]		/áuto/ [áu̯to] *auto*	

Figura 57. Fonemas y alófonos de vocales.

Fuente: Qullis, A. (1993). Tratado de fonología y fonética españolas. Madrid: Gredos.

[66] Revista de Filología Española

67 Letras o signos gráficos que representan sonidos

Consonantes Oclusivas Sordas.

/p/	/p/	[p]	[p]	*p*	/kápa/ [kápa] *capa*
/t/	/t/	[t]	[t]	*t*	/páta/ [páta] *pata*
/k/	/k/	[k]	[k]	*c + a, o, u*	/kóka/ [kóka] *coca*
				qu + e, i	/késo/ [késo] *queso*
				k	/kílo/ [kílo] *kilo*

Figura 58. Fonemas y alófonos de consonantes oclusivas sordas.

Fuente: Qullis, A. (1993). Tratado de fonología y fonética españolas. Madrid: Gredos.

Consonantes Oclusivas Sonoras.

/b/	/b/	[b]	[b]	*v, b*	/bóNba/ [bómba] *bomba*
		[β]	[b̥]		AFI: /la bóba/ [la βóβa] *la boba*
					RFE: /la bóba/ [la b̥óba] *la boba*
/d/	/d/	[d]	[d]	*d*	/dóNde/ [dóⁿde] *dónde*
					/tóldo/ [tól̪do] *toldo*
		[ð]	[d̥]		AFI: /ése dédo/ [ése ðéðo] *ese dedo*
					RFE: /ése dédo/ [ése d̥éd̥o] *ese dedo*
/g/	/g/	[g]	[g]	*g + a, o, u*	/gáNga/ [gáŋga] *ganga*
				gu + e, i	/gíso/ [gíso] *guiso*
					/gér̃a/ [gér̃a] *guerra*
		[ɣ]	[g̞]		AFI: /béga/ [béɣa] *vega*
					RFE: /béga/ [bég̞a] *vega*

*Figura 59.*Fonemas y alófonos de consonantes oclusivas sonoras.

Fuente: Qullis, A. (1993). Tratado de fonología y fonética españolas. Madrid: Gredos.

Consonantes Fricativas.

/f/	/f/	[f]	[f]	*f*	/fófo/ [fófo] *fofo*
/θ/	/θ/	[θ]	[θ]	*c + e, i* *z + a, o, u*	[θeθína/ [θeθína] *cecina* /aθuθáR/ [aθuθár] *azuzar*
/s/	/s/	[s]	[s]	*s*	/sóso/ [sóso] *soso*
/ǰ/	/y/	[ǰ]	[y]	*y;* *hi-* + vocal	AFI: /máǰo/ [májo] *mayo* RFE: /máyo/ [máyo] *mayo*
		[d͡ʒ]	[ŷ]		AFI: /ǰó/ [d͡ʒó] *yo* /kóNǰuxe/ [kón‿d͡ʒuxe] *cónyuge* /el ǰélo/ [el‿d͡ʒélo] *el hielo* RFE: /yó/ [ŷó] *yo* /kóNyuxe/ [kón‿ŷuxe] *cónyuge* /el yélo/ [el‿ŷélo] *el hielo*
/x/	/x/	[x]	[x]	*g + e, i* *j + a, e, i, o, u*	/xitáno/ [xitáno] *gitano* /xosé/ [xosé] *José*
/t͡ʃ/	/ĉ/	[t͡ʃ]	[ĉ]	*ch*	AFI: /mut͡ʃát͡ʃo [mut͡ʃát͡ʃo] *muchacho* RFE: /muĉáĉo/ [muĉáĉo] *muchacho*

Figura 60. Fonemas y alófonos de consonantes fricativas.

Fuente: Qullis, A. (1993). Tratado de fonología y fonética españolas. Madrid: Gredos.

Consonantes Nasales.

/ɲ/	/n̰/	[ɲ]	[n̰]	*ñ*	AFI: /káɲa/ [káɲa] *caña* RFE: /káɲ̰a/ [káɲ̰a] *caña*
/-N/	/-N/	[-m]	[-m]	*n, m*	/úN bóNbo/ [ű̃m bómbo] *un bombo*
		[-ŋ]	[-m̰]	*n*	AFI: /úN faról/ [ű̃ŋ faról] *un farol* RFE: /úN faról/ [ű̃m̰ farol] *un farol*
		[-n̰]	[-n̰]	*n*	/úN diéNte/ [ű̃n̰ djén̰te] *un diente*
		[-n̰]	[-n̰]	*n*	/úN θíne/ [ű̃n̰ θíne] *un cine*
		[-n]	[-n]	*n*	/úN sól/ [ű̃n sól] *un sol*
		[n̻]	[n̻]	*n*	AFI: /úN t͡ʃíko/ [ű̃n̻ t͡ʃíko] *un chico* RFE: /úN ĉíko/ [ű̃n̻ ĉíko] *un chico*
		[-ŋ]	[-ŋ]	*n*	/úN kóNgo/ [ű̃ŋ kóŋgo] *un Congo*

Figura 61. Fonemas y alófonos de consonantes nasales.

Fuente: Qullis, A. (1993). Tratado de fonología y fonética españolas. Madrid: Gredos.

Consonantes Róticas.

/ɾ/	/ɾ/	[ɾ]	[ɾ]	*r*	/péro/ [péro] *pero*
/r̄/	/r̄/	[r̄]	[r̄]	*r-; -rr-; n, l + r*	/pér̄o/ [pér̄o] *perro*

Figura 60. Fonemas y alófonos de consonantes oclusivas róticas.

Fuente: Qullis, A. (1993). Tratado de fonología y fonética españolas. Madrid: Gredos.

Consonantes Laterales.

/l/	/l/	[l]	[l]	*l*	/lílas/ [lílas] *lilas;* /el sól/ [el sól] *el sol*
		[l̪]	[l̪]		/el tóldo/ [el̪ tol̪do] *el toldo*
		[l̪]	[l̪]		/el θíne/ [el̪ θíne] *el cine*
		[l̪]	[l̪]		AFI: /el t͡ʃíko/ [el̪ t͡ʃíko] *el chico*
					RFE: /el ĉíko/ [el̪ ĉíko] *el chico*
/ʎ/	/l̬/	[ʎ]	[l̪]	*ll*	AFI: /káʎe/ [kál̪e] *calle*
					RFE: /kál̬e/ [kál̪e] *calle*

Figura 63. Fonemas y alófonos de consonantes laterales.

Fuente: Qullis, A. (1993). Tratado de fonología y fonética españolas. Madrid: Gredos.

Apéndice B: Instalación de Sistema Operativo y Pure Data en el Sistema Embebido, Programación en Python y Arranque y Apagado del Sistema.

Instalación de Sistema Operativo.

La imagen del sistema operativo de la Raspberry Pi se monta en una tarjeta SD en un computador, para esto se deben seguir los pasos de la Figura 39.

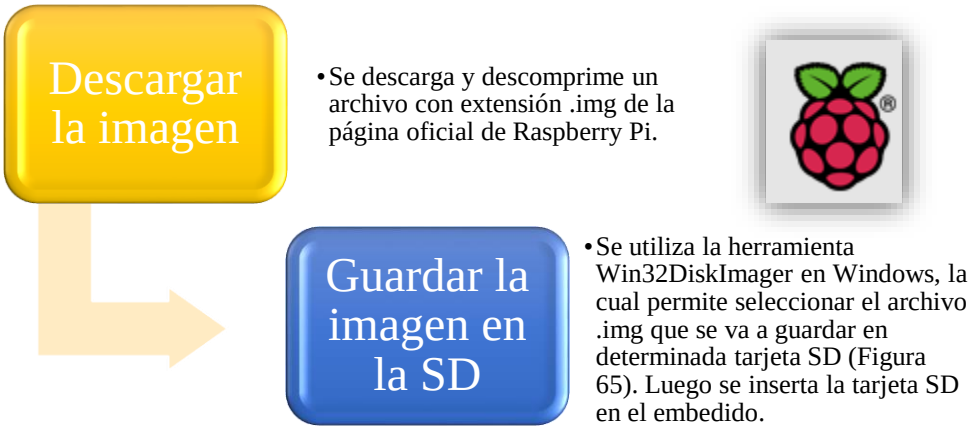

Figura 64. Pasos para instalar sistema operativo en Raspberry Pi.

Fuente: Propia.

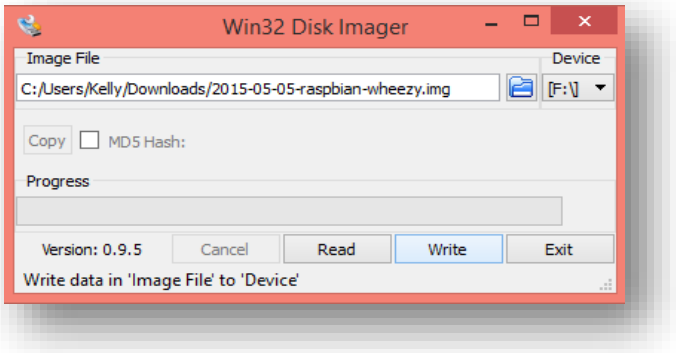

Figura 65. Instalación de Sistema Operativo en SD mediante Win32DiskImager.

Fuente: Propia

El sistema operativo descargado es Raspbian[68] porque está optimizado para el hardware del embebido Raspberry Pi y proporciona más de 35.000 paquetes y software precompilado con facilidad de instalación (Raspberry Pi Foundation, 2015). El entorno de escritorio de Raspbian se muestra en la Figura 66.

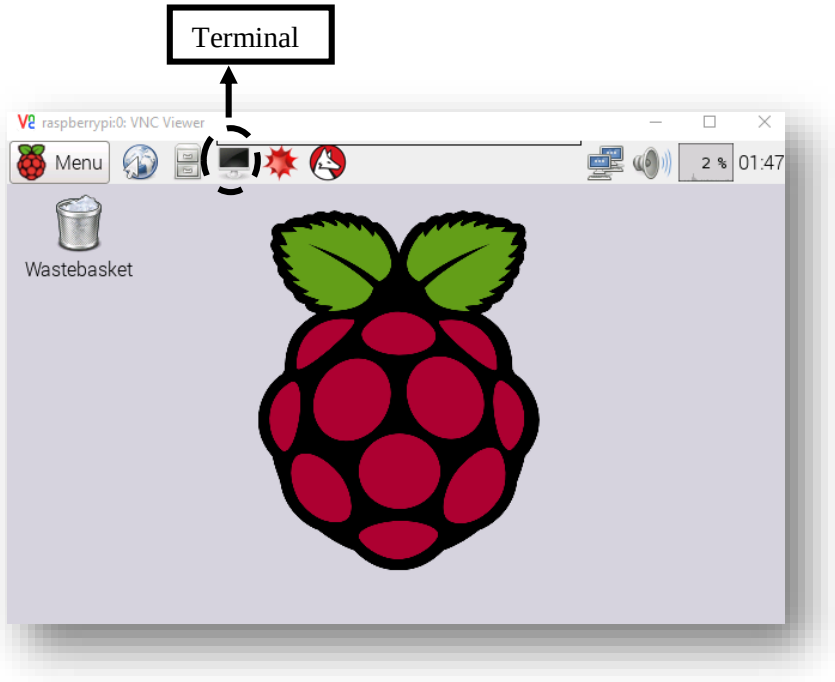

Figura 66. Escritorio de Raspbian.

Fuente: Propia

Instalación de Pure Data.

La instalación de Pure Data se realizó siguiendo el proceso de la Figura 67. Para descargar Pure Data se conectó el cable Ethernet que provee el acceso a internet.

[68] Sistema operativo basado en Debian.

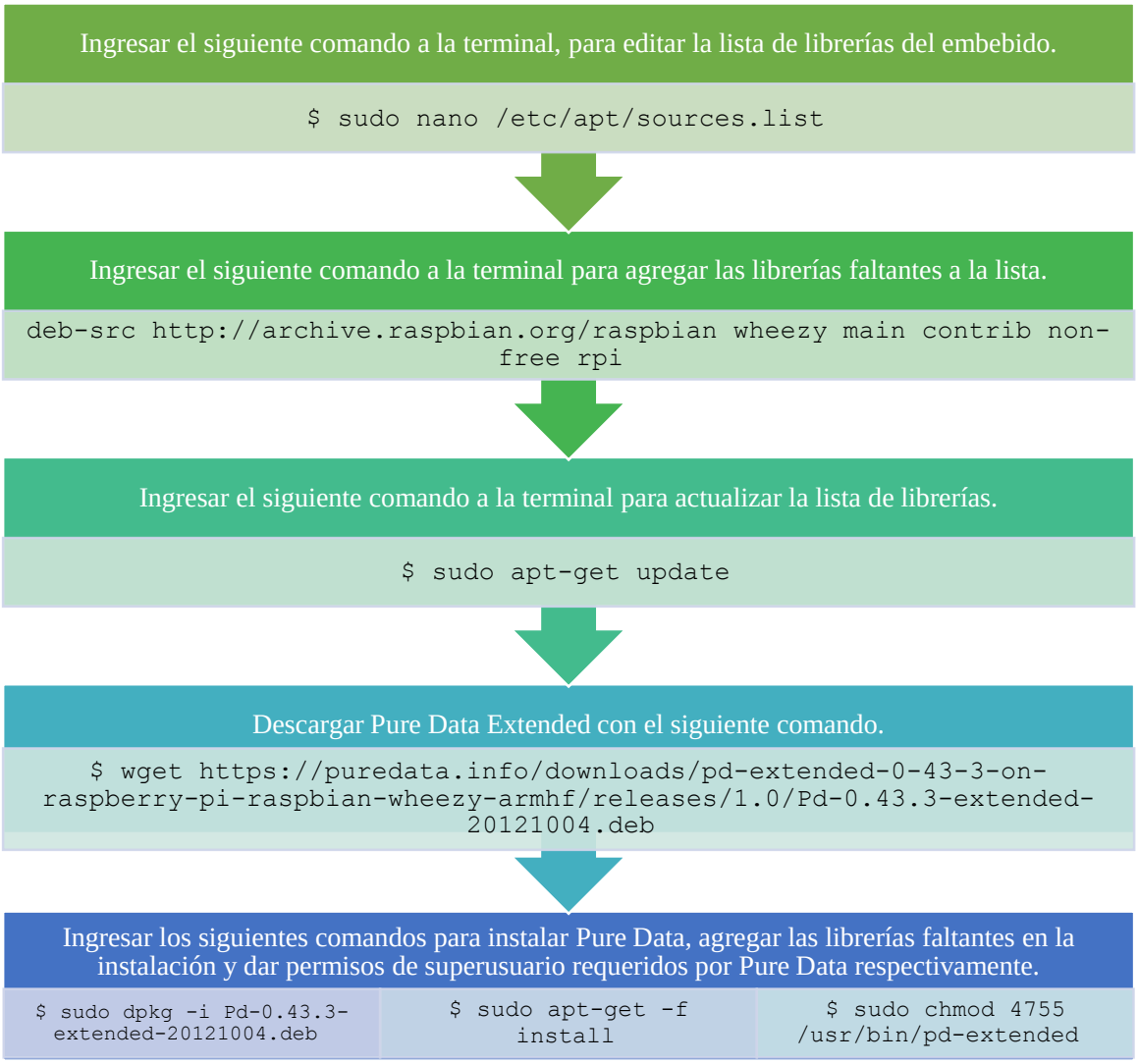

Figura 67. Proceso de instalación de Pure Data en el sistema embebido.

Fuente: Propia

Programación en Pyhton para la Pantalla LCD.

Para inicializar la LCD se utiliza:

```
lcd = LCD.Adafruit_CharLCDPlate()
```

Al oprimir los botones, se ingresan “eventos” al sistema utilizando la librería *evdev*, la cual permite leer “eventos” de entrada en Linux. Se conoce como “evento” un movimiento de

mouse, un toque en una pantalla táctil o la acción de oprimir una tecla. El siguiente comando importa algunas funciones de la librería *evdev*.

```
from evdev import InputDevice, categorize, ecodes
```

Inputdevice lee los eventos de dispositivos externos, *categorize* identifica el tipo de evento (en el caso del teclado, scancodes[69]) y *ecodes* realiza, para este caso, un mapeo numérico de los eventos provenientes del teclado. En los siguientes comandos se puede ver la sintaxis de estas funciones.

```
dev = InputDevice('/dev/input/event0')
    for event in dev.read_loop():
if event.type == ecodes.EV_KEY:
    data = categorize(event)
```

Estos comandos permiten identificar el teclado como el dispositivo externo que genera eventos, reconocer cuando se oprime y se suelta una determinada tecla y asignar un número entero a cada una. En la Figura 68 se muestran los códigos ASCII de cada tecla.

```
scancodes = {
    # Scancode: ASCIICode
    0:'', 1: u'ESC', 2: u'1', 3: u'2', 4: u'3', 5: u'4', 6: u'5', 7: u'6', 8: u'7', 9: u'8',
    10: u'9', 11: u'0', 12: u'-', 13: u'=', 14:'' , 15: u'TAB', 16: u'Q', 17: u'W', 18: u'E', 19: u'R',
    20: u'T', 21: u'Y', 22: u'U', 23: u'I', 24: u'O', 25: u'P', 26: u'[', 27: u']', 28: u'CRLF', 29: u'LCTRL',
    30: u'A', 31: u'S', 32: u'D', 33: u'F', 34: u'G', 35: u'H', 36: u'J', 37: u'K', 38: u'L', 39: u';',
    40: u'"', 41: u'`', 42: u'LSHFT', 43: u'\\', 44: u'Z', 45: u'X', 46: u'C', 47: u'V', 48: u'B', 49: u'N',
    50: u'M', 51: u',', 52: u'.', 53: u'/', 54: u'RSHFT', 56: u'LALT', 57: ' ', 100: u'RALT'
}
```

Figura 68. Códigos ASCII del teclado en Python.

Fuente: Propia.

Con el siguiente condicional se obtiene el código ASCII de la tecla oprimida y se muestra en la LCD.

[69] Códigos que envía el teclado al ordenador para indicar si una tecla esta oprimida o no.

```
if data.keystate == 1:
            key_lookup = scancodes.get(data.scancode)
            lcd.set_cursor(0,1)
            lcd.message('%s'% key_lookup)
```

Con *get(data.scancode)* se obtiene el código ASCII de la matriz de la Figura 68, *lcd.set_cursor* posiciona el cursor y *lcd.message* muestra la información en la pantalla.

Arranque y Apagado del Sistema.

Al encenderse, el sistema embebido debe iniciar el patch *TTS.pd* y el código de Pyhton, con el fin de dar comienzo a la ejecución del sistema TTS. Para esto, se ingresan los siguientes comandos en la terminal:

```
sudo nano ~/.config/lxsession/LXDE/autostart
        @sudo /home/pi/./python.sh
          @sudo /home/pi/./pd.sh
```

Los archivos *pd.sh* y *python.sh* contienen comandos guardados en un bloc de notas con la extensión *.sh*, la cual indica la ejecución de comandos de la terminal. El comando de *python.sh* es el siguiente:

```
#!/bin/bashpython /home/pi/tp.py
```

tp.py es el archivo de Python que contiene la programación de la pantalla LCD. El comando *#!/bin/bash* se utiliza para leer los comandos de la terminal y */home/pi/tp.py* es la ruta en donde se encuentra el archivo de Python. El comando de *pd.sh* es:

```
#!/bin/bashsleep 1pd-extended -noadc -blocksize 512
                /home/pi/pd/PATCH/TTS.pd
```

El comando *sleep 1* ejecuta un retraso de 1 segundo con el fin de darle este tiempo a Pure Data, para que se inicie después del arranque de Pyhton. *Noadc* cierra la entrada análoga de Pure Data. *blocksize 512* es la configuración de Pure Data que permite tomar bloques de 512 muestras para procesar la información. La opción *Retardo (ms)* de las preferencias de audio de Pure Data se configuró en 500 ms.

Al iniciar el sistema, se muestra en la pantalla "*HOLA SOY TESPEECON"* y "*PRESIONA ENTER PARA EMPEZAR"* utilizando el siguiente código de Pyhton:

```
lcd.set_cursor(5,0)
lcd.message('HOLA!')
lcd.set_cursor(1,1)
lcd.message('SOY TESPEECON')
time.sleep(5)
lcd.clear()

lcd.set_cursor(1,0)
lcd.message('PRESIONA ENTER')
lcd.set_cursor(2,1)
lcd.message('PARA INICIAR')
```

El sistema operativo del embebido se apaga oprimiendo la tecla *esc*, se muestra en la pantalla *"¡ADIOS!!"*, mediante el siguiente código:

```
lcd.set_cursor(5,0)
lcd.message('¡ADIOS!')
lcd.set_cursor(4,0)
lcd.message('APAGANDO')
time.sleep(23)
lcd.clear()
```

Apéndice C: Encuestas.

UNIVERSIDAD DE SAN BUENAVENTURA SECCIONAL BOGOTÀ
ENCUESTA PARA EVALUAR DISPOSITIVO DE CONVERSIÓN TEXTO A VOZ

La siguiente encuesta tiene el objetivo de evaluar el dispositivo de conversión texto a voz presentado, el cual está dirigido a personas con discapacidad para hablar con el fin de facilitar la comunicación mediante algunas frases y palabras básicas. El funcionamiento del dispositivo consiste en ingresar palabras en el teclado inalámbrico, las cuales podrán ser vistas en la pantalla y ser escuchadas por las demás personas mediante un altavoz. El procedimiento de evaluación se divide en dos partes. La **PRIMERA PARTE** consiste en escuchar **UNA VEZ,** 18 frases ingresadas al teclado por la persona encargada y escribirlas en la Tabla 1. La **SEGUNDA PARTE** consiste en ingresar palabras o frases de la Tabla 2[1] y responder las preguntas de la Tabla 3 y 4, las cuales están relacionadas con los siguientes parámetros: funcionamiento, comodidad y utilidad. El funcionamiento del dispositivo se evaluará mediante las preguntas 1, 2, 3, 4 y 7; la comodidad se relaciona con el peso físico, tamaño y facilidad de uso del teclado y será evaluada mediante las preguntas 5, 6 y 8; la utilidad del dispositivo para personas con discapacidad para hablar se evaluará mediante la pregunta 9. Las indicaciones adicionales serán mencionadas antes de iniciar la prueba.

PRIMERA PARTE

Tabla 1. Primera Prueba

Frase 1	
Frase 2	
Frase 3	
Frase 4	
Frase 5	
Frase 6	
Frase 7	
Frase 8	
Frase 9	
Frase 10	
Frase 11	
Frase 12	
Frase 13	
Frase 14	
Frase 15	
Frase 16	
Frase 17	
Frase 18	

[1] Las palabras de la Tabla 2, están basadas en la lista de frases y expresiones importantes en conversaciones cotidianas realizada por la oficina de turismo de Austria.

Figura 61. Encuesta realizada a personas oyentes, primera parte.

Fuente: Propia.

SEGUNDA PARTE

Tabla 2. Palabras a Generar en el Sistema.

Sí	Perdón	Muy Bien	Mañana
No	Buenos días	Quien	Bueno
Por favor	Buenas tardes	Cuál	Ayuda
Gracias	Adiós	¿Dónde está el baño?	Mal
Eso	Hola, ¿Cómo estás?	¿Cuándo?	Soy Feliz
¿Cómo Dice?	¿Qué tal?	¿Cuánto tiempo?	Triste
No entiendo	Dolor	¿Habla español?	Hambre
Con permiso	¿Cuánto es?	Hoy	Sed
Vale	Tengo	Buenas noches	Estoy

Tabla 3. Preguntas.

	Pregunta	Si	No
1	¿Todas las palabras son entendibles?		
2	Si respondió *No* a la pregunta anterior, ¿qué palabra(s) no entendió?		
3	¿Escuchó el acento en todas las palabras?		
4	Si respondió *No* a la pregunta anterior, ¿qué palabra(s) no escuchó con acento?		

Califique de 1 a 5 las siguientes preguntas, siendo 1 muy malo y 5 excelente:

Tabla 4. Preguntas.

		1	2	3	4	5
5	¿Considera que el peso físico del dispositivo es apropiado para su función?					
6	¿Considera que el tamaño del dispositivo es apropiado para su función?					
7	¿Se ven claramente las palabras en la pantalla?					
8	¿Es fácil oprimir las teclas del teclado?					
9	¿Considera que el dispositivo es útil para personas con discapacidad para hablar?					

Califique su experiencia utilizando el dispositivo.

Muy Buena___ Buena___ Aceptable___ Mala___ Muy Mala___

¡Gracias por su Colaboración!

Figura 62. Encuesta realizada a personas oyentes, segunda parte.

Fuente: Propia.

UNIVERSIDAD DE SAN BUENAVENTURA SECCIONAL BOGOTÁ
ENCUESTA PARA EVALUAR DISPOSITIVO DE CONVERSIÓN TEXTO A VOZ

La siguiente encuesta tiene el objetivo de evaluar el dispositivo de conversión texto a voz presentado, el cual está dirigido a personas con discapacidad para hablar con el fin de facilitar la comunicación mediante algunas frases y palabras básicas. El funcionamiento del dispositivo consiste en ingresar palabras en el teclado inalámbrico, las cuales podrán ser vistas en la pantalla y ser escuchadas por las demás personas mediante un altavoz. El procedimiento de evaluación consiste en ingresar palabras o frases de la Tabla 1[2] y responder las preguntas de la Tabla 2, las cuales están relacionadas con los siguientes parámetros: funcionamiento, comodidad y utilidad. El funcionamiento del dispositivo se evaluará mediante la pregunta 3; la comodidad se relaciona con el peso físico, tamaño y facilidad de uso del teclado y será evaluada mediante las preguntas 1, 2 y 4; la utilidad del dispositivo para personas con discapacidad para hablar se evaluará mediante la pregunta 5.

Tabla 1. Palabras a Generar en el Sistema.

Buenos días	Soy feliz	¿Cómo dice?	Adiós
Hola	No entiendo	Tengo dolor	Mal
¿Dónde está el baño?	¿Quién es?	Si	Mañana
Tengo hambre	Con permiso	Muy bien	Hoy
¿Cuánto tiempo?	Ayuda	Gracias	¿Qué es eso?
No sé	Perdón	Bueno	¿Habla español?
Tengo sed	Estoy triste	¿Cómo estás?	Por favor
Buenas tardes	¿Cuánto vale eso?	Buenas noches	¿Qué tal?

Califique de 1 a 5 las siguientes preguntas, siendo 1 muy malo y 5 excelente

Tabla 2. Preguntas.

	PREGUNTA	1	2	3	4	5
1	¿Considera que el peso físico del dispositivo es apropiado para su función?					
2	¿Considera que el tamaño del dispositivo es apropiado para su función?					
3	¿Se entienden las palabras en la pantalla?					
4	¿Es fácil oprimir las teclas del teclado?					
5	¿Considera que el dispositivo es útil para personas con discapacidad para hablar?					

Califique su experiencia utilizando el dispositivo.

Muy Buena___ Buena___ Aceptable___ Mala___ Muy Mala___

¡Gracias por su Colaboración!

[2] Las palabras de la Tabla 1, están basadas en la lista de frases y expresiones importantes en conversaciones cotidianas realizada por la oficina de turismo de Austria.

Figura 63. Encuesta realizada a personas con discapacidad para hablar.

Fuente: Propia.

Apéndice D: Análisis Estadístico

Tabla 20. *Análisis estadístico. Prueba con personas sordas. Pregunta 1.*

Variable: ¿Considera que el peso físico del dispositivo es apropiado para su función?				
Escala	**Frecuencia**	**Porcentaje (%)**	**Media**	**Moda**
1	0	0	4,3	5
2	1	7,1		
3	1	7,1		
4	4	28,5		
5	8	57,1		

Fuente: Propia.

Tabla 21. *Análisis estadístico. Prueba con personas sordas. Pregunta 2.*

Variable: ¿Considera que el tamaño del dispositivo es apropiado para su función?				
Escala	**Frecuencia**	**Porcentaje (%)**	**Media**	**Moda** 2-4
1	1	7,1	3,3	
2	4	28,5		
3	2	14,3		
4	3	21,4		
5	4	28,6		

Fuente: Propia.

Tabla 22. *Análisis estadístico. Prueba con personas sordas. Pregunta 3.*

Variable: ¿Se entienden las palabras en la pantalla?				
Escala	**Frecuencia**	**Porcentaje (%)**	**Media**	**Moda**
1	0	0	4,8	5
2	0	0		
3	0	0		
4	2	14,3		
5	12	85,7		

Fuente: Propia.

Tabla 23. *Análisis estadístico. Prueba con personas sordas. Pregunta 4.*

Variable: ¿Es fácil oprimir las teclas?				
Escala	**Frecuencia**	**Porcentaje (%)**	**Media**	**Moda**
1	1	7,1	4,1	5
2	0	0		
3	3	21,4		
4	2	14,3		
5	8	57,1		

Fuente: Propia.

Tabla 24. *Análisis estadístico. Prueba con personas sordas. Pregunta 5.*

Variable: ¿Considera que el dispositivo es útil para personas con discapacidad para hablar?				
Escala	**Frecuencia**	**Porcentaje (%)**	**Media**	**Moda**
1	2	14,2	3,1	3
2	1	7,1		
3	6	42,8		
4	3	21,4		
5	2	14,2		

Fuente: Propia.

Tabla 25. *Análisis estadístico. Prueba con personas sordas.*

Variable: Califique su experiencia utilizando el dispositivo					
Escala	**Valores**	**Frecuencia**	**Porcentaje (%)**	**Media**	**Moda**
Muy Buena	5	5	35,7	4,2	5
Buena	4	7	50		
Aceptable	3	2	14,2		
Malo	2	0	0		
Muy Malo	1	0	0		

Fuente: Propia.

Tabla 26. *Análisis estadístico. Prueba con personas oyentes. Pregunta 1.*

Variable: ¿Considera que el peso físico del dispositivo es apropiado para su función?				
Escala	**Frecuencia**	**Porcentaje (%)**	**Media**	**Moda**
1	0	0	4,9	5
2	0	0		
3	0	0		
4	2	8		
5	23	92		

Fuente: Propia.

Tabla 27. *Análisis estadístico. Prueba con personas oyentes. Pregunta 2.*

Variable: ¿Considera que el tamaño del dispositivo es apropiado para su función?				
Escala	**Frecuencia**	**Porcentaje (%)**	**Media**	**Moda** 5
1	0	0	4,5	
2	0	0		
3	3	12		
4	5	20		
5	17	68		

Fuente: Propia.

Tabla 28. *Análisis estadístico. Prueba con personas oyentes. Pregunta 3.*

Variable: ¿Se entienden las palabras en la pantalla?				
Escala	**Frecuencia**	**Porcentaje (%)**	**Media**	**Moda** 5
1	0	0	5	
2	0	0		
3	0	0		
4	0	0		
5	25	100		

Fuente: Propia.

Tabla 29. *Análisis estadístico. Prueba con personas oyentes. Pregunta 4.*

Variable: ¿Es fácil oprimir las teclas?				
Escala	**Frecuencia**	**Porcentaje (%)**	**Media**	**Moda**
1	0	0	5	5
2	0	0		
3	0	0		
4	0	0		
5	25	100		

Fuente: Propia.

Tabla 30. *Análisis estadístico. Prueba con personas oyentes. Pregunta 5.*

Variable: ¿Considera que el dispositivo es útil para personas con discapacidad para hablar?				
Escala	**Frecuencia**	**Porcentaje (%)**	**Media**	**Moda**
1	0	0	4,8	5
2	0	0		
3	1	4		
4	4	16		
5	20	80		

Fuente: Propia.

Tabla 31. *Análisis estadístico. Prueba con personas oyentes.*

			Variable:	Califique su experiencia utilizando el dispositivo	
Escala	**Valores**	**Frecuencia**	**Porcentaje (%)**	**Media**	**Moda**
Muy Buena	5	22	88	4,8	5
Buena	4	3	12		
Aceptable	3	0	0		
Malo	2	0	0		
Muy Malo	1	0	0		

Fuente: Propia.

Apéndice E: Especificaciones Técnicas.

Tabla 32. *Especificaciones Raspberry Pi Modelo B.*

Características	
BRCM2835 SoC	Yes
Standard SoC Speed	700Mhz
RAM	512MB*
Storage	Full SD
Ethernet 10/100	Yes
HDMI output port	Yes
Composite video output	Yes
Number of USB2.0 ports	2
Expansion header	26
Number of available GPIO	17
3.5mm audio jack	Yes
Number of camera interface ports (CSI-2)	1
Number of LCD display interface ports (DSI)	1
Power (bare, approx, 5v)	700mA, 3.5W
Size	85 x 56 x 17mm

Fuente: https://www.raspberrypi.org/documentation/hardware/raspberrypi/models/specs.md

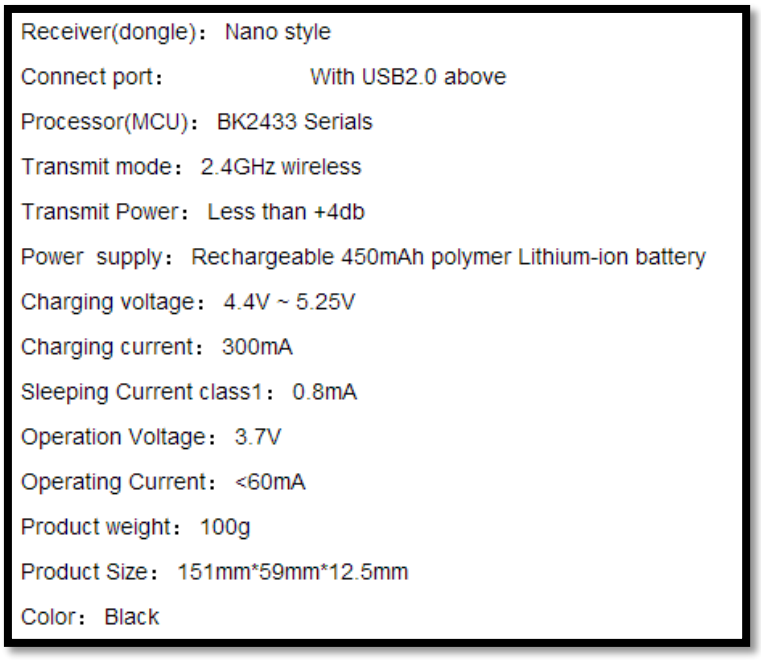

Figura 64. Especificaciones teclado inalámbrico Rii X1

Fuente: http://www.riitek.com/goods/detail/58.htm

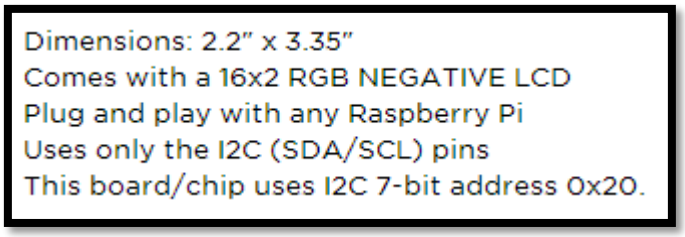

Figura 65. Especificaciones pantalla LCD Adafruit i2C.

Fuente: https://www.adafruit.com/products/1110

Element	Externally-polarized (DC bias) condenser
Polar patterns	Cardioid, Omnidirectional, Figure-of-eight
Frequency response	20-18,000 Hz
Low frequency roll-off	80 Hz, 12 dB/octave
Open circuit sensitivity	–36 dB (15.8 mV), re 1V at 1 Pa
Impedance	100 ohms
Maximum input sound level	149 dB SPL, 1 kHz at 1% T.H.D.; 159 dB SPL, with 10 dB pad (nominal)
Noise[1]	17 dB SPL
Dynamic range (typical)	132 dB, 1 kHz at Max SPL
Signal-to-noise ratio[1]	77 dB, 1 kHz at 1 Pa
Phantom power requirements	48V DC, 4.2 mA typical
Switches	Polar selection; Flat, roll-off; 10 dB pad (nominal)
Weight	510 g (18.0 oz)
Dimensions	188.0 mm (7.40") long, 53.4 mm (2.10") maximum body diameter
Output connector	Integral 3-pin XLRM-type
Audio-Technica case style	R1
Accessories furnished	AT8449 shock mount for 5/8"-27 threaded stands; microphone dust cover; protective carrying case

Figura 66. Especificaciones Micrófono Audiotechnica AT4050.

Fuente: http://www.audio-technica.com/cms/resource_library/literature/dc82a6109f196750/p52150_at4050_spec_sheet.pdf

- Power: DC: 3.6v – 5.5V
- Output power: 3W +3 W (4 in Europe)
- Signal-to-noise ratio: 90dB
- Efficiency:> 90%
- Size: 24 x 15.5 x mm
- positioning holes size: Spacing 20mm
- aperture: 2mm

Figura 67. Especificaciones de amplificador *PAM8403*

Fuente: http://www.elecfreaks.com/store/pam8403-super-mini-digital-amplifier-board-p-531.html

Apéndice F: Anexo Digital

El siguiente archivo se encuentra en el CD anexado al documento:

- Video de prueba del dispositivo de conversión texto a voz. “Tespeecon.mp4”.