UNIVERSIDAD AUTÓNOMA METROPOLITANA

Unidad Iztapalapa

DIVISIÓN DE CIENCIAS BÁSICAS E INGENIERÍA
MAESTRÍA EN CIENCIAS Y TECNOLOGÍAS DE LA INFORMACIÓN

"Sistema de texto-a-habla expresivo en español"

Idónea comunicación de resultados que para obtener el grado de:
Maestra en Ciencias y Tecnologías de la Información

Presenta:
Ing. Natividad Felisa Navarrete Gómez

Asesora:
M. en C. Alma Edith Martínez Licona

Jurado calificador:

Presidente: **Dra. Mariko Nakano Miyatake**
Secretario: **M. en C. Alma Edith Martínez Licona**
Vocal: **M. en IB. Fabiola Margarita Martínez Licona**

México, CDMX, a 29 de Octubre de 2018

Resumen

En el presente trabajo se desarrolla un Sistema de Texto a Habla Expresivo en Español o TTS-EE que convierte una entrada de texto a un sonido de habla natural e inteligible y con emoción, simulando la interacción humana es decir una relación de comunicación entre una o más personas.

Están involucrados en el desarrollo de este sistema programas como Python, MBROLA, eSpeak, el alfabeto fonético SAMPA, Easy Align e espeak. Se da un panorama general que va de lo histórico a lo más actual y también se describe detalladamente en qué consiste un Sistema de Texto a Habla.

Una parte fundamental de este TTS-EE es la Transcripción ortográfica-fonética con la aplicación de reglas de acentuación, reglas de silabificación y reglas de transcripción basadas en SAMPA, se desarrollan módulos en código Python para tal fin. Además se realiza una síntesis de voz a través del sintetizador MBROLA y el desarrollo de código python para esta síntesis.

Otra parte medular del TTS-EE es la creación de una base de datos de habla en español de México similar a una base de habla en español de España que ya se tiene, de las cuales se obtienen datos estadísticos de la duración y el pitch de cada fonema utilizado por medio de una segmentación con el programa PRAAT e Easy Align como herramientas.

Se cuenta con un corpus del español del cual se obtuvo una tasa de sílabas por cada grupo de audios que se resumen en la Tabla 1 además de obtenerse una tasa de fonemas que se resumen en la Tabla 2.

Finalmente se integra cada parte, la transcripción ortográfica-fonética, el sintetizdor MBROLA para obtener una voz con emoción a partir del texto de entrada pre-procesado, a través de la interfaz gráfica del TTS-EE.

Tasa de sílabas español ES y MX audios 1-184					
Audio	Estructura	Español	Alegría	Neutral	Tristeza
1-150	Oraciones largas cortas, interrogativas y párrafos	ES	7	6.5	5.5
		MX	6.2	6	6.4
1-100	Afirmativas largas y cortas	ES	6.7	6.5	5.2
		MX	5.5	6	5.8
101-134	Interrogantes y oraciones con énfasis	ES	7.6	6.5	5.6
		MX	6.2	6.5	6
135-150	Párrafos	ES	6.9	6.8	5.6
		MX	6.19	6.19	6.39
151-184	Dígitos y palabras aisladas	ES	4.5	5	4.1
		MX	3	4.9	3.9

Tabla 1. Resumen de la tasa de sílbas de español ES y MX audios 1-184.

Tasa de fonemas español ES y MX audios 1-184					
Audio	Estructura	Español	Alegría	Neutral	Tristeza
1-150	Oraciones largas cortas, interrogativas y párrafos	ES	15	16	12.5
		MX	6	6.2	6.4
1-100	Afirmativas largas y cortas	ES	15	15.9	12.1
		MX	6	5.5	5.8
101-134	Interrogantes y oraciones con énfasis	ES	15	18	12.5
		MX	6.5	6.2	6
135-150	Párrafos	ES	15.5	16	12.9
		MX	6.19	6.19	6.39
151-184	Dígitos y palabras aisladas	ES	12	11.9	9.9
		MX	4.9	3	3.9

Tabla 2. Resumen de la tasa de sílbas de español ES y MX audios 1-184.

Para el emodata y para el correcto funcionamiento del TTS-EE se obtuvieron resultados interesantes. Las emosiones son expresadas de manera correcta en el TTS respecto a Emofilt, gracias al silabeo, a la tasa de fonemas, a la duración en los fonemas, duración de pausas y pitch, para la emosión tristeza las diferencias para las voces finales tienen un promedio de 0.2s y 2.3s para palabras cortas y frases largas respectivamente; para la emoción neutral, las diferencias son menores a comparación de la emoción tristeza y alegría, para sus voces finales un promedio de 0.0028s y 0.52s para palabras cortas y frases largas y finalmente para la alegría las diferencias para voces finales tienen un promedio de 0.13s y 1.16s para palabras cortas y frases largas, resultados que se mostraran más a detalle a lo largo de este trabajo de investigación.

Agradecimientos

A la Universidad Autónoma Metropolitana Unidad Iztapalapa UAM-I.

Al Consejo Nacional de Ciencia y Tecnología-CONACYT.

Muy especialmente a mis asesores el Dr. John y la M.C. Alma, por su tiempo y paciencia.

A mi madre Felisa, mi padre Felipe, mis hermanos Richard, Juanito y Dany.

También a mís queridas amigas Dul y Zula.

Con cariño a mi mejor amigo **César Iranyr.**

Y sobre todo a Dios por nunca abandonarme.

Índice

Lista de tablas

Lista de figuras

Lista de cuadros

Lista de gráficas

1. Introducción

1.1 Síntesis de Voz

La síntesis de voz es la producción artificial del habla humana, surgen para este propósito los llamados sintetizadores de voz y pueden ser implementados tanto en hardware como en software [1]. El habla sintetizada se genera concatenando segmentos de grabaciones que se encuentran almacenadas en una base de datos, los sistemas que almacenan fonemas y **difonemas** (la transisión entre dos fonos consecutivos) proveen un rango de salida más amplio [2].

1.1.1 Tipos de síntesis

La síntesis de voz consta de dos características importantes: la naturalidad que describe qué tanta es la semejanza del sonido generado al habla humana y la inteligibilidad que se refiere a la facilidad con la cual se entiende el significado del habla generada, por lo que un sintetizador de voz tiene la tarea de tratar de maximizar ambas particularidades.

Las principales tecnologías para generar una voz sintética son la síntesis concatenativa y síntesis formante [3].

La síntesis formante utiliza un modelo acústico, parámetros como la frecuencia fundamental, fonación (voicing) y niveles de ruido que se analizan para crear una onda acústica de habla artificial.

La síntesis concatenativa trata de la unión de segmentos de habla grabados produciendo un sonido más natural en la síntesis de voz. Debido a que las diferencias entre las variaciones naturales de la voz y la naturaleza de las tecnologías para automatizar la fragmentación de las ondas sonoras resultan en una salida imperfecta, surgieron tres tipos de síntesis concatenativa:

La selección de unidades utiliza unidades como fonemas, sílabas, morfemas, palabras, frases y oraciones, para esta síntesis se utiliza una gran base de datos de habla grabada. Esta técnica provee más naturalidad debido a que aplica poco procesamiento de señales (DSP) a las grabaciones aunque algunos sintetizadores si utilizan el DSP en la concatenación de las grabaciones para suavizar la onda acústica. [4]

En la sistensis por difonemas se utiliza una base de datos mínima, la cual contiene todas las transiciones de sonido a sonido del lenguaje que se desee sintetizar, es decir, un ejemplo de cada difonema, la información prosódica de una oración es impuesta sobre estas unidades mínimas por medio de técnicas de procesamiento digital de señales como LPC (Linear Predictive Coding) [5].

Dominio específico: Son sintetizadores que concatenan palabras o frases pregrabadas para generar nuevas expresiones. Se aplican a calculadoras o relojes donde la variedad de los textos de salida del sintetizador se limita a un dominio en particular. El nivel de naturalidad es alto porque el número de oraciones almacenadas es reducido y se asemejan a la entonación y pronunciación de las grabaciones originales [6].

Para el TTS-EE se eligió MBROLA que tiene el tipo de síntesis de concatenación de difonos y que se describe en la siguiente sección.

1.1.2 MBROLA

Para la síntesis de voz se utilizó MBROLA, que es un proyecto iniciado por el laboratorio TCTS de la facultad politécnica de Mons en Bélgica. El objetivo de este sintetizador es manejar el habla para una gran cantidad de voces, idiomas y dialectos, lo cual es útil en el sentido que contiene voces como Mx1, Mx2 para español de México y Es1, Es2, y Es4 para español de España, además impulsa la generación de prosodia, el cual es uno de los más grandes retos para los sintetizadores de texto a habla [5]. Los archivos ejecutables de éste sintetizador son de uso libre compatibles con sistemas operativos como Windows, Linux y Mac OS.

MBROLA tiene sus orígenes a partir del sintetizador MBROLA 2.0, un sintetizador que se basa en la concatenación de difonos. La primera base de datos de difonos que se elaboró fue para una voz masculina en el idioma francés. Aunque ahora cuenta con los siguientes idiomas o voces, véase *Tabla 3.*

Idiomas del sintetizador MBROLA			
Africano	Checo	Islandés	Portugués
Inglés americano	Holandés	Indonesio	Rumano
Árabe	Estonio	Irán	***Español***
Portugués brasileño	Alemán	Italiano	***Español mexicano***
Bretón	Griego	Japonés	Español venezolano
Inglés británico	Hebreo	Lituano	Sueco
Francés canadiense	Hindi	Malay	Tegulu
Croata	Húngaro	Polaco	Turco

Tabla 3. Idiomas del sintetizador MBROLA.

Para MBROLA se crean bases de datos de difonos que son necesarias para correr el sintetizador, estás bases han sido proporcionadas por diferentes laboratorios de investigación y compañías. Lo que se usa como entrada de MBROLA 2.0 es una lista de alófonos con información prosódica y muestras de habla lineal de 16 bits de salida.

MBROLA no es un sistema de texto a habla; es por eso que en el proyecto TTS-EE se incorpora el sintetizador MBROLA como su back end que se encarga de la transcripción ortográfica fonética.

Algoritmo MBROLA. El programa MBROLA 2.00 utiliza una técnica llamada adición de traslape de resíntesis multicapa con la que se produce el habla por concatenación de difonos además se planea tener disponibles trífonos o polífonos.

MBROLA comparte una habilidad de MBR-PSOLA para suavizar las discontinuidades espectrales en el dominio del tiempo para mejorar la fluidez de la señal y permitir la codificación de la base de datos. Así MBROLA acumula flexibilidad y la relación de compresión de datos de modelos de habla paramétricos y simplicidad computacional de los sintetizadores de dominio de tiempo, por lo que para MBROLA:

1. La complejidad computacional es de bajo costo.
2. La capacidad de suavizar las discontinuidades espectrales que surgen en los puntos de unión de difonos produce una voz sintética de alta calidad.
3. La base de datos de difonos esta simplificada.
4. La base de datos de difonos es codificada de manera eficiente con un costo de menor porcentaje de la carga computacional requerida por síntesis.

El archivo de entrada para MBROLA contiene una lista de fonemas con información prosódica, es decir, duración y pitch y el archivo de salida es un archivo de audio, los

formatos soportados son Raw, Wav, Au, Aiff. El pitch o frecuencia fundaméntal es la frecuencia más baja de una forma de onda periódica.

El formato de archivo es muy simple como se muestra en la *Tabla 4*, la primera columna muestra los fonemas en trascripción SAMPA, la siguiente columna es para la duración de cada fonema en milisegundos, las últimas dos columnas son para el pitch y son dos valores, el primer valor es la posición relativa del punto de pitch en porcentaje de la duración de cada fonema y el segundo el valor de pitch en Hertz.

En la palabra *plantaría* el énfasis se encuentra en donde esta acentuada en la nomenclatura del archivo para MBROLA la duración del fonema/i/ es de 80ms, un pitch de 100Hz iniciando desde el 90% de su reproducción.

El pitch permite dibujar una curva de entonación lineal por tramas La sonoridad se codifica en la base de datos pero los sonidos sordos no se modifican en el pitch en el momento de la síntesis.

Para el funcionamiento de MBROLA se necesita crear una base de datos de difonos que son unidades de habla que inician en medio de la parte estable de un fonema y finalizan en la mitad de la siguiente, gracias a esto se minimizan los problemas de concatenación en la síntesis y es donde se da la mayor parte de las transiciones y coarticulaciones entre fonemas. Como primer paso se crea un corpus de texto, para lo cual es necesario mezclar una lista de todos los fonos del idioma, esta lista se obtiene enumerando los alófonos y aunque no es necesario contar con todas las variaciones alofónicas para crear un sintetizador inteligible, pocos alófonos pueden afectar la naturaleza del habla sintética.

Una vez que se consigue la lista de fonos y en lo posible todos los alófonos se obtiene al instante una lista de difonos.

En FR1 una voz francesa, por ejemplo, no se consideran los alófonos en absoluto. Como resultado se dan algunos fenómenos alofónicos, como ensordecimiento de /R/ cuando es precedido de oclusivas sordas (b,v, d y g)

Plantaría un árbol			
Fonema	Duración (ms)	Porcentaje de inicio del fonema	Pitch
p	100		
l	80		
a	90		
n	80		
t	85		
a	90		
r	50		
i	80	90	100
a	90		
_	50		
u	90		
n	80		
_	50		
a	108	30	130
r	50		
b	60		
o	90		
l	80		
_	50		

Tabla 4. Nomenclatura del archivo.

La grabación del corpus es por un hablante profesional entrenado, lee el corpus que se graba y guarda digitalemte esta lectura se hace con una entonción monótona para mantener una F0 constante donde se utlizan dispositivos de audio de alta calidad como pre-amplificadores y convertidor A/D que evitan agregar ruido a la grabación aunque la mejor manera de evitarlo es realizar la grabación dentro de una sala insonorizada profesional

Y finalmente para la segmentación todos los difonos deben ser vistos manualmente con herramientas de visualización de señal o de manera automática a través de algoritmos de segmentación.

1.1.3 Emofilt

Un sistema importante para la comparación de resultados es Emofilt [7] el cual es un programa de código abierto que simula la activación emocional con síntesis de voz que se basa en el motor de síntesis de MBROLA, así que hay que tenerlo instalado para poder utilizar Emofilt. Es de uso libre, es decir no comercial, tampoco como en el caso de MBROLA es un sistema de texto a habla completo pero actúa como un convertidor entre la fonemización y el componente de generación de habla.

Fue desarrollado originalmente en la Universidad Técnica de Berlín en 1998 y retomado en 2005 como un proyecto de código abierto y reescrito en el lenguaje java. Corre bajo una máquina virtual como Linux (Suse), Mac (OS X 1.6) y Windows XP.

La simulación emocional es mejorada gracias a la manipulación de los siguientes aspectos de una señal del habla, restringida por las limitaciones que un enfoque de la concatenación de difonemas presenta:

- Cambios en el pitch
- Cambios en la duración
- Calidad de voz (simulación de la vibración y soporte de la base de datos de múltiples calidades de voces)
- Articulación (remplazo de vocales centrales y descentralizadas con sus antagonistas)

Emofilt tiene dos interfaces principales:

Desarrollador emofilt es un editor gráfico para archivos XML que describen la emoción con retroalimentación acústica y visual. Véase *Figura 1*. *La interfaz etiquetador de historia* consiste en un editor de texto con la posibilidad de marcar el texto con las emociones como lo muestra la *Figura 2.* Y Emofilt en sí mismo toma archivos donde se describe la emoción como una entrada que actúa como un filtro en el framework de MBROLA.

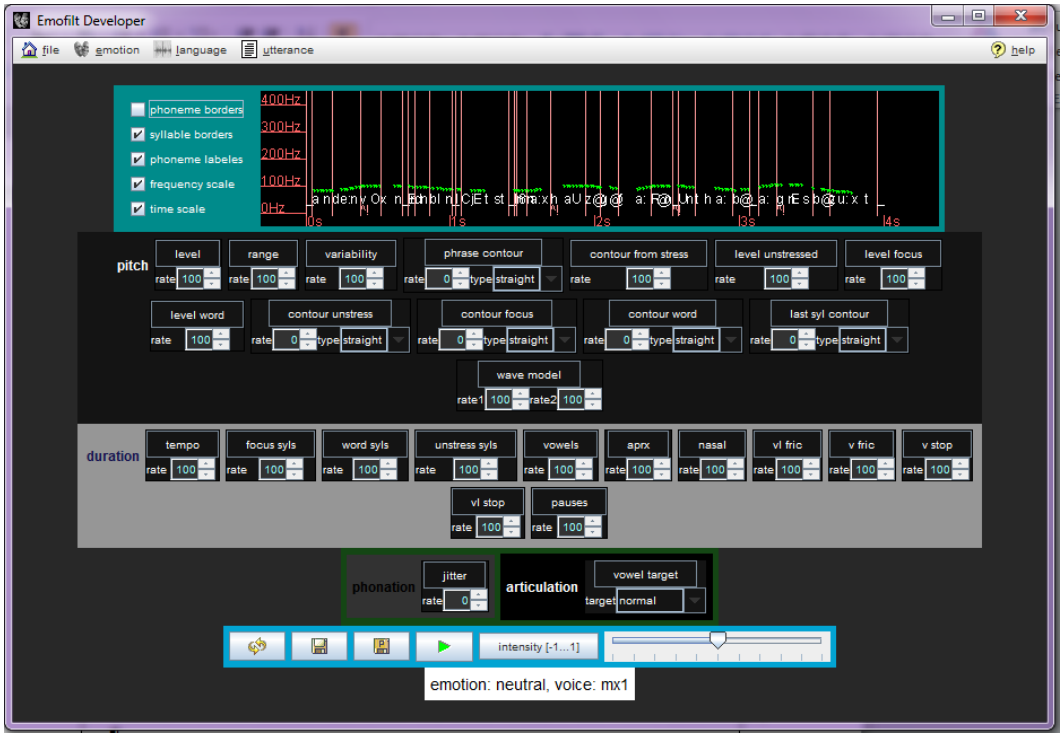


Figura 1. Desarrollador Emofilt.

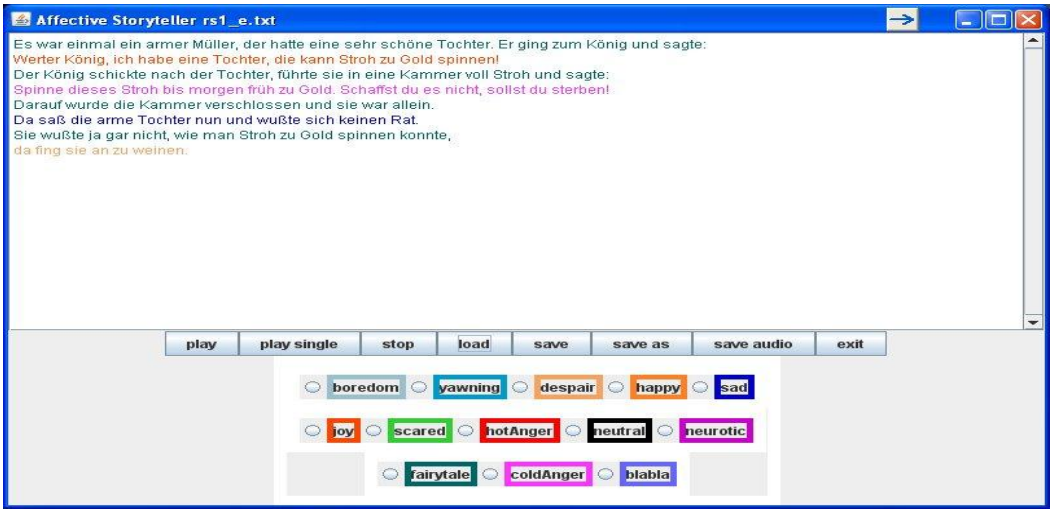


Figura 2. StoryTagger interface.

2. Marco Teórico

2.1 Historia de los Sistemas de Texto a habla

El primer sistema de texto a habla completo fue hecho por Noriko Umeda en 1968 en el laboratorio Electrotechnical en Japón, posteriormente junto con Cecil Coker y Catherine Browman crean el primer TTS para los laboratorios Bell en el año de 1973. En ese mismo año los laboratorios Haskins no se quedaron atrás y también crearon su propio sistema de texto a habla.

Para el año de 1976, Raymond Kurzweil desarrolló el Kurzweil una máquina de lectura para los invidentes. Un sistema muy económico llamado Votrax Type-n-Talk fue elaborado por Richard Gagnon en 1978. Un año después en 1979 aparece el sistema M.I.T. MITalk, sus desarrolladores fueron Jonathan Allen, Sheri Hunnicut y Dennis Klatt

El año de 1982 contó con varios sistemas, entre ellos, uno que manejó la concatenación de difonos y que fue nombrado Echo low-cost, además surgió el sistema Infovox multi-lenguaje de Rolf Carlson, Bjorn Granstrom y Sheri Hunnicut y finalmente el Speech Plus Inc. “Prose-2000” un sistema comercial.

Al año siguiente el sistema Klattalk de M.I.T de Deniss Klatt fue la base para el sistema comercial DECtalk de la corporación Digital Equiptment. El DECtalk contenía muchas voces y hablaba cerca de 300 palabras por minuto.

Le sigue el primer sintetizador para un sistema operativo el Macin Talk lanzado por Apple computer´s en 1984 y un año después surge AmigaOS un segundo sistema operativo que en su sistema de síntesis de voz incorpora voces masculinas y femeninas e índices de stress. En ese mismo año de 1985 los laboratorios Bell de AT&T también contaron con su sistema de texto a habla [8] [9].

En 2001 eSpeak TTS permite la manipulación de parámetros como tasa y volumen utilizando la síntesis por formantes.

Interactive Loquendo TTS Demo-Naunce en 2005 permite el habla de cualquier texto, en diferentes idiomas y regiones manejando la selección de unidades.

En 2011 AT&T Labs Research crea natural voices que produce sonidos naturales a través de selección de unidades y de Modelo de Ruido más Amónico.

Cepstral LLC en 2012 y puede hablar cualquier texto con la voz que se elija la prosodia se obtiene de material pregrabado.

2.2 Antecedentes

Los sistemas de texto a habla o TTS tienen un buen número de desarrollos con diversas características y enfoques, mejoras e idiomas; principalmente ha habido muchos avances en el idioma inglés [10] como lo muestra la *Tabla 5 y Tabla 6.*

Sistema	Idioma(s)
YORK TALK	Inglés
TTS-University of Birmingham	Inglés europeo y americano
Dec Talk	Inglés
Ipox	Alemán
Eurovocs	Japonés, Inglés, Alemán, Español y Francés

Tabla 5. Desarrollos de sistemas TTS [10]

Sistema	Idioma	Tecnología/ Enfoque	Destreza Lingüística	Aplicación	Disp.
Español interactivo/en marcha	E	G/R	A, L, E, O	EL	Com
GETARUN/SLIM	I, It	S, RH	A, Pron	EL	Inv
Hwe-1997	E	G/R	Pron, Fon	EL	Inv
I SEE	I	PT, S	V	EL	Inv
Learn to Speak Span	E	RH	O, Pron, V, G	EL	Com
Pron. y Fonét. 2.0	E	G/R, vis, órg. art.	Pron, Fon	EL	Com
Pronto	E	RH	Pron	EL	Inv
SPACE	N	S/RH	L	EL	Inv
TAIT	E, A	RH	O, Pron	EL	Inv
Versant	I, E, Ar	RH	O, Pron	TE	Com
	E=español, I=inglés, It=italiano, N=neerlandés, A=alemán, Ar=árabe.	G/R=grabación R=reproducción, S=síntesis de voz, RH=reconocimiento del habla, PT=pantalla táctil, vis=visualización de espectrograma, órg. art. =órganos articulatorios.	A=comprensión auditiva; E=producción escrita, L=comprensión lectora, O=destrezas orales, Pron=pronunciación , Fon=aprendizaje de la fonética y la fonología, V=vocabulario y G=gramática.	EL=enseñanza de lenguas	Com=comercial, Inv.=Prototipo o proyecto de investigación. [11]

Tabla 6. Características de los sistemas y los programas analizados para la enseñanza de la lengua [10]

La creación de los TTS en inglés han servido de inspiración para que a partir de ellos se relicen varios esfuerzos; TTS en español de España, Venezuela, Colombia un español latinoamericano.

Algunos TTS completos se muestran en la *Tabla 7*, los cuales manejan diversos tipos de síntesis como son por formantes, selección de unidades, concatenación. También implementan nuevos algoritmos con el caso de AT&T el HMN o Modelos de ruido más armónico y además cuentan con voces de casi todas las partes del mundo como inglés americano, inglés británico, alemán, italiano, francés, chino, español de varias regiones de Latinoamérica.

Sistemas de Texto a habla						
Afiliación	**Año**	**Descripción**	**Idiomas**	**Hablante**	**Tecnología**	**Disp.**
AT&T Labs – Research.[12]	2011	Natural voices es un sistema que toma texto y produce sonidos naturales, una voz sintetizada en una gran variedad de voces e idiomas.	Español	Hombre Mujer	Selección de unidades Modelo de Ruido más Armónico (HMN siglas en inglés)	Comercial
Cepstral LLC. [13]	2012	Puede hablar cualquier texto que se les da, con la voz que usted elija.	Inglés y español	Hombre Mujer	Obtención de prosodia natural del material grabado	Comercial
Interactive Loquendo TTS Demo-Naunce [14]	2005-Act.	Habla cualquier texto con diferentes idiomas, regiones y hablantes, su salida es un .wav	Español, Inglés, Italiano, Francés, etc.	Hombre Mujer	Selección de unidades	No comercial
eSpeak [15]	2001	TTS que permite la manipulación de parámetros como la tasa, volumen, formato de sonido y da como salida un .wav	Español Inglés	Hombre Mujer	Síntesis por formantes	No comercial

Tabla 7. Sistemas de texto a habla.

En cuestión de síntesis se tienen varios desarrollos, como en el caso de Emofilt, MBROLA, uno por parte de GTH. Los que serán de ayuda para este proyecto son MBROLA y Emofilt por sus características de síntesis concatenativa que da un habla más natural como se muestra en la *Tabla 8*.

Sistemas de síntesis de voz no comerciales							
Autor	**Sistema**	**Afiliación**	**Año**	**Descripción**	**Idiomas**	**Ventajas**	**Tipo de Síntesis**
Roberto Barra Chicote[16]	Grupo de Tecnología del Habla.	Polytechnic University of Madrid	1980 - 2010	Síntesis del habla emocional basada en HMM. Ej. Uso prosódico emocional y modelos fuente en datos de voz codificados en HMM.	Español.	Identificación automática de la emoción en el habla.	Formantes, concatenativa, HMM.
Felix Burkhardt [17]	Emofilt.	Dautsche Telekom Laboratories.	2005	Basado en la simulación (prosodia) con MBROLA.	Voces de MBROLA.	Utilización de las voces de MBROLA.	Concatenativa.
F. Malfrere [18]	Mbroling	MBROLIGN from TCTS Lab, Mons. TCTS Lab	1999	Síntesis con prosodia basada en datos con MBROLA (el algoritmo de árboles de decisión fue entrenado en una base de datos hablada con un efecto apropiado).	Castellano Español	Etiquetado fonético y prosódico.	Concatenativa.
VOQUAL group [18]	MBROLA.	Facultad Politécnica de Mons (Bélgica).	1996 – 2006	Generación de prosodia	Español	Generación de prosodia.	Concatenativa.

Tabla 8. Sistemas de síntesis de voz no comerciales.

2.3 Sistema de Texto a habla

Un sistema de texto a habla es una tecnología lingüística que debe ser capaz de que una máquina lea un texto y lo transforme automáticamente a su correspondiente forma sonora [19] y [20].

La *Figura 3* muestra el diagrama funcional de un sintetizador TTS de manera general [24], el cual comprende de un módulo de procesamiento del lenguaje natural (NPL por sus siglas en inglés) o también llamado *front end*, capaz de producir una transcripción fonética de la lectura del texto junto con la entonación y el ritmo deseado, denominado frecuentemente como prosodia.

El módulo NPL tiene como primera tarea convertir el texto, símbolos, números y abreviaciones, a su equivalente en palabras escritas, esto llamado normalización del texto, pre-procesamiento o señalización. La segunda tarea es asignar transcripciones fonéticas a cada palabra y dividirlas en unidades prosódicas como frases, cláusulas y oraciones, este proceso se llama conversión texto a fonema o lo que es lo mismo de grafema (representación gráfica de las letras) a fonema [21]. La salida del bloque NLP es una representación lingüística formada por transcripciones fonéticas y por información prosódica.

El bloque de proceso de síntesis o back end transforma la información simbólica que recibe del bloque anterior, en una voz de salida. Este es el sintetizador que transforma la representación lógica lingüística en sonido.

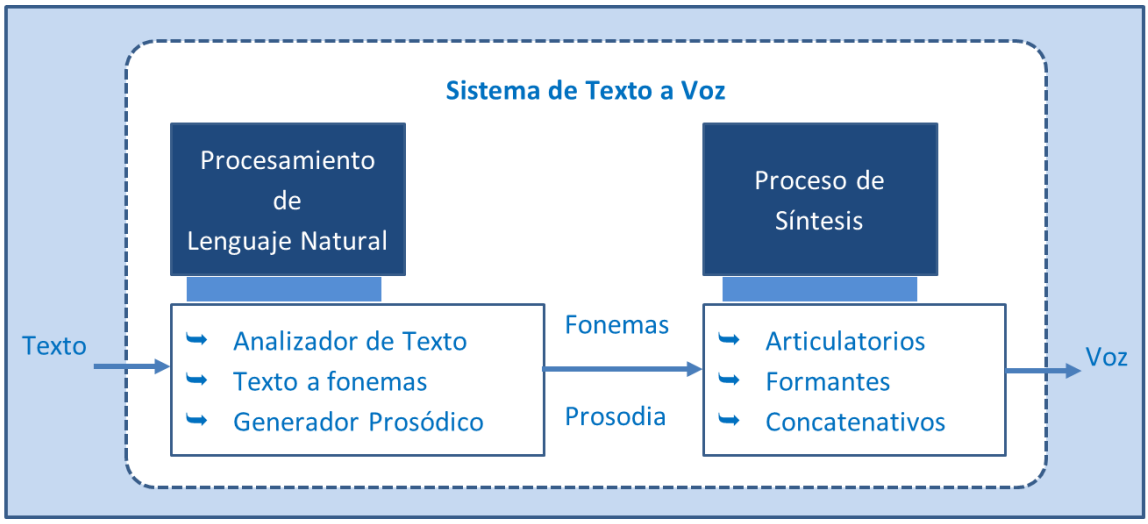


Figura 3. Estructura general de un sistema TTS. [33]

2.4 Transcripción ortográfica fonética

La función del sistema de texto a habla es transcribir automáticamente texto ortográfico en una cadena de simbolos de tipo fonético.

Los grafemas son las unidades mínimas de escritura de un idioma que correponden a las letras del alfabeto; el transcriptor ortográfico fonético trabaja con una serie de reglas que indican cómo se deben de transcribir estos grafemas a unidades de tipo fonético atendiendo al contexto en que se presentan. La utilidad de este transcriptor ortográfico-fonético es principalmente para reconocimiento de voz [22].

Existen dos aproximaciones a la conversión de grafemas a fonemas: los sistemas basados en reglas y los métodos inductivos, que intentan aprender automáticamente las reglas fonológicas a partir de ejemplos.

La transcripción de grafemas ha tenido mucha más atención en otras lenguas tales como el inglés. Los sistemas de síntesis utilizan normalmente transcriptores basados en reglas y un diccionario de excepciones (como, por ejemplo, el sistema MITalK), aunque también se han propuesto aproximaciones inductivas, como el sistema conexionista NETtalk de Sejnowski y Rosenberg.

2.4.1 Sistema de producción vocal

La comunicación oral es la expresión de los pensamientos que se da a través de la palabra hablada con fines comunicativos, la cual tiene como ventajas:

Facilidad: Mecanismo natural de la comunicación humana.
Aprendizaje: Mecanismo más precoz de comunicación.
Expresión: Se transmite información extra-lingüística generada por manifestaciones fisiológicas de estados anímicos o patologías.

La voz es el resultado de un proceso físico voluntario del aparato fonador, es una señal acústica o sonido, una onda de presión longitudinal formada por la compresión y expansión de las moléculas del aire que se transmite en dirección paralela a la aplicación de la energía.

Anatomía del aparato fonador

El aparato fonador es el responsable de la producción de la voz humana, para esto intervienen los órganos del sistema respiratorio y digestivo controlados por el sistema nervioso central.

Un flujo de aire sale de los pulmones, pasa por la laringe donde se encuentran las cuerdas vocales, la faringe, las cavidades, oral y nasal y finalmente los elementos articulatorios, labios, dientes, alvéolo, paladar, velo del paladar y la lengua, así es conformado el aparato fonador que se muestra en la *Figura 4*, el cual se divide en tres partes:

El sistema de generación. Integrado por los músculos abdominales y torácicos que hacen presión en los pulmones, los cuales producen una gran cantidad de aire que sale por los bronquios, la tráquea y finalmente llega a la laringe donde sufre una excitación por parte del sistema de vibración.

El sistema de vibración. En este están involucradas las cuerdas vocales, las superiores y las inferiores que son las que participan en la producción de voz. Cuando se encuentran juntas y tensas el aire choca haciendo que se produzcan sonidos diversos.

El sistema resonante. Conformado de las cavidades faríngea, oral y nasal. Los sonidos provenientes del sistema anterior se desplazan hacia las cavidades donde son modificados y amplificados para ser expulsados hacia el exterior.

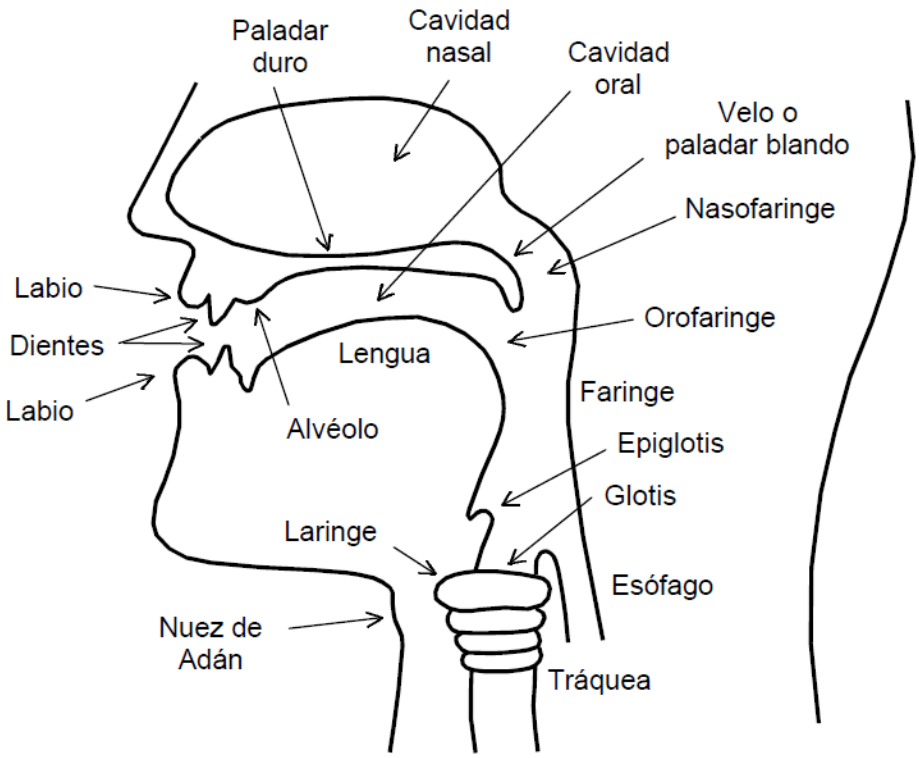


Figura 4. Corte esquemático del aparato fonador.

El grupo de órganos que intervienen en la fonación se dividen en los siguientes tres grupos:

Cavidades infraglóticas. Constan de los órganos para la respiración, es decir pulmones, bronquios y tráquea. En el proceso los pulmones toman aire, baja el diafragma y agranda

la cavidad torácica, esta espiración, crea la energía necesaria para generar una onda de presión acústica que llegará a los órganos fonadores superiores.

Cavidad laríngea. Modifica el flujo de aire que es generado por los pulmones y así convertirlo o no en una señal. El cricoides, último cartílago de la tráquea, forma la base de la laringe, siendo su órgano principal las cuerdas vocales, los cuales son dos pares de repliegues compuestos de ligamentos y músculos. Están formadas por un par inferior o cuerdas vocales verdaderas, que se juntan o separan por medio de la acción de los músculos crico-aritenoides lateral y posterior, estas están protegidas en su parte anterior por el cartílago tiroides y la parte superior de la laringe está unida al hueso hioides.

La figura 5 presenta un corte esquemático de la laringe en plano horizontal, las cuerdas vocales están en posición extrema abierta y la apertura de estas es llamada glotis. La cavidad laríngea termina en la epiglotis que es un cartílago en forma de cuchara que cierra la apertura de la laringe en el acto de deglución.

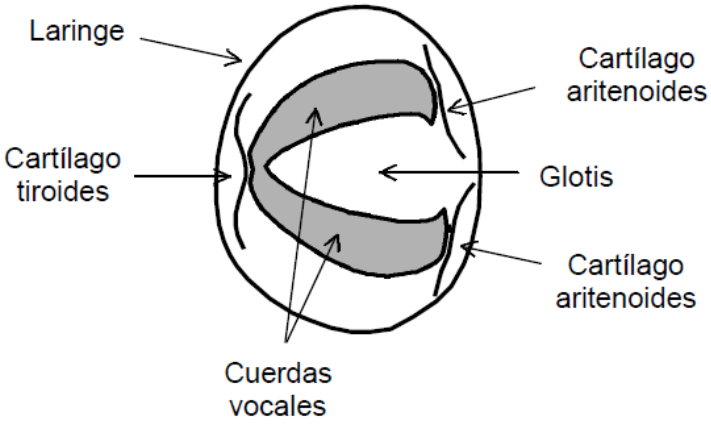


Figura 5. Corte esquemático de la laringe en plano horizontal.

Cavidades supraglóticas. O tracto vocal, están integradas por la faringe, cavidad oral y nasal. En la fonación perturba el flujo de aire procedente de la laringe, donde da lugar a la señal acústica generada a la salida de la nariz y la boca. La faringe de forma tubular, une la laringe con las cavidades bucal y nasal, se divide en tres: faringe laríngea, faringe bucal y faringe nasal, estas dos últimas separadas por el velo paladar. Mientras el volumen de la faringe laríngea es modificado por los movimientos de la laringe, lengua y epiglotis, el volumen de la faringe bucal es modificado por la lengua.

En el tracto vocal existen varios lugares de articulación como lo muestra la *Figura 6* donde los sonidos sufren cambios temporales, los cuales se relacionan directamente con la salida de la voz.

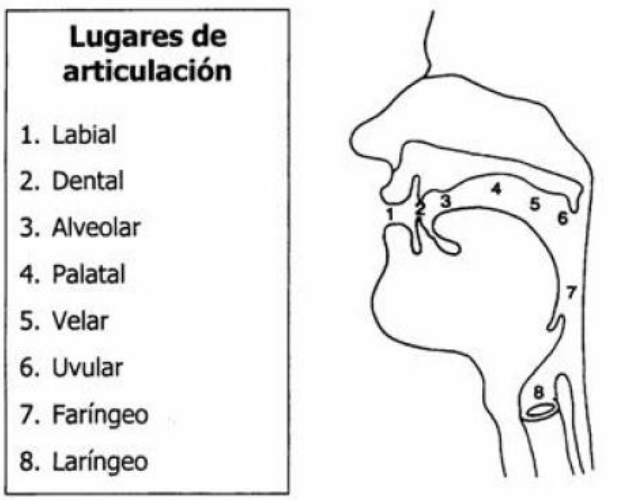


Figura 6. Lugares de articulación.

El aparato fonador emite algunos sonidos que pueden clasificarse de acuerdo a ciertos aspectos del fenómeno de emisión, los cuales son:

- Carácter vocálico o consonántico
- Oralidad y nasalidad
- Tonalidad
- Lugar de articulación
- Modo de articulación
- Posición de los órganos de articulación
- Duración

Carácter vocálico y consonántico. Las vocales son sonidos emitidos en una sola vibración de las cuerdas vocales donde no existe ningún obstáculo entre la laringe, y las cavidades oral y nasal, son sonidos de carácter tonal de espectro discreto. En cuanto a las consonantes se emiten teniendo como obstáculo algún elemento articulatorio, los sonidos pueden ser tonales en consecuencia de la vibración de las cuerdas vocales.

Oralidad y nasalidad. Los fonemas que se producen por medio de la boca son llamados orales y los que el aire pasa por la cavidad nasal, nasales.

Tonalidad. Los fonemas *tonales* o *sonoros* se producen cuando hay vibración de la cuerdas vocales, todas las vocales son tonales y algunas consonante como “b”, “d”, “m” entre otras. En cambio los fonemas *sordos* carecen de vibraciones glotales más bien se

producen por la turbulencia del aire pasando velozmente por un espacio reducido, ejemplo de ello “f”, “j” “s”, “z”.

Lugar de articulación. La articulación consiste en que algún miembro del aparato fonador funcione como un obstáculo para la circulación del flujo de aire. Dependiendo del lugar de articulación se producen los fonemas (*Tabla 9)*:

Alveolares: Por oposición de la punta de la lengua con la región alveolar.
Bilabiales: Por oposición de ambos labios.
Glotales: Por articulación en la propia glotis.
Labiodentales: Por la oposición de los dientes superiores con el labio inferior.
Linguodentales: Por oposición de la punta de la lengua con los dientes superiores.
Palatales: Por oposición de la lengua con el paladar duro.
Velares: Por oposición de la parte posterior de la lengua con el paladar blando.

La realización de la articulación se puede efectuar de diferente manera produciendo los siguientes fonemas que se presentan en la *Tabla 10*:

Africados: Oclusión seguida por fricación.
Aproximantes: La obstrucción muy estrecha que no llega a producir turbulencia.
Fricativos: El aire sale atravesando un espacio estrecho.
Laterales: La lengua obstruye el centro de la boca y el aire sale por los lados.
Oclusivos: La salida del aire se cierra momentáneamente por completo.
Vibrantes: La lengua vibra cerrando el paso del aire intermitentemente.

Lugar de articulación	**Modo de articulación**								
	Oral								**Nasal**
	Oclusiva		**Fricativa**		**Africada**	**Lateral**	**Vibrante**	**Aproximante**	**Sonora**
	Sorda	**Sonora**	**Sorda**	**Sonora**	**Sorda**	**Sonora**	**Sonora**	**Sonora**	
Bilabial	p	b, v		b, v				w	m
Labiodental			f						
Linguodental			z	d					
Alveolar	t	d	s	y	ch	l	r, rr		n
Palatal				(y)	(ch)	ll		i	ñ
Velar	k	g	j	g					
Glotal			h						

Tabla 9. Clasificación de las consonantes de la lengua castellana según el lugar y el modo de articulación y la sonoridad.

Posición vertical	Tipo de vocal	Posición horizontal (avance)		
		Anterior	Central	Posterior
Alta	Cerrada	i		u
Media	Media	e		o
Baja	Abierta		a	

Tabla 10. Clasificación de las vocales castellanas según la posición de la lengua.

2.4.2 SAMPA

El alfabeto computacional Speech Assessment Methodology Phonetic Alphabet (SAMPA) fue desarrollado como parte del proyecto ESPRIT 1541 entre 1987 y 1989 por un grupo internacional de fonetistas. Los primeros idiomas que manejó fueron: danés, francés, inglés, holandés e italiano para el año de 1989; después se elaboraron las versiones para el noruego y el sueco en 1992, y posteriormente para el español, griego y portugués en 1993.

Entre uno de sus propósitos, SAMPA busca simplicidad en la transcripción y la relativa facilidad de uso por parte de personas con poca formación en fonética. El principio que rige el alfabeto SAMPA es básicamente fonológico, al igual que el Alfabeto Fonético Internacional AFI o IPA por sus siglas en inglés, empleándose solamente símbolos distintos en el caso de segmentos con valor diferencial. La *Tabla 11* muestra el alfabeto computacional SAMPA para el español, que será de gran utilidad para la transcripción de texto a voz en español de México. Se puede ver en la primera columna el modo de articulación de los símbolos SAMPA o fonemas, la lista de todos los fonemas SAMPA consonantes, semivocales y vocales, la palabra original sin transcribir donde aparece el fonema en su forma de grafema y la transcripción SAMPA de cada fonema de la lista de palabras [23].

Modo de articulación	Símbolo SAMPA	Palabra original	Transcripción
Consonantes			
Plosivas	p	**p**adre	"**p**aDre
	b	**v**ino	"**b**ino
	t	**t**omo	"**t**omo
	d	**d**onde	"**d**onde
	k	**c**asa	"**k**asa

	g	**g**ata	"**g**ata
Africadas	tS	mu**ch**o	"mu**tS**o
	jj	**hi**elo	"**jj**elo
Fricativas	f	**f**ácil	"**f**aTil
	B	ca**b**ra	"ka**B**ra (= **/b/**)
	T	**c**inco	"**T**inko
	D	na**d**a	"na**D**a (= **/d/**)
	s	**s**ala	"**s**ala
	x	mu**j**er	mu"**x**er
	G	lue**g**o	"lwe**G**o (= **/g/**)
Nasales	m	**m**ismo	"**m**ismo
	n	**n**unca	"**n**unka
	J	a**ñ**o	"a**J**o
Líquidas	l	**l**ejos	"**l**exos
	L	caba**ll**o	ka"ba**L**o (o como **jj**)
	r	pu**r**o	"pu**r**o
	rr	to**rr**e	"to**rr**e
Semivocales	j	re**i**	rre**j**
		p**i**e	p**j**e
	w	de**u**da	"de**w**Da
		m**u**y	m**w**i
Vocales	i	p**i**co	"p**i**ko
	e	p**e**ro	"p**e**ro
	a	v**a**lle	"b**a**Le
	o	t**o**ro	"t**o**ro
	u	d**u**ro	"d**u**ro

Tabla 11. Alfabeto computacional SAMPA.

2.4.3 Las reglas de acentuación

El proceso de acentuación parte de una palabra correctamente tildada, si la palabra lleva tilde no necesita ser acentuada debido a que el golpe de voz cae en la sílaba tildada [22], estas son las reglas de acentuación ortográficas, de acuerdo a la colocación de la sílaba tónica y la terminación de la palabra como lo muestra la *Tabla 12* [24], un ejemplo de estas son las sobresdrújulas y esdrújulas ya que todas se acentúan de forma escrita y no necesitan ser procesadas, por ejemplo las palabras *es**crí**beselas* y ***rá**pido*.

Sin embargo las palabras que no llevan tilde se acentuaran con otras reglas complementarias:

Agudas: Son agudas aquellas palabras que terminan en consonante exceptuando n, s, se aplicará acento prosódico en la última sílaba.

Graves o llanas: Se aplicará la acentuación prosódica en la penúltima sílaba a las palabras que terminan en n, s o vocal.

Las monosílabas no se acentúan.

Sobresdrújulas	Esdrújulas	Graves o llanas	Agudas
			_ _ _ *
		_ _ * _	
	_ * _ _ _		
* _ _ _ _			
¿Cuáles se acentúan de forma escrita o se tildan?			
Todas	Todas	Terminadas en consonante, menos n, s o vocal.	Terminadas en vocal o consonantes n, s.
Ejemplos			
O-tór-ga-se-lo A-ví-sa-me-lo Es-crí-be-se-las	Má-gi-co Rá-pi-do Lá-gri-mas	Héc-tor A-ní-bal Váz-quez	Ru-bí Ven-drás Es-tán

Tabla 12. Reglas de acentuación ortográfica.

Al detectar la sílaba tónica, se necesita resolver el caso existente de cuando hay más de una vocal y/o sílaba de manera que se acentúe la vocal adecuada, para lo cual se aplican reglas de diptongos e hiatos.

Diptongo: Vocal fuerte(a, e, o) + vocal débil (i, u), el acento fonético recae en la vocal fuerte: bElla, ciErra, puEsto.

La combinación de vocal débil + vocal débil el acento recae en la segunda letra: ruIdo, fuImos y viUda.

Hiato: Dos vocales fuertes no pueden compartir sílaba: ma-Es-tro, con-tra-Er y siguen las normas generales.

Las palabras que no pasan por estas reglas llevan acento ortográfico.

Las palabras que terminan en el sufijo *mente* llevan el acento en la primera *e* más el acento de la sílaba tónica de la palabra, es decir tienen dos sílabas tónicas, el resto de las palabras solo tienen una.

Triptongo: La unión de tres vocales contiguas, una fuerte en medio de dos débiles, forman un triptongo si llevaran acento prosódico u ortográfico se hará en la vocal fuerte como ParaguAy y apreciÉis.

2.4.4 Silabeo

El alfabeto o abecedario del *Cuadro 1* del español se compone por 27 letras, las cuales debido a su pronunciación se clasifican en dos grupos: vocales y consonantes.

A	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**
J	**K**	**L**	**M**	**N**	**Ñ**	**O**	**P**	**Q**
R	**S**	**T**	**U**	**V**	**W**[1]	**X**	**Y**	**Z**

Cuadro 1. Abecedario del español.

Fonéticamente, el grupo de las vocales se conforma por seis letras con sonido propio, es decir pueden pronunciarse solas, sin apoyarse de ningún otro sonido.

Las vocales se clasifican en abiertas, semiabiertas o cerradas, debido a que la boca actúa como una caja de resonancia que al pronunciar una vocal se realiza una apertura de menor a mayor grado permitiendo la salida del aire sin dificultad.

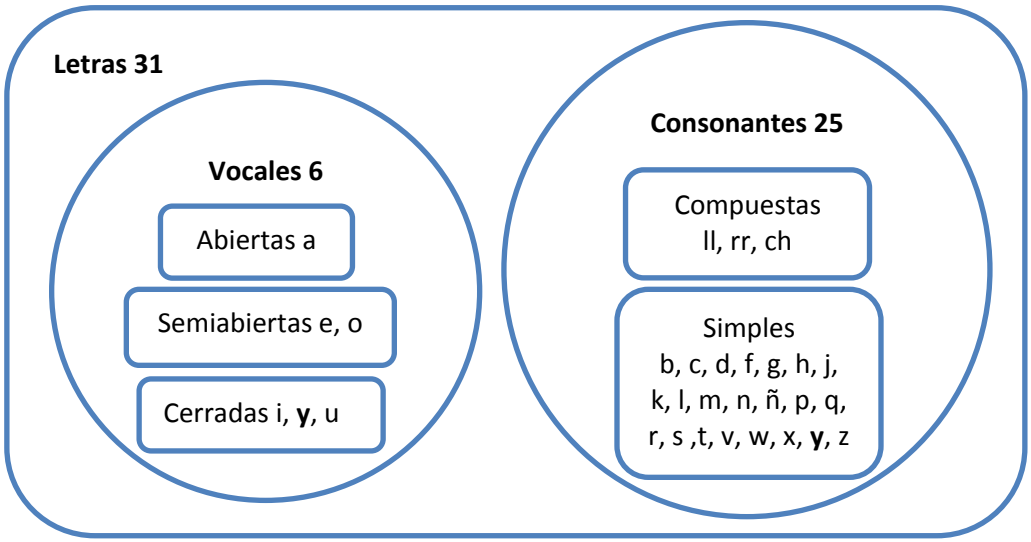
El grupo de las consonantes, está formado por veinticinco letras, éstas solo se pronuncian si van apoyadas por lo menos de una vocal. Existen las consonantes compuestas que son 3, éstas son simples en su pronunciación pero dobles en su escritura. Las restantes son consonantes simples en su pronunciación y escritura, véase el *Cuadro 2*.

Los grupos de letras que se pronuncian en una sola emisión de voz se llaman sílabas. En castellano existen silabas formadas por:

Un solo sonido, siempre es una vocal: va-mos **a**, **a**-par-ta-do, **ú**-ni-co.
Dos sonidos, una vocal y una consonante, no importa el orden: **al**-ma, **re**-**to**.
Tres sonidos: **her**-mo-sa, fa-mi-**lia**.

[1] La W se emplea en voces extranjeras

Cuatro sonidos, **siem**-pre, en-cua-der-na-**ción**.
Más de cuatro sonidos: **triun**-fo, **trans**-la-ción.



***Cuadro* 2. Clasificación de las letras.**

Como se puede notar el grafema "y" aparece como vocal y consonante, ya que en ciertas palabras suena como "i" y en otras como "ll", por lo que toma lugar en ambos grupos.

Las palabras están formadas por varias silabas, y de acuerdo con esto es el nombre que reciben:

Monosílabas, tienen una sola sílaba: dar, pan, sol.
Bisílabas o *disílabas*, con dos sílabas: can-ción, a-mor.
Trisílabas, están formadas por tres: a-mi-go, á-ni-mo.
Tetrasílabas, constan de cuatro: des-hi-dra-tar, di-ná-mi-ca.
Polisílabas, se llaman así las que tienen cinco o más sílabas: fe-rro-ca-rri-le-ro, tra-ba-ja-do-ras.

No se pueden separar las letras que forman una sílaba. Las letras dobles correspondientes a un solo sonido se anotan, sin dividir, con las sílabas a la que pertenecen. A partir de esto surgen varias reglas para realizar el silabeo *Tabla* 13 [10].

#	Regla de silabificación	Grupos	Ejemplo
1	En las sílabas tiene que haber al menos un vocal, sin vocal no hay sílaba.		**ú**-ni-co.
2	Cada elemento del grupo de *consonantes inseparables*, no puede ser separado al dividir una palabra en sílabas.	bl, br, ch, cl, cr, dr, fl, fr, gl, gr, kr, ll, pl, pr, rr, tr, tl.	
3	Cuando una consonante se encuentra entre dos vocales, se une a la segunda vocal.		u**n**e, u-**n**e
4	Cuando hay dos consonantes entre dos vocales, cada vocal se une a una consonante. Exceptuando el grupo de consonantes inseparables.		co**mp**oner, co**m**-**p**o-ner a**pr**ender, a-**pr**en-der.
5	Si son tres las consonantes colocadas entre dos vocales, las dos primeras consonantes se asociarán con la primera vocal y la tercera consonante con la segunda vocal. Aunque esta regla no se cumple cuando la segunda y tercera consonante forma parte del grupo de consonantes inseparables.		tra**nsp**orte, tra**ns**-**p**or-te. cu**mpl**e, cu**m**-**pl**e
6	Las palabras que contienen una h precedida o seguida de otra consonante, se dividen separando ambas letras.		a**nh**elo, a**n**-**h**e-lo
7	El diptongo es la *unión inseparable* de dos vocales , existen tres tipos de diptongos: Vocal abierta + vocal cerrada Vocal cerrada + vocal abierta Vocal cerrada + vocal cerrada La unión de dos vocales abiertas o semiabiertas, no forma diptongo, pueden quedar solas o unidas a una consonante.	ai, au, ei, eu, io, ou, ia, ua, ie, ue, oi, uo, ui, iu, ay, ey,	j**au**la, j**au**-la. **aé**reo, **a**-**e**-re-o.
8	La h entre dos vocales, no destruye un diptongo.		**ahu**yentar, **ahu**-yen-tar
9	La acentuación sobre la vocal cerrada de un diptongo provoca su destrucción.		Mar**ía**, Ma-r**í**-**a**

10	La *unión de tres vocales* forma un triptongo, vocal cerrada + (vocal abierta o vocal semiabierta) + vocal cerrada. Las combinaciones de vocales forman los siguientes triptongos.	iai, iei, uai, uei, uau, iau, uay, uey.	

Tabla 13. Reglas de silabificación.

2.4.4.1 Estructura de las sílabas

Con base en las reglas mencionadas se deduce una estructura de las sílabas mostrada en el *Cuadro 3* [10].

(V)	vocal, 1 o 2 vocales
(VC)	vocal (1 o 2 vocales) + consonante (1 consonante)
(CV)	consonante (1 ó 2 consonantes) + vocal (de 1 a 3 vocales)
(CVC)	consonante (1ó 2 consonantes)+vocal (de 1 a 3 vocales)+consonante (1 ó 2 consonantes)

Cuadro 3. Estructura de las sílabas.

Existen tres casos dentro de las combinaciones de las sílabas en el idioma español que se desarrollan a continuación [10].

Caso 1. Inicio de sílaba es *vocal*

Las posibles combinaciones de inicio de sílaba con vocal se muestran en la *Tabla 14*, con sus respectivos ejemplos y la regla de aplicada para la separación de la palabra en sílabas.

Combinación	Ejemplo	Número de regla aplicada
V	A	1
VC	Ar+co	4
VV	Au+tónomo	7
VVC	Aus+tero	7,4

Tabla 14. Inicio de sílaba: vocal.

Caso 2.- Inicio de sílaba es *consonante + vocal*

La *Tabla 15* muestra combinaciones, ejemplos y reglas para el caso 2.

Combinación	Ejemplo	Número de regla aplicada

C	Y	1
CV	La	1
CVC	Las	1
CVCC	Cons+tantinopla	5
CVV	Jau+ría	7
CVVC	Cuan+to	4
CVVV	Cuau+titlan	10
CVVVC	Cuauh+temoc	10,6

Tabla 15. Inicio de sílaba es: consonante + vocal.

Caso 3. Inicio de sílaba: *consonante + consonante*

Y la Tabla 16 muestra este último caso.

Combinación	Ejemplo	Número de regla aplicada
CCV	Tra + mo	2
CCVC	Tras + te	2,4
CCVCC	Trans + plante	2,5
CCVV	Trau + ma	2,7
CCVVC	Claus + tro	2,7,5

Tabla 16. Inicio de sílaba: consonante + consonante.

2.4.5 Reglas de transcripción

Las reglas de trascripción son las que se utilizan para la conversión de grafemas a fonemas y éstas han sido obtenidas de [19]. La correspondencia de la transcripción en algunos casos es de 1 a 1 de grafema a fonemas, por ejemplo: *b* siempre se pronuncia como /b/, bicicleta -> bisiklEta.

Aunque la mayoría de los casos las reglas son dependientes del contexto, por ejemplo para el grafema *c*: se pronuncia como /k/ antes de *a*, *o*, *u*; antes de *e*, *i* se pronuncia como /T/ o /s/; y antes del grafema *h* como /tS/. Ejemplos: caramelo -> karamElo, cinco -> sInko o TInko, chocolate -> tSokolAte.

Una característica importante de la transcripción es que necesitan ser aplicadas en cierto orden, como en el caso del grafema *ch* el cual se debe convertir antes a /tS/ y después aplicar a la palabra la regla en donde se elimina la h por ser muda, por ejemplo al transcribir *chocolate* debería obtenerse tSokolAte si se procesa antes la *h* se obtendría una transcripción errónea kokolAte. La *Tabla 17* muestra de manera detallada la transcripción

para español de España y México, con algunos ejemplos de cómo queda la transcripción con algunas diferencias entre ambos idiomas.

GRAFEMA	REGLAS DE TRANSCRIPCIÓN ESPAÑOL ESPAÑA A SAMPA	EJEMPLOS	REGLAS DE TRANSCRIPCIÓN ESPAÑOL MÉXICO A SAMPA	EJEMPLOS
A	**A**	ala: "Ala	**A**	azúcar: **a**sUkar
B	Después de una pausa: **b** Después <m> o <n>:**b** Otros casos: **B**	comba: "kom**b**a labio: "la**B**jo	<b> siempre representa el fonema /**b**/ bilabial sonoro de barco, beso, blusa o abuelo.	barco: **b**Arko
C	seguida de <e> o <i>:**T** en la posición final seguida por <l> o <r>:**G** seguida por <b>, <d>,<g> (precedida <a>, <o>, <u>,<m>,<n>,<ñ> or <v>):**G** otros casos: **K**	celo: "**T**elo acné: a**G**"ne tacto:"ta**k**to coro: "**k**oro tecla: "te**k**la	ante vocales <e> o <i>: **s** ante <a>, <o>, <u>, ante consonante y en posición final de sílaba o de palabra: **k**	celo: **s**Elo acné: a**k**nE coro: **k**Oro
Ch	**tS**	chelo: "**tS**elo	**tS**	chelo: **tS**Elo
D	Después de una pausa: **d** después <l>,<m> o <n>: **d** otros casos: **D**	caldo:"kal**d**o codo:"ko**D**o	**d**	caldo: kAl**d**o
E	**e**	elefante: **e**lefAnte	**e**	elefante: **e**lefAnte
F	**f**	cofia: "ko**f**ja	**f**	cofia: kO**f**ia
G	Después de una pausa y seguida por <r>, <l>,<a>, <o> o <u>: **g** Después de <m> o <n> y seguida por <a>, <o> o <u>: **g** seguida por <i> o <e>: **x** otros casos: **G**	tongo: "ton**g**o genio: "**x**enjo tigre: "ti**G**re lago: "la**G**o	Ante <a>, <o> o <u>, posición final de sílaba, y agrupado con otra consonante: **g** Dígrafo <gu> ante <e> e <i>: **g** Sonido independiente <gü> ante <e>, <i>: **g** Ante <i> o <e>:**x**	golosina: **g**olosina digno: dI**g**no guitarra: **g**uitarra antigüedad: anti**g**wedAd girasol: **x**irasOl
H	En posición inicial de palabra seguida por <ie>: **jj** otros casos: **No suena**	hierba: "**jj**erBa halo: "alo	En posición inicial de palabra seguida por <ie>: **jj** Otros casos: **No suena**	hola: Ola hierba: **jj**Erba
I	En posición nuclear in la sílaba: **i** en posición no nuclear en la sílaba: **j**	tipo: "t**i**po cielo: "T**j**elo	**i**	isla: **I**sla
J	**X**	jarana: "**j**arana	**x**	jalea: **x**alEa
K	**k**	kiosko: "**k**josko	**k**	kiosko: **k**iOsko
L	**l**	lote: "**l**ote	**l**	leer: **l**eEr

Ll	**L**	tallo: "ta**L**o	**dZ**	llanta: **dZ**Anta
M	**m**	arma: "ar**m**a	**m**	mamá: **m**amA
N	Seguida por <p>,<b>,<v>,<m> o <f>: **m** en otros casos: **n**	ánfora: "a**m**fora cono: "ko**n**o	**n**	natividad: **n**atibidAd
Ñ	**J**	uña: "u**J**a	**J**	niña: n**J**a
O	**O**		**o**	oro: Or**o**
P	**p**	perro: "**p**erro	**p**	papá: **p**apA
Q	Siempre seguida por <u>: **k**	queso: "**k**eso	Siempre seguida por <u>:**k** Algunas voces científicas, locuciones latinas <qu> ante <a>, <u>: **ku**	queso: **k**Eso quo: **k**uo
R	En posición inicial de palabra: **rr** precedida por <l>, <n> o <s>: **rr** otros casos: **r**	rama: "**rr**ama honra: "on**rr**a arpa: "a**r**pa trampa: "**t**rampa pera: "pe**r**a amor: a"mo**r**	En posición inicial de palabra y después de una consonante que no pertenezca a la misma sílaba: **rr** otros casos: **r**	rosa: **rr**Osa honra: On**rr**a
Rr	**rr**	carro: "ka**rr**o	**rr**	perro: pE**rr**o
S	**s**	rasgo: "rra**s**Go casa: "ka**s**a trasto: "tra**s**to	**s**	sauce: **s**Awse
T	En la posición de final de silaba: **D** Otros casos: **t**	atleta: a**D**"leta toro: "**t**oro	**t**	tucán: **t**ukAn
U	Sin diéresis precedida por <g> o <q>: **no suena** En posición no nuclear en la silaba: **w** En la posición nuclear en la silaba: **u**	queso: "keso cigüeña: Ti"G**w**eJa lujo: "l**u**xo	Sin diéresis precedida por <g> o <q>: **no suena** En posición no nuclear en la silaba: **w** En la posición nuclear en la silaba: **u**	agüita: ag**w**Ita puerto: p**w**Er to apúrate: ap**U**rate
V	Después de una pausa: **b** después <m> o <n>:**b** otros casos: **B**	con velo: kom "**b**elo calvo: "kal**B**o	<b> siempre representa el fonema **/b/** bilabial sonoro de barco, beso, blusa o abuelo.	vaca: **b**Aka
W	En palabras extranjeras: Gu, gü o como una <v>: **w**	whisky: "**gw**iski kiwi: ki**B**i	En palabras extranjeras: **Gu, gü**.	waffle: gwAffle
X	En posición no inicial de palabra y seguido por una vocal: **Gs** Otros casos: **s**	examen: e**G**"**s**amen externo: e**s**"terno	Intervocálica o final de palabra: **ks** (gs posición relajada) Posición inicial: **s** **Topónimos: x**	examen: e**ks**Amen relax: rrelA**ks** xoconostle: **s**okonOstle México: mE**x**iko

Y	En posición inicial de la sílaba con dos o más sonidos: **jj** Otros casos: será procesada como **i**	yunque: "**jj**unque cónyugue: "kon**jj**uGe dos y dos: dos **i** "Dos muy: mw**i**	En posición inicial de la silaba con dos o más sonidos, intervocálica: **jj** <y> y fin de palabra: será procesada como **<i>**	mui: mw**i** yoyo: **jj**Ojjo
Z	**T**	zarza: **T**ar**T**a tizne: "ti**T**ne	**s**	zorro: **s**Orro

Tabla 17. Reglas de transcripción para el español.

3. Metodología

El habla es la forma más natural e importante de comunicación en la interacción humana frente a frente. Debido al surgimiento de sistemas de información más sofisticados, se reconoce la creciente importancia de la interacción hombre-máquina [25], la cual puede ser observada comúnmente a través de los dispositivos de entrada y salida, como teclado, monitor, mouse, etc.

El presente proyecto se desarrolla a partir de la importancia que hay en la interacción hombre-máquina, de la simulación de un diálogo análogo a la forma de la comunicación humana y de la realización de tareas en común en la interacción hombre-máquina que se encuentra dentro del ámbito de las **tecnologías del habla**, dando paso a la creación de un ***Sistema de texto-a-habla expresivo en español*** (o TTS por sus siglas en inglés, Text to Speech). Un sistema TTS convierte arbitrariamente una entrada de texto a un sonido de habla natural e inteligible, simulando la interacción humana [26]. Para hacer esto posible, en el habla o voz obtenida debe ser perceptible alguna emoción, aunque actualmente los sintetizadores no incorporan la emoción [27] debido probablemente a la complejidad que presenta la expresión vocal humana.

Para realizar esta función el TTS está compuesto por dos etapas: al principio un módulo, llamado *front end*, que analiza el texto y lo convierte en una especificación lingüística, y posteriormente otro módulo llamado *back end* que toma esta especificación lingüística y la cambia a una forma de onda sintetizada. La especificación lingüística contiene información fonémica y prosódica, usualmente dando como salida una lista de los fonemas que aparecen en el texto y un conjunto de detalles sobre cómo se deben pronunciar. Estos detalles vienen en forma de valores de ciertos parámetros importantes como son la frecuencia fundamental F0 (llamada pitch), la duración, la calidad de voz y la articulación. Las especificaciones obtenidas se pasan a un sintetizador de voz para generar la señal a escuchar.

Se elige la F0 como un parámetro debido a que su variación le da diferentes matices al enunciado hablado, proporcionándole un énfasis por lo que ayuda a tener más claro de

qué emoción se trata. La duración de las frases o palabras y las pausas también contribuyen a conocer si se está hablando con cierta emoción ya que se tiende a hablar más rápido cuando se está enojado o más lento si se está triste. Estos parámetros son requeridos por el sintetizador Mbrola para realizar la síntesis de voz.

La logitud del tracto vocal del hablante humano su acento y dialecto son carcterísicas que pudiesen afectar la frecuencia fundamental de la voz por lo que se consideran los parámetros de la calidad de voz y articulación cuidando la selección del hablante humano y las condiciones de grabación.

Existe una gran variedad de sistemas TTS de diferente calidad que dependen de la técnica de síntesis con la que son elaborados como síntesis por formantes, concatenativa, Modelos Ocultos de Markov (HMM), por mencionar algunas. Estos sistemas TTS ofrecen salidas de habla en diferentes idiomas como inglés, alemán y francés y aunque también existe la versión en español, el avance ha sido muy pobre con respecto al del idioma inglés. Los tipos de archivos de la voz salida son .mp3, .wav, .pho, entre otros. Algunos ejemplos de los TTS son eSpeak, Festival y Mary TTS.

Los sintetizadores de voz utilizados para este proyecto son MBROLA y Emofilt los cuales no incoporan la primera etapa *front end* como lo hace el TTS-EE desarrollado en este trabajo de investigación.

A la fecha, los sistemas TTS han sido utilizados en muchas aplicaciones [19], algunas son enlistadas a continuación:

Sistemas de información telefónica

- Información meteorológica.
- Información ciudadana.
- Noticias.

Acceso telefónico a textos escritos

- Consulta a distancia de bases de datos.
- Mensajería vocal.
- Lectura del correo electrónico.
- Lectura de mensajes.
- Asistentes virtuales.
- Asistente interactivo digital para teléfonos móviles.

Aplicaciones a los invidentes

- Lectura de textos en formato electrónico.
- Escáner, reconocimiento óptico de caracteres, conversión de texto en habla.

Aplicaciones a los disminuidos vocales

- Prótesis vocales.

Los sistemas TTS y sus aplicaciones prometen formar parte importante en una interacción hombre-máquina más natural, su investigación y las potenciales mejoras de esta tecnología son imprescindibles.

3.1 Objetivo general

El objetivo principal de este proyecto es desarrollar un Sistema Texto-a-Habla Expresivo en Español de México.

3.1.1 Objetivos específicos

a. Implementar el módulo Front end del TTS en español de México en el lenguaje de programación Python.

b. Obtener parámetros prosódicos importantes como duración y pitch, calidad de la voz y articulación para lograr la incorporación de emociones en un TTS.

c. Integrar los puntos a y b **para formar** un sintetizador con emoción.

3.2 Justificación

La finalidad de este proyecto es obtener un Sistema de Texto a Habla Expresivo en Español, el cuál proporcionará una forma diferente de comunicación entre la computadora y el ser humano. Se tendrá como entrada un texto entendible para la computadora y se obtendrá como resultado un habla sintetizada compresible para el hombre la cual expresará emoción.

El principal problema al que se enfrenta el TTS-EE es la producción de habla “natural”, esto es que debe presentar sonidos naturales como la acentuación, coarticulación entre sílabas, pausas entre una sílaba u otra, entre otros elementos. El hecho de aplicar alguna

emoción al habla sintetizada como alegría, tristeza, o cualquier otra, favorecerá la experimentación con la emulación del habla y así hacerla más agradable en su apreciación [7], además ofrecerá la posibilidad de generar una señal de voz para fines de reconocimiento de emoción en el habla [11] y [30]

A partir de la revisión teórica se encontró que existen un sin número de TTS, particularmente en el idioma inglés con muy buena calidad y en los que se han utilizado una gran cantidad de parámetros para hacerlos más agradables y entendibles; por otro lado hay un menor número de desarrollos de TTS en español castellano y sobre todo en el de México del cual existen pocos esfuerzos; la contribución es la parte de TTS en español de México con la implementación de la ***x*** y palabras nativas de origen indígena y con el léxico de México; la grabación de un corpus oral para este español del cual no existe; se tiene un corpus completo con el cual se puede obtener un vocabulario ilimitado de palabras u oraciones a diferencia de algún sistema con frases pregrabadas.

En el sentido de la estructura de las palabras del idioma español se puede notar que la divisibilidad es de una o más sílabas, este tipo de segmentación permite tener un conjunto de sonidos pronunciados en una sola emisión de voz, también se trabaja con segmentación a nivel fonema aunque al unir dos fonemas la coarticulación de estos no es tan buena como en las sílabas [22], es por eso que se aplica el silabeo.

Es un trabajo integral donde se unen varios programas como SAMPA, Emofilt, Mbrola, Python, Praat-Easy Align; para la transcripción ortográfica fonética, reconocimiento de voz, silabeo, segmentación automática y manual y aplicación de emociones.

En el ámbito de la aplicación el proyecto de investigación puede ser útil para lectura automática de correos electrónicos, narración, asistente virtual online, eAprendizaje, ayudar a personas con capacidades diferentes como débiles visuales, invidentes, personas con problemas de comunicación o alguna actividad laboral.

3.3 Hipótesis

H1. SAMPA es el mejor alfabeto fonético para realizar la transcripión ortográfica-fonética ya que es comprensible para la PC, su simplicidad en la transcripción y la relativa facilidad de uso por parte de personas con poca formación fonética a diferencia de AFI.

H2. Las 32 reglas de transcripción son las que se utilizan para la conversión de grafemas a fonemas aplicadas en cierto orden darán una transcripción correcta. 32 por cada grafema Tabla 17.

H3. El silabeo proporcionará una mayor naturalidad en el habla sintetizada de cualquier español debido a la coarticulación de las sílabas.

H4. Los mejores parámetros para dar expresión a la emoción se ven reflejados en la duración y el pitch del fonema y la pausa entre las sílabas.

H5. MBROLA es el sintetizador más adecuado para trabajar el habla ya que cuenta con las voces necesarias de Español México y España para la investigación.

H6. Las voces con emoción obtenidas a través del TTS-EE serán comparables con las emociones de Emofilt.

3.4 Estructura general

El proyecto Sistema de texto-a-habla expresivo en español o TTS-EE se resume básicamente en el diagrama mostrado en la *Figura 7.* El cual se irá describiendo a continuación, cabe mencionar que este diagrama está basado en un trabajo de investigación similar para el idioma árabe [28].

El TTS-EE inicia con un **texto pre-procesado en español de México** que parte de la idea de que la entrada de texto a convertir o transcribir se presenta sin problemas de siglas, palabras extranjeras, abreviaturas, números, fechas, horas, símbolos, etcétera, este proyecto no profundiza en el tema del pre-procesamiento el cual inclusive puede llegar a ser un tema de investigación aparte.

Una vez teniendo este texto depurado se puede proceder a realizar la **transcripción ortográfica fonética** al español principalmente de México aunque también se realiza para el de España o castellano. Como es un sistema computacional se hace uso del alfabeto fonético SAMPA, para aplicar ciertas reglas de transcripción de grafemas a fonemas del idioma español, también se verifica la acentuación del texto.

Se selecciona el alfabeto fonético SAMPA porque los símbolos son comprensibles para la computadora y es más eficiente trabajar con los fonemas que contiene.

Se aplica un silabeo que cuenta con ciertas reglas y se utiliza esta técnica ya que la estructura de las palabras en el idioma español está conformada por división de sílabas que son conjuntos de sonidos que pueden ser pronunciados en una sola emisión de voz. La transcripción podría ser solo a nivel fonema pero al unir dos o más fonemas es importante la coarticulación entre ellos para una mejor calidad de voz.

Se dispuso de una **base de datos de audios en español** de México y de España la cual es necesaria para obtener datos como duraciones y pitch de cada uno de los fonemas tanto sonoros como sordos, a través de la aplicación de una **segmentación automática** con el programa Praat basado en **htk**, y se realizó un ajuste manual de la segmentación con el mismo programa.

Los audios están grabados en diferentes emociones como alegría, tristeza y una neutra como punto de referencia, estas emociones se incorporan al habla sintetizada a través de un archivo de formato .pho el cual es compatible con el sintetizador MBROLA, lo que favorece la experimentación para emular un habla más agradable [29].

El sistema de texto-a-habla expresivo en español es desarrollado en el lenguaje de programación python, se utilizó pySripter como IDE y wxFormBuilder como GUI.

A continuación se describe cómo está constituida la estructura general del ***Sistema de Texto a Habla Expresivo en Español*** o ***TTS-EE*** que se muestra en la *Figura 7*, así como los pasos que se siguieron para su realización, problemas y soluciones.

Figura 7. Sistema de texto a habla expresivo en español.

Paso 1. Texto pre-procesado en español

Inicialmente para el TTS-EE se trabaja con un texto pre-procesado para realizar la transcripción ortográfica-fonética, se utiliza un texto simple, palabras o una frase como entrada. El pre-procesamiento se refiere a la señalización del énfasis o acentuación de cada palabra, se diferencia una vocal acentuada transformándola a una vocal mayúscula, por ejemplo la palabra *aquí* se convertiría en *aqUI, hola* en *hOla,* etc. Se puede notar que no importa si es acento escrito o prosódico ya que ambos se toman en cuenta y se transforman a través de las reglas generales de acentuación. En el apartado *Transcripción ortográfica fonética* sección: *Las reglas de acentuación* se pueden ver más a detalle las reglas aplicadas.

El tratamiento de los signos de puntuación como lo son los signos de interrogación, exclamación, comas, punto, punto o coma; siglas, palabras extranjeras, abreviaturas, números, fechas, horas, símbolos es parte del pre-procesamiento del texto; en este trabajo por su complejidad no se toca a profundidad. La complejidad es como tratar cada símbolo o signo de puntuación, como transcribirlo o el como aplicarle más tiempo en la oración si es por ejemplo una coma.

Esta parte está programada en el módulo *acentuador100713.py* del TTS-EE.

Paso 2. Transcripción ortográfica-fonética

Este paso se refiere a cómo se convierte ese texto pre-procesado a su forma fonética. Es decir la conversión de un grafema a un fonema, un grafema es la unidad mínima en un lenguaje, es decir la grafía o las letras del abecedario, de un sonido. Por otro lado un fono es cualquiera de las realizaciones posibles de un fonema, y éste es la abstracción (imagen mental) de los sonidos del habla humana. Por otra parte un sonido es un fenómeno producido al vibrar las cuerdas vocales del aparato fonador[2].

Teniendo esto en mente se puede dar un primer ejemplo de cómo queda esta transcripción ortográfica-fonética: las palabras *año viejo* se transcribirían a *AJo biExo,* de correspondencia 1 a 1, de grafema a fonema, es decir a -> /A/, ñ ->/J/, o -> /o/, v ->/b/, i ->/i/, e ->/E/, j ->/x/ y o ->/o /. Se puede notar que el fonema se representa con dos

[2] http://www.gramaticas.net/2011/05/fonemas-definicion-y-ejemplos.htm

diagonales (//). Pero, ¿de dónde se obtienen estos fonemas?, se obtienen del alfabeto fonético SAMPA que contiene una lista de cada uno de los fonemas del español, así como de otros idiomas, en el apartado *Transcripción ortográfica fonética* sección *SAMPA* se puede ver a detalle esta lista.

Ahora conociendo la representación de los grafemas a fonemas, se aplican las reglas de transcripción que dicen cómo se va a realizar la conversión, ya que no solo es correspondencia 1 a 1 cómo i->/i/, esa es la parte más básica, pero hay correspondencias más complejas por ejemplo la palabra *cinco* se transcribiría a *Tinko,* ya la letra *c* no corresponde al fonema ficticio */c/* por ejemplificar, sino al fonema /T/ y también al fonema /k/ y en otros casos al fonema /s/.

El grafema x tiene varias realizaciones de fonemas /ks/, /gs/, /x/ y /s/ que es parte importante del español de México y que se pueden reflejar en palabras de origen indígena como lo son Xohimilco, xoconostle y algunas otras como éxito y léxico lo cuál representa una aportación en la creación del TTS-EE.

Así es que existen varios grafemas con diferentes representaciones fonéticas es por eso que son de suma importancia las reglas de transcripción para realizar una conversión adecuada y precisa, es aquí donde se obtiene la representación fonética de la entrada de texto. Esta es una parte medular del Sistema de Texto a Habla Expresivo en español y que se encuentra programado en el módulo *OrtFonsSyllSp220713.py*

A la par también se realiza una silabificación del texto de entrada, gracias a las reglas de silabificación se puede realizar esta acción, se cuenta con 40 reglas para este propósito, basadas en un trabajo de investigación[10] y un par desarrolladas para las necesidades del este proyecto.

También se programó esta parte y se encuentra plasmada en los módulos *OrtFonsSyllSp220713.py*, *syl2fon.py*, *reglas100713.py* y *silprog230713.py,* todos estos programas también se describirán en mayor medida más adelante.

Paso 3. Base de datos español México

La base de español de España sirve de referencia para grabar el corpus o base de español de México con los textos, frases, palabras y números que conforman. Se cuidaron los aspectos de grabación y hablante, esto es fundamental ya que el habla debe representar una emoción de manera eficaz.

La función es proveer datos para alimentar el TTS-EE como duración y pitch de cada fonema y sus realizaciones, así como duraciones de pausas, que son datos importantes para representar cada emoción. Se toman 3 emociones las más básicas que se pueden utilizar, *alegría*, *tristeza* y una *neutral* que en sí ésta última es un habla que carece de emoción. Todo esto se trata a profundida en el tema "La base de datos".

Paso 4. Anotaciones y segmentación automática

A partir de la base de datos del habla en español, para ambas bases se realiza el proceso de anotación prosódica que está relacionada con el alineamiento del sonido y texto, este alineamiento también se conoce como segmentación.

Para este paso se hace uso del programa Praat basado en HTK y del complemento EasyAlign, estas dos herramientas ayudan a realizar la segmentación automática de manera más práctica en una primera instancia, la segmentación consiste en que una sección del texto corresponda a una parte del audio, la división es fonemas, sílabas, palabras e incluso la frase completa.

Obviamente el interés está en los fonemas y en algunos casos las sílabas, al tener la segmentación hay que analizar y escuchar que realmente el fonema representado en texto y en sonido por una señal de onda correspondan entre sí, al no corresponder se realizó una minuciosa alineación manual.

Aquí a través de algunos programas de Praat es donde realmente se obtienen los datos de duración y pitch o frecuencia fundamental (F0). También en este trabajo se tiene el capítulo correspondiente a *La segmentación* donde se describe ampliamente como se realizó la segmentación de la base de datos.

Paso 5. Asignación de la duración y el pitch

En el Sistema de Texto a Habla Expresivo en Español o TTS-EE existe un módulo llamado *sampapho.py*, el cuál manda llamar archivos .txt que contienen los datos estadísticos de duraciones y pitch para cada fonema, arrojados de la segmentación automática y manual, también se pueden apreciar datos de las pausas que existen en cada palabra y entre palabras, estas pausas solo tienen duración y no pitch debido a que no hay excitación alguna en las cuerdas vocales.

Algunas dificultades son en cuanto a fonemas correspondientes a las vocales, ya que en ellas se representa el énfasis o acentuación, y se tienen valores diferentes en duración y pitch para un mismo fonema entonces se tiene que reconocer esos fonemas con énfasis para asignarles su duración y pitch adecuada.

También las duraciones de las pausas son variadas dependiendo de qué fonema las precede y sucede, hay que hacer también un reconocimiento de estos fonemas para asignar la duración precisa a cada pausa.

Es importante mencionar que la duración corresponde tanto a fonemas sordos y sonoros y el pitch solo a fonemas sonoros, la sonoridad sucede a partir de la vibración de las cuerdas vocales.

Paso 6. Voz con emoción

Con todo este procedimiento se hace un acercamiento a un habla con emoción y de dos tipos de español, de México y de España. Finalmente se pueden generar varios audios de formato .wav y .pho este último compatible con el sintetizador MBROLA, con alguna emoción elegida como alegría, tristeza y neutral. Y con un par de voces en español de México y otro par de España.

Aquí en esta sección están involucrados los programas Python y PyScripter para la programación del TTS-EE, MBROLA para obtener las voces de los dos tipos de español las cuales están basadas en el alfabeto fonético SAMPA, el programa eSpeak para un par de ejemplos más de voces en español, wxPython y wxFormBuilder para la realización de la interfaz gráfica del sistema y que se representan en los programas *SpGuiNewMiVent.py* y *Spguinew.py*

También existe una sección dónde se describe el sintetizador *MBROLA* y el *TTS-EE* con mayor precisión, más adelante.

4. La base de datos

4.1 Base de datos del español de España, características del corpus.

a) Información general

El título de la base de datos en español de España utilizada es **Base de datos de grabaciones de síntesis del habla emocional.**

Los autores de esta base son: Catherine Tchong, Jacques Toen, Zdravko Kacic, Asunción Moreno y Albino Nogueiras.

El status de esta base de datos es público. La base contiene varios archivos como *INDEX_SP\contents.lst,* el cual contiene información sobre los nombres de archivo y el contenido de cada archivo en la base de datos.

El archivo *INDEX_SP\lexicon.tbl* contiene el vocabulario y la transcripción fonética SAMPA de cada palabra en la base.

El archivo *TABLE_SP\script.txt* contiene el indicador del hablante que aparece durante las grabaciones.

Existen varias nomenclaturas del archivo que a continuación se describen en la *Tabla 18*.

LL	Dos letras de código español sp = español.
S	Estilo emocional T = Tristeza, J = Alegría, N = Neutral/Normal.
n	Sesión de grabación (1 o 2)
X	Género del hablante (M o F)
yyy	Identificador del elemento (001 a 175)
Sf	Frecuencia de muestreo (16 = 16kHz)

Tabla 18. Nomenclatura del archivo.

En cuanto a la transcripción de los archivos todas las transcripciones pueden ser encontradas en el archivo *INDEX_SP\script.txt.*

b) *Condiciones de grabación*

Las grabaciones fueron hechas usando un micrófono AKG 320. Las señales del habla se grabaron a 16Khz. El ambiente de grabación se realizó en condiciones de calma. La base de datos española fue grabada en una habitación en silencio con una pared de cristal que divide la sala en dos partes. El hablante lee las frases que aparecen directamente desde la PC. Para evitar ruido, la pantalla, PC y el sistema de grabación, se colocaron en un lado de la habitación y el hablante en silencio en el otro lado.

Dos operadores supervisaron las grabaciones en el momento de la grabación, uno de ellos comprueba los enunciados para que correspondan exactamente con el texto para ser leído. El otro operador comprueba el sistema de grabación.

c) Contenido del corpus

El corpus contiene 184 oraciones diferentes. Incluye números, palabras, oraciones afirmativas, exclamaciones, interrogativas y párrafos, como se muestra en la *Tabla 19*.

Identificador de elementos	Contenido de corpus
1 a 100	Oraciones afirmativas incluyendo largas y cortas
101 a 134	Interrogativas y (5) oraciones con énfasis.
135 a 150	Párrafos
151 a 160	Dígitos
161 a 184	Palabras aisladas

Tabla 19. Contenido del corpus.

La lista completa de palabras es dada abajo en la *Tabla 20*:

Cero	No	Basta ya
Uno	Sí	Arriba
Dos	Ayer	Gracias
Tres	Hola	Por favor
Cuatro	Hoy	De nada
Cinco	Abajo	Izquierda
Seis	Fuera	Derecha
Siete	Adiós	Dentro
Ocho	Nunca	Pronto
Nueve	Lento	Rápido
	Tarde	Detrás
	Antes	Delante

Tabla 20. Lista de dígitos y palabras aisladas para la base de datos en español España.

La base de datos arroja también un recuento de la frecuencia de los fonemas en el corpus español que se muestra en la siguiente *Tabla 21.*

SAMPA	Conteo	Porcentaje
g	13	0,15
tS	22	0,25
J	26	0,30
L	33	0,38
b	34	0,39
jj	39	0,45
N	43	0,50
x	62	0,71
rr	68	0,78
f	80	0,92
G	92	1,06
d	93	1,07
z	111	1,28
w	127	1,46
T	164	1,89
B	190	2,19
p	216	2,49
j	228	2,63
u	239	2,75
m	299	3,44

k	317	3,65
D	364	4,19
i	371	4,27
t	404	4,65
l	428	4,93
r	465	5,35
s	516	5,94
n	528	6,08
o	816	9,40
e	1115	12,84
a	1181	13,60
Total	8684	100

Tabla 21. Recuento de la frecuencia de los fonemas en el corpus español España.

El corpus fue grabado dos veces en MPEG en las emociones T=tristeza, J=Alegría y N= Neutral/Normal.

Los contenidos de los archivos del habla fueron supervisados en el momento de la grabación. La mala pronunciación u otra desviación del script fueron detectadas. No se expresan la mala pronunciación.

Los contenidos y transcripciones de la base de datos completa están incluidos en el directorio INDEX.

4.2 Base de datos del español de México, características del corpus.

a) La base de datos que fue utilizada es la de grabaciones para síntesis de habla emocional, en idioma español de México, en este caso solo se hace uso de las emociones principales alegría, tristeza y una neutral de un hablante masculino, grabada en una primera sesión como se detalla a continuación *Tabla 22*:

Base de datos de grabaciones para síntesis de habla emocional	
Idioma	Español México
Emociones	Alegría, Tristeza, Neutral
Hablante	Masculino
Sesión	Sesión 1
Tipo de archivos	.ort .wav

Tabla 22. Base de datos de grabaciones para síntesis del habla emocional.

b) Los tipos de archivos utilizados en esta base de datos son los .ort que contienen oraciones, palabras, números; además los audios originalmente tienen el formato .l16 pero por conveniencia en el manejo de la información se transformaron a .wav.

c) La base de datos se encuentra constituida como muestra la siguiente *Tabla 23*:

Rango	**Contenido del corpus**
1 a 100	Oraciones afirmativas largas y cortas
101 a 134	Preguntas y oraciones acentuadas
135 a 150	Párrafos
151 a 160	Dígitos
161 a 184	Palabras aisladas

Tabla 23. Contenido del corpus español México.

Los textos de esta base de datos son seleccionados haciendo referencia a la base de datos de español de España que esta previamente grabada.

5. Sistema de Texto a Habla Expresivo en español (TTS-EE)

5.1 Descripción del Sistema de Texto a habla expresivo en español (TTS-EE).

El TTS-EE está desarrollado en el lenguaje de programación Python, a través del IDE PyScripter, cuenta con 11 módulos que son desarrollados a continuación.

5.1.1 Módulo acentuador (acentuador100713.py)

El módulo acentuador aplica todas las reglas de acentuación al texto ingresado, su finalidad es sustituir el acento prosódico u ortográfico por mayúsculas, así se identifica que una vocal tiene énfasis, este módulo tiene varias funciones como las que se describen a continuación.

La función **esvocal** verifica si la letra que va pasando es una vocal.

La función **numvocales** cuenta el número de vocales que no lleva acento.

La función **yaLlevacento** devuelve un 1 si ya está acentuada, eso se refiere que si la entrada es una letra ‘áéíóú’ el programa las convierte en ‘AEIOU’ y devuelve un cero (0) si no se realiza la conversión es decir que no está acentuada la vocal.

La función **acento_pos** devuelve la vocal convertida a mayúscula dependiendo si la palabra es:

1) Aguda, que no acabe en n, s o vocal.
2) Llana, que acabe en consonante menos n, s.

La función **acentua_rec** aplica el acento tónico a las monosílabas ya que todas estas se acentúan. Se encarga también de acentuar las palabras terminadas en MENTE, este tipo

de palabras tienen 2 puntos de acentuación, el acento de la propia palabra y el del sufijo mEnte acentuado también en la primera E.

Por otro lado la función **pon_cosas** elimina la letra u, la cual no genera ningún sonido, para simplificar la palabra y manejarla más fácilmente.

que	Qe
qui	Qi
gue	Ge
gui	Gi

La función **quita_cosas** regresa las palabras a su forma original.

Q	qu
G	gu

Finalmente la función **acentua** es la encargada de llamar a todas las funciones anteriormente mencionadas para que el módulo acentuador aplique al texto entrante.

5.1.2 Módulo cambia acentos y notación (*cambioacentos110713.py*)

Este módulo cambiará los acentos de las vocales por mayúsculas de una línea de texto con formato unicode.

Iniciando con la función **cambioacevow** la cual cambia todas las vocales ya estén acentuadas o no, por su notación que es la letra 'V', lo que significa que es una vocal.

u'á' ->'V'

La función **cambiocon** cambia todas las consonantes por su notación, la letra 'C' que significa precisamente eso consonante.

u'b' -> 'C'

La función **elimacens** se encarga de cambiar las vocales acentuadas por sus correspondientes letras capitales, y también sustituye la ü o Ü con diéresis por M y la ñ o Ñ por una N para no tener ningún problema con esos caracteres.

u'á' ->'A'
u'Ü' ->'M'
u'ñ' ->'N'

Función **regresaacent** devuelve los acentos a las vocales, transforma la nomenclatura que corresponde a la diéresis y a la letra ñ, para que se puedan imprimir en pantalla.

A -> uá
E -> ué
I -> uí
O -> uó
U -> uú
M -> uü
N -> uñ

5.1.2 Módulo crea y escucha un archivo .wav (mbrolamod.py)

Este módulo permite crear y escuchar un archivo .wav que se genera a partir de un archivo de tipo .pho compatible con el sintetizador Mbrola.

Para este módulo se crean dos carpetas una temporal para alojar a los archivos que se van escuchando y una carpeta dónde se depositaran los archivos .wav y .pho creados.

En el caso que se requiera escuchar el texto deseado se requieren las voces de Mbrola correspondientes al español de México y de España. En la ventana principal al seleccionar el idioma los encontramos de la siguiente forma el idioma 0 es para la voz MX1, el 1 para MX2, el puesto 2 para ES1, el 3 para ES2 y finalmente el 4 para ES4.

Para la función **crear_wav** se elige alguna voz en español del 0 al 4 y el ejecutable de Mbrola.

Y con la función **escuchalo** permite escuchar el audio en un tiempo determinado.

5.1.3 Módulo Transcripción Ortográfica-Fonética (OrtFonsSyllSp220713.py)

Este Módulo se basa en el reporte de Castro y artículos de Llisterri [35]. Es una parte fundamental debido a que en este se realiza la transcripción ortográfica-fonética aplicando las reglas de transcripción.

La función ***findfons*** busca la posición de la cadena *bucastr* que está contenida en la palabra o línea. Los valores devueltos representan la posición de str en sils una línea que

tiene espacios al principio y al final, así las palabras que contengan una ‘u’ no se verán afectadas como por ejemplo ‘que’ o ‘gue’.

El módulo transcriptor o ***trans_ort2phon*** realiza la transformación de los grafemas a fonemas

5.1.3.1 El caso del grafema X

La letra X es muy importante para el español de México, se tienen muchas palabras de origen indígena que tienen sonidos específicos, por ejemplo en el caso de xoconostle sonará el grafema x como el fonema /s/ por situarse al principio de la palabra, no así con la palabra Ximena donde también se localiza el grafema x al principio de palabra pero el sonido es como una jota y se representa fonéticamente como /x/, para la palabra México el fonema /x/ es intervocálico es decir está situado entre vocales, para la x igualmente suena como jota y se simboliza /x/. Pero por otro lado la palabra éxito, también el grafema x es intervocálico y no suena como jota más bien suena como /ks/ o en algunos casos dependiendo de la literatura como /gs/. Es por eso que en los casos dónde no se aplica una regla en específico se creó una lista de palabras con x que básicamente suenan como jota y quedaran con el fonema /x/, el resto de las palabras con x se les aplican las reglas de transcripción establecidas [31].

En el apartado de ***Reglas de transcripción*** se puede ver con mayor detalle las reglas aplicadas en este módulo.

5.1.4 Módulo reglas de silabificación (reglas100713.py)

El módulo de las reglas de silabificación se divide en tres segmentos primero para las palabras que inicien en vocal y se identifican por la nomenclatura ‘V’**.**

Para esto se utiliza la función **segmentV** y se aplican las siguientes reglas.

Donde la V es igual a vocal, C es igual a consonante las letras minúsculas son los grafemas mismos, istrip es triptongo y isdip es diptongo y ND es vocal fuerte + vocal fuerte SA acentuada en vocal fuerte y E acento en e.

1	VhVV && istrip -> VhVV
2	VhV && isdip -> VhV
3	VE, VCV, V[CC]V, V(WA) -> V

4	VCE, VCCV, VC[CC]V -> VC
5	VCCE, VCCCV, VCC[CC]V -> VCC
6	(VVE, VVCV, VV[CC]V) && ND -> V
7	(VVE, VVCV, VV[CC]V) && Not ND -> VV
8	(VVCE, VVCCV, VVC[CC]V) && ND -> V
9	(VVCE, VVCCV, VVC[CC]V) && Not ND -> VVC

Luego para las palabras que inicien con 'CV' se tiene la función **segmentCV** y se aplican otras reglas.

10	CVhVV && istrip -> CVhVV
11	CVhV && isdip -> CVhV
12	CVE, CVCV, CV[CC]V -> CV
13	CVCE, CVCCV, CVC[CC]V -> CVC
14	CVCCE, CVCCCV, CVCC[CC]V -> CVCC
15	(CVVE, CVVCV, CVV[CC]V) && ND -> CV
16	(CVVE, CVVCV, CVV[CC]V) && Not ND -> CVV
17	(CVVCE, CVVCCV, CVVC[CC]V) && ND -> CV
18	(CVVCE, CVVCCV, CVVC[CC]V) && Not ND -> CVVC
19	(CVVVE, CVVVCV, CVVV[CC]V) && istrip -> CVVV
20	(CVVVE, CVVVCV, CVVV[CC]V) && gui(SA) ->
21	CVVVE && gui -> CVVV
22	(CVVVCV, CVVV[CC]V) && gui -> CVVVCV
23	CVVVE, CVVVCV, CVVV[CC]V -> CV
24	(CVVVCE, CVVVCCV, CVVVC[CC]V) && istrip -> CVVVC
25	CVVVCE && gui(SA) -> CVV
26	(CVVVCE, CVVVCCV, CVVVC[CC]V) && gui -> CVVVC
27	CVVVCE, CVVVCCV, CVVVC[CC]V -> CV
28	CVVCsE -> CV

Y finalmente para las palabras que inicien con CC se encuentra la función **segmentCC** y las siguientes reglas.

29	CCVhVV && istrip -> CCVhVV
30	CCVhVCE && isdip -> CCVhVC
31	CCVhVC && isdip -> CCVhV
32	CCVE, CCVCV, CCV[CC]V -> CCV
33	CCVCE, CCVCCV, CCVC[CC]V -> CCVC

34	CCVCCE, CCVCCCV, CCVCC[CC]V -> CCVCC
35	(CCVVE, CCVVCV, CCVV[CC]V) && ND -> CCV
36	(CCVVE, CCVVCV, CCVV[CC]V) && Not ND -> CCVV
37	(CCVVCE, CCVVCCV, CCVVC[CC]V) && ND -> CCV
38	(CCVVCE, CCVVCCV, CCVVC[CC]V) && Not ND -> CCVVC
39	CCVVVC && istrip -> CCVVVC
40	CCVVVC -> CCV

Cabe señalar que una vez que se divide la palabra en silabas el resto de las sílabas de la misma palabra puede caer en cualquiera de los tres segmentos.

También se tienen otras funciones como la función que identifica si es un trinomio es **istrip**.

La función para reconocer que una palabra tiene un diptongo es **isdip**

La función **irred** encuentra los pares irreducibles en este caso: br bl, cr, cl, dr, fr, fl, gr, gl, kr, ll, pr, pl, tr, rr, ch, tl.

Y finalmente la función **nd** identifica si la palabra no tiene diptongo.

5.1.5 Módulo crea archivo .pho (sampapho.py)

En este módulo a partir de la transcripción SAMPA se va generar un archivo compatible con Mbrola, con la estructura que marca este sintetizador para poder ser leído.

Primero se tiene la función **langescribe_pho** aquí dependiendo del tipo de español ya sea de México o de España se irán haciendo los cambios correspondientes en algunos fonemas por ejemplo el fonemas /dZ/ se aplica en el español México y el fonema /L/ se aplica para el español de España, esto corresponde al grafema ll.

Posteriormente se tiene la función **escribe_pho** que es una de las funciones medulares del TTS-EE aquí es donde se obtienen los datos de duración y pitch de cada uno de los fonemas, estos datos están alojados en archivos de texto, que se mandan a llamar para llenar el archivo .pho con toda la información necesaria. Archivos que están clasificados por tipo de español y por emoción como lo son alegría tristeza y neutral.

España-neutral / esneutral.txt
España-alegría / esalegria.txt

España-tristeza/ estristeza.txt
México-neutral / mxneutral.txt
México-alegría/ mxalegría.txt
México-tristeza/ mxtristeza.txt

También en ésta función se han agregado apartados del promedio de la duración de las pausas, que también dan énfasis en si es una emoción de alegría, tristeza o neutral y también están divididos por tipo de español. Las pausas pueden ser entre palabras o suceder en la misma palabra entre las sílabas.

5.1.6 Módulo silabificación (silprog230713.py)

Este módulo recibe una lista de palabras y devuelve su silabificación, su función **silabeo** llama a **tamsegmento** que este a su vez aplica las reglas de silabificación dependiendo si inicia la palabra de la forma 'V', 'CV' o 'CC'. Aquí es donde también se manda a llamar a **trans_ort2phon** para realizar la transcripción ortográfica fonética.

5.1.7 Módulo funciones de la interfaz (SpGuiNewMiVent.py)

En éste módulo se utiliza la clase del mismo nombre **SpGuiMiVent** que contienen las funciones siguientes:

OnConver esta función es para el botón ***Transcripción*** dependiendo del tipo de voz elegida se realiza la transcripción ortográfica-fonética. Como se puede ver en la figura 8 siguiente.

Figura 8. TTS-EE Descripción de las funciones de la interfaz.

La función **OnGuardar** guarda el audio de la oración trascrita dependiendo de la voz elegida en un archivo .wav o .pho los cuales se guardan en la carpeta emowavs.

La función **OnEscuchar** crea un archivo temporal llamado prueba.pho y prueba.wav que al dar click en el botón escuchar se aprecia la oración con emoción.

La función **OnButtonClick** reinicia las operaciones para una nueva conversión.

Con la función **salirOnButtonClick** se sale del sistema.

La función **AyudaOnMenuSelection** se puede ver la ayuda del sistema.

La función **AcercadeOnMenuSelection** muestra la información acerca del sistema.

Y finalmente la función **NomArch** agrega el nombre del archivo dependiendo de qué voz se elige.

5.1.8 Módulo syl2fon.py

Importa todos los módulos, y cuenta con la función **sil2fon** que quita el final de la línea, convierte la línea en minúscula, transforma a V o C las letras de todas las palabras de la línea.

Cambia a capitales todas las palabras acentuadas de la línea acentúa palabra por palabra y elimina la puntuación.

5.1.9 Módulo principal (mainnew.py)

Este es el módulo con el que se inicia el sistema desde PyScripter, se manda a llamar **SpGuiNewMiVent.py** con la función **onInit.**

5.1.10 Módulo Interfaz (Spguinew.py)

El código de este módulo fue creado automáticamente al diseñar la interfaz gráfica del TTS-EE, a través de la aplicación de diseño GUI llamada ***wxFormBuilder.***

5.1.10.1 wxFormBuilder

wxFormBuilder es una aplicación de diseño GUI de código abierto que se utiliza con las herramientas wxWidgets, que permite la creación de aplicaciones multiplataforma. Una interfaz ágil y fácil que permite un desarrollo más rápido y un mantenimiento más rápido del software la cual está desarrollada en el lenguaje de programación C + +.

wxFormBuilder no es solo una herramienta de desarrollo visual, sino que también permite incluir componentes no gráficos. Se pueden emitir programas en C + +, Python, PHP y el código XRC, el cual no se puede editar directamente en el programa, para editar se necesita en este caso el editor PyScripter.

Para la interfaz gráfica tenemos los siguientes elementos:

1. Título de la ventana (***TTS-EE***)

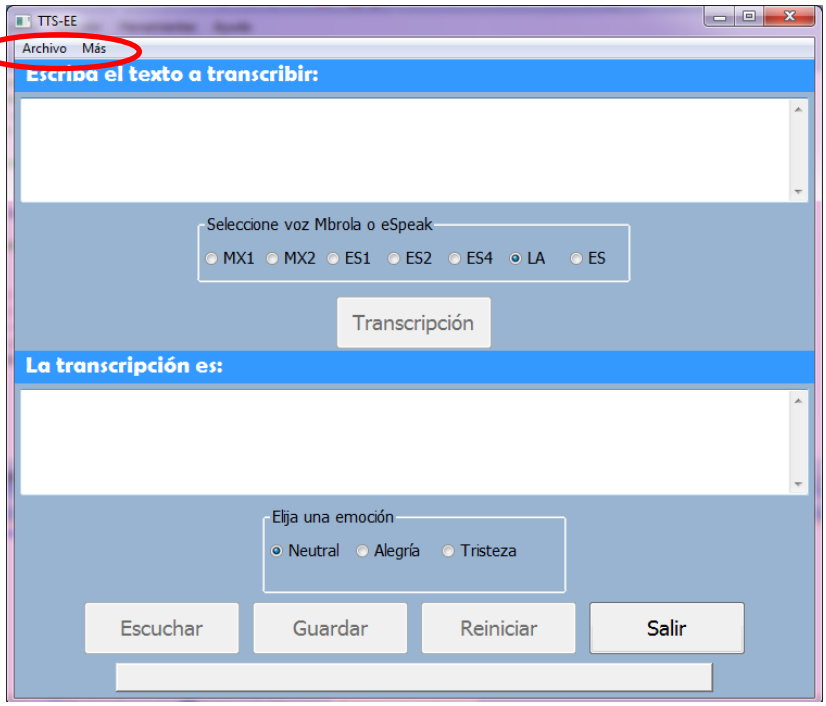


Figura 9. Título de la ventana.

2. El *menú* donde se tiene la opción ***Archivo*** y luego ***Salir***.

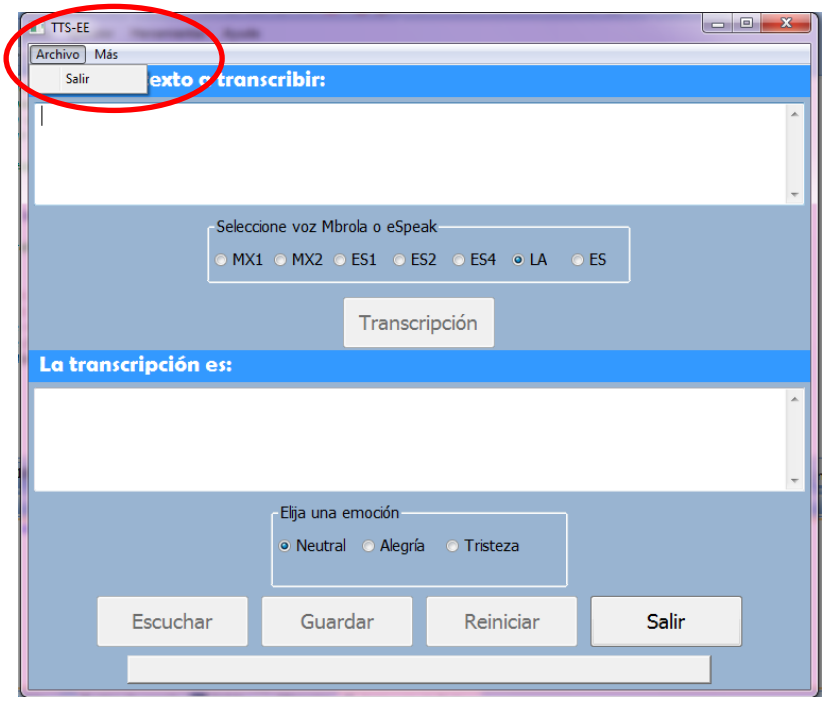


Figura 10. Menú-Archivo-Salir.

3. La opción ***Más, Ayuda*** y ***Acerca de…***

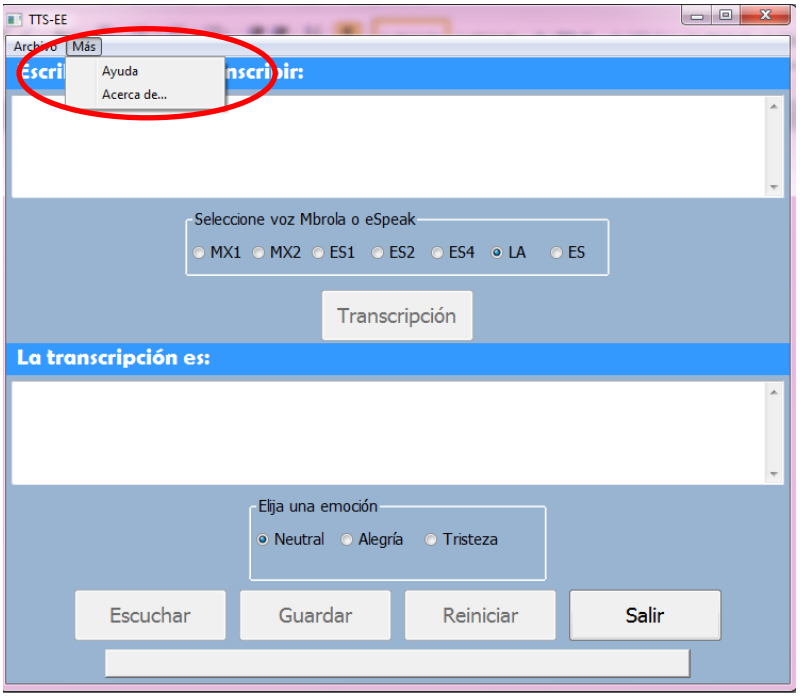


Figura 11. Más-Ayuda-Acerca de...

En ***Ayuda*** se tiene lo siguiente:

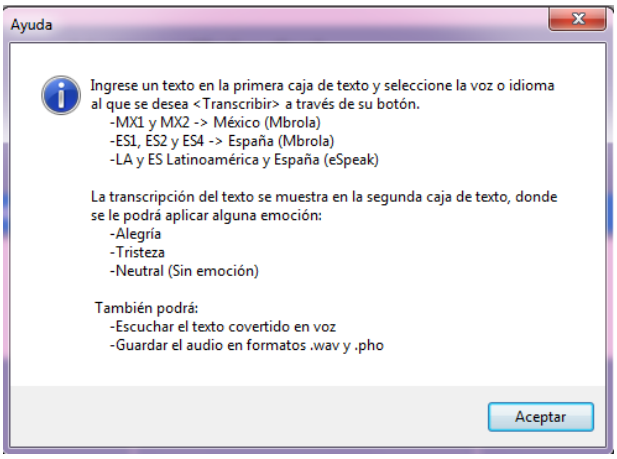


Figura 12. Ayuda.

El apartado ***Acerca de…*** muestra el siguiente mensaje:

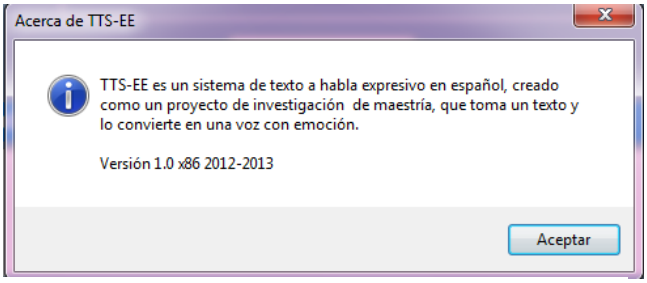


Figura 13. Acerca de...

4. La interfaz muestra dos apartados unos para escribir el texto y el otro para obtener el resultado.

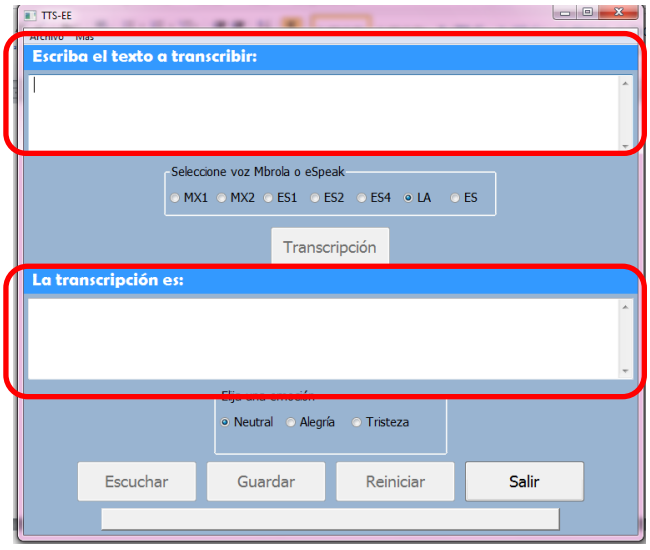


Figura 14. Cuadros de texto.

5. Al escribir el texto a transcribir, se tiene que seleccionar la voz Mbrola o eSpeak y dar click en el botón **Transcripción**.

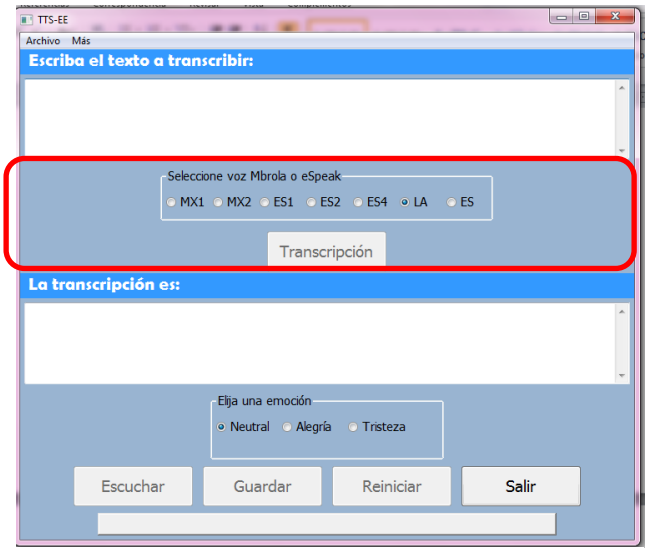


Figura 15. Voces Mbrola-eSpeak y botón Transcripción.

6. También hay un apartado donde se elige alguna de las emociones, Neutral, Alegría y Tristeza.

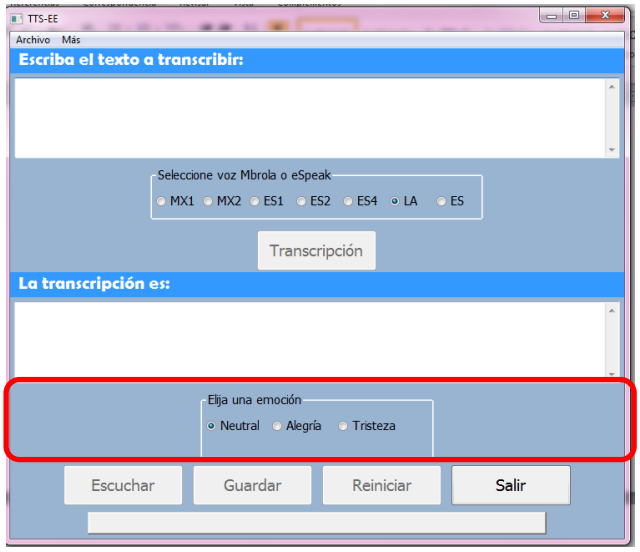


Figura 16. Voces Mbrola-eSpeak y botón Transcripción.

7. En esta figura se muestra la sección de los botones para Escuchar, Guardar, Reiniciar y Salir.

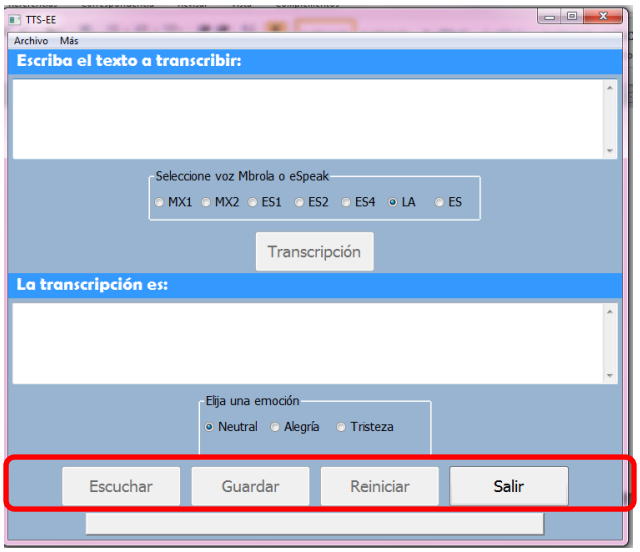


Figura 17. Escuchar, Guardar, Reiniciar y Salir.

8. Y finalmente en esa sección se muestran las acciones que se van realizando.

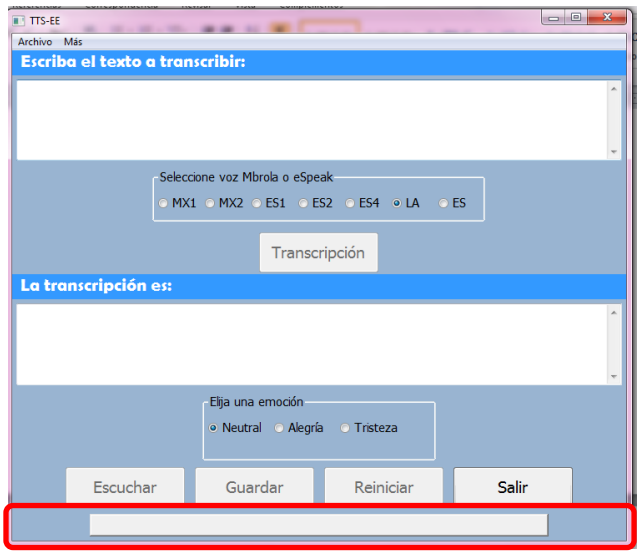


Figura 18. Barra de Estado.

6. Resultados

6.1 Tasa de sílabas español de España.

A través del programa *calculate_tempo.praat* se obtienen datos como son tasa de fonemas y tasa de sílabas.

La *Figura 19* muestra la tasa de sílabas de los audios 1 a 150 que son oraciones largas, cortas, interrogativas y párrafos. Para la emoción neutral se tiene una tasa de alrededor de 6.5 sílabas, para alegría de 7 y para tristeza 5.5 sílabas eso quiere decir que al hablar con emoción, una emoción de tristeza refleja lentitud, la neutral se encuentra en medio y la alegría va más rápido.

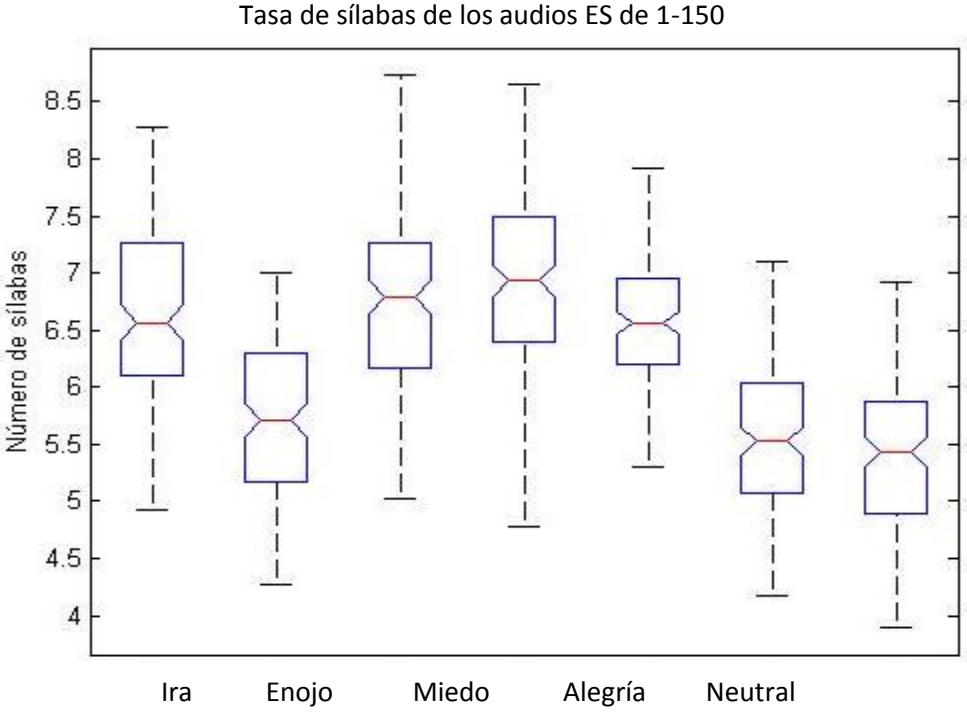


Figura 19. Tasa de sílabas de los audios ES de 1-150.

La *Figura 20* muestra la tasa de sílabas de los audios 1 a 100 que son las oraciones afirmativas largas y cortas. Para la emoción neutral se tiene una tasa de alrededor de 6.5 sílabas, para alegría 6.7 y para tristeza 5.2, donde se observa el mismo comportamiento.

La *Figura 21* muestra la tasa de sílabas de los audios 101 a 134 que son interrogantes y oraciones con énfasis. Para la emoción neutral se tiene una tasa de alrededor de 6.5 sílabas, para alegría 7.6 y para tristeza 5.6.

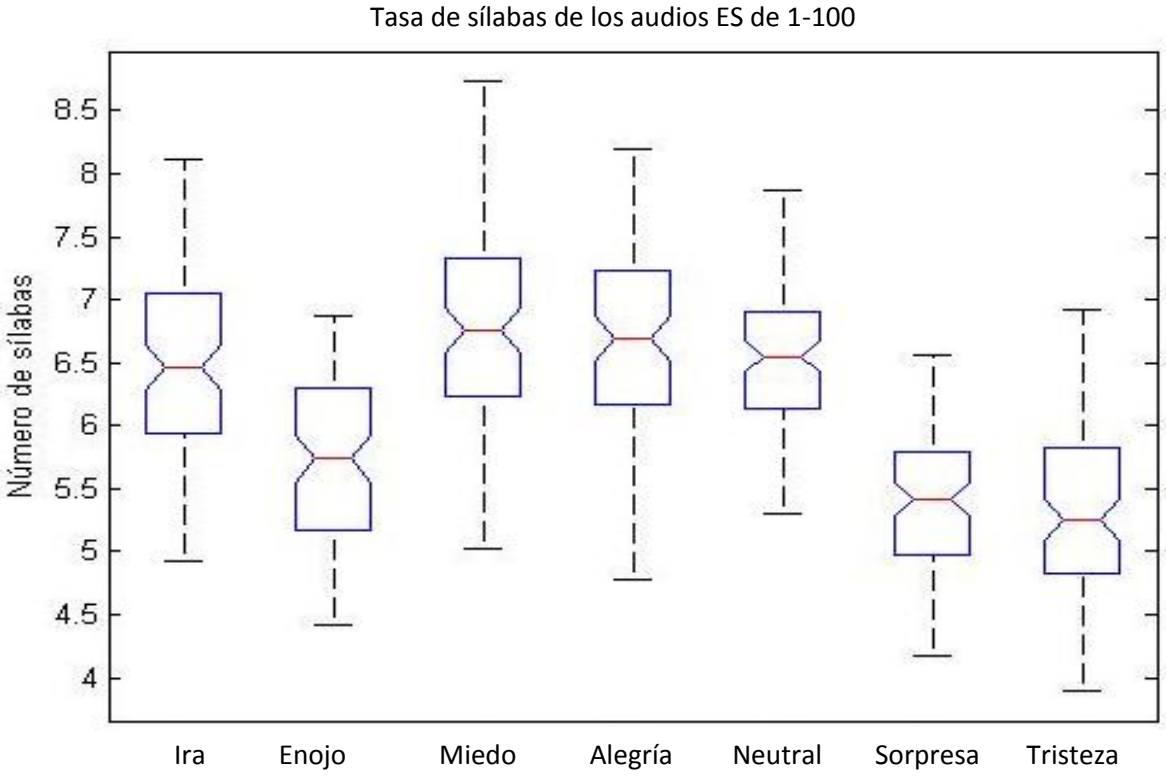


Figura 20. Tasa de sílabas de los audios de 1-100.

La *Figura 22* muestra la tasa de sílabas de los audios 135 a 150 que son párrafos. Para la emoción neutral se tiene una tasa de alrededor de 6.8 sílabas, para alegría 6.9 y para tristeza 5.6.

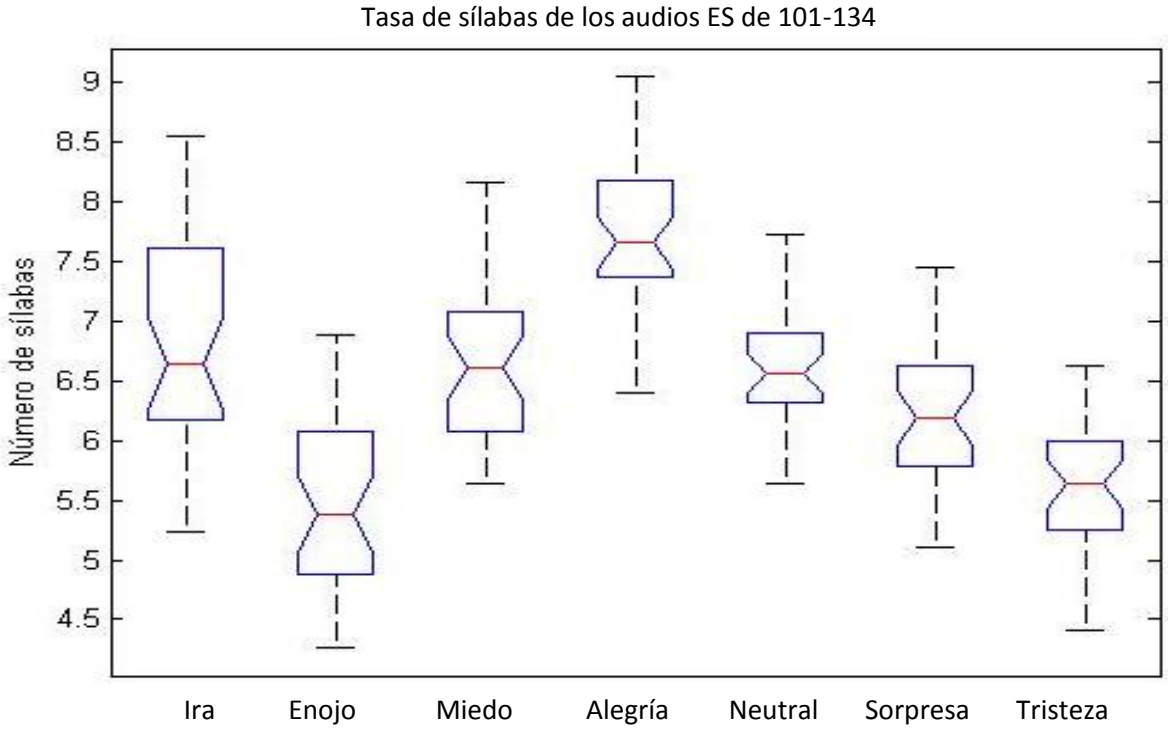


Figura 21. Tasa de sílabas de los audios ES de 101-134.

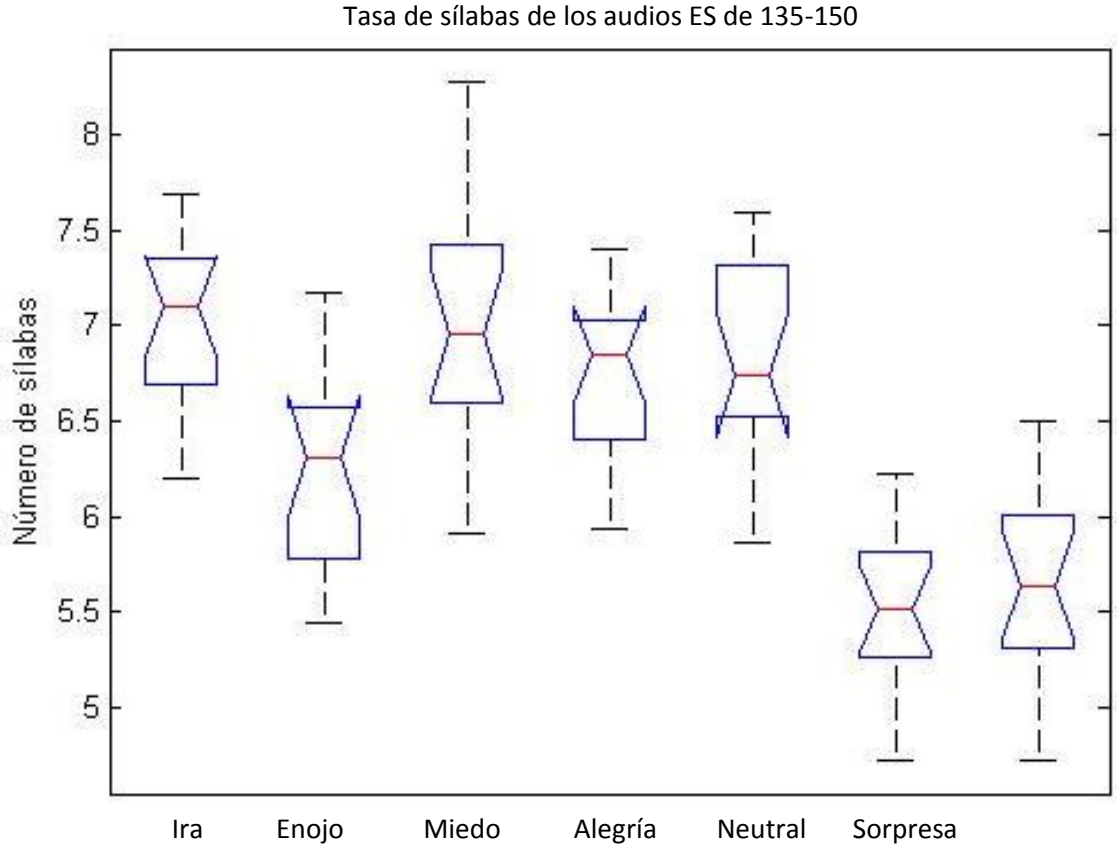


Figura 22. Tasa de sílabas de los audios de ES 135-150.

La *Figura 23* muestra la tasa de sílabas de los audios 151 a 184 que son dígitos y palabras aisladas. Para la emoción neutral se tiene una tasa de alrededor de 5 sílabas, para alegría 4.5 y para tristeza 4.1 sílabas.

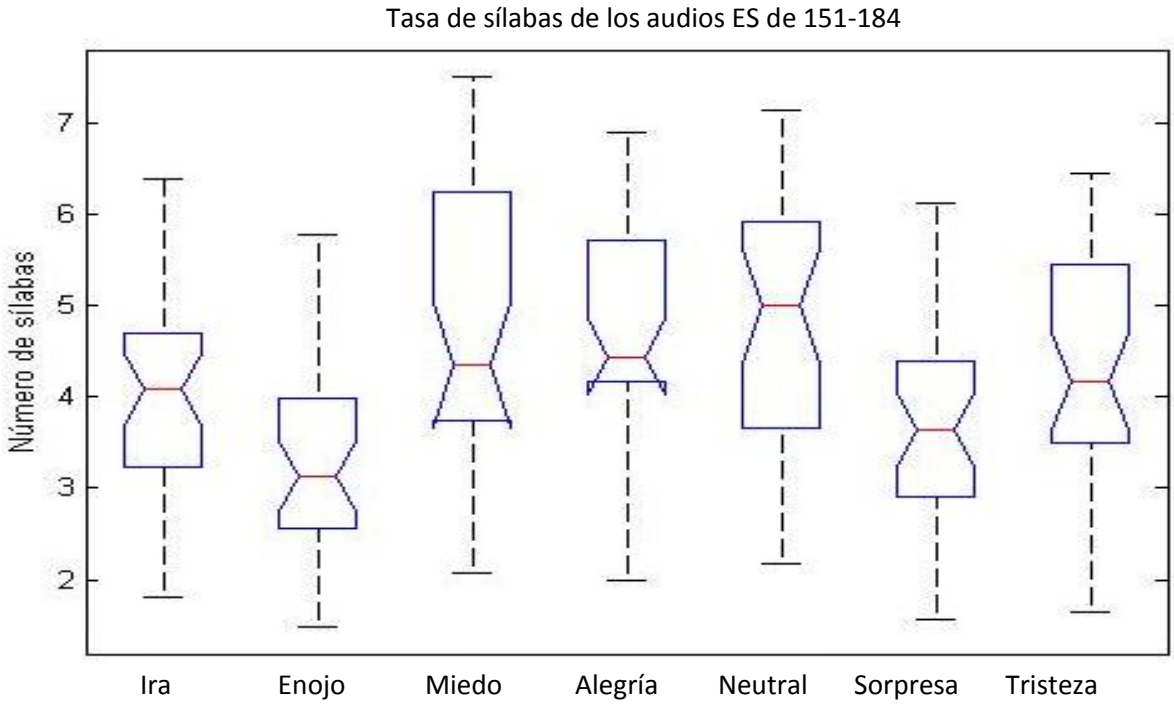


Figura 23. Tasa de sílabas de los audios ES de 151-184.

6.2 Tasa de fonemas español de España.

La *Figura 24* muestra ahora la tasa de fonemas de los audios 1 a 150 que son oraciones largas, cortas, interrogativas y párrafos. Para la emoción neutral se tiene una tasa de alrededor de 15 fonemas, para alegría 16 y para tristeza 12.5 fonemas eso quiere decir que al hablar la emoción de tristeza refleja lentitud, y la alegría rapidez con respecto a la neutral.

La *Figura 25* muestra la tasa de fonemas de los audios 1 a 100 que son las oraciones afirmativas largas y cortas. Para la emoción neutral se tiene una tasa de alrededor de 15 fonemas, para alegría 15.9 y para tristeza 12.1 fonemas.

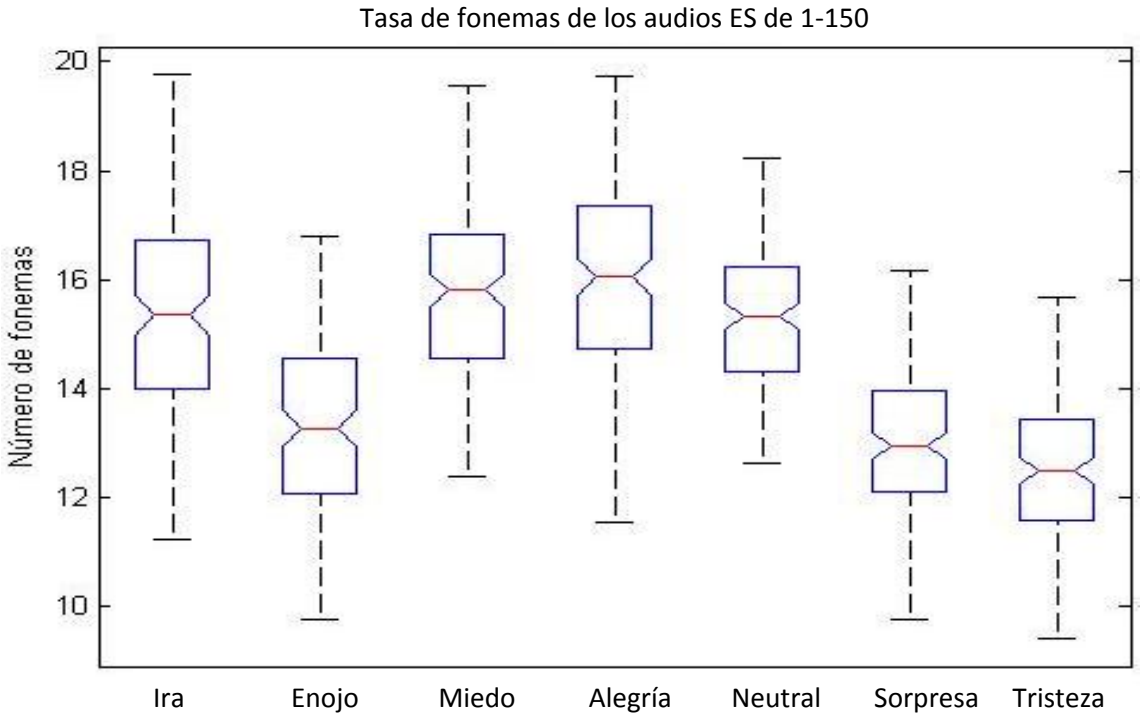


Figura 24. Tasa de fonemas de los audios ES de 1-150

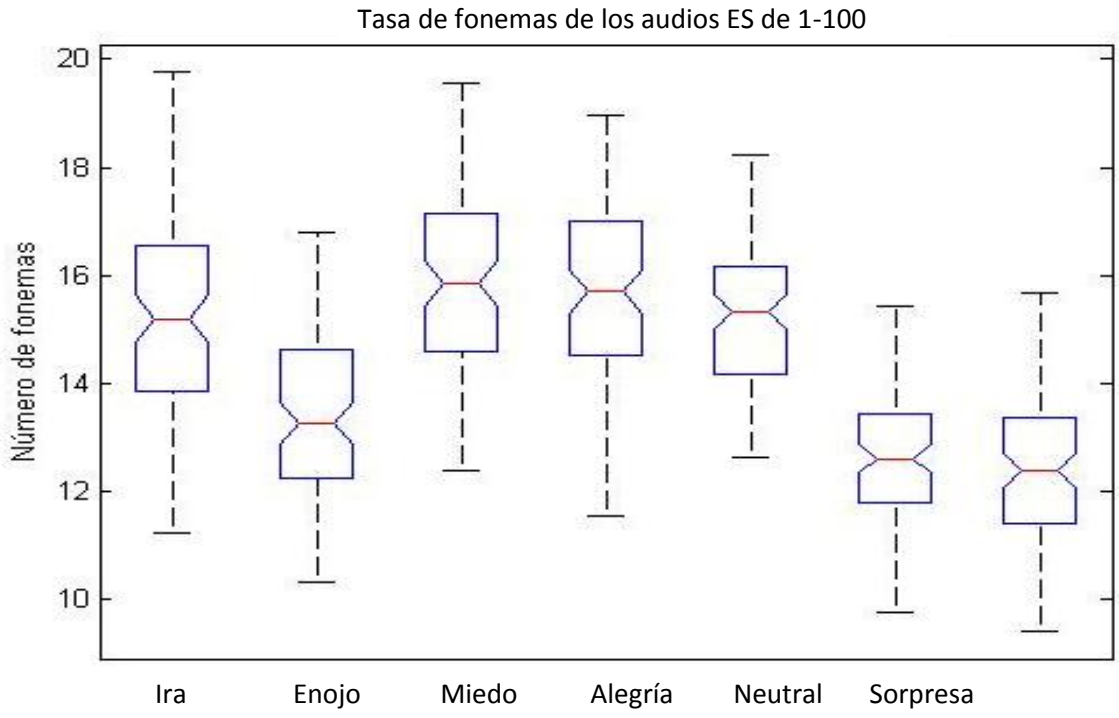


Figura 25. Tasa de fonemas de los audios ES de 1-100.

La *Figura 26* muestra la tasa de fonemas de los audios 101 a 134 que son interrogantes y oraciones con énfasis. Para la emoción neutral se tiene una tasa de alrededor de 15 fonemas, para alegría 18 y para tristeza 12.5.

Tasa de fonemas de los audios ES de 101-134

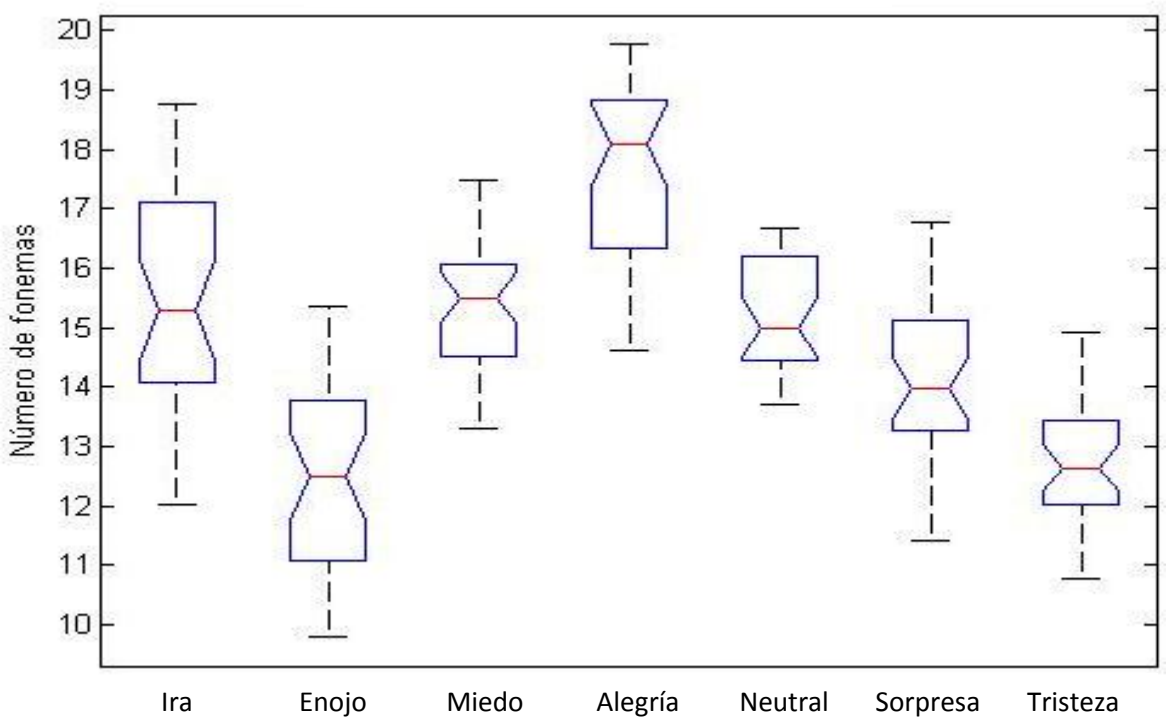


Figura 26. Tasa de fonemas de los audios ES de 101-134

Tasa de fonemas de los audios ES de 135-150

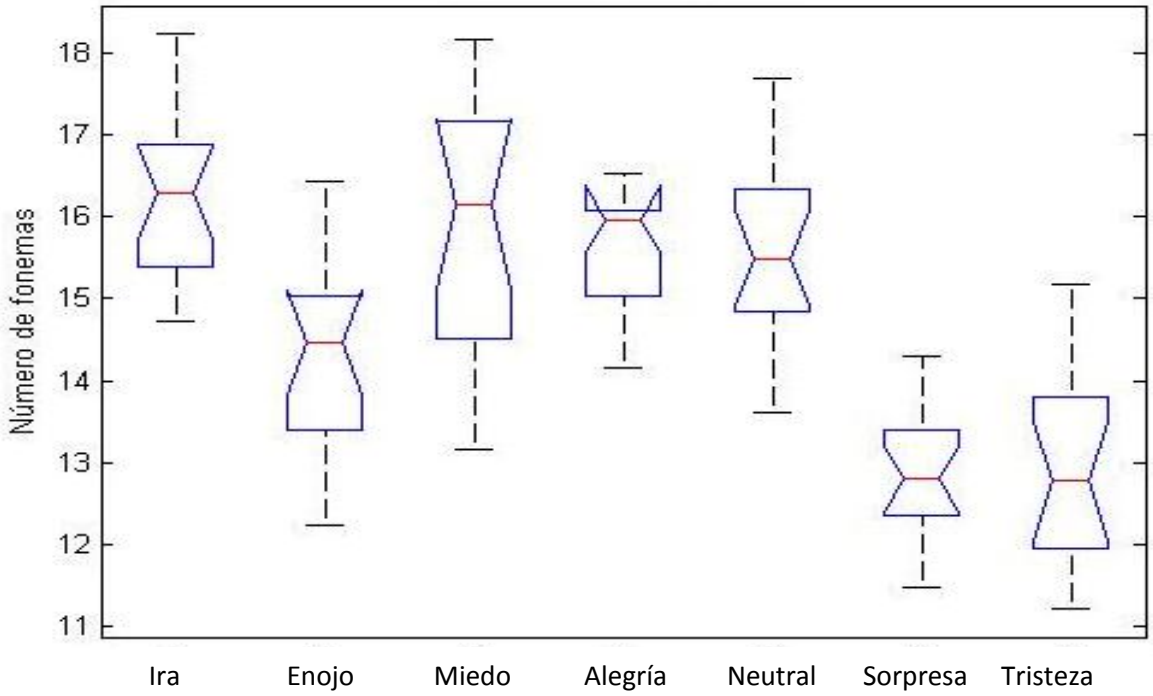


Figura 27. Tasa de fonemas de los audios ES de 135-150

La *Figura 27* muestra la tasa de fonemas de los audios 135 a 150 que son párrafos. Para la emoción neutral se tiene una tasa de alrededor de 15.5 fonemas, para alegría 16 y para tristeza 12.9.

La *Figura 28* muestra la tasa de fonemas de los audios 151 a 184 que son dígitos y palabras aisladas. Para la emoción neutral se tiene una tasa de alrededor de 12 fonemas, para alegría 11.9 y para tristeza 9.9 aquí existe una ligera variación entre las emociones.

6.3 Duración y pitch de los fonemas /a/ y /e/ España

Estos fonemas /a/ y /e/ son los que más frecuencia tienen dentro en el habla española la *Figura 29* muestra la duración del fonema /a/. Para la emoción neutral se tiene una duración de alrededor de 0.05, para alegría 0.06 y para tristeza 0.04.

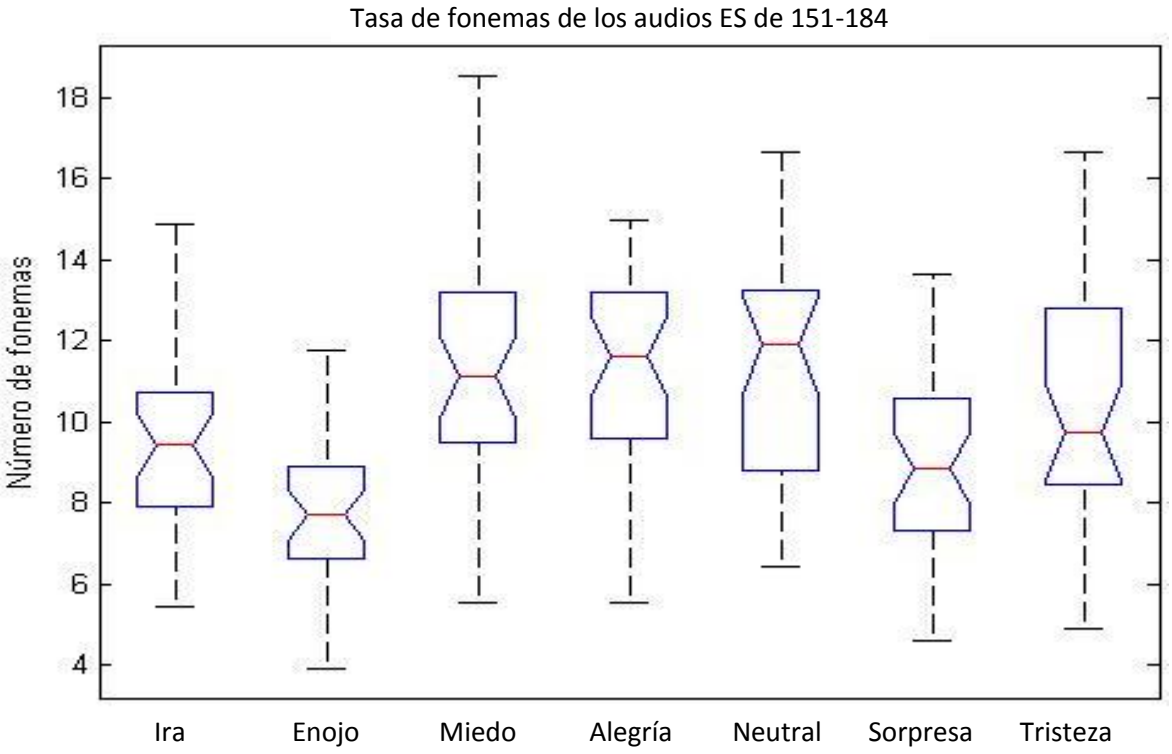


Figura 28. Tasa de fonemas de los audios ES de 151-184

La *Figura 30* muestra la duración del fonema /e/. Para la emoción neutral se tiene una duración de alrededor de 0.045, para alegría 0.05 y para tristeza 0.05 donde existe una variación en la tristeza.

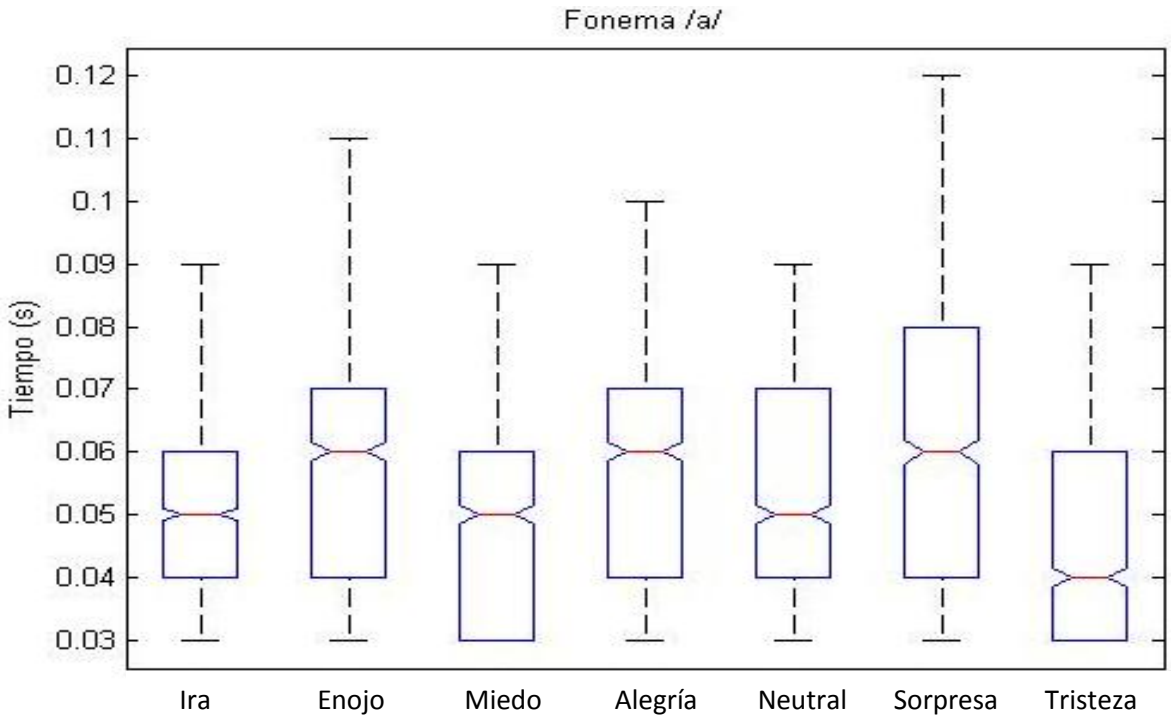


Figura 29. Duración fonema /a/.

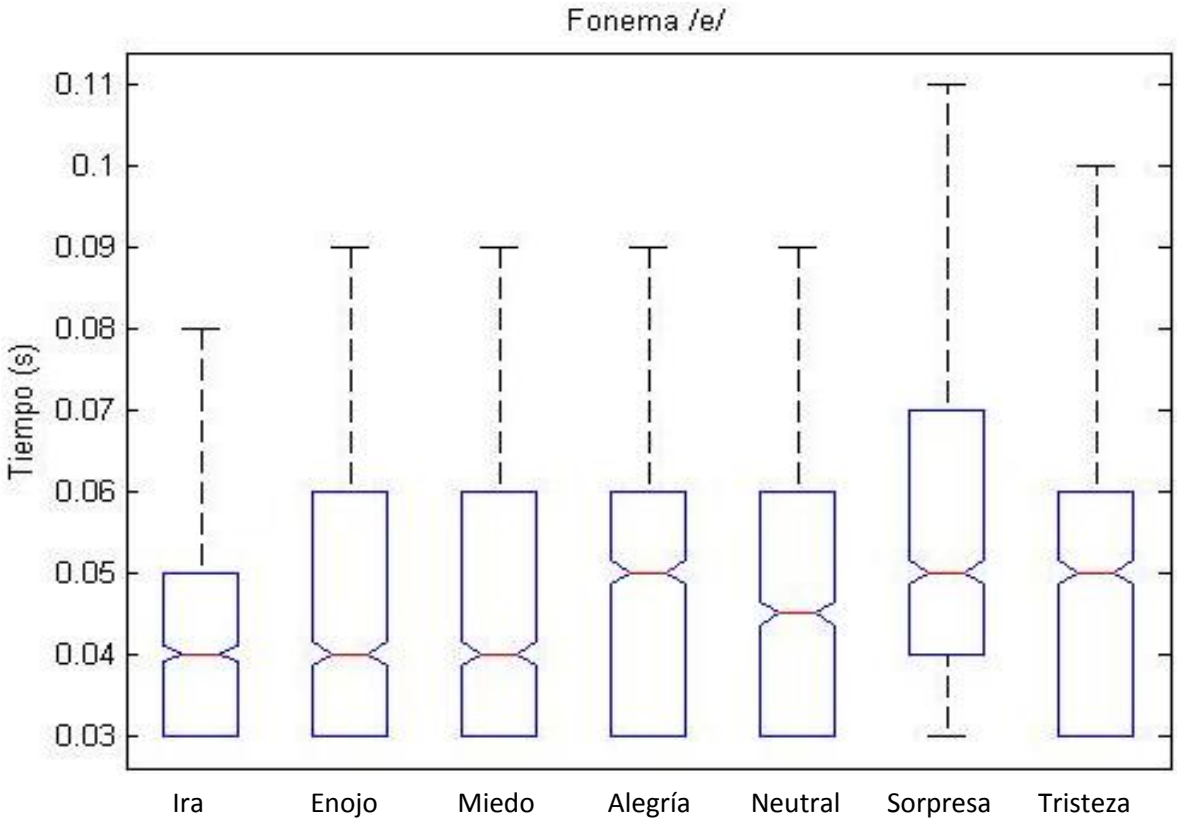


Figura 30. Duración fonema /e/.

La *Figura 31* muestra el pitch del fonema /a/. Para la emoción neutral se tiene un pitch de alrededor de 130, para alegría 230 y para tristeza 130 donde la tristeza tiene el valor de la emoción neutral, y es comprensible ya que la tristeza es muy cercana al rango de la emoción neutral según el Plano del espacio tridimensional de emociones, un comportamiento similar se observa pasa con la *Figura* 32 que muestra el pitch del fonema /e/. Para la emoción neutral se tiene un pitch de alrededor de 130, para alegría 230 y para tristeza 130.

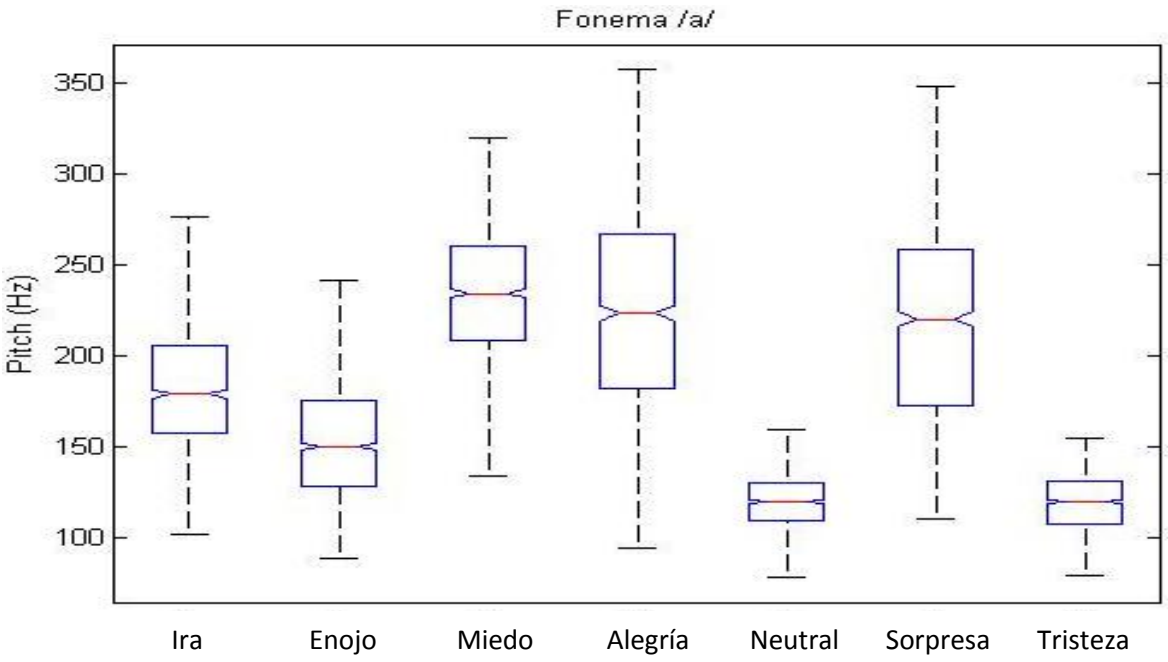


Figura 31. Pitch fonema /a/.

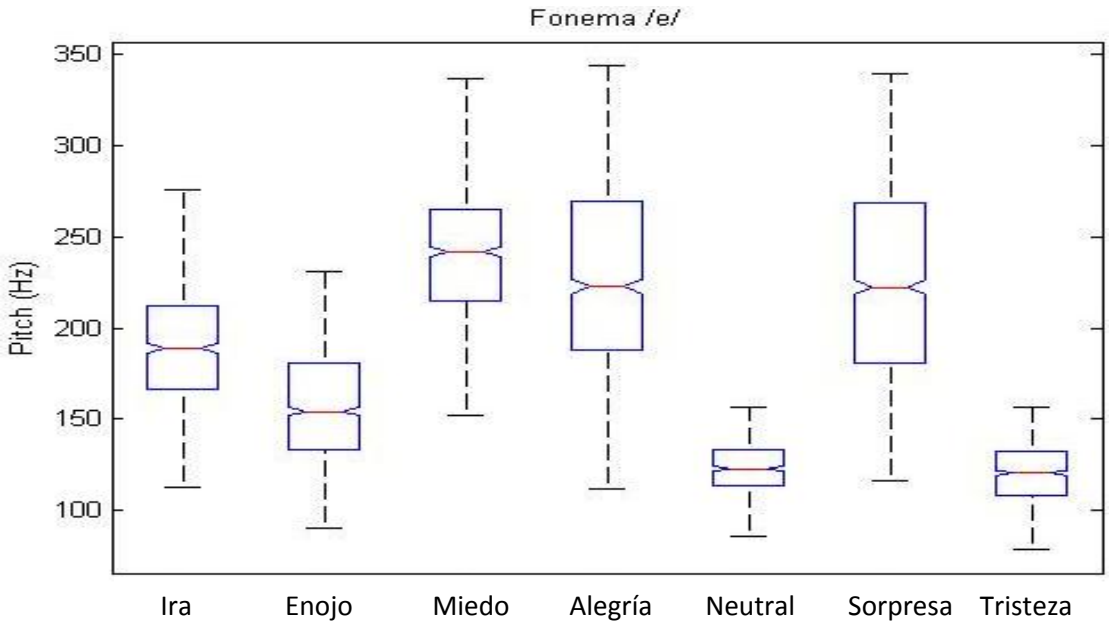


Figura 32. Pitch fonema /e/.

6.4 Tasa de sílabas español de México.

La *Figura 33* muestra la tasa de sílabas de los audios del 1 a 150 que son oraciones largas, cortas, interrogativas y párrafos. Para la emoción neutral se tiene una tasa de alrededor de 6 sílabas, para alegría 6.2 y para tristeza 6.4 sílabas aquí se puede observar una variación con la tristeza.

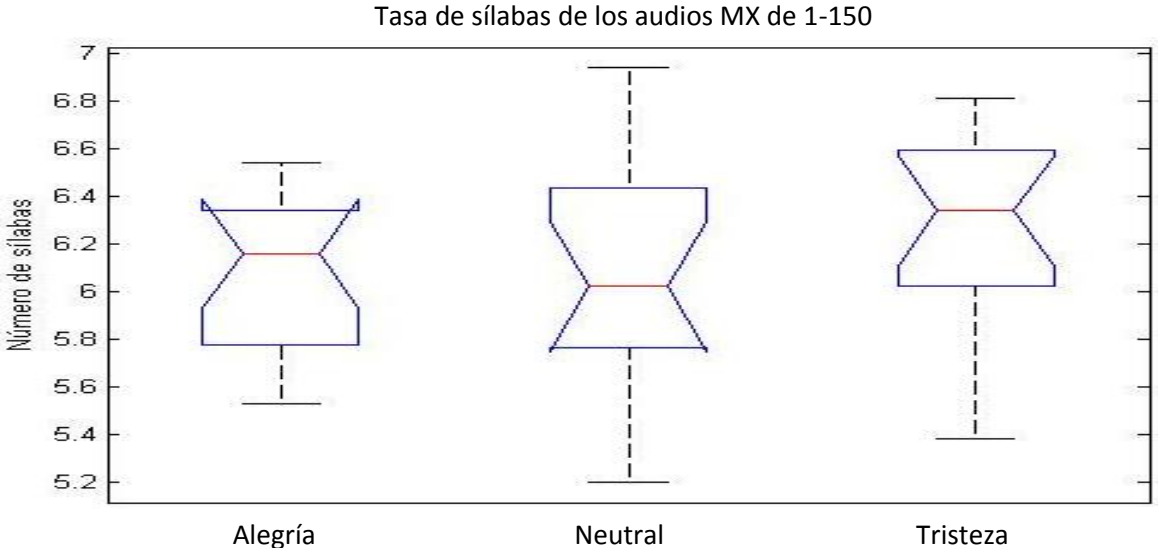


Figura 33. Tasa de sílabas de los audios MX de 1-150

La *Figura 34* muestra la tasa de sílabas de los audios 1 a 100 que son las oraciones afirmativas largas y cortas. Para la emoción neutral se tiene una tasa de alrededor de 6 sílabas, para alegría 5.5 y para tristeza 5.8 sílabas donde nuevamente se refleja un poco elevada con respecto de la neutral.

La *Figura 35* muestra la tasa de sílabas de los audios 101 a 134 que son interrogantes y oraciones con énfasis. Para la emoción neutral se tiene una tasa de alrededor de 6.5 sílabas, para alegría 6.2 y para tristeza 6 sílabas.

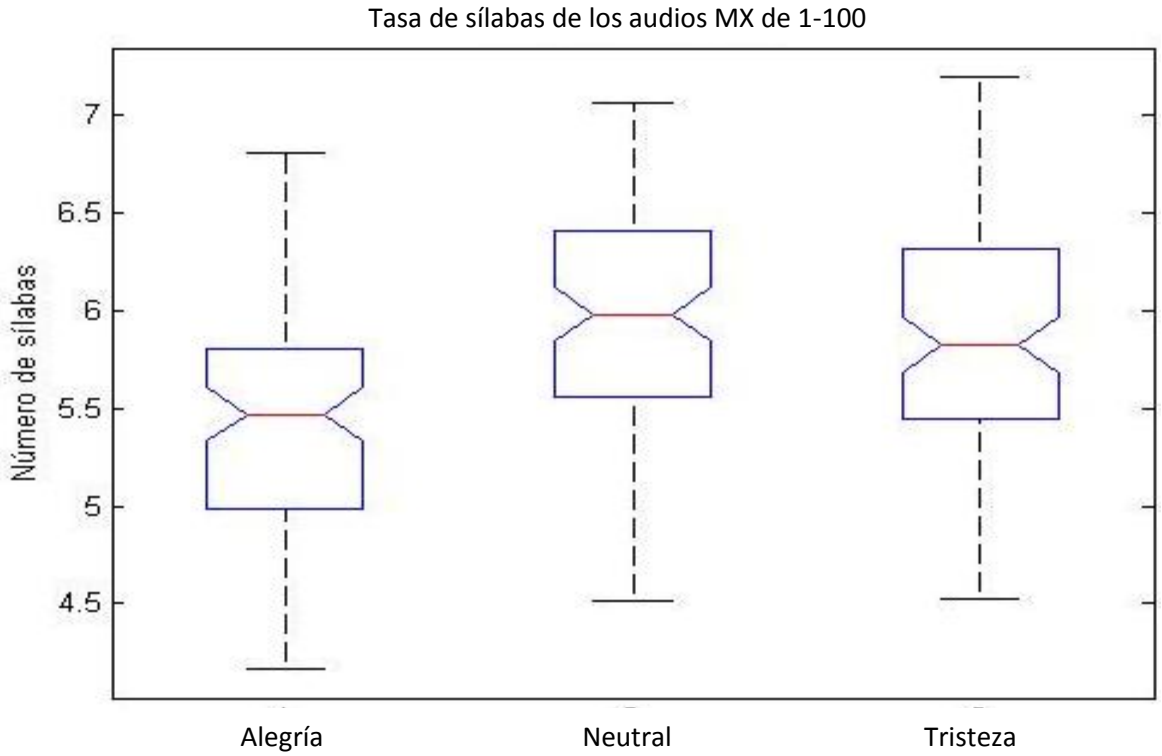


Figura 34. Tasa de sílabas de los audios MX de 1-100

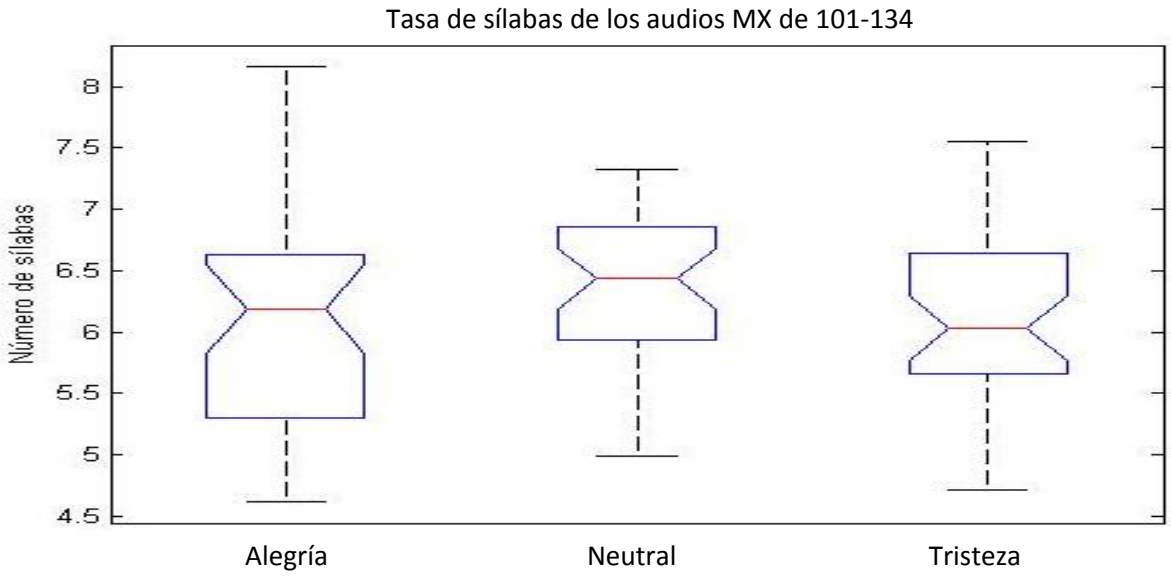


Figura 35. Tasa de sílabas de los audios MX de 101-134

La *Figura 36* muestra la tasa de sílabas de los audios 135 a 150 que son párrafos. Para la emoción neutral se tiene una tasa de alrededor de 6.19 sílabas, para alegría 6.19 y para tristeza 6.39 sílabas.

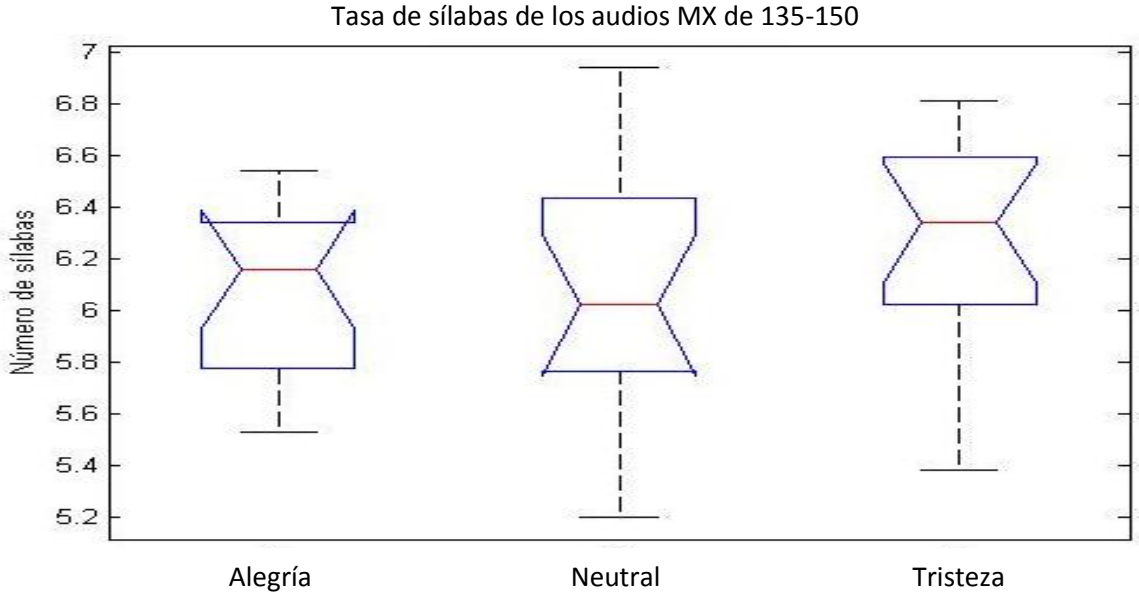


Figura 36. Tasa de sílabas de los audios MX de 135-150

La *Figura 37* muestra la tasa de sílabas de los audios 151 a 184 que son dígitos y palabras aisladas. Para la emoción neutral se tiene una tasa de alrededor de 4.9 sílabas, para alegría 3 y para tristeza 3.9 sílabas.

6.5 Tasa de fonemas español de México.

La *Figura 38* muestra la tasa de fonemas de los audios 1 a 150 que son oraciones largas, cortas, interrogativas y párrafos. Para la emoción neutral se tiene una tasa de alrededor de 14.1 fonemas, para alegría 13 y para tristeza 13.9 fonemas.

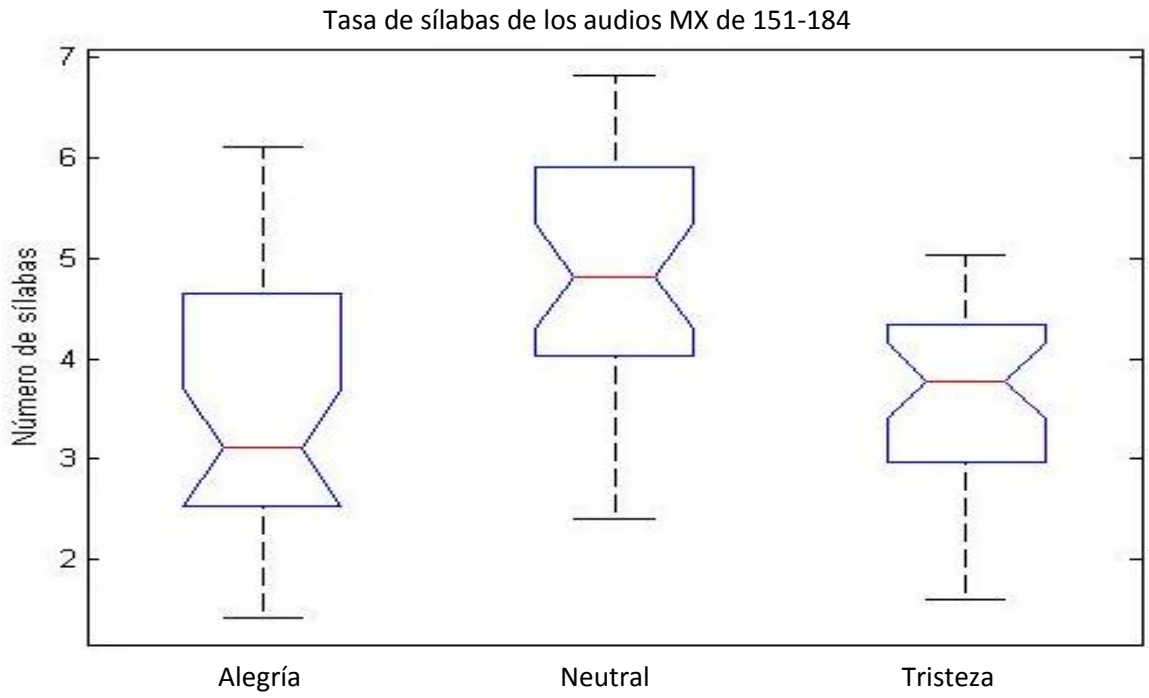


Figura 37. Tasa de sílabas de los audios MX de 151-184

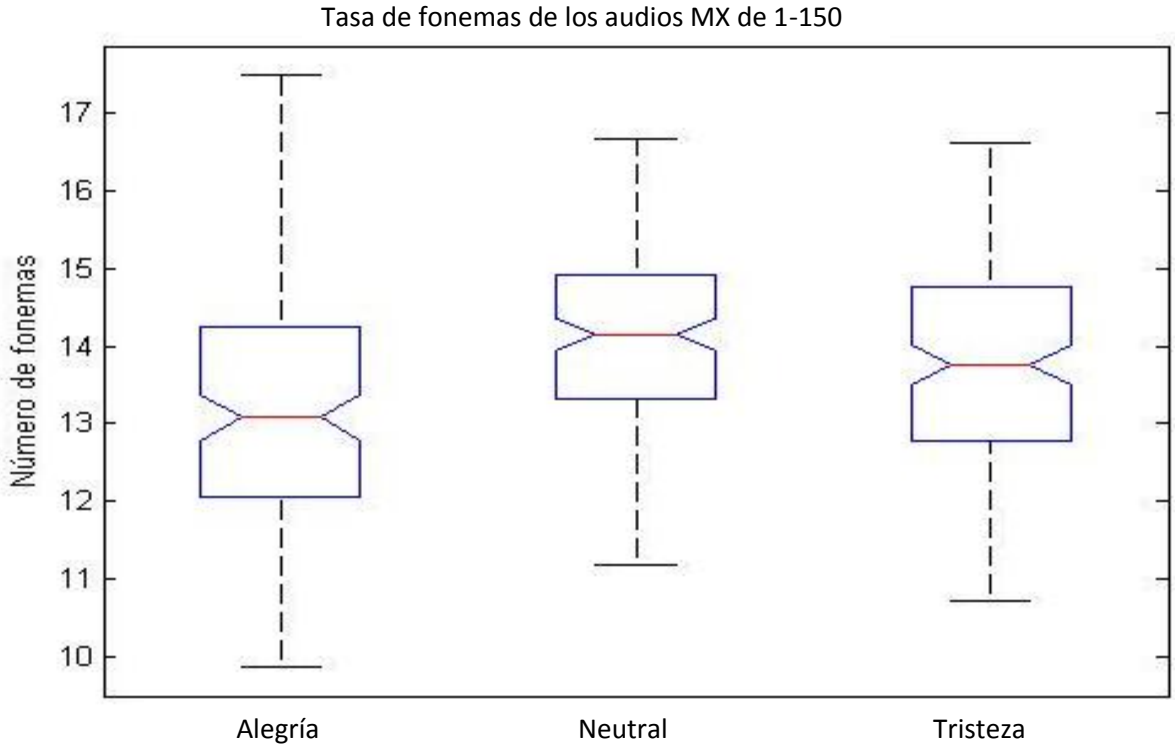


Figura 38. Tasa de fonemas de los audios MX de 1-150.

La *Figura 39* muestra la tasa de fonemas de los audios 1 a 100 que son las oraciones afirmativas largas y cortas. Para la emoción neutral se tiene una tasa de alrededor de 14 fonemas, para alegría 12.5 y para tristeza 13.5 fonemas.

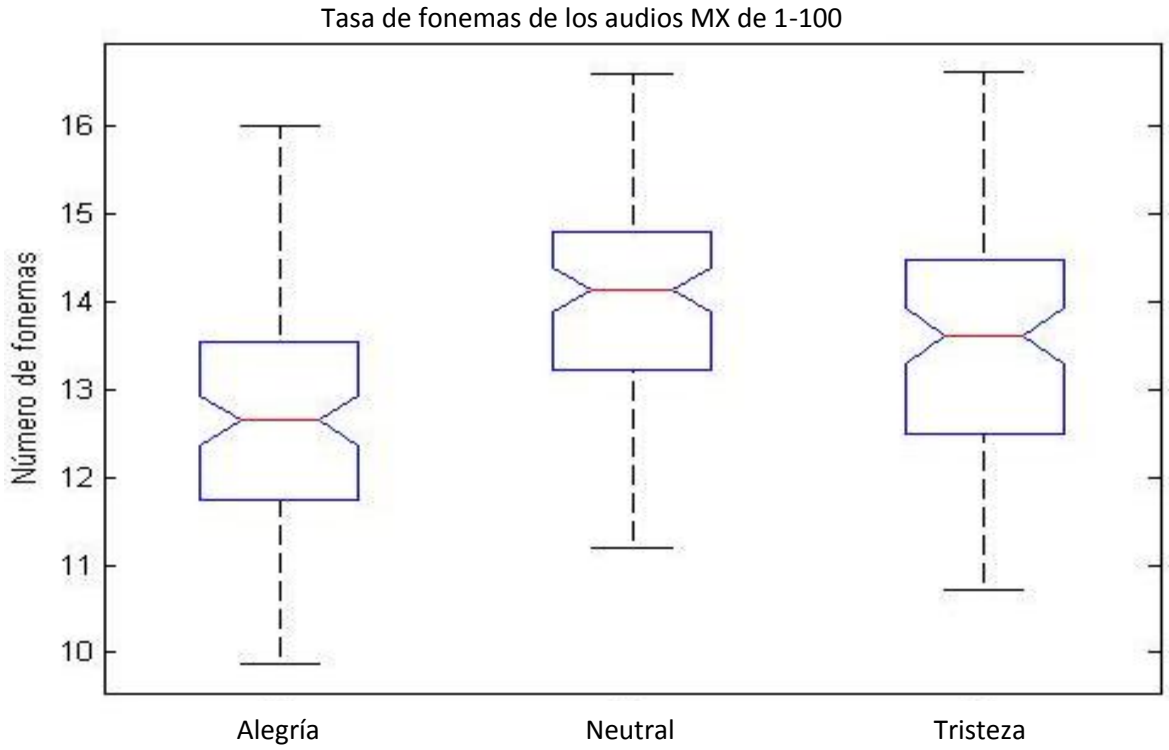


Figura 39. Tasa de fonemas de los audios MX de 1-100.

La *Figura 40* muestra la tasa de fonemas de los audios 101 a 134 que son interrogantes y oraciones con énfasis. Para la emoción neutral se tiene una tasa de alrededor de 14.5 fonemas, para alegría 14.1 y para tristeza 14.

La *Figura 41* muestra la tasa de fonemas de los audios 135 a 150 que son párrafos. Para la emoción neutral se tiene una tasa de alrededor de 13.9 fonemas, para alegría 14.3 y para tristeza 14.5 fonemas.

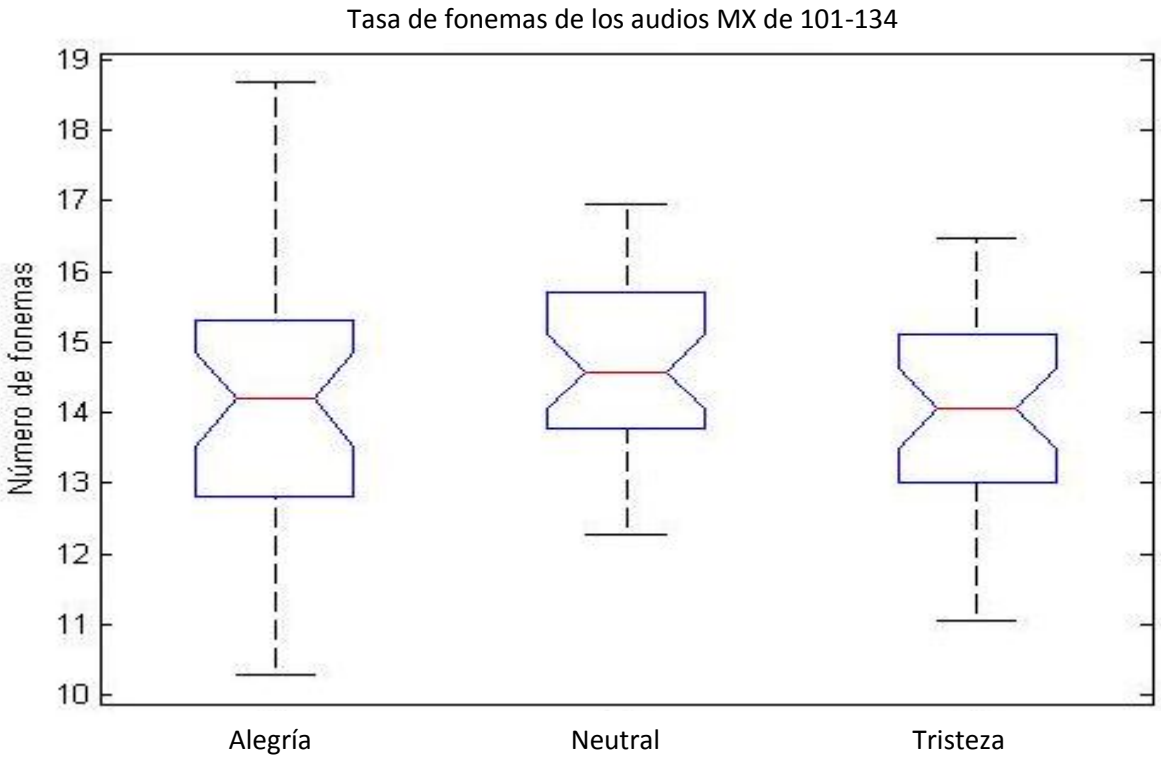


Figura 40. Tasa de fonemas de los audios MX de 101-134

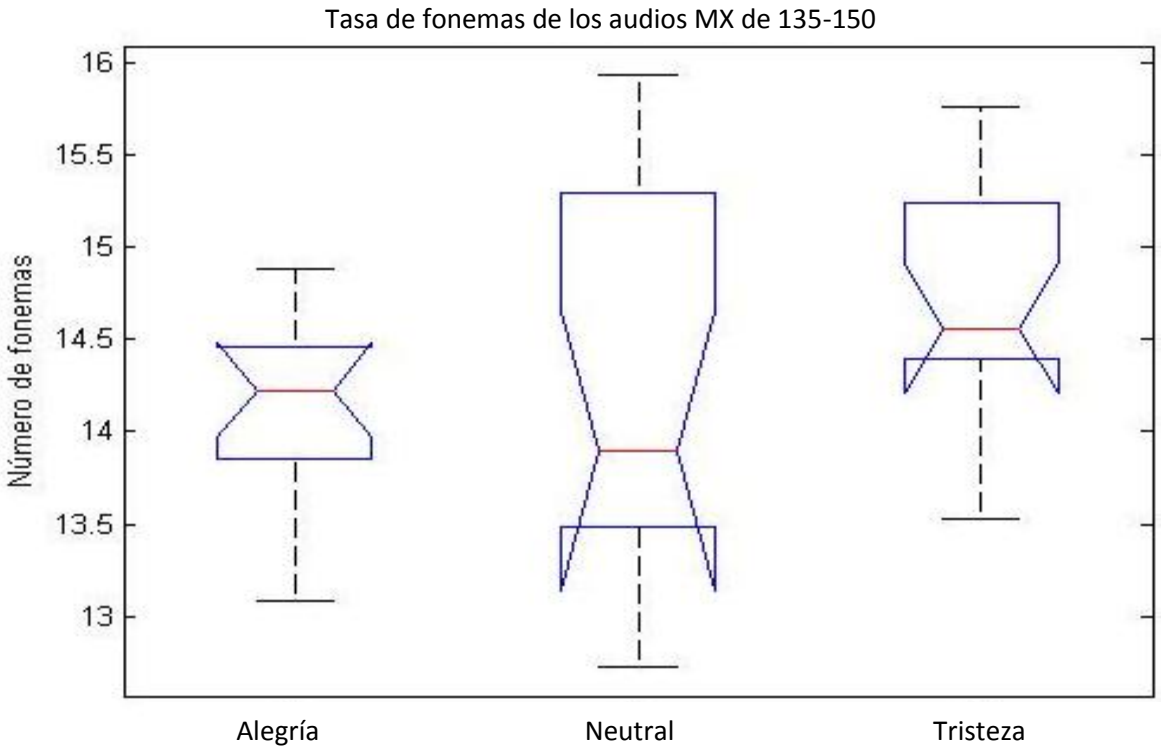


Figura 41. Tasa de fonemas de los audios MX de 135-150.

La *Figura 42* muestra la tasa de fonemas de los 151 a 184 que son dígitos y palabras aisladas. Para la emoción neutral se tiene una tasa de alrededor de 11.1 fonemas, para alegría 7 y para tristeza 8 fonemas.

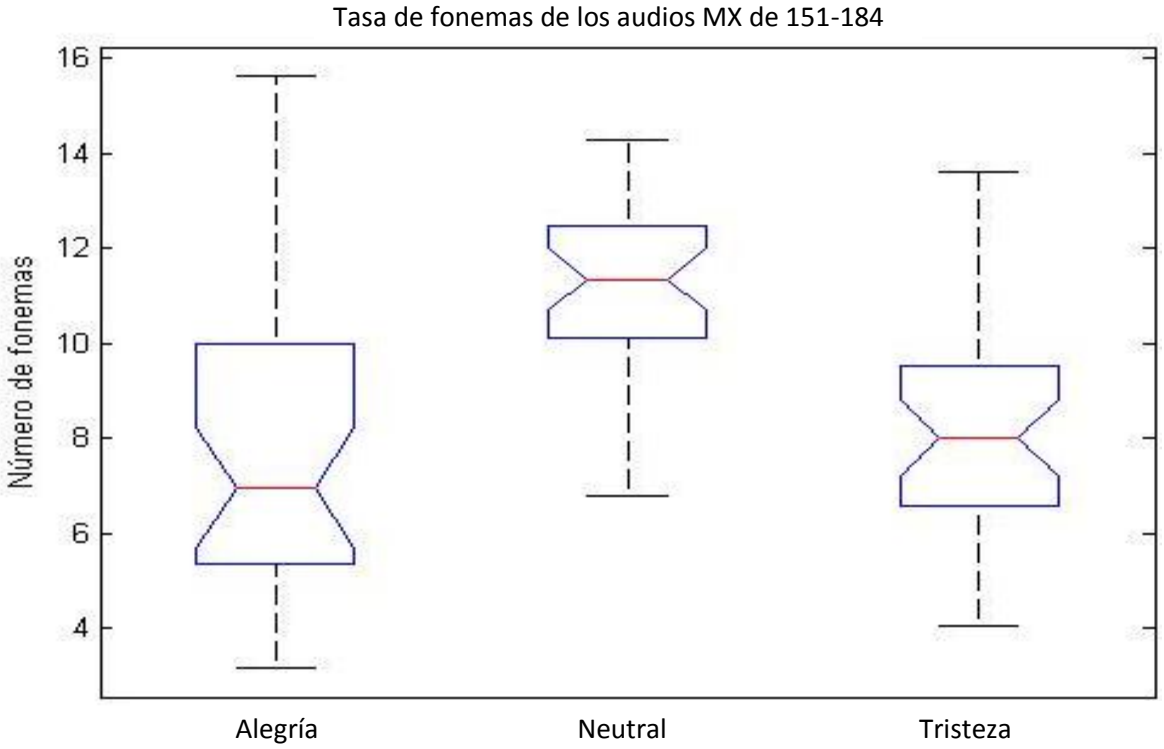


Figura 42. Tasa de fonemas de los audios MX de 151-184

6.6 Duración y pitch de los fonemas /a/ y /e/ México

La *Figura 43* muestra el fonema /a/. Para la emoción neutral se tiene una duración de alrededor de 0.06, para alegría 0.1 y para tristeza 0.08.

La *Figura 44* muestra el fonema /e/. Para la emoción neutral se tiene una duración de alrededor de 0.05, para alegría 0.07 y para tristeza 0.06.

La *Figura 45* muestra el fonema /a/. Para la emoción neutral se tiene un pitch de alrededor de 100, para alegría 230 y para tristeza 110.

La *Figura 46* muestra el fonema /e/. Para la emoción neutral se tiene una pitch de alrededor de 100, para alegría 230 y para tristeza 120.

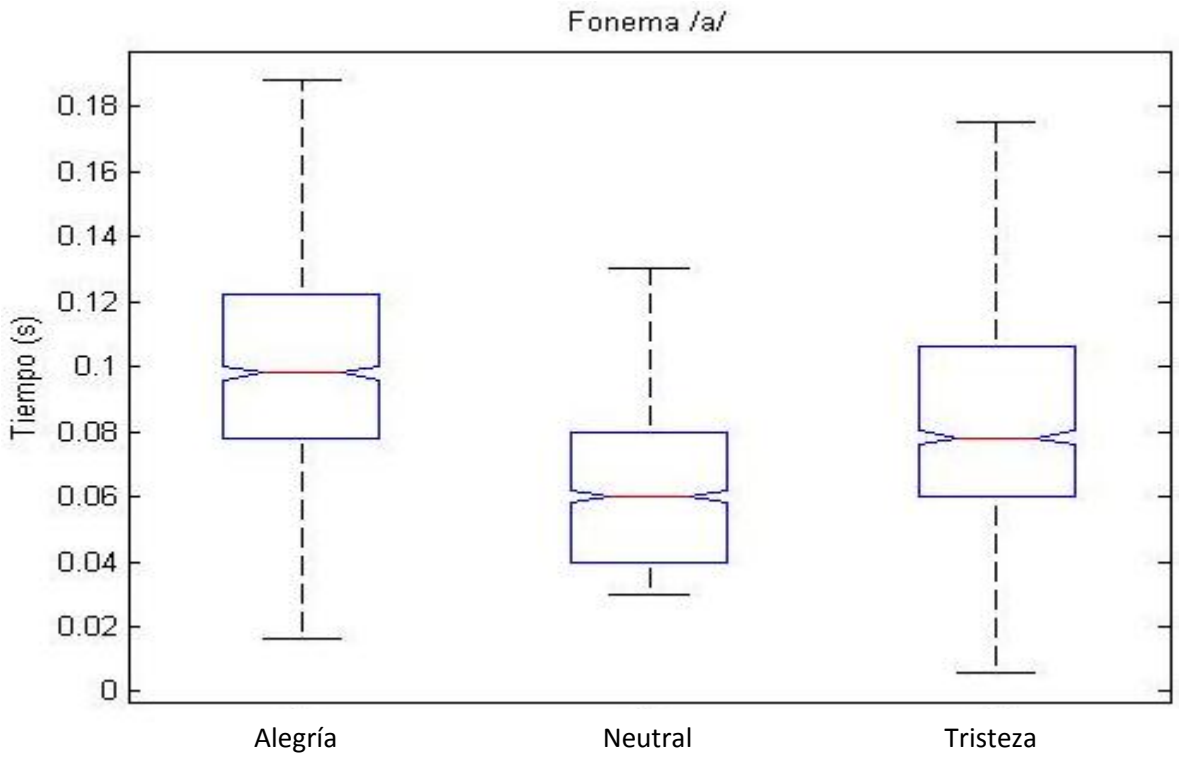


Figura 43. Duración Fonema /a/.

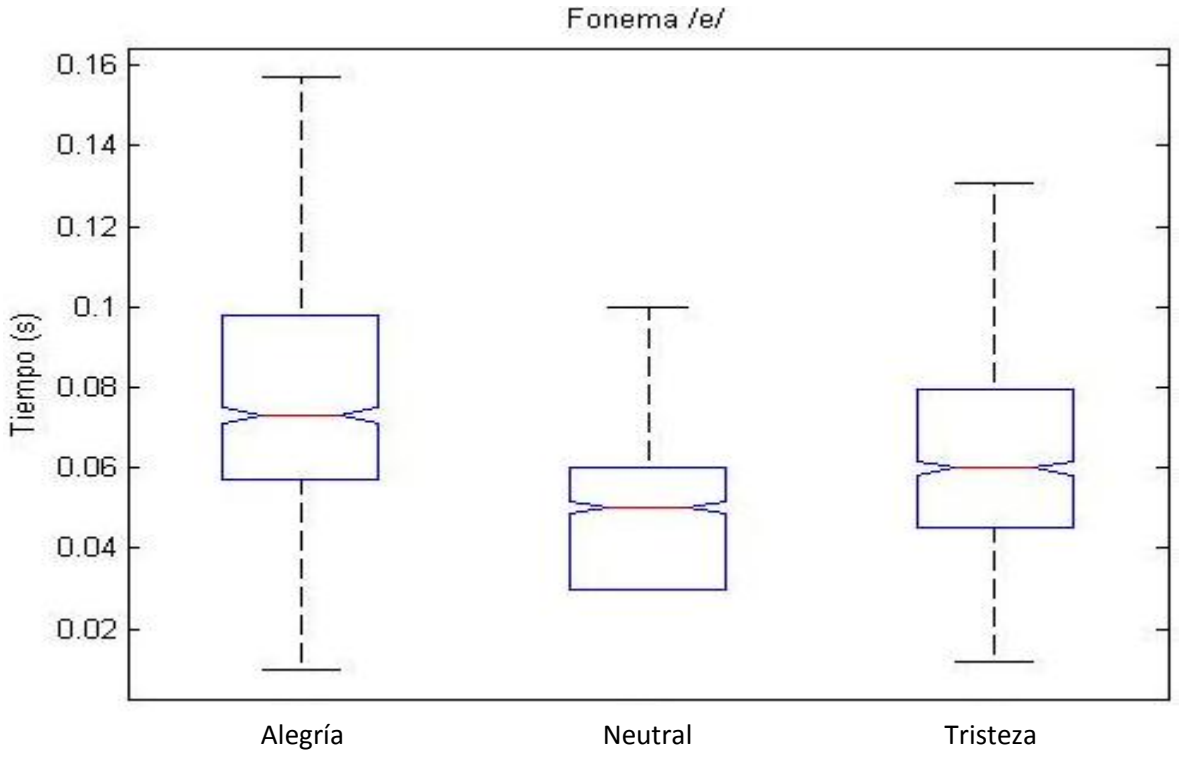


Figura 44. Duración Fonema /e/.

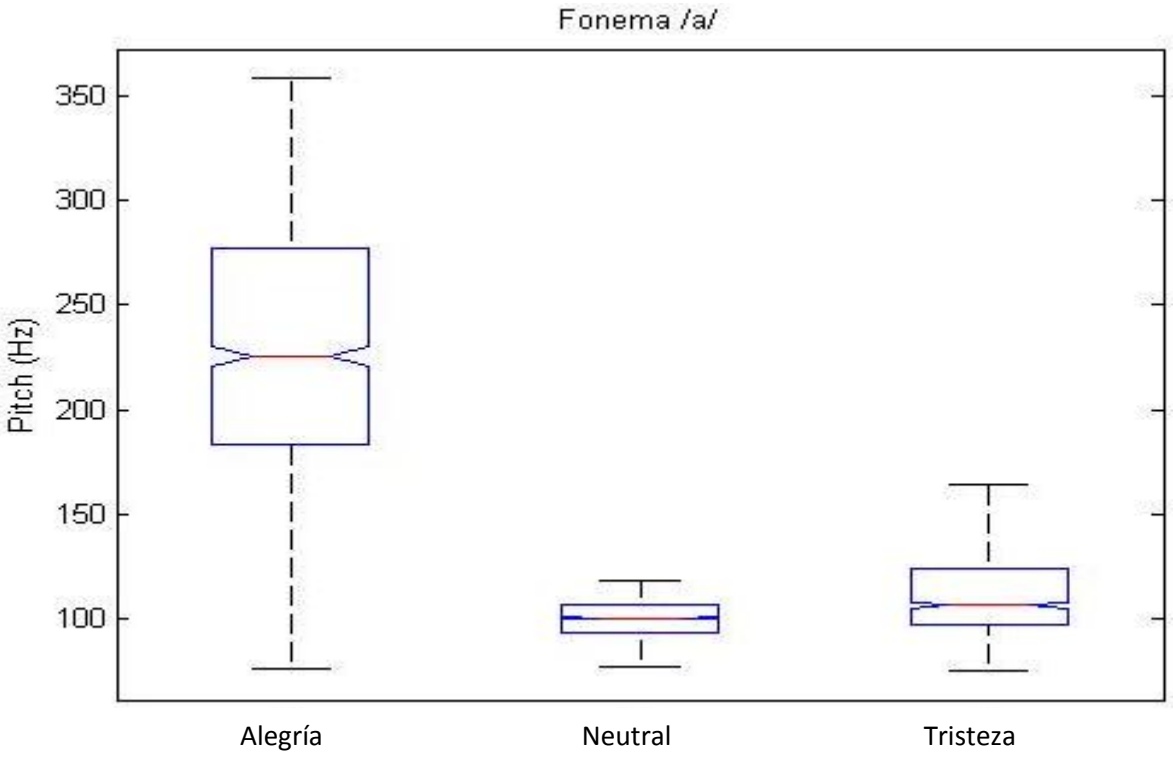


Figura 45. Pitch Fonema /a/.

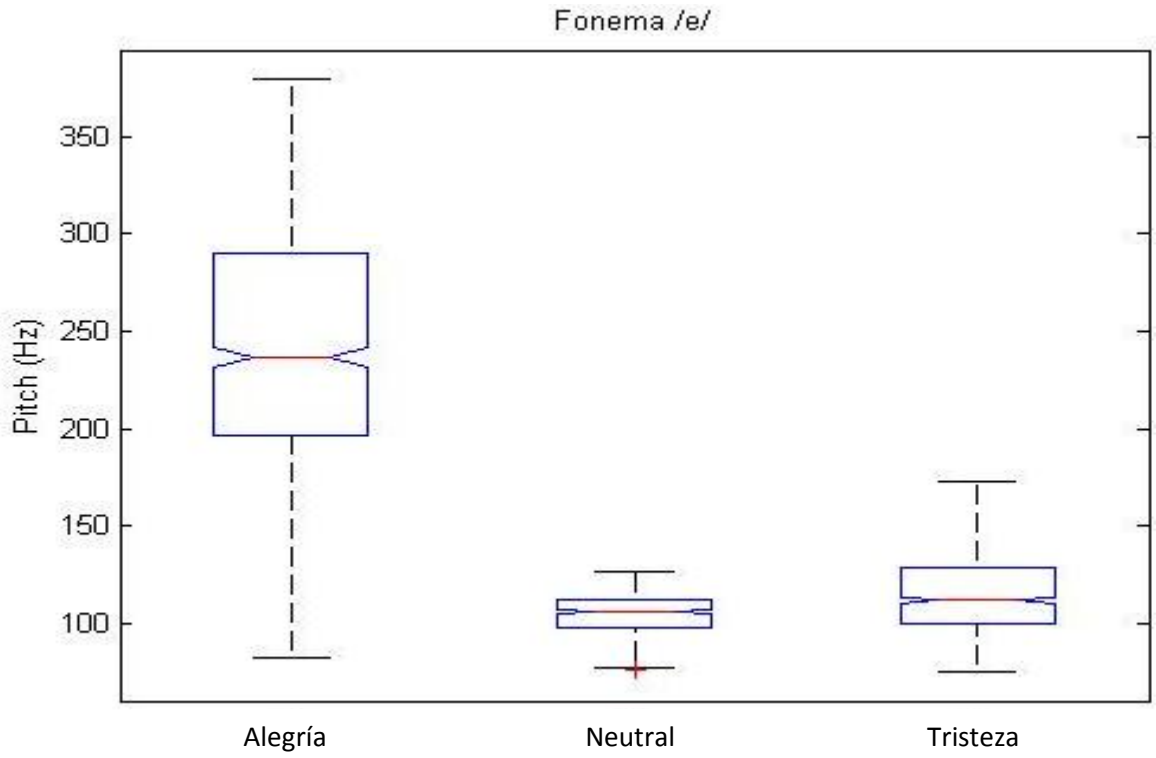


Figura 46. Pitch Fonema /e/.

6.7 Comparación tasa de sílabas del español de España y México.

Las siguientes *Tablas 24 y 25* reúnen los datos de **número de sílabas y fonemas** en las diferentes estructuras de audios del español de España y México.

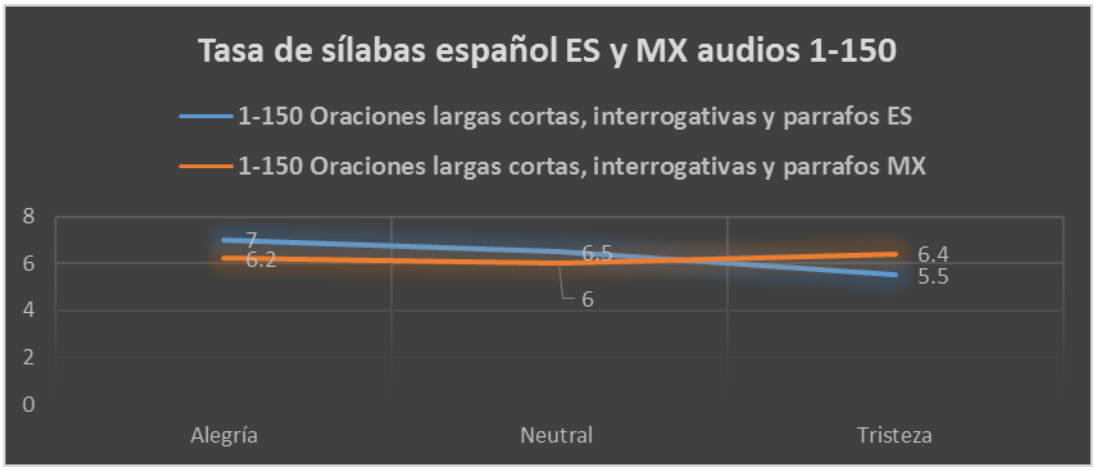
Tasa de sílabas español ES y MX audios 1-184					
Audio	Estructura	Español	Alegría	Neutral	Tristeza
1-150	Oraciones largas cortas, interrogativas y párrafos	ES	7	6.5	5.5
		MX	6.2	6	6.4
1-100	Afirmativas largas y cortas	ES	6.7	6.5	5.2
		MX	5.5	6	5.8
101-134	Interrogantes y oraciones con énfasis	ES	7.6	6.5	5.6
		MX	6.2	6.5	6
135-150	Párrafos	ES	6.9	6.8	5.6
		MX	6.19	6.19	6.39
151-184	Dígitos y palabras aisladas	ES	4.5	5	4.1
		MX	3	4.9	3.9

Tabla 24. Tasa de sílaba del español de España y México.

Tasa de fonemas español ES y MX audios 1-184					
Audio	Estructura	Español	Alegría	Neutral	Tristeza
1-150	Oraciones largas cortas, interrogativas y párrafos	ES	15	16	12.5
		MX	6	6.2	6.4
1-100	Afirmativas largas y cortas	ES	15	15.9	12.1
		MX	6	5.5	5.8
101-134	Interrogantes y oraciones con énfasis	ES	15	18	12.5
		MX	6.5	6.2	6
135-150	Párrafos	ES	15.5	16	12.9
		MX	6.19	6.19	6.39
151-184	Dígitos y palabras aisladas	ES	12	11.9	9.9
		MX	4.9	3	3.9

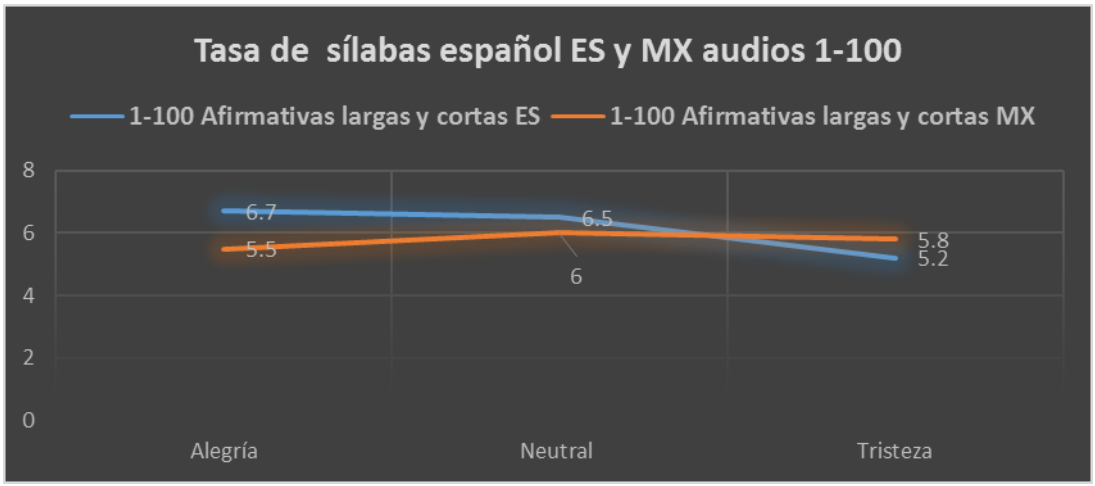
Tabla 25. Tasa de fonemas del español de España y México.

La *Gráfica 1*, muestra la comparación entre español de España y México para los primeros 150 audios. En la de España se nota la tendencia de la emoción más alta de 7 a 5.5, en español de México va de 6.2 a 6.4, la variación



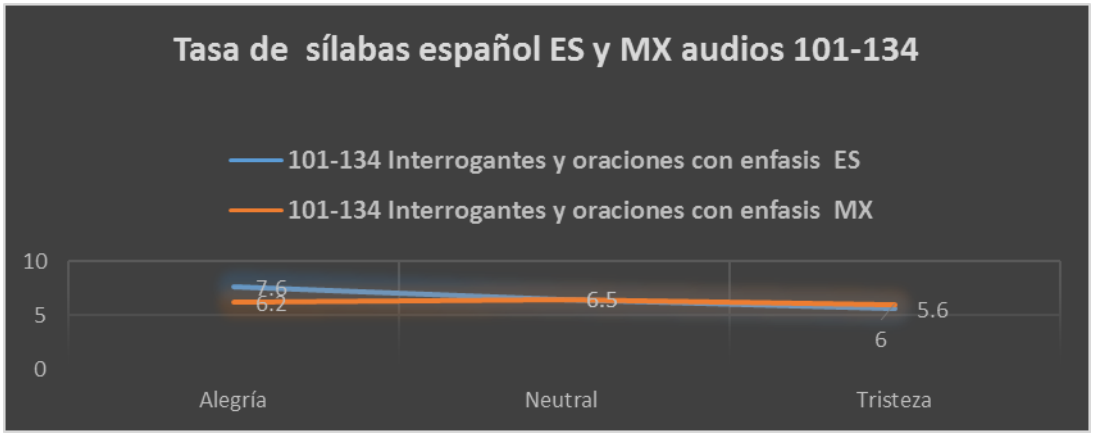
Gráfica 1. Tasa de sílabas español ES y MX audios 1-150.

La misma tendencia se puede observar en las *Gráficas* de la *2* a la *5*, se recuerda que las condiciones de grabación, equipo y lugar además el hablante no son las mismas y estas podrían ser las razones por las cuales existen esas diferencias.



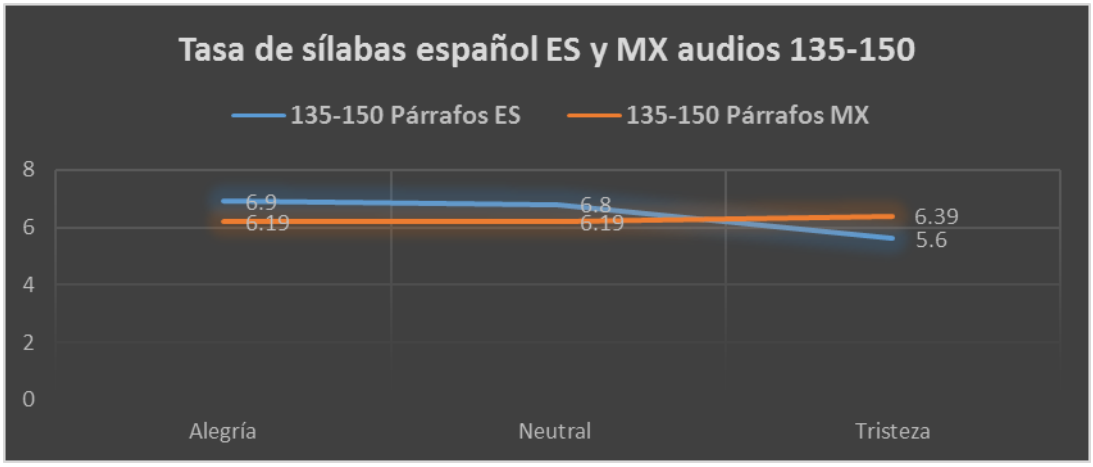
Gráfica 2. Tasa de sílabas español ES y MX audios 1-100.

Pero en la gráfica 3 para el español de México es interesante observar que se mantiene la tendencia de emoción alta 6.2, 6.5 neutral y 6 tristeza, son interrogantes y oraciones con énfasis que es lo que se necesita para este proyecto, ese énfasis en la emoción.

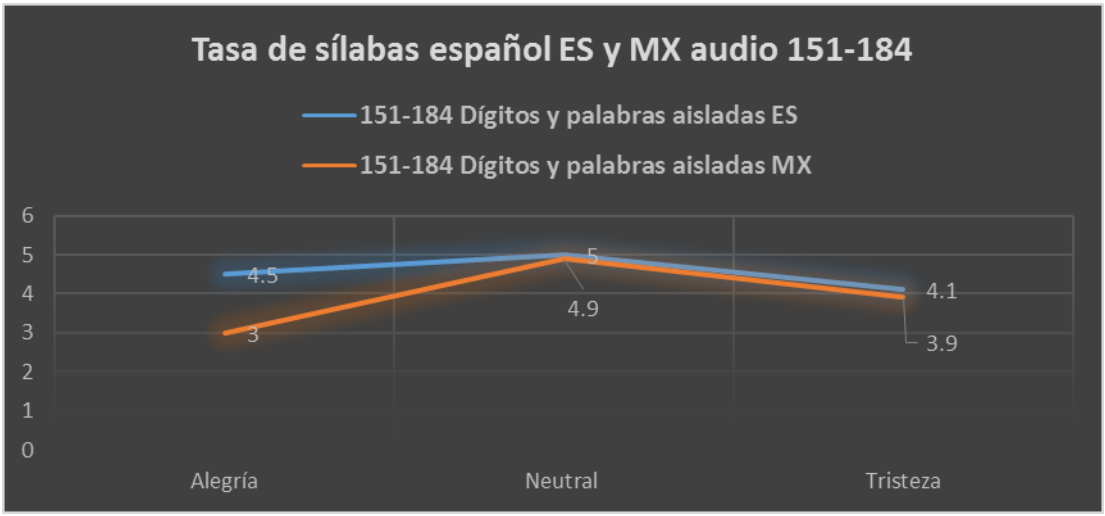


Gráfica 3. Tasa de sílabas español ES y MX audios 101-134.

En la gráfica 4 tambien existen ligeras variación con una tristeza de español de México a la alza.



Gráfica 4. Tasa de sílabas español ES y MX audios 135-150.

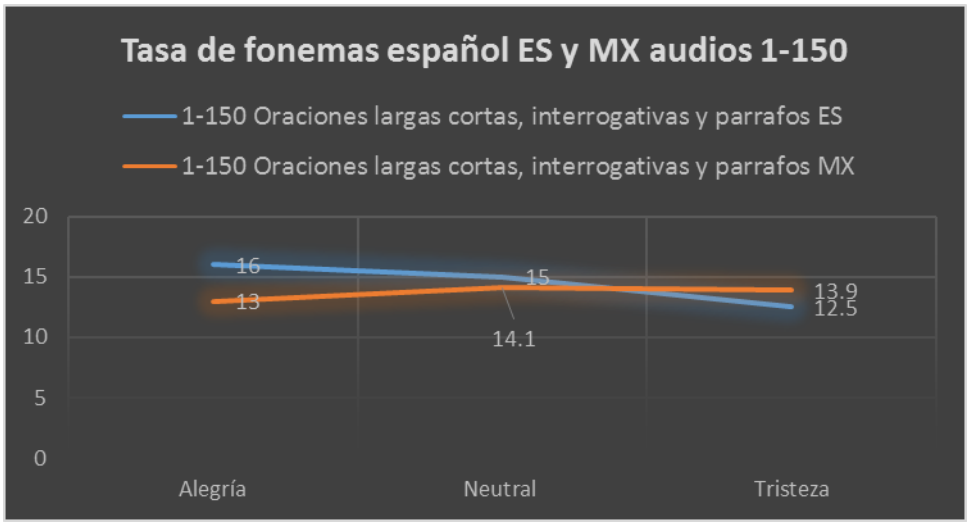


Gráfica 5. Tasa de sílabas español ES y MX audios 151-184.

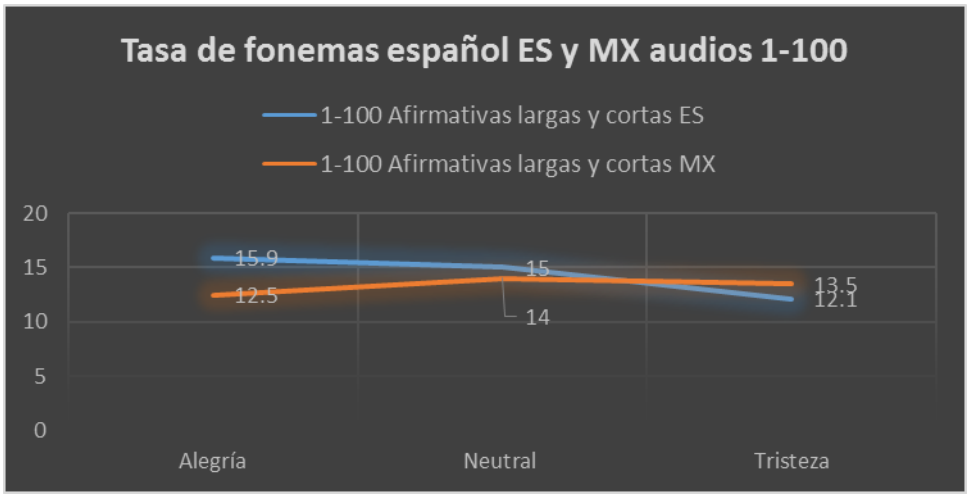
La tendencia del español de España de acuerdo a la tasa de silabas se refleja de manera regular alegría como una emoción alta, neutra como media y tristeza como baja.

6.8 Comparación tasa de fonemas del español de España y México.

En esta comparación ocurre algo similar que en la tasa de sílabas del punto anterior, la tasa de fonemas de ambos españoles muestra la misma tendencia en las *Gráficas* de la *6* a la *13* se puede observar esto.

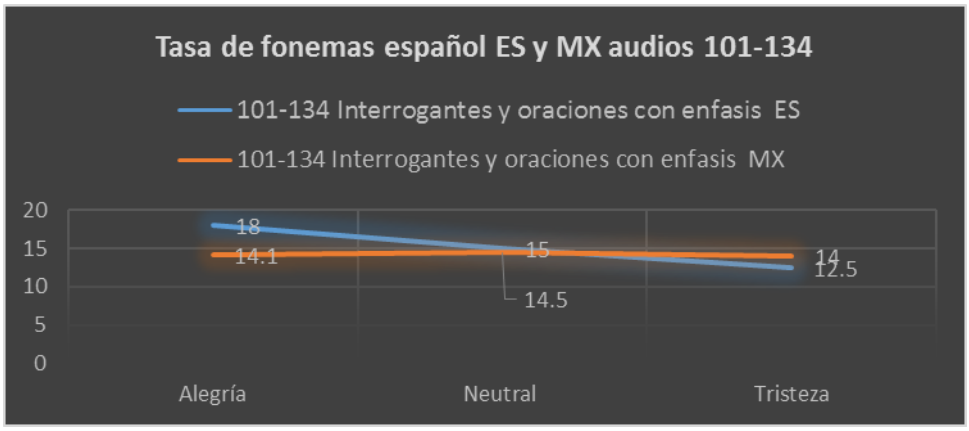


Gráfica 6. Tasa de fonemas español ES y MX audios 1-150.



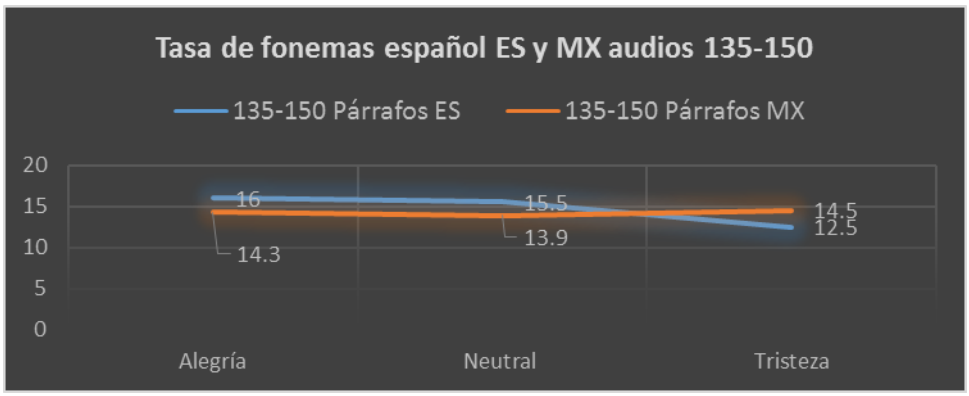
Gráfica 7. Tasa de fonemas español ES y MX audios 1-100.

La Gráfica 8 muestra la tasa de fonemas de corpus de 101 a 134 interrogantes y oraciones con énfasis.



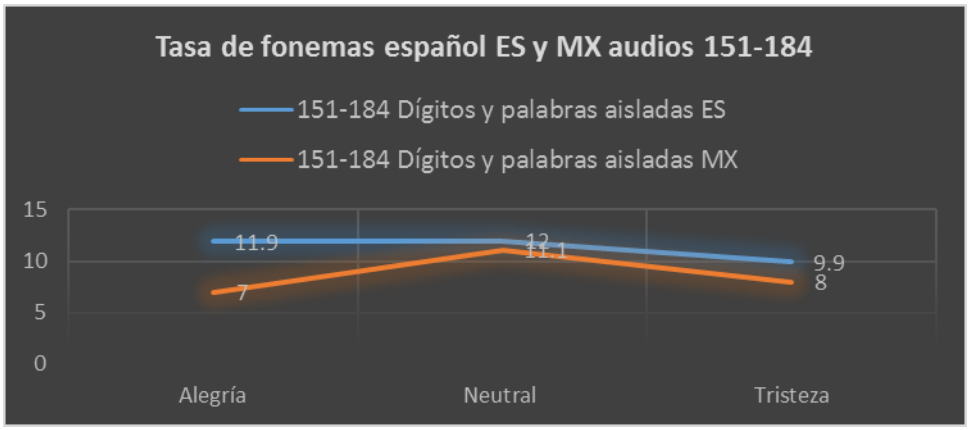
Gráfica 8. Tasa de fonemas español ES y MX audios 101-134.

La Gráfica 9 muestra la tasa de fonemas de corpus de 135 a 150 párrafos.



Gráfica 9. Tasa de fonemas español ES y MX audios 135-150.

La Gráfica 10 muestra la tasa de fonemas de corpus de 151 a 184 digitos y palabras aisladas.



Gráfica 10. Tasa de fonemas español ES y MX audios 151-184.

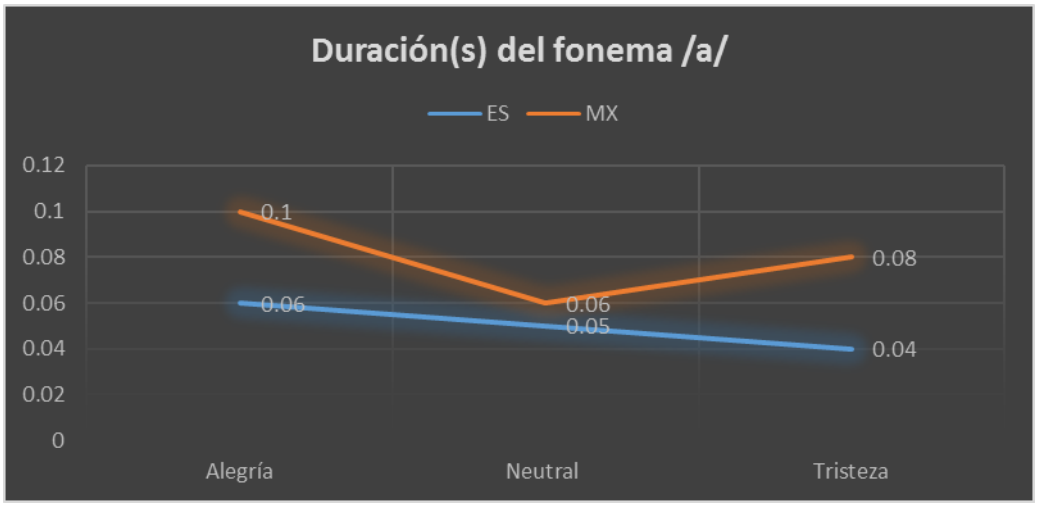
Lla emoción neutral y tristeza toman un comportamiento adecuado para el español de México. En España sigue un buen comportamiento y tendencia las tres emociones. La emoción neutral y tristeza comparando entre ES y MX son similares.

6.9 Comparación de la duración de fonemas /a/ y /e/ del español de España y México.

En cuanto a duración por fonema /a/ en la *Gráfica 11 y Tabla 26*, en ES la tendencia de alta a baja es más adecuada que en la MX aunque alegría se mantiene alta y tristeza baja.

Duración fonema /a/			
Español	Alegría	Neutral	Tristeza
ES	0.06	0.05	0.04
MX	0.1	0.06	0.08

Tabla 26. Duración fonema /a/

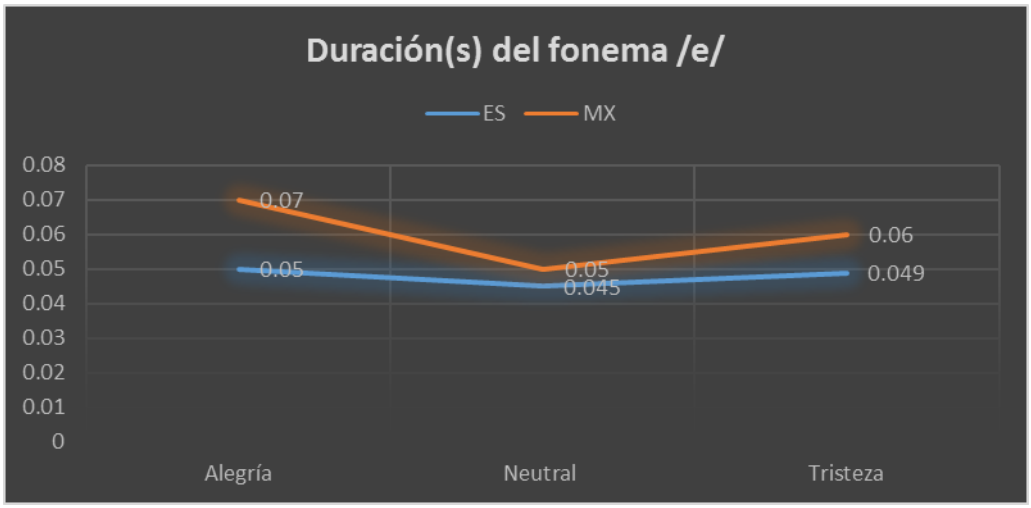


Gráfica 11. Duración(s) del fonema /a/

En la duración del fonema /e/ aunque las cifras no son las mismas la tendencia es similar. Lo mismo sucede en pitch del fonema /e/ y /a/ donde los números se acercan mucho en ambos españoles. *Gráfica 12 y Tabla 27*,

Duración fonema /e/			
Español	Alegría	Neutral	Tristeza
ES	0.05	0.045	0.049
MX	0.07	0.05	0.06

Tabla 27. Duración fonema /e/

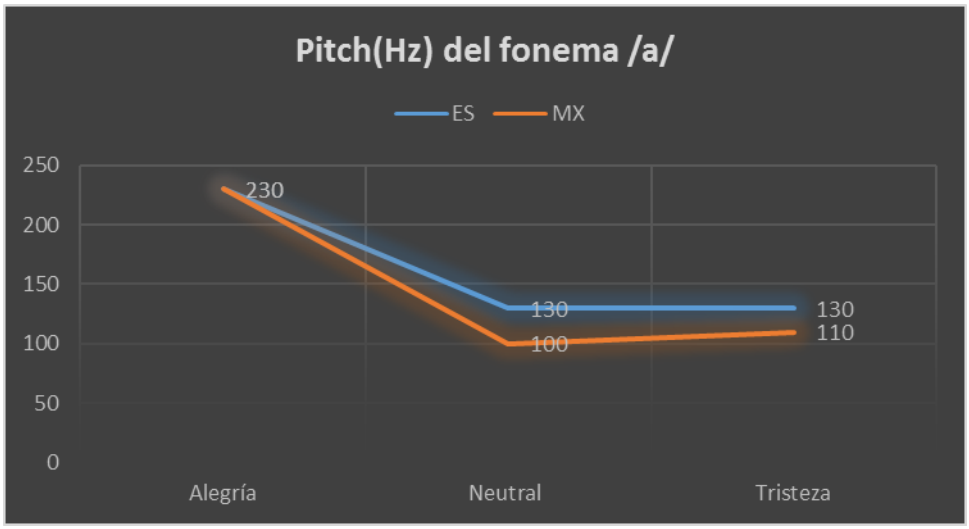


Gráfica 12. Duración(s) del fonema /e/

El pitch de el fonema /a/ tiene una buena tendencia con respecto a las emociones. Ver Tabla 28 y Gráfica 13.

Pitch(Hz) fonema /a/			
Español	Alegría	Neutral	Tristeza
ES	230	130	130
MX	230	100	110

Tabla 28. Pitch fonema /a/

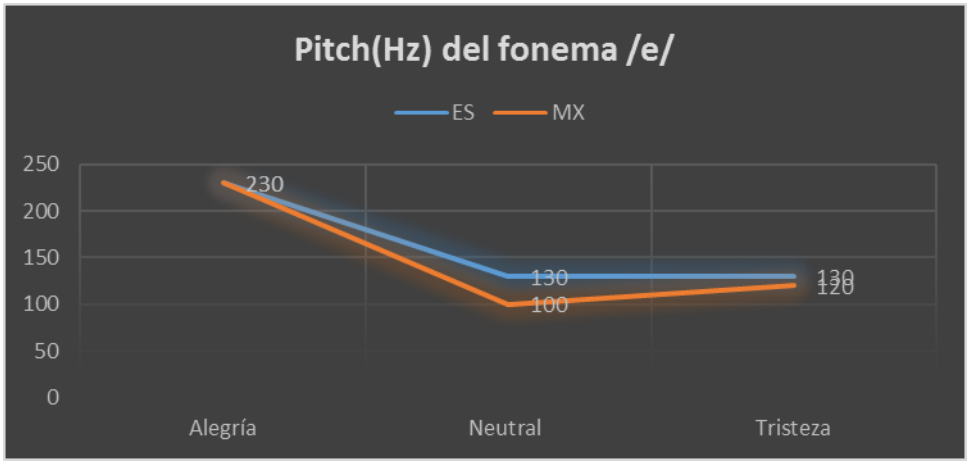


Gráfica 13. Pitch(Hz) del fonema /a/

El pitch de el fonema /e/ también tiene una buena tendencia con respecto a las emociones. Ver Tabla 29 y Gráfica 14.

Pitch(Hz) fonema /e/			
Español	Alegría	Neutral	Tristeza
ES	230	130	130
MX	230	100	120

Tabla 29. Pitch fonema /e/



Gráfica 14. Pitch(Hz) del fonema /e/

Los resultados que a continuación se presentan se obtuvieron del *Sistema de Texto a habla expresivo en español* y se comparan con los audios generados del sintetizador *Emofilt* en el cual se tiene que configurar la emoción y la voz.

Emofilt ofrece un abanico de emociones pero la comparación se hace entre alegría, tristeza y neutral además de las voces para España Es1, Es2 y Es4 y para el español de México Mx1 y Mx2 con las que cuenta el TTS-EE.

Las figuras a continuación presentadas la primera refleja la señal natural y la otra sección es el espectro de Emofil o del TTS-EE.

Se muestran las señales de onda de los audios de la frase "*¿Cómo se va a aceptar que la mujer tome la iniciativa?",* primero la *Figura 47* muestra esta frase de la emoción ***alegría***, de la voz ***Es1***, generada de ***Emofilt*** con una duración de ***5.25s.*** Todas las duraciones se obtienen de su respectiva señal de onda en la parte superior se puede ver la línea de tiempo de la frase.

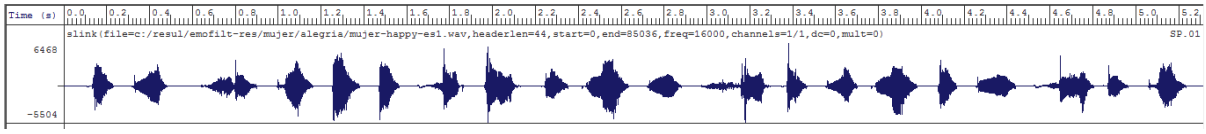


Figura 47. ¿Cómo se va a aceptar que la mujer tome la iniciativa? Emoción: alegría. Es1 Emofilt.

Ahora la *Figura 48* muestra la señal de onda de la misma frase para ***alegría***, de la voz ***Es1*** generada del ***TTS-EE***, la duración es de ***3.71s.*** Con una diferencia de ***1.54s.***, siendo más rápida ***Emofilt.***

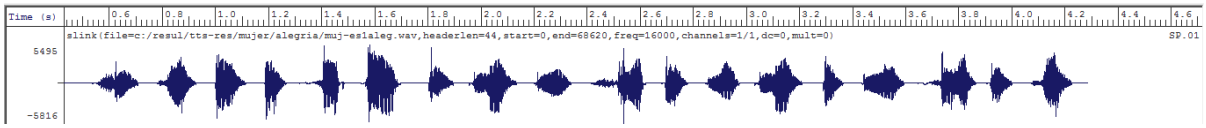


Figura 48. ¿Cómo se va a aceptar que la mujer tome la iniciativa? Emoción: alegría Es1 TTS-EE.

La *Figura 49 muestra la señal de onda de la emoción* ***alegría****, pero ahora de la voz* ***Es2*** *generada de* ***Emofilt*** *que presenta una duración de* ***5.14s.***

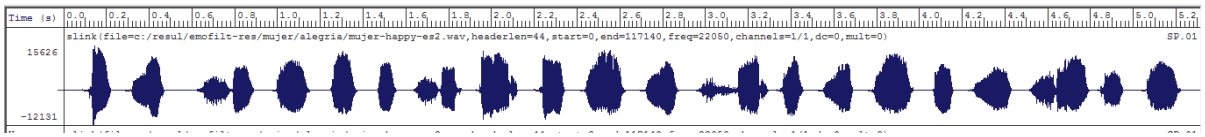


Figura 49. ¿Cómo se va a aceptar que la mujer tome la iniciativa? Es2 alegría Emofilt.

Y para ***TTS-EE*** de ***alegría*** para la voz ***Es2*** la duración es de ***4.16s***., Ver *Figura 50*. La diferencia entre estos dos es de ***0.98s.***

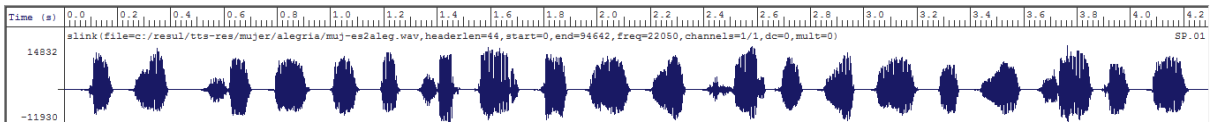


Figura 50. ¿Cómo se va a aceptar que la mujer tome la iniciativa? Es2 alegría TTS-EE.

La *Figura 51* muestra la señal de onda para la emoción ***alegría***, pero ahora de la voz ***Mx1,*** es decir español México generada en ***Emofilt*** y que presenta una duración de ***6.38s.***

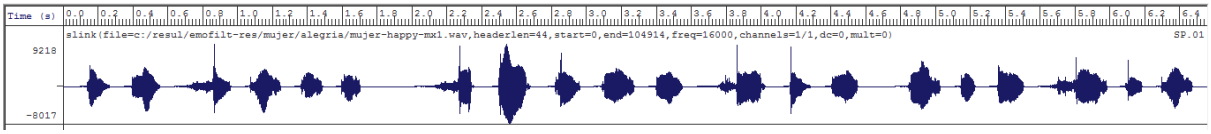


Figura 51. ¿Cómo se va a aceptar que la mujer tome la iniciativa? Mx1 alegría Emofilt.

La *Figura 52* muestra la señal de onda para la emoción ***alegría***, de la voz ***Mx1,*** generada por el ***TTS-EE*** y que presenta una duración de ***5.13s.***

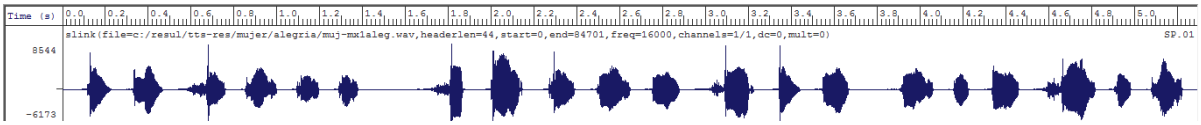


Figura 52. ¿Cómo se va a aceptar que la mujer tome la iniciativa? Mx1 alegría TTS-EE.

Se tiene una duración similar para ***alegría***, de la voz ***Mx2*** generada por ***Emofilt***, la *Figura 53* muestra su señal de onda con una duración de ***6.38s***.

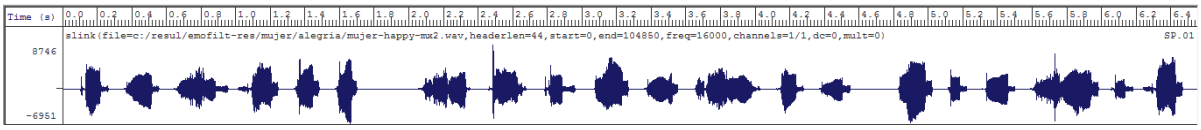


Figura 53. ¿Cómo se va a aceptar que la mujer tome la iniciativa? Mx2 alegría Emofilt.

Y finalmente para ***alegría***, de la voz ***Mx2*** generada por el ***TTS-EE***, la *Figura 54* muestra su señal de onda con una duración de ***5.12s***.

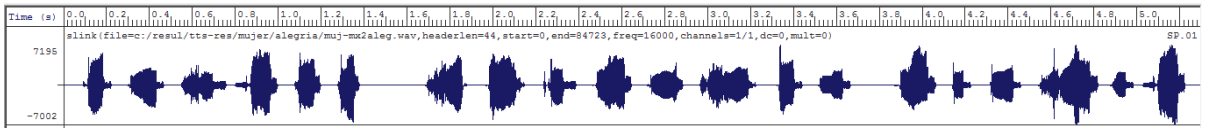
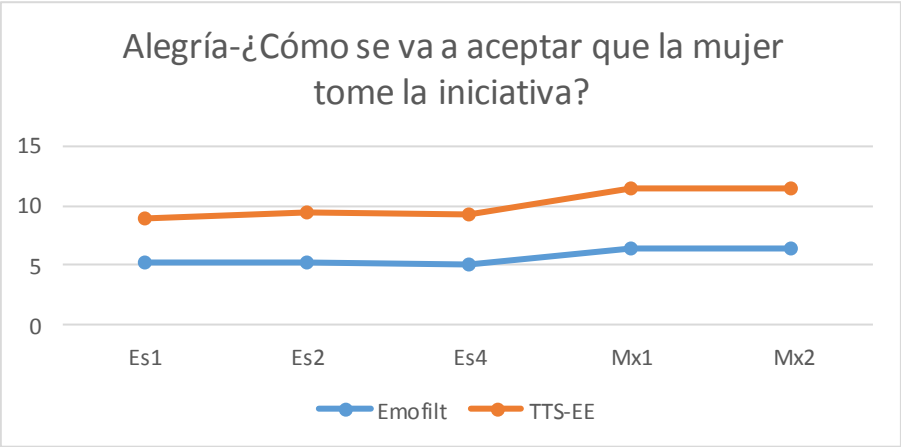


Figura 54. ¿Cómo se va a aceptar que la mujer tome la iniciativa? Mx2 alegría TTS-EE.

La *Tabla 30* muestra el resumen de estos resultados, una comparación entre ***Emofilt*** y el ***TTS-EE*** para la emoción ***alegría***, con todas las voces y la diferencia entre estos. Y lo que se refleja en la Gráfica 15 es que aunque las cifras no son iguales los comportamientos de la voz son similares.

Alegría-¿Cómo se va a aceptar que la mujer tome la iniciativa?			
Voz	**Emofilt**	**TTS-EE**	**Diferencia**
Es1	5.15s	3.71s	1.44s
Es2	5.14s	4.22s	0.92s
Es4	5.09s	4.16s	0.93s
Mx1	6.38s	5.13s	1.25s
Mx2	6.38s	5.12s	1.26s

Tabla 30. Comparación Alegría.



Gráfica 15. Alegría ¿Cómo se va a aceptar que la mujer tome la iniciativa?

Siguiendo con la misma frase *"¿Cómo se va a aceptar que la mujer tome la iniciativa?"*, pero ahora con la emoción ***neutral*** de español para la voz ***Es1*** generada de ***Emofilt***, la duración es de ***4.14s.***, véase *Figura 55*.

Y para ***neutral*** de la voz ***Es1*** generada del ***TTS-EE***, la duración es de ***4.04s.*** Con una diferencia de ***0.1s.,*** véase *Figura 56.*

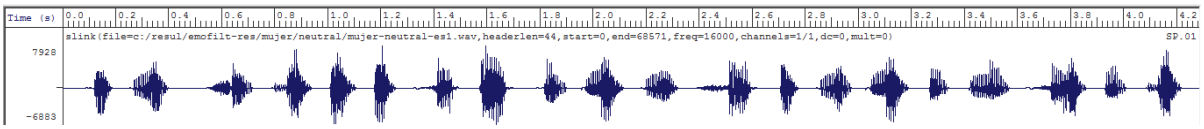


Figura 55. ¿Cómo se va a aceptar que la mujer tome la iniciativa? Es1 neutral Emofilt.

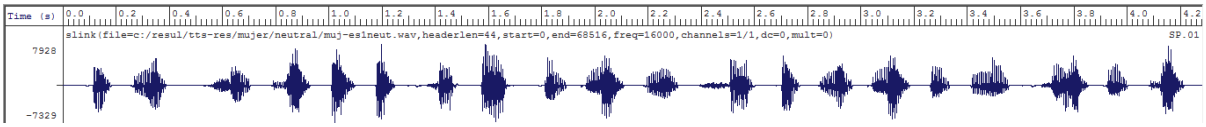


Figura 56. ¿Cómo se va a aceptar que la mujer tome la iniciativa? Es1 neutral TTS-EE.

La *Figura 57* muestra la señal de onda de la frase con emoción ***neutral*** de la voz ***Es2*** generada de ***Emofilt***, la duración es de ***4.15s.***

Figura 57. ¿Cómo se va a aceptar que la mujer tome la iniciativa? Es2 neutral Emofilt.

La señal de onda con emoción ***neutral***, de voz ***Es2*** generada del ***TTS-EE*** es mostrada en la *Figura 58*, teniendo una duración de ***4.15s***.

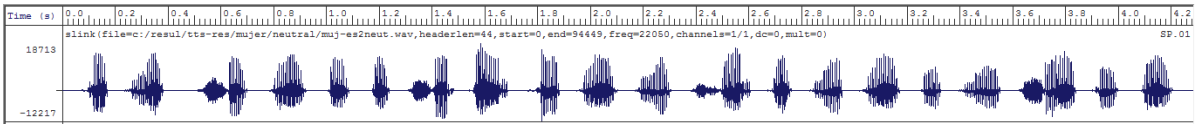


Figura 58. ¿Cómo se va a aceptar que la mujer tome la iniciativa? Es2 neutral TTS-EE.

Para la misma frase pero con la voz ***Es4*** generada por ***TTS-EE*** para la emoción ***neutral*** la duración es de ***4.13s.*** Ver *Figura 59.*

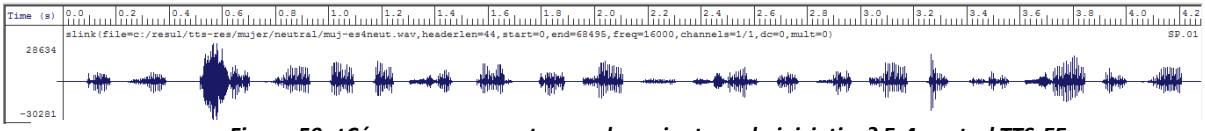


Figura 59. ¿Cómo se va a aceptar que la mujer tome la iniciativa? Es4 neutral TTS-EE.

Ahora pasando a la voz ***Mx1***, ***neutral*** generada por ***Emofilt*** tiene la duración de ***5.38s***., y se puede ver en la *Figura 60.*

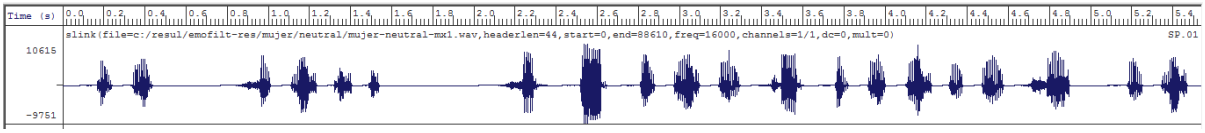


Figura 60. ¿Cómo se va a aceptar que la mujer tome la iniciativa? Mx1 neutral Emofilt.

Y para ***TTS-EE***, de la misma voz ***Mx1*** con emoción ***neutral*** con una duración de ***5.31s.*** Con una diferencia de ***0.07s.,*** ver *Figura 61.*

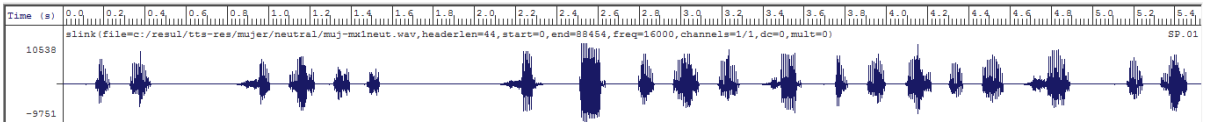


Figura 61. ¿Cómo se va a aceptar que la mujer tome la iniciativa? Mx1 neutral TTS-EE.

Ahora para la voz ***Mx2*** con la emoción ***neutral*** generada por ***Emofilt,*** la duración es de ***5.38s.***, véase la *Figura 62.*

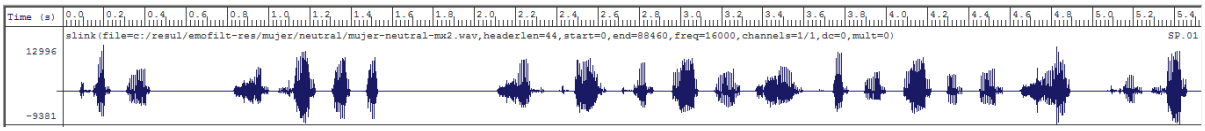


Figura 62. ¿Cómo se va a aceptar que la mujer tome la iniciativa? Mx2 neutral Emofilt.

Y finalizando con la emoción ***neutral*** de la voz ***Mx1*** generada con el ***TTS-EE*** tiene una duración de ***5.4s.***, y su diferencia es de ***0.02s.***, véase *Figura 63.*

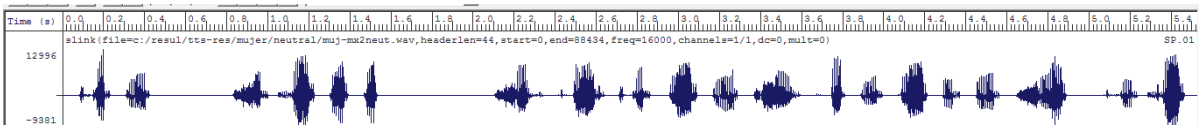
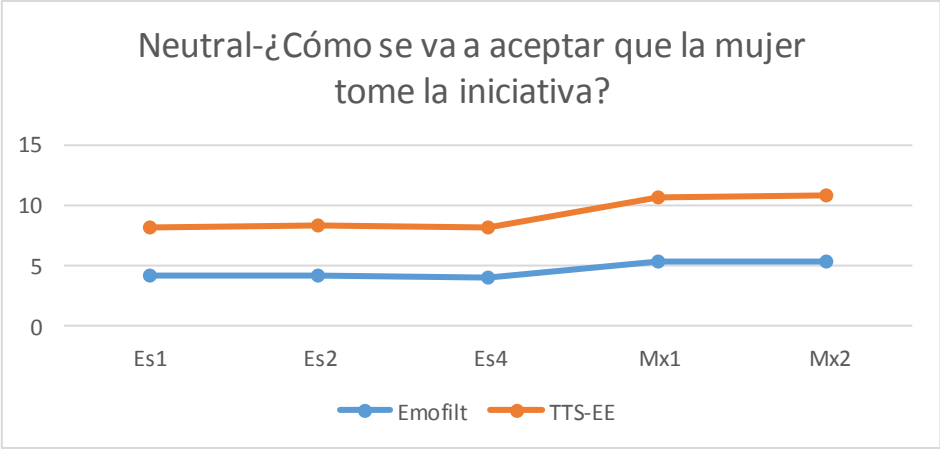


Figura 63. ¿Cómo se va a aceptar que la mujer tome la iniciativa? Mx2 neutral TTS-EE.

El resumen de estos datos se encuentra en la *Tabla 31*, para la emoción ***neutral***, de todas las voces desde ***Es1*** a ***Mx2***, la comparación entre ***Emofilt*** y el ***TTS-EE***. La *Gráfica 16* ayuda a visualizar que el comportamiento de las voces es muy similar y las diferencias son mínimas en la emoción neutral.

Neutral-¿Cómo se va a aceptar que la mujer tome la iniciativa?			
Voz	**Emofilt**	**TTS-EE**	**Diferencia**
Es1	4.14s	4.04s	0.1s
Es2	4.15s	4.15s	0s
Es4	4.06s	4.13s	0.07s
Mx1	5.38s	5.31s	0.07s
Mx2	5.38s	5.4s	0.02s

Tabla 31. Comparación Neutral.



Gráfica 16. Neutral ¿Cómo se va a aceptar que la mujer tome la iniciativa?

Finalmente se muestran los resultados para esta frase con la emoción tristeza y con las diferentes voces del español.

La *Figura 64*, muestra la frase con emoción ***tristeza*** con la voz ***Es1*** generada por ***Emofilt***, la duración es de ***8s.***

Luego en la *Figura 65* se puede notar la misma voz con ***tristeza*** pero ahora generada con el ***TTS-EE***, la duración de ***5.72s.***, con una diferencia de ***2.28s.***

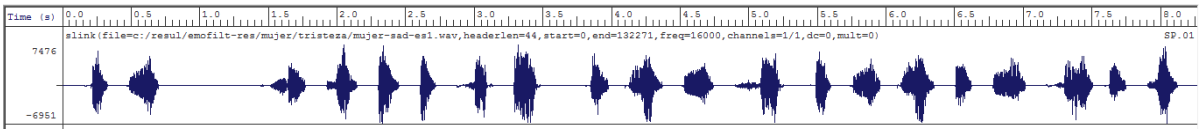


Figura 64. ¿Cómo se va a aceptar que la mujer tome la iniciativa? Es1 tristeza Emofilt.

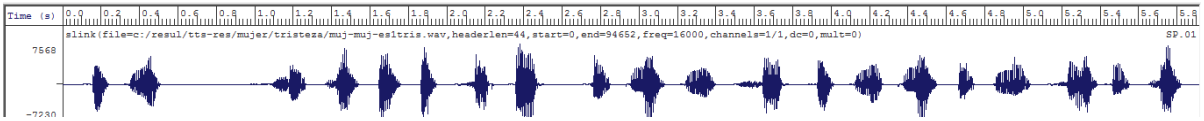


Figura 65. ¿Cómo se va a aceptar que la mujer tome la iniciativa? Es1 tristeza TTS-EE.

Para ***Es2*** de ***tristeza*** generada a partir de ***Emofilt***, la duración es de ***7.95s.*** y la contraparte resultante del ***TTS-EE*** es de ***5.72s.,*** con una diferencia entre ellos de ***2.23s.***, se pueden observar estas señales de onda en las *Figuras 66 y 67.*

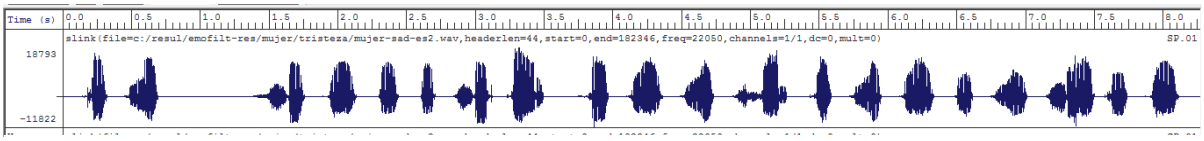


Figura 66. ¿Cómo se va a aceptar que la mujer tome la iniciativa? Es2 tristeza Emofilt.

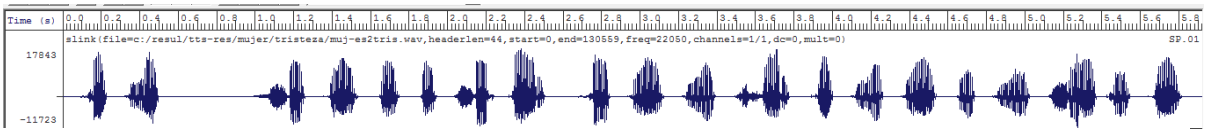


Figura 67. ¿Cómo se va a aceptar que la mujer tome la iniciativa? Es2 tristeza TTS-EE.

La *Figura 68* muestra la señal de onda de la frase para la voz ***Es4*** de la emoción ***tristeza*** con una duración de ***5.74s.***

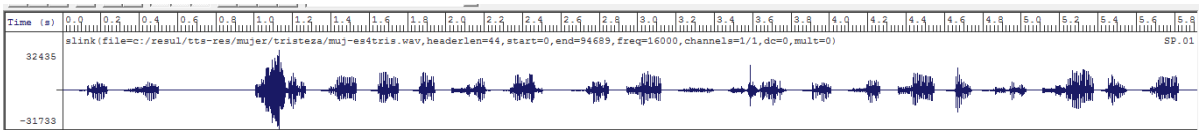


Figura 68. ¿Cómo se va a aceptar que la mujer tome la iniciativa? Es4 tristeza TTS-EE.

Para la voz ***Mx1*** con ***tristeza*** de ***Emofilt*** es de ***8.45s.***, y la generada por ***TTS-EE*** es de ***6.04s***., la diferencia entre estas es de ***2.41s.,*** se puede observar esto en la *Figura 69* y la *Figura 70.*

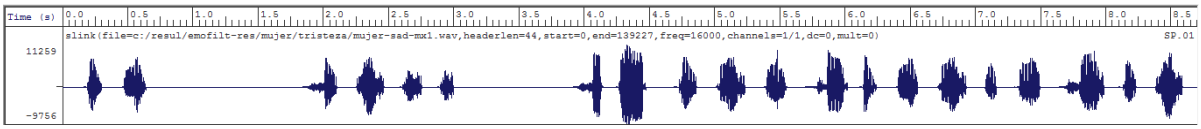


Figura 69. ¿Cómo se va a aceptar que la mujer tome la iniciativa? Mx1 tristeza Emofilt.

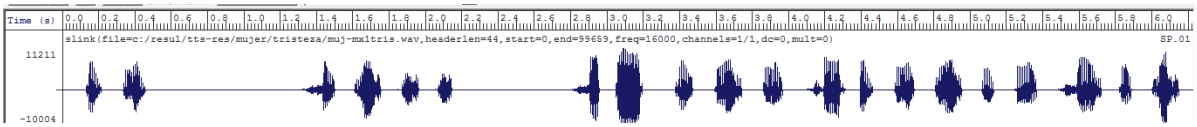


Figura 70. ¿Cómo se va a aceptar que la mujer tome la iniciativa? Mx1 tristeza TTS-EE.

Para la siguiente voz ***Mx2*** generada de ***Emofilt*** la duración es de 8.59s, se puede notar en la *Figura 71.*

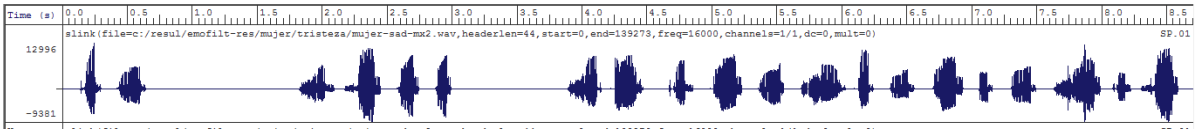


Figura 71. ¿Cómo se va a aceptar que la mujer tome la iniciativa? Mx2 tristeza Emofilt.

Y finalmente la voz ***Mx2*** pero ahora generada por el ***TTS-EE,*** su duración es de ***6.06s.***, con la diferencia de ***2.53s.,*** véase *Figura 72.*

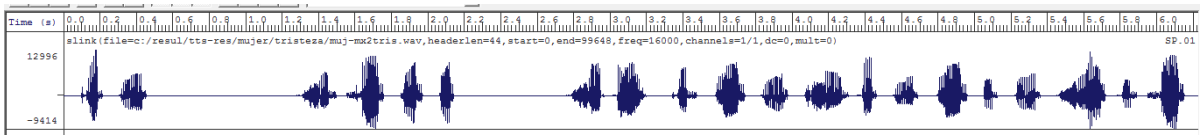
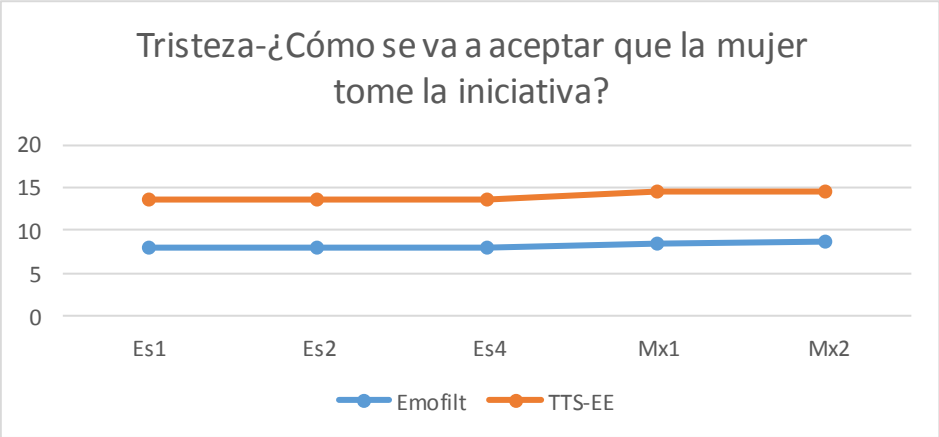


Figura 72. ¿Cómo se va a aceptar que la mujer tome la iniciativa? Mx2 tristeza TTS-EE.

Finalmente en la *Tabla 32* se pueden ver estos resultados, las diferencias en duración tanto de ***Emofilt*** como de ***TTS-EE*** entre las diferentes voces para la emoción ***tristeza. La*** *Gráfica 17* muestra una tendencia similar entre Emofilt y el TTS-EE aunque con una diferencia ligeramente mayor que las emociones anteriores.

Tristeza-¿Cómo se va a aceptar que la mujer tome la iniciativa?			
Voz1	**Emofilt**	**TTS-EE**	**Diferencia**
Es1	8	5.72	2.28s
Es2	7.95	5.72	2.23s
Es4	7.88	5.74	2.14s
Mx1	8.45	6.04	2.41s
Mx2	8.59	6.06	2.53s

Tabla 32. Comparación Tristeza.



Gráfica 17. Tristeza ¿Cómo se va a aceptar que la mujer tome la iniciativa?

Ahora se analiza la palabra “Siete”, también para las emociones alegría, neutral y tristeza, para las voces de español de España Es1 Es2, Es4, y de español de México Mx1 y Mx2.

La *Figura 73* muestra la palabra ***Siete*** de la voz ***Es1 alegría*** generado por ***Emofilt***, la duración es de ***465ms.***

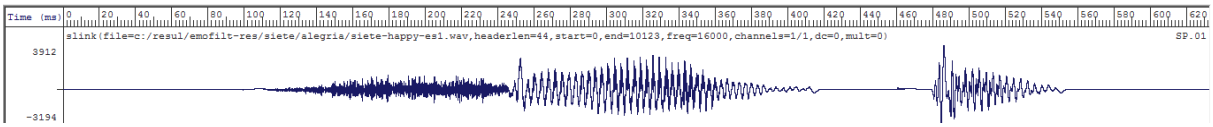


Figura 73. Siete Es1 alegría Emofilt.

La *Figura 74* muestra la misma palabra para la voz ***Es1***, de ***alegría*** generado por el ***TTS-EE*** con una duración de ***350ms.*** La diferencia es de ***106ms.***

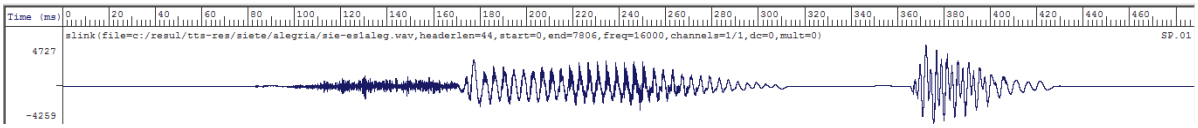


Figura 74. Siete Es1 alegría TTS-EE.

La siguiente *Figura 75* muestra la señal de onda para la palabra "*Siete"* para la voz ***Es2*** de ***alegría*** generada de ***Emofilt***, la duración es ***461ms.***

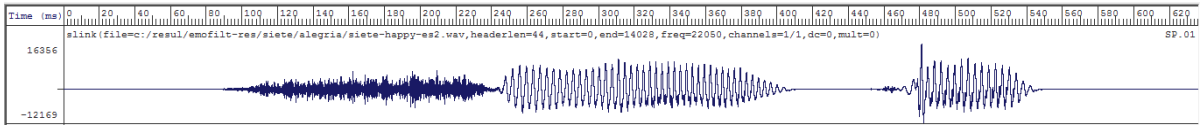


Figura 75. Siete Es2 alegría Emofilt.

Para la voz ***Es2*** de ***alegría*** generada por el ***TTS-EE***, la duración es de ***367ms,*** la diferencia es de ***94ms.*** Véase *Figura 76.*

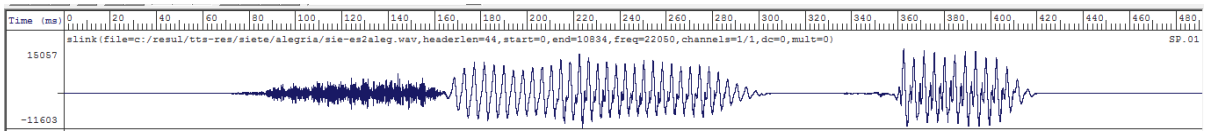


Figura 76. Siete Es2 alegría TTS-EE.

La voz ***Es4*** de ***alegría*** generada por el ***Emofilt*** tiene una duración ***466ms.*** Ver *Figura 77.*

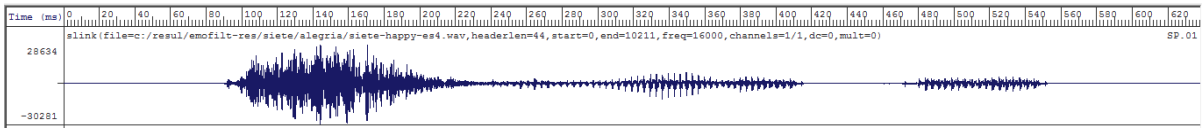


Figura 77. Siete Es4 alegría Emofilt.

La *Figura 78* para la voz ***Es4 alegría*** genera ***TTS-EE*** la duración es de ***351ms.*** La diferencia es de ***115ms.***

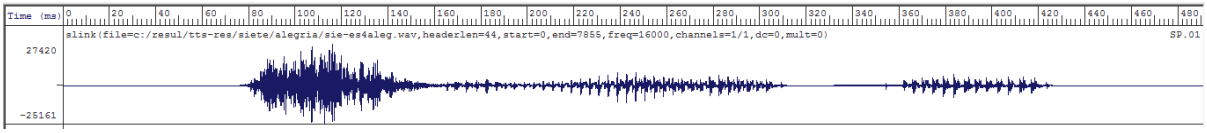


Figura 78. Siete Es4 alegría TTS-EE.

La señal de voz de ***alegría*** para ***Mx1*** del ***Emofilt*** es de ***601ms.*** Ver *Figura 79*. Para la misma voz y de la misma emoción pero generada por ***TTS-EE 432ms.*** La diferencia es de ***169ms.*** Véase la *Figura 80.*

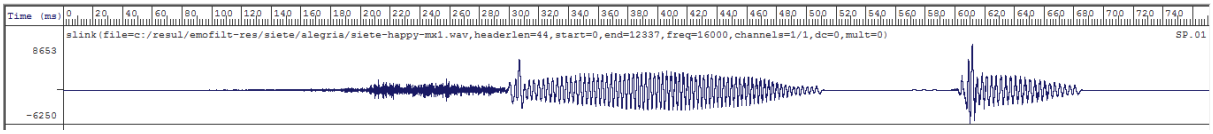


Figura 79. Siete Mx1 alegría Emofilt.

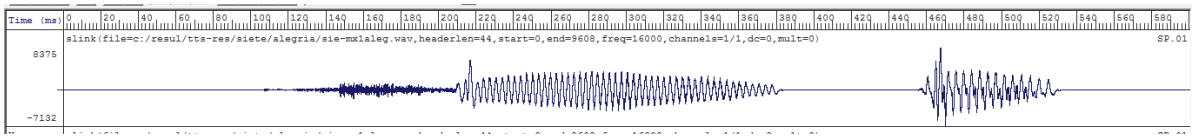


Figura 80. Siete Mx1 alegría TTS-EE.

Para la misma palabra "*Siete"* de la voz ***Mx2*** de ***Emofilt*** la duración es de ***615ms.*** y para ***Mx2*** de ***alegría*** de ***TTS-EE*** la duración es de ***438ms.*** y la diferencia es de ***177ms.*** *Véase Figura 81 y 82.*

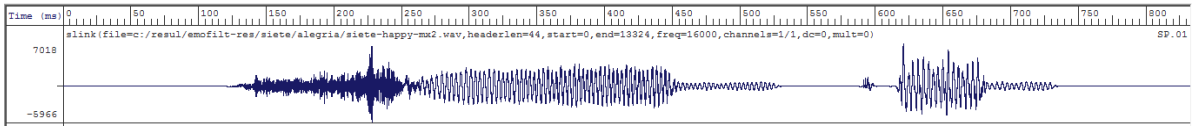


Figura 81. Siete Mx2 alegría Emofilt.

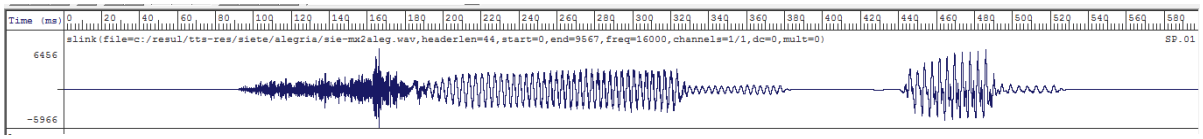
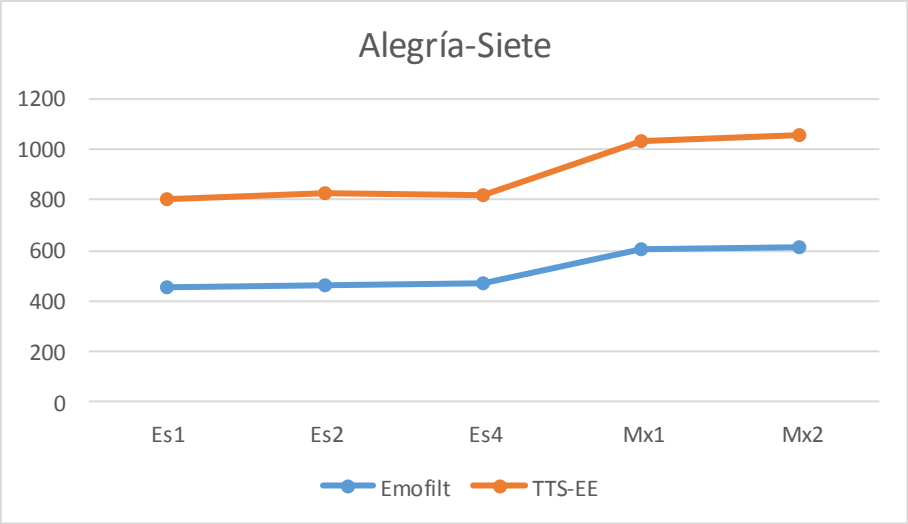


Figura 82. Siete Mx2 alegría TTS-EE.

La siguiente *Tabla 33* tiene los resultados de la palabra *"Siete"* de la emoción ***alegría*** de las voces ***Es1, Es2, Es4, Mx1 y Mx2,*** para el sintetizador ***Emofilt*** y el ***TTS-EE.*** Que se puede ver una similitud en la *Gráfica 18*.

Alegría-Siete			
Voz	**Emofilt**	**TTS-EE**	**Diferencia**
Es1	456ms.	350ms.	106ms.
Es2	461ms.	367ms.	94ms.
Es4	466ms.	351ms.	115ms.
Mx1	601ms.	432ms.	169ms.
Mx2	615ms.	438ms.	177ms.

Tabla 33. Comparación Alegría-Siete.



Gráfica 18. Alegría-Siete.

La siguiente *Figura 83*, muestra una señal de onda de la palabra "*Siete"* de la emoción ***neutral***, de la voz ***Es1*** resultante de ***Emofilt***, la duración es de ***355ms.***

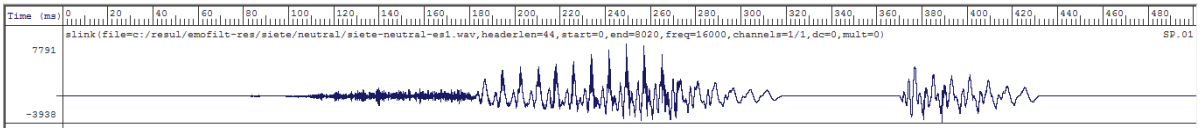


Figura 83. Siete Es1 neutral Emofilt.

Y para la señal de onda generada por el ***TTS-EE*** para la voz ***Es1***, ***neutral*** de la misma palabra. Tiene una duración de ***360ms***. Con una diferencia de ***10ms.***Ver *Figura 84.*

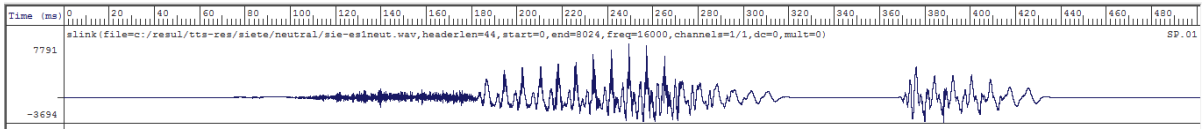


Figura 84. Siete Es1 neutral TTS-EE.

En la *Figura 85* se muestra la señal de onda de "*Siete"* de ***Es2*** de la emoción ***neutral*** generada por ***Emofilt*** con una duración de ***352ms*** y la siguiente *Figura 86* es para la voz generada por ***TTS-EE*** la duración es la misma con una obvia diferencia de ***0ms.***

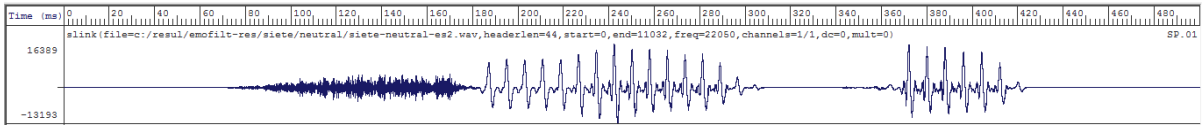


Figura 85. Siete Es2 neutral Emofilt.

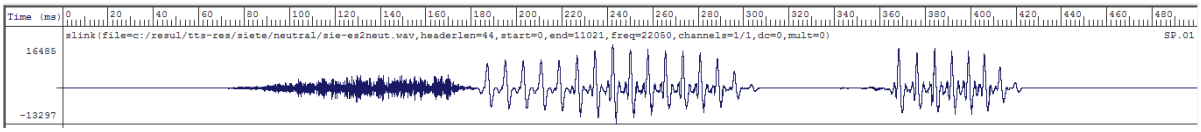


Figura 86. Siete Es2 neutral TTS-EE.

En la *Figura 87* se muestra la señal de onda de "*Siete*" para la voz ***Es4*** de la emoción ***neutral*** obtenida del ***Emofilt*** con una duración de ***350ms.*** y para la señal de voz obtenida del ***TTS-EE*** la duración es de ***352ms.*** con una diferencia de ***2ms.*** Ver *Figura 88*.

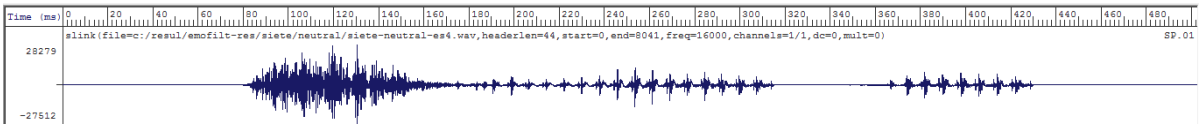


Figura 87. Siete Es4 neutral Emofilt.

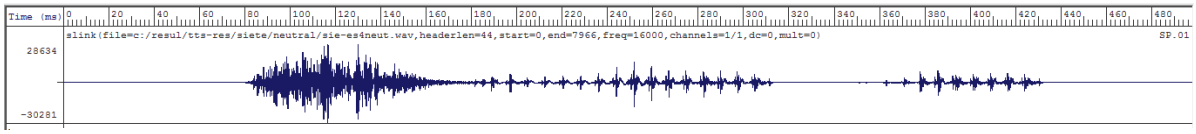


Figura 88. Siete Es4 neutral TTS-EE.

La señal de onda siguiente es para la palabra *"Siete"* de la voz ***Mx1*** para la emoción neutral resultante de ***Emofilt*** la duración es de ***482ms***. y para el ***TTS-EE*** la duración es ***84ms.*** con una diferencia de ***2ms.*** Véanse *Figuras 89* y *90.*

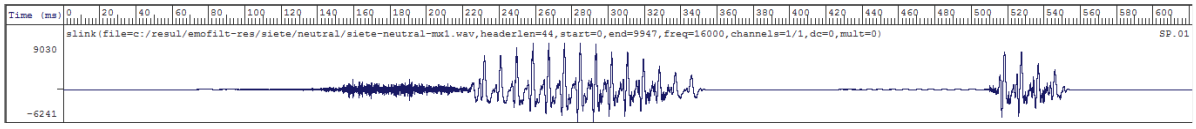


Figura 89. Siete Mx1 neutral Emofilt.

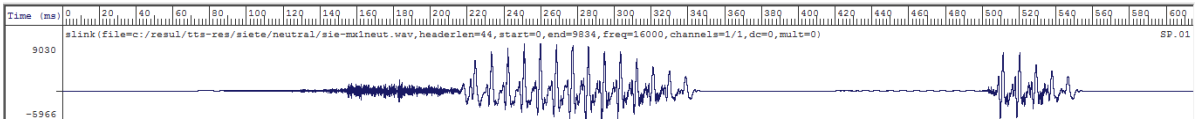


Figura 90. Siete Mx1 neutral TTS-EE.

Finalmente para esta misma palabra para la emoción ***neutral*** con la voz ***Mx2*** generada de ***Emofilt*** la duración es de ***462ms.*** y la generada por el ***TTS-EE*** es la misma duración donde no hay diferencia. Véase *Figura 91* y *92.*

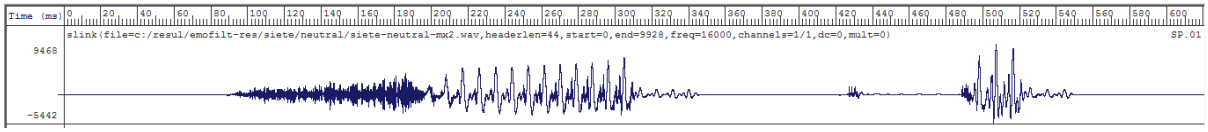


Figura 91. Siete Mx2 neutral Emofilt.

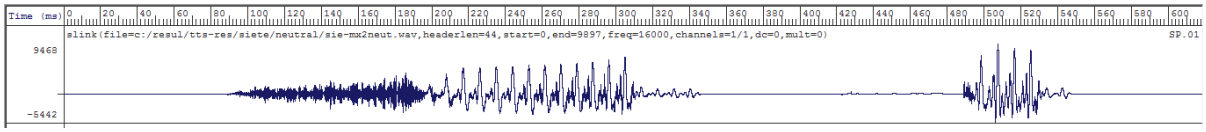
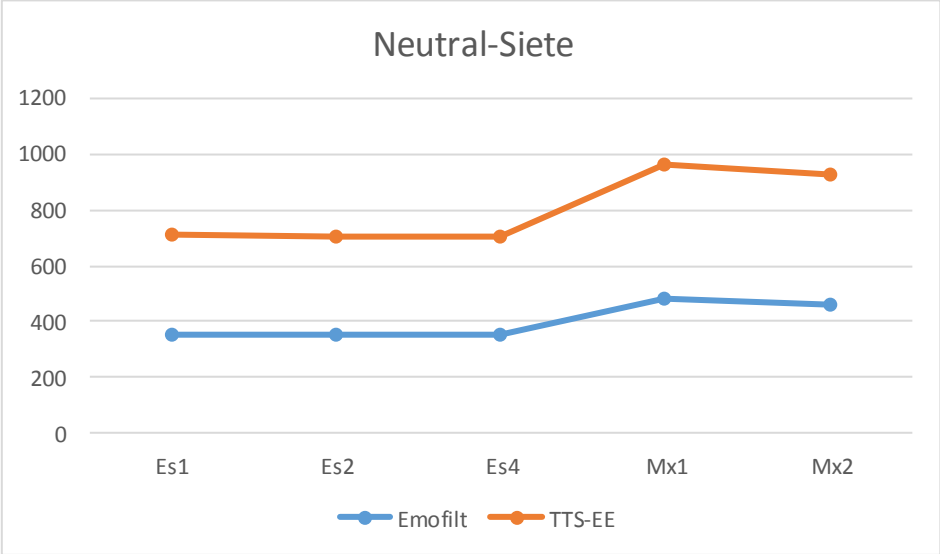


Figura 92. Siete Mx2 neutral TTS-EE.

La *Tabla 34* muestra todos estos resultados entre *Emofilt* y el *TTS-EE* para todas las voces del español de España y México utilizadas para la emoción ***neutral.*** Que tienen una tendencia similar y una mínima diferencia que se observa en la *Gráfica 19.*

Siete-Neutral			
Siete	**Emofilt**	**TTS-EE**	**Diferencia**
Es1	350ms.	360ms.	10ms.
Es2	352ms.	352ms.	0ms.
Es4	350ms.	352ms.	2ms.
Mx1	482ms.	484ms.	2ms.
Mx2	462ms.	462ms.	0ms.

Tabla 34. Comparación Neutral-Siete.



Gráfica 19. Neutral-Siete.

Ahora se muestra la señal de onda de la palabra "*Siete*" con la voz ***Es1*** para la emoción ***tristeza*** generada por ***Emofilt*** con una duración de ***610ms.*** Véase *Figura 93*.

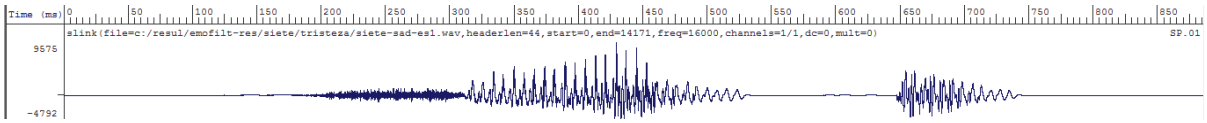


Figura 93. Siete Es1 tristeza Emofilt.

La *Figura 94* muestra la señal de la misma palabra para la emoción ***tristeza***, la voz ***Es1*** obtenida del ***TTS-EE*** con una duración de ***414ms.*** y una diferencia con el anterior de ***196ms.***

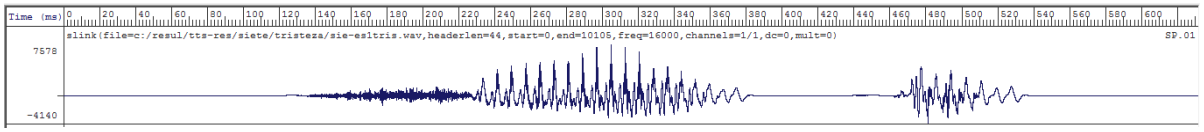


Figura 94. Siete Es1 tristeza TTS-EE.

Ahora para la misma palabra pero para la voz ***Es2***, de la emoción ***tristeza*** obtenida del ***Emofilt*** la duración es de ***595ms.*** Ver *Figura 95*.

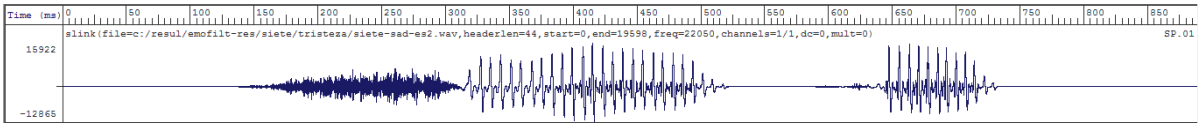


Figura 95. Siete Es2 tristeza Emofilt.

La *Figura 96* muestra la señal de onda generada por el ***TTS-EE*** de la misma voz y emoción que la anterior con una duración de ***424ms.***, con una diferencia de ***171ms.***

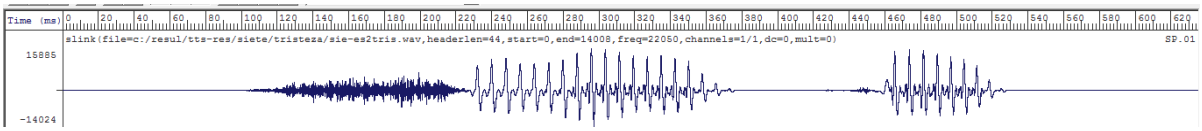


Figura 96. Siete Es2 tristeza TTS-EE.

La señal de onda para la palabra *"Siete"* para la voz ***Es4*** con emoción ***tristeza*** generada por el ***Emofilt*** tiene una duración de ***590ms.*** Ver *Figura 97.*

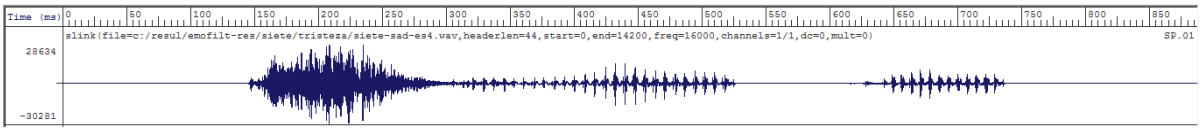


Figura 97. Siete Es4 tristeza Emofilt.

Para la misma voz y emoción pero generada con el ***TTS-EE*** la duración es de ***424ms.*** con una diferencia de ***166ms.*** Véase *Figura 98.*

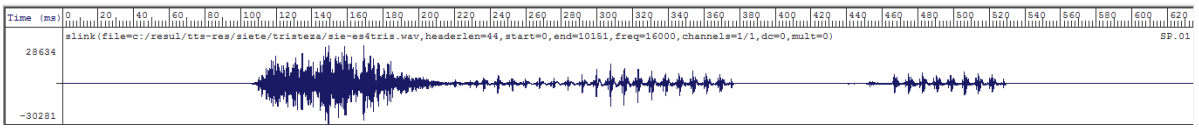


Figura 98. Siete Es4 tristeza TTS-EE.

Ahora para la voz ***Mx1*** con emoción ***tristeza*** generada por ***Emofilt*** la duración es de ***590ms.*** Ver *Figura 99.*

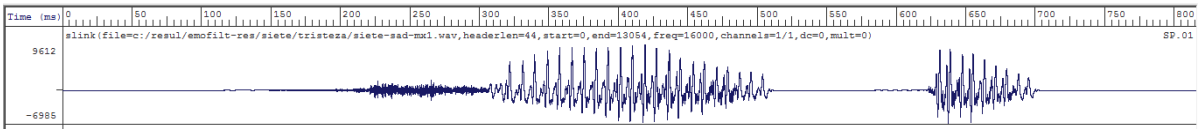


Figura 99. Siete Mx1 tristeza Emofilt.

Y la señal de onda generada por el ***TTS-EE*** tiene una duración de ***389ms.*** Con una diferencia del anterior de ***201ms.*** Ver *Figura 100.*

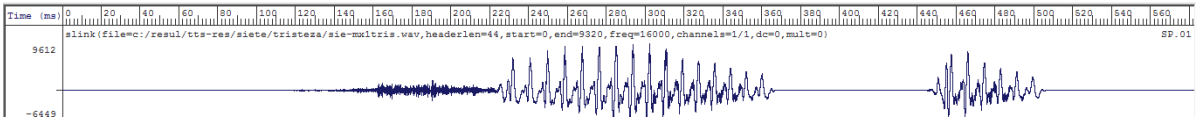


Figura 100. Siete Mx1 tristeza TTS-EE.

La *Figura 101* muestra la señal de onda para la palabra *"Siete"* con la voz ***Mx2*** con emoción ***tristeza*** que tiene una duración de ***408ms.***

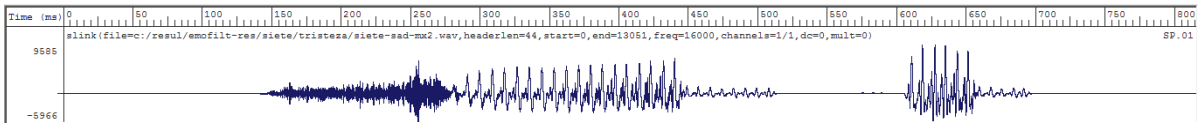


Figura 101. Siete Mx2 tristeza Emofilt.

Y en la *Figura 102* se muestra la misma palabra con la emoción ***tristeza*** y la voz ***Mx2*** con una duración de ***555ms.*** y una diferencia de ***147ms.***

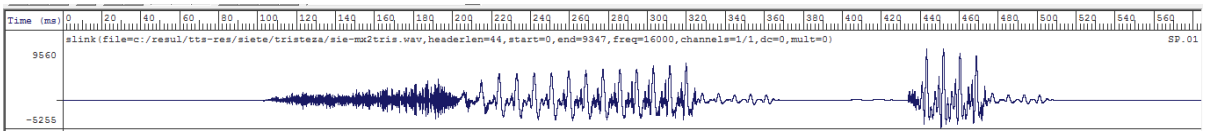
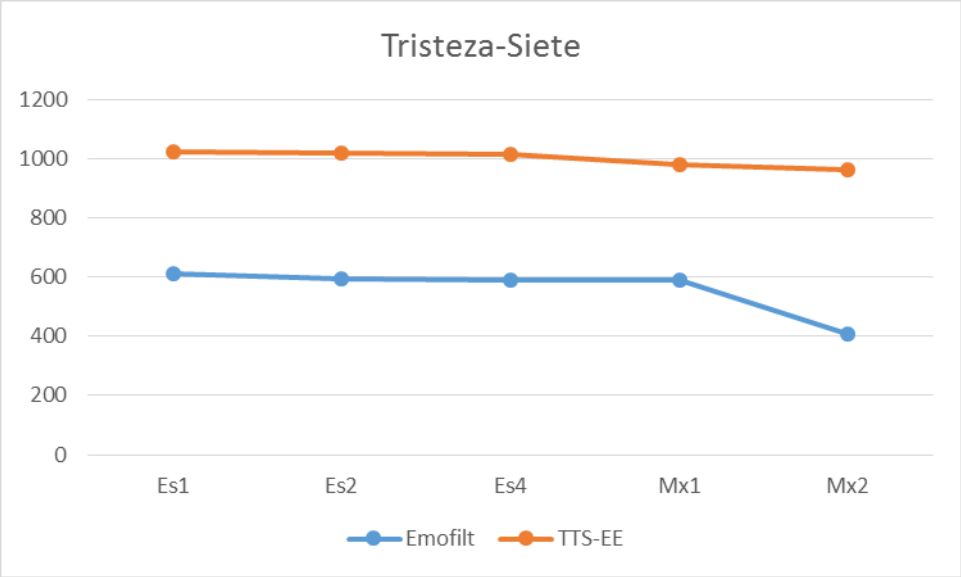


Figura 102. Siete Mx2 tristeza TTS-EE.

La *Tabla 35* muestra los resultados de la palabra *"Siete"* con la emoción ***tristeza*** para las voces del español de España y México entre el sintetizador ***Emofilt*** y el ***TTS-EE.*** Gráficamente se observa una tendencia similar en la *Grafica 20* aunque con una pequeña diferencia.

Tristeza-Siete			
Voz	**Emofilt**	**TTS-EE**	**Diferencia**
Es1	610	414	196ms
Es2	595	424	171ms
Es4	590	424	166ms
Mx1	590	389	201ms
Mx2	408	555	147ms

Tabla 35. Comparación Tristeza-Siete.



Gráfica 20. Tristeza-Siete.

7. Discusión sobre el TTS-EE

En el siguiente apartado se analiza y se discute los resultados obtenidos en la sección anterior, con el objetivo de contrastar las hipótesis de trabajo presentadas en el apartado 3.3.

H1. SAMPA es el mejor alfabeto fonético para realizar la transcripión ortográfica-fonética ya que es comprensible para la PC, su simplicidad en la transcripción y la relativa facilidad de uso por parte de personas con poca formación fonética a diferencia de AFI.

Es cierta ya que para generar el Sistema TTS-EE no hubo ninguna dificultad al aplicar SAMPA en código Python en la tabla 36 se puede ver la transcripción en la columna TTS-EE-México donde no se observa algún carácter extraño a la PC.

H2. Las 32 reglas de transcripción son las que se utilizan para la conversión de grafemas a fonemas aplicadas en cierto orden darán una transcripción correcta 32 por cada grafema.

Para la parte de la transcripción ortográfica-fonética el TTS-EE se aplican las 32 reglas de transcripción del alfabeto fonético SAMPA se puede observar una correcta transcripción con la siguiente *Tabla 36* que es una comparativa entre los ejemplos del alfabeto fonético SAMPA y los resultados del TTS-EE por lo que se cumple la segunda hipótesis.

GRAFEMA	Ejemplo SAMPA	TTS-EE-España	Ejemplo SAMPA	TTS-EE-México
A	ala: "Al**a**	ala: A la	azúcar: **a**sUkar	azúcar: a sU kar
B	comba: "kom**b**a labio: "la**B**jo	comba: kom bA labio: lA bio	barco: **b**Arko	barco: bAr ko
C	celo: "**T**elo acné: a**G**"ne tacto:"ta**k**to coro: "**k**oro	celo: TE lo acné: ak nE tacto: tAk to coro: kO ro	celo: **s**Elo acné: a**k**nE coro: **k**Oro	sE lo ak nE kO ro

	tecla: "te**k**la	tecla: tE kla		
Ch	chelo: "**tS**elo	chelo: tSE lo	chelo: **tS**Elo	tSE lo
D	caldo:"kal**d**o codo:"ko**D**o	caldo: kAl do codo: kO do	caldo: kAl**d**o	kAl do
E	elefante: **e**lefAnte	elefante: e le fAn te	elefante: **e**lefAnte	e le fAn te
F	cofia: "ko**f**ja	cofia: kO fia	cofia: kO**f**ia	kO fia
G	tongo: "ton**g**o genio: "**x**enjo tigre: "ti**G**re lago: "la**G**o	tongo: tOn go genio:xe niO ti grE la gO	golosina: **g**olosina digno: dI**g**no guitarra: **g**uitarra antigüedad: anti**g**wedAd girasol: **x**irasOl	go lo sI na dIg no gi tA rra an ti gwe dAd xi ra sOl
H	hierba: "**jj**erBa halo: "alo	jjer bA a lO	hola: Ola hierba: **jj**Erba	O la jjEr ba
I	tipo: "t**i**po cielo: "T**j**elo	ti pO Tie lO	isla: **I**sla	Is la
J	jarana: "**j**arana	xa ra nA	jalea: **x**alEa	xa le A
K	kiosko: "**k**josko	kios kO	kiosko: **k**iOsko	kios kO
L	lote: "**l**ote	lo tE	leer: **l**eEr	le Er
Ll	tallo: "ta**L**o	ta LO	llanta: **dZ**Anta	dZAn ta
M	arma: "ar**m**a	ar mA	mamá: **m**amA	ma mA
N	ánfora: "a**m**fora cono: "ko**n**o	Am fo ra ko nO	natividad: **n**atibidAd	na ti bi dAd
Ñ	uña: "u**J**a	u JA	niña: nI**J**a	ni JA
O			oro: Or**o**	o rO
P	perro: "**p**erro	pe rrO	papá: **p**apA	pa pA
Q	queso: "**k**eso	ke sO	queso: **k**Eso quo: **k**uo	ke sO qwo
R	rama: "**rr**ama honra: "on**rr**a arpa: "a**r**pa trampa: "t**r**ampa pera: "pe**r**a amor: a"mo**r**	rra mA on rrA ar pA tram pA pe rA a mOr	rosa: **rr**Osa honra: On**rr**a	rro sA On rra
Rr	carro: "ka**rr**o	ka rrO	perro: pE**rr**o	pe rrO
S	rasgo: "rra**s**Go casa: "ka**s**a trasto: "tra**s**to	rras gO ka sA tras tO	sauce: **s**Awse	saw sE

T	atleta: a**D**"leta toro: "**t**oro	a tle tA to rO	tucán: **t**ukAn	tu kAn
U	queso: "keso cigüeña: Ti"G**w**eJa lujo: "l**u**xo	ke sO Ti gwe JA lu xO	agüita: ag**w**lta puerto: p**w**Er to apúrate: ap**U**rate	a gwi tA pwer tO a pU ra te
V	con velo: kom "**b**elo calvo: "kal**B**o	kOn be lO kal bO	vaca: **b**Aka	ba kA
W	whisky: "**gw**iski kiwi: ki**B**i	guIs ki ki gwI	waffle: gwAffle	gwaf flE
X	examen: e**G**"**s**amen externo: e**s**"terno	e ksa mEn eks ter nO	examen: e**ks**Amen relax: rrelA**ks** xoconostle: **s**okonOstle México: mE**x**iko	e ksa mEn rre lAks so ko nos tlE mE xi ko
Y	yunque: "**jj**unque cónyugue: "kon**jj**uGe dos y dos: dos **i** "Dos muy: mw**i**	jjun kE kOn jju ge dOs i dOs mwi	mui: mw**i** yoyo: **jj**Ojjo	mwI jjo jjO
Z	zarza: **T**ar**T**a tizne: "ti**T**ne	Tar TA tiT nE	zorro: **s**Orro	so rrO

Tabla 36. Comparación SAMPA y TTS-EE transcripción ortográfica fonética.

H3. El silabeo proporcionará una mayor naturalidad en el habla sintetizada de cualquier español debido a la coarticulación de las sílabas.

Para dar una mejor expresión en la emoción se aplica el silabeo en las palabras, que proporciona una mayor naturalidad en el habla sintetizada debido a la coarticulación de las sílabas, cubriendo la tercera hipótesis**.**

Como resultado del TTS-EE en la siguiente frase ejemplo ¿Cómo se va aceptar que la mujer tome la iniciativa?

El silabeo permite que no se lea de corrido la frase presentado una mayor naturalidad a la voz ya que es como normalmente hablamos ***kO mo sE bA a sep tAr kE lA mu xEr tO me lA i ni sia ti bA***, entre silabas existe una duración la cual esta aplicada por emoción de acuerdo a los resultados obtenidos en la segmentación automática y realizando un promedio. Para alegría el promedio de la duración es 81ms *Figura103*.

esalegria: Bloc de notas

Archivo Edición Formato Ver Ayuda

```
{'p':(23, 0, 217),'b':(35, 0, 214),'t':(21, 0, 242),'d':(30, 0, 224),\
'k':(22, 0, 218),'g':(41, 0, 220),'s':(50, 0, 225),\
'Q':(68, 0, 207),'f':(70, 0, 201),'B':(41, 0, 223),\
'T':(63, 0, 210),'D':(44, 0, 231),'s':(79, 0, 218),\
'x':(102, 0, 211),'G':(52, 0, 209),'m':(58, 0, 225),\
'n':(74, 0, 230),'J':(76, 0, 214),'l':(53, 0, 220),\
'L':(75, 0, 207),'r':(34, 0, 221),'R':(73, 0, 226),\
'j':(57, 0, 228),'w':(51, 0, 230),'i':(74, 0, 228),\
'e':(56, 0, 227),'a':(73, 0, 225),'o':(68, 0, 234),\
'u':(68, 0, 233),'A':(103, 0, 249),'E':(82, 0, 251),\
'I':(123, 0, 259),'o':(75, 0, 251),'U':(111, 0, 258),\
'_':(81,)}
```

Figura 103. Promedio duración y pitch alegría.

Para la emoción neutral el promedio de la duración son 83ms. *Figura 104*

esneutral: Bloc de notas

Archivo Edición Formato Ver Ayuda

```
{'p':(21, 0, 121),'b':(58, 0, 123),'t':(21, 0, 138),'d':(45, 0, 126),\
'k':(26, 0, 128),'g':(56, 0, 118),'s':(58, 0, 135),\
'Q':(74, 0, 123),'f':(82, 0, 120),'B':(44, 0, 119),\
'T':(81, 0, 120),'D':(47, 0, 123),'s':(92, 0, 123),\
'x':(110, 0, 127),'G':(54, 0, 117),'m':(60, 0, 122),\
'n':(72, 0, 123),'J':(77, 0, 123),'l':(51, 0, 119),\
'L':(76, 0, 119),'r':(39, 0, 122),'R':(75, 0, 123),\
'j':(58, 0, 126),'w':(54, 0, 124),'i':(70, 0, 124),\
'e':(55, 0, 124),'a':(68, 0, 121),'o':(66, 0, 126),\
'u':(62, 0, 127),'A':(83, 0, 122),'E':(71, 0, 127),\
'I':(96, 0, 129),'o':(61, 0, 127),'U':(89, 0, 125),\
'_':(83,)}
```

Figura 104. Promedio duración y pitch neutral.

Y la emoción tristeza el promedio de la duración son 115ms. Figura

estristeza: Bloc de notas

Archivo Edición Formato Ver Ayuda

```
{'p':(31, 0, 131),'b':(54, 0, 120),'t':(29, 0, 137),'d':(43, 0, 123),\
'k':(35, 0, 129),'g':(39, 0, 112),'s':(60, 0, 148),\
'Q':(82, 0, 117),'f':(95, 0, 126),'B':(51, 0, 120),\
'T':(90, 0, 125),'D':(51, 0, 121),'s':(104, 0, 120),\
'x':(136, 0, 130),'G':(59, 0, 121),'m':(82, 0, 120),\
'n':(87, 0, 122),'J':(92, 0, 127),'l':(62, 0, 118),\
'L':(77, 0, 115),'r':(49, 0, 123),'R':(91, 0, 124),\
'j':(70, 0, 125),'w':(61, 0, 124),'i':(82, 0, 126),\
'e':(62, 0, 122),'a':(75, 0, 121),'o':(72, 0, 123),\
'u':(71, 0, 130),'A':(103, 0, 125),'E':(82, 0, 126),\
'I':(123, 0, 133),'o':(75, 0, 128),'U':(111, 0, 126),\
'_':(115,)}
```

Figura 105. Promedios duración y pitch tristeza.

H4. Los mejores parámetros para dar expresión a la emoción se ven reflejados en la duración y el pitch del fonema y la pausa entre las sílabas.

Usando de referencia las Figuras 103, 104 y 105 se puede notar para cada fonema su duración y pitch que fueron extraidos de la segmentación automática fonética obteniendo un promedio por cada emoción y alimentan estos archivos para español de México y España.

Auditivamente al obtener una voz con emoción se aprecia la variación entre tristeza, alegría y neutral. El promedio de la duración de la tristeza se dispara a 82ms y neutral y tristeza se comportan directamente 78ms y 94ms respectivamente. Los promedio de cada fonema de su duración y pitch intevienen de manera importante para tener esta apreciasión los podemos observar en la figura 106, 107 y 108.

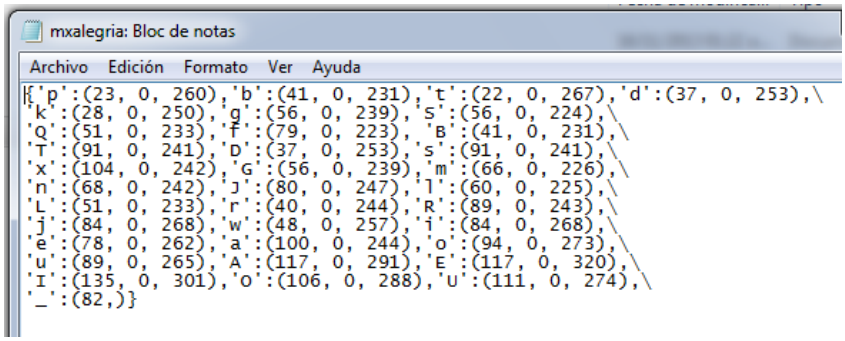


Figura 106. Promedios duración y pitch alegría MX.

Y la emoción tristeza el promedio de la duración son 115ms. Figura

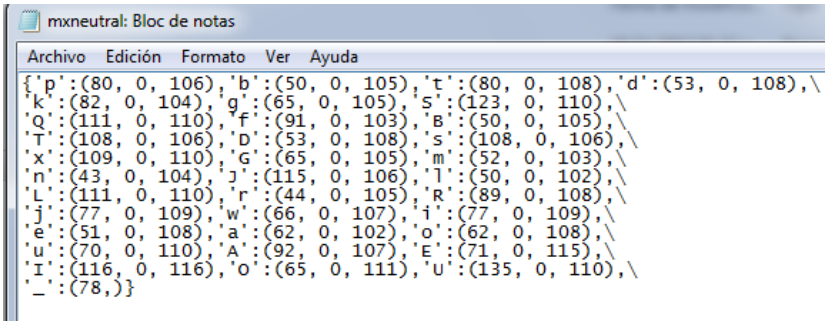


Figura 107. . Promedios duración y pitch neutral MX.

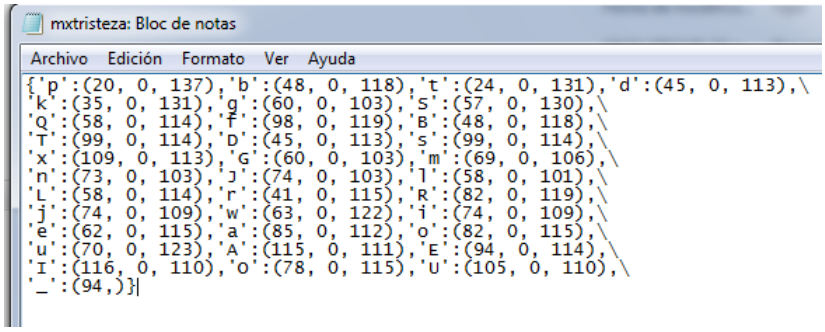


Figura 108. . Promedios duración y pitch tristeza MX.

El 70 por ciento de la transcripción hecha por el TTS-EE para español de España según los ejemplos es correcto el otro 30% tiene ligeras variaciones pero el 100% de las realizaciones es comprensible y perceptible audivamente en el TTS-EE. Esto se puede notar en la *Tabla 36* por ejemplo en comparativa con las columnas Ejemplo SAMPA y TTS-EE-México.

H5. MBROLA es el sintetizador más adecuado para trabajar el habla ya que cuenta con las voces necesarias de Español México y España para la investigación.

El sistema finalmente cuenta con MX1, MX2, ES1, ES2 y ES4 como lo muestra la Figura 109. Para cruzarlos con las tres emociones Neutral, Alegría y Tristeza.

Figura 109. TSS-EE.

H6. Las voces con emoción obtenidas a través del TTS-EE serán comparables con las emociones de Emofilt.

Como se puede observar en las tablas siguientes para la tabla 37 muestra en promedio la diferencia de 1.09s para español de España para sus 3 voces y 1.25s para español de México para sus dos voces para emoción alegría de una frase larga.

Alegría-¿Cómo se va a aceptar que la mujer tome la iniciativa?				
Voz	**Emofilt**	**TTS-EE**	**Diferencia**	**Promedio de diferencia**
Es1	5.15s	3.71s	1.44s	1.09s
Es2	5.14s	4.22s	0.92s	
Es4	5.09s	4.16s	0.93s	
Mx1	6.38s	5.13s	1.25s	1.25s
Mx2	6.38s	5.12s	1.26s	

Tabla 37. Alegría-¿Cómo se va a aceptar que la mujer tome la iniciativa?

Para la tabla 38 muestra en promedio la diferencia de 0.05s para español de España para sus 3 voces y 0.04s para español de México para sus dos voces para emoción Neutral de una frase larga.

Neutral-¿Cómo se va a aceptar que la mujer tome la iniciativa?				
Voz	**Emofilt**	**TTS-EE**	**Diferencia**	**Promedio de diferencia**
Es1	4.14s	4.04s	0.1s	0.05s
Es2	4.15s	4.15s	0s	
Es4	4.06s	4.13s	0.07s	
Mx1	5.38s	5.31s	0.07s	0.04s
Mx2	5.38s	5.4s	0.02s	

Tabla 38. Neutral-¿Cómo se va a aceptar que la mujer tome la iniciativa?

Para la tabla 39 muestra en promedio la diferencia de 2.21s para español de España para sus 3 voces y 2.47s para español de México para sus dos voces para emoción Tristeza de una frase larga.

Tristeza-¿Cómo se va a aceptar que la mujer tome la iniciativa?				
Voz1	**Emofilt**	**TTS-EE**	**Diferencia**	**Promedio de diferencia**
Es1	8	5.72	2.28s	2.21s
Es2	7.95	5.72	2.23s	
Es4	7.88	5.74	2.14s	
Mx1	8.45	6.04	2.41s	2.47
Mx2	8.59	6.06	2.53s	

Tabla 39. Tristeza-¿Cómo se va a aceptar que la mujer tome la iniciativa?

Para la tabla 40 muestra en promedio la diferencia de 105ms para español de España para sus 3 voces y 173ms para español de México para sus dos voces para emoción Alegría de una palabra.

Alegría-Siete				
Voz	**Emofilt**	**TTS-EE**	**Diferencia**	**Promedio de diferencia**
Es1	456ms.	350ms.	106ms.	105ms
Es2	461ms.	367ms.	94ms.	
Es4	466ms.	351ms.	115ms.	
Mx1	601ms.	432ms.	169ms.	173ms
Mx2	615ms.	438ms.	177ms.	

Tabla 40. Comparación Alegría-Siete.

Para la tabla 41 muestra en promedio la diferencia de 4ms para español de España para sus 3 voces y 1ms para español de México para sus dos voces para emoción Neutral de una palabra.

Siete-Neutral				
Siete	**Emofilt**	**TTS-EE**	**Diferencia**	**Promedio de diferencia**
Es1	350ms.	360ms.	10ms.	4ms
Es2	352ms.	352ms.	0ms.	
Es4	350ms.	352ms.	2ms.	
Mx1	482ms.	484ms.	2ms.	1ms
Mx2	462ms.	462ms.	0ms.	

Tabla 41. Comparación Neutral-Siete.

Para la tabla 42 muestra en promedio la diferencia de 177.66ms para español de España para sus 3 voces y 174ms para español de México para sus dos voces para emoción Tristeza de una palabra.

Tristeza-Siete				
Voz	**Emofilt**	**TTS-EE**	**Diferencia**	**Promedio de diferencia**
Es1	610	414	196ms	177.66ms
Es2	595	424	171ms	
Es4	590	424	166ms	
Mx1	590	389	201ms	174ms
Mx2	408	555	147ms	

Tabla 42. *Comparación Tristeza-Siete.*

Son muy cercanas en la emoción neutral respecto a sus promedios, la tristeza y alegría se encuentran más separadas respecto a sus promedios entre el TTS-EE y Emofilt. Aunque recordemos que el TTS-EE tiene el plus de la trascripción ortográfica-fonética.

8. Conclusiones y trabajo futuro

En el presente trabajo, se desarrolló un Sistema de Texto a Habla Expresivo en Español, existen en el sistema partes fundamentales y se tienen varias conclusiones.

La transcripción ortográfica-fonética entre SAMPA y el TTS para el español de España y de México es del 100% ya que se realizan las 32 reglas de transcripción.

De acuerdo a resultados las emociones son expresadas de manera correcta en el TTS respecto a Emofilt, gracias al silabeo, a la tasa de fonemas, a la duración en los fonemas, duración de pausas y pitch, para la emosión tristeza las diferencias para las voces finales tienen un promedio de 0.2s y 2.3s para palabras cortas y frases largas respectivamente, para la emoción neutral con las diferencias menores a comparación de la emoción tristeza y alegría, para sus voces finales un promedio de 0.0028s y 0.52s para palabras cortas y frases largas y finalmente para la alegría las diferencias para voces finales tienen un promedio de 0.13s y 1.16s para palabras cortas y frases largas.

MBROLA es el sintetizador adecuado para trabajar el habla ya que cuenta con las voces necesarias para el español de México y España para la investigación; ES1, ES2, ES4, MX1 y MX2.

La contribución del TTS- EE a diferencia de Emofilt que solo es un sintetizador, es que aplica eSpeak para LA o español Latinoamericano y ES un español castellano que tienen otra nomenclatura en transcripción.

Los promedios de duraciones y pitch recolectados de los corpus a través de la segmentación automática de cada fonema para cada español alimentan el emodata del TTS-EE de manera adecuada.

Trabajo futuro del TTS-EE

La finalidad del TTS-EE obtener una voz con emoción a través de una entrada de texto preprocesada.

De todos los temas que toca el TTS-EE para el trabajo futuro se tiene lo siguiente:

- Tomar otros para metrps expresivos para obtener un resultado más fino en cuanto a las emociones.
- Mejorar la transcripción ortográfica fonética al 100% para el español de España.
- Mejorar la entrada y salida de la información del TSS-EE.
- Ampliar el rango de emociones.
- Utilizar un hablante femenino.
- Aplicar los principios del TTS-EE para otro idioma.

9. Anexos

A1. Glosario

Alófono: Cada una de las variantes que se dan en la pronunciación de un mismo fonema, según la posición de este en la palabra o sílaba, según el carácter de los fonemas vecinos, etc.; p. ej., la b oclusiva de tumbo y la fricativa de tubo son alófonos del fonema /b/.

Corpus: Conjunto lo más extenso y ordenado posible de datos o textos científicos, literarios, etc., que pueden servir de base a una investigación.

Difonos: Representa el sonido que abarca desde la mitad de la realización de un fonema hasta la mitad de la realización del fonema siguiente. El propósito de esta unidad de sonido es incorporar a la unidad de síntesis la transición de sonido entre fonemas. Segmento acústico que incluye la transición entre dos fonos consecutivos, formado por la parte estacionaria del primero, la transición del primero al segundo y la parte estacionaria del segundo.

Diptongo: Conjunto de dos vocales diferentes que se pronuncian en una sola sílaba; p. ej., aire, puerta, fui.

Diptongo creciente: Diptongo cuya segunda vocal constituye el núcleo silábico.

Diptongo decreciente: Diptongo cuya primera vocal constituye el núcleo silábico

Fonema*:* (Del gr. φώνημα, sonido de la voz). Cada una de las unidades fonológicas mínimas que en el sistema de una lengua pueden oponerse a otras en contraste significativo; p. ej., las consonantes iniciales de pozo y gozo, mata y bata; las interiores de cala y cara; las finales de par y paz; las vocales de tan y ten, sal y sol, etc. Dentro de cada fonema caben distintos alófonos.

Grafema: Unidad mínima e indivisible de la escritura de una lengua.

Laxo: Que no tiene la tensión que naturalmente debe tener.

Transcripción: Acción y efecto de transcribir.

Tecnologías del habla: Se denomina tecnologías del habla a aquellas que utilizan al lenguaje oral para la comunicación hombre-máquina.

A2. Programa en matlab para la conversión de .l16 a .wav

```
function arch2wav(directorio)
%5/9/12 código para convertir los archivos de S0329 a .wav
% los archivos de habla están guardados como secuencias de 16-bits, 16khz
% sin encabezada ni compresión (Linear PCM, Intel Byte format).
%cada archivo corresponde a una palabra aislada o a una oración.
%Los archivos tienen extensión (.l16)
%eval(['cd']c:)
%D=dir;

eval(['cd ' directorio]);
fs=16000;
numero=0;
D=dir;
for i = 2:size(D,1)
    if(size(strfind(D(i).name,'.l16'),1))
        numero = numero+1;
        % abrir archivo del habla
        fid = fopen(D(i).name, 'r');

        speech=fread(fid, inf, 'int16', 0, 'ieee-le');
        speech=speech/max(abs(speech));
        fclose(fid);

        % escribir como wav
        nombre=D(i).name(1:strfind(D(i).name, '.l16'));
        nombre=[nombre 'wav'];
        wavwrite(speech,fs,nombre);
        fprintf('Numero de archivos cambiados: %d \n', numero);
    end
end
```

A3. Script para la creación de TextGrid

```
#29/10/12 Toma archivos de sonidos del español y produce el TextGrid con fonemas sílabas y palabras.
# using easyalign
directory$ = "C:\Segmentacion\seg"
#directory$ = "C:\Segmentacion\seg"
Create Strings as file list... wavfileList 'directory$'\*.wav
noOfwavs = Get number of strings
Create Strings as file list... txtfileList 'directory$'\*.txt
noOftxts = Get number of strings
if noOfwavs <> noOftxts
  exit  Number of wavs and texts is different
else
  echo Number of wavs and texts is 'noOfwavs'
endif

for ifile to noOfwavs
  select Strings wavfileList
  wavfileName$ = Get string... 'ifile'
  printline 'wavfileName$'
  select Strings txtfileList
  txtfileName$ = Get string... 'ifile'
  printline 'txtfileName$'
  if left$(wavfileName$,7) <> left$(txtfileName$,7)
    exit  Wrong wav and text files
  endif
  Read from file... 'directory$'\'wavfileName$'
  Read Strings from raw text file... 'directory$'\'txtfileName$'
  select all
  nameSound$ = selected$ ("Sound",-1)
  select Sound 'nameSound$'
  plus Strings 'nameSound$'
  execute C:\plugin_easyalign\utt_seg2.praat ortho no
  select Sound 'nameSound$'
  plus TextGrid 'nameSound$'
  execute C:\plugin_easyalign\phonetize_orthotier2.praat ortho phono spa yes no
  select Sound 'nameSound$'
  plus TextGrid 'nameSound$'
  execute C:\plugin_easyalign\align_sound.praat ortho phono yes spa }-';(),.?¿ no yes no 90 yes no
  textgrid = selected("TextGrid")
  select 'textgrid'
```

```
  #Save as text file... 'directory$'\'textgrid'.TextGrid
  Save as text file... 'directory$'\'nameSound$'.TextGrid
  #remove some of the objects
  #pause continue or not
  select all
  minus Strings wavfileList
  minus Strings txtfileList
  Remove
  #pause continue
endfor
select all
Remove
echo Done
```

A4. Segmentación

A4.1 El uso de Praat en la investigación del corpus

Praat es un programa de computadora para analizar, sintetizar y manipular el habla y otros sonidos, es de código abierto y de disponibilidad libre para la mayoría de las plataformas de computadoras (MacOs, Windows, Linux), está disponible para ambos sistemas operativos como el de 32 bits y 64 bits, y puede ser descargado de praat.org.

Un corpus del habla típicamente consiste de un grupo de archivos de sonido, cada uno de los cuales es alineado con un archivo de anotación, e información de metadatos. Las fortalezas de Praat son en el análisis acústico de los sonidos individuales, en la anotación de esos sonidos y en la navegación de sonidos múltiples y archivos de anotaciones a través del corpus. Con el análisis acústico de ancho de corpus, se pueden obtener tablas de datos listas para el análisis estadístico, que puede ser realizada por secuencias de comandos

A4.2 EasyAlign en español: Una herramienta de segmentación automática bajo Praat.

El EasyAlign es una herramienta que se utiliza para la segmentación fonética automática del habla. Es un complemento gratuito de Praat que tiene como principal ventaja facilitar la segmentación del habla a partir de una transcripción ortográfica. A través de algunas etapas manuales se obtiene una anotación con varias filas dónde cada una representa la frase de manera fonética, silábica, léxica y ortográfica dentro un TextGrid. Se estudiaron tres aspectos en una evaluación de EasyAlign los cuales son las fronteras de los segmentos, su duración y el segmento. Los resultados muestran pocas diferencias entre la segmentación automática y humana, EasyAlign está disponible para varias lenguas entre ellas el castellano que es de interés para el proyecto.

La alineación fonética o segmentación fonética tiene el propósito de determinar la posición del tiempo del fonema, sílaba, entre otros límites de la palabra en un corpus de habla cualquiera que sea su duración con base en la grabación del audio y su transcripción ortográfica.

Una segmentación manual precisa requeriría 800 veces de tiempo real, por ejemplo 13 horas para una grabación de un minuto, lo cual es una desventaja especialmente cuando se enfrenta con un largo corpus de habla espontáneo, por lo que ésta herramienta de

alineamiento fonético automático es altamente deseable. Aunque también existen muchas variaciones fonéticas pueden disminuir la exactitud del proceso además de que con herramientas computacionales precias y preparación de datos, los sistemas automáticos pueden cometer errores que un ser humano no haría.

EasyAlign está basado en HTK, un conjunto de herramientas HMM conocido, es como una capa amigable bajo Praat, esto facilita el proceso de alineación debido a que cuenta con un sistema de fonetización y modelos acústicos entrenados.

A4.3 Descripción de EasyAlign

El sistema EasyAlign está hecho de scripts Praat, incluye dos componentes externos: un sistema de conversión grafema fonema y una herramienta de segmentación para la alineación a nivel fonema. La distribución es tipo auto-instalable o plug-in y contiene los modelos acústicos entrenados de los fonemas.

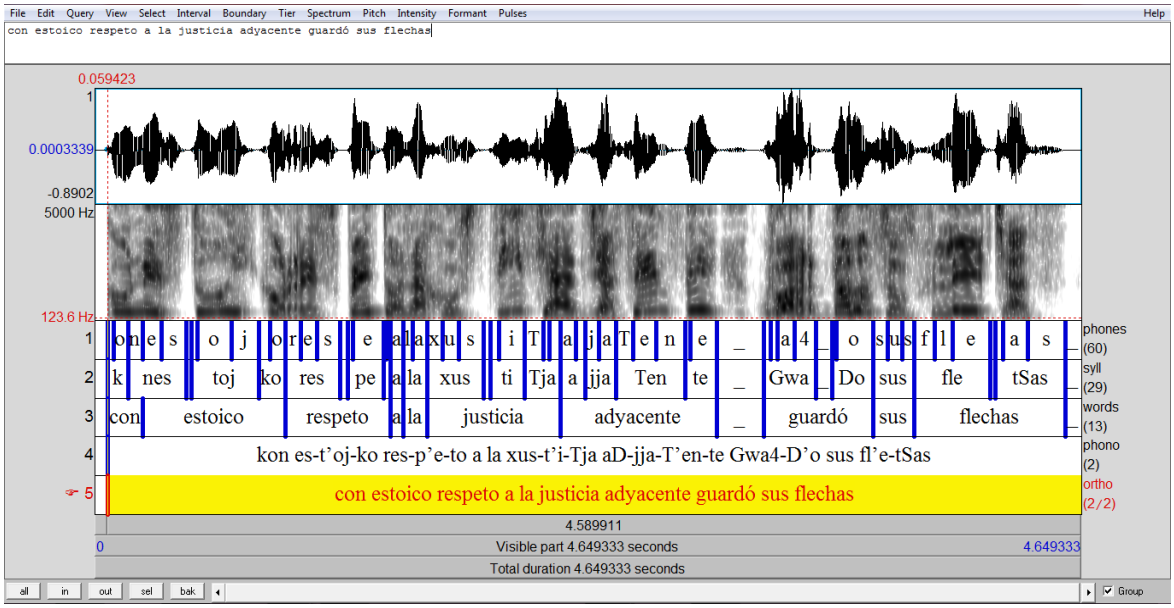


Figura 110. TextGrid resultante con 5 niveles de abajo hacia arriba: ortho, phono, words, syll, phones del enunciado con "estoico respeto a la justicia adyacente guardó sus flechas".

El proceso para la segmentación del archivo es la siguiente: tener un archivo de audio y su correspondiente transcripción ortográfica en un archivo de texto, el usuario tiene que pasar por verificaciones manuales y ajustes para asegurar una mejor calidad. Los tres pasos que debe realizar el usuario son macro-segmentación a nivel oración, conversión grafema fonema, y segmentación del fonema, como resultado se obtiene un TextGrid de

varios niveles, el formato de anotación dentro de Praat, con fonemas, sílabas, palabras y segmentación del enunciado como se muestra en la *Figura 10*.

Por otro lado la *Figura 11* presenta los pasos necesarios para utilizar EasyAlign y para entrenarlo.

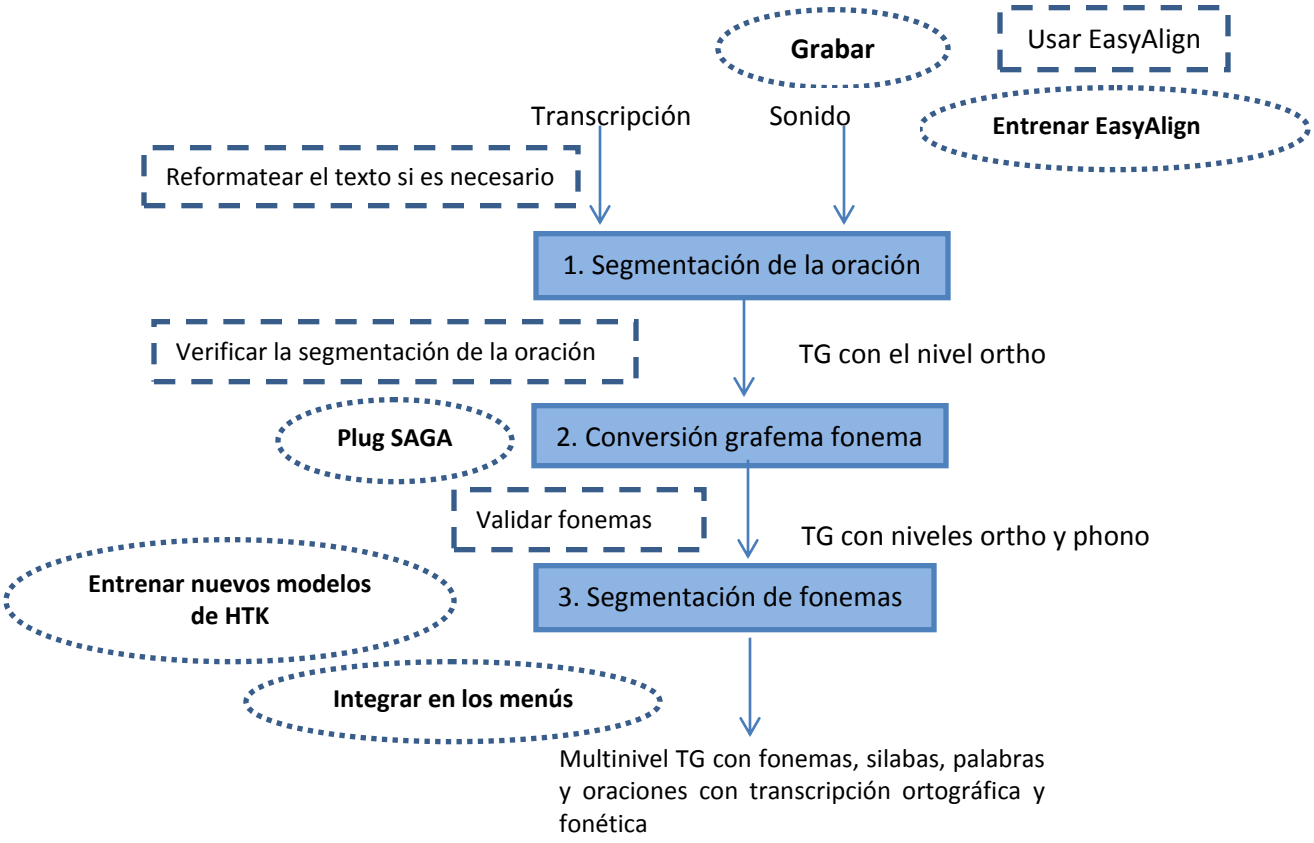


Figura 111. EasyAlign proceso habitual en cajas cuadradas y las medidas de adaptación en formas ovaladas.

A4.4 Instalación de Praat

Para la instalación de Praat se accede a http://www.fon.hum.uva.nl/praat/ después en el apartado de Download Praat se selecciona en este caso Windows y se descarga el archivo 32-bit edition: praat5368_win32.zip, se descomprime y se coloca en la carpeta deseada, en este caso, mis documentos, se corre la aplicación y Praat está listo para utilizarse, para la segmentación se requiere también de EasyAlign.

A4.5 Instalación de EasyAlign

Para la instalación de EasyAlign previamente se necesita tener Praat instalado, ahora se va a la página de EasyAlign en la siguiente dirección: http://latlcui.unige.ch/phonetique/easyalign.php y en el apartado de Download se da click en EasyAlign.exe para iniciar la descarga, se guarda el archivo, se ejecuta y se siguen los pasos de instalación como muestran las siguientes figuras, al tener instalada esta herramienta de Praat se puede comenzar con la segmentación.

Paso 1. Inicio de la instalación dónde se da click en el botón *Next* como muestra la *Figura 12*.

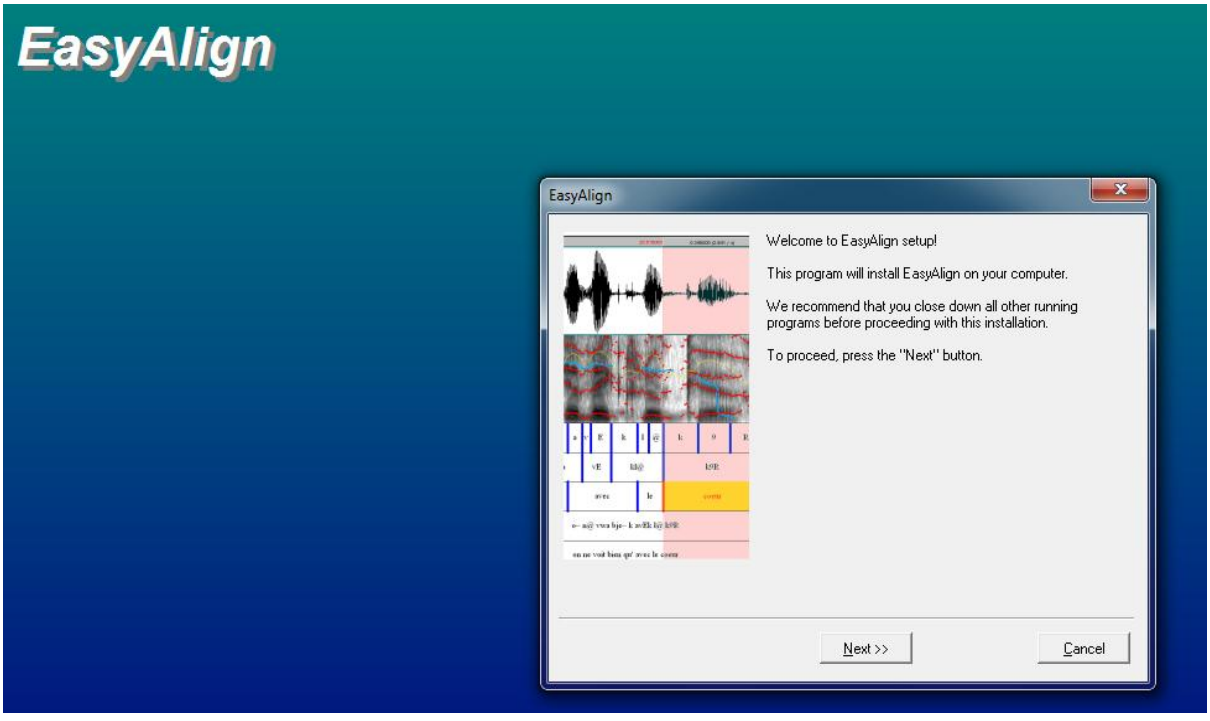


Figura 112. Inicio de la instalación de EasyAlign.

Paso 2. Se selecciona la carpeta dónde se instala el EasyAlign, como se ve en la *Figura 13*.

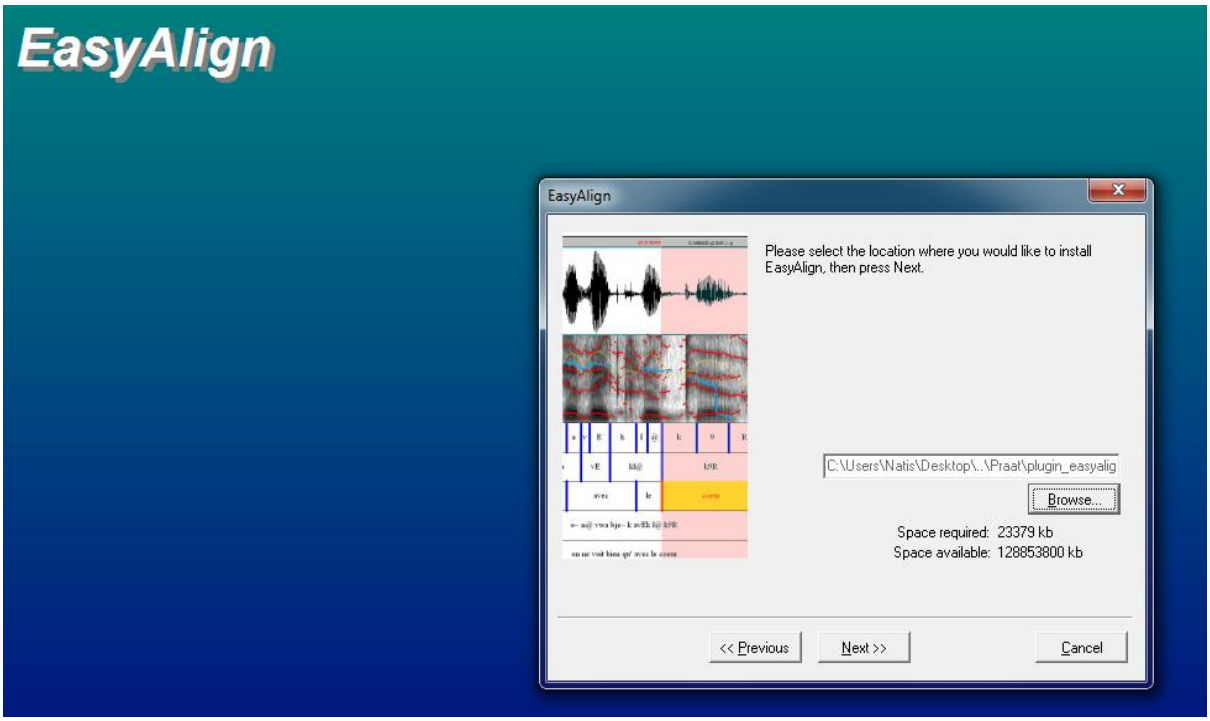


Figura 113. Lugar de instalación del EasyAlign.

Paso 3. Se puede observar en la *Figura 14* que se ejecuta la instalación.

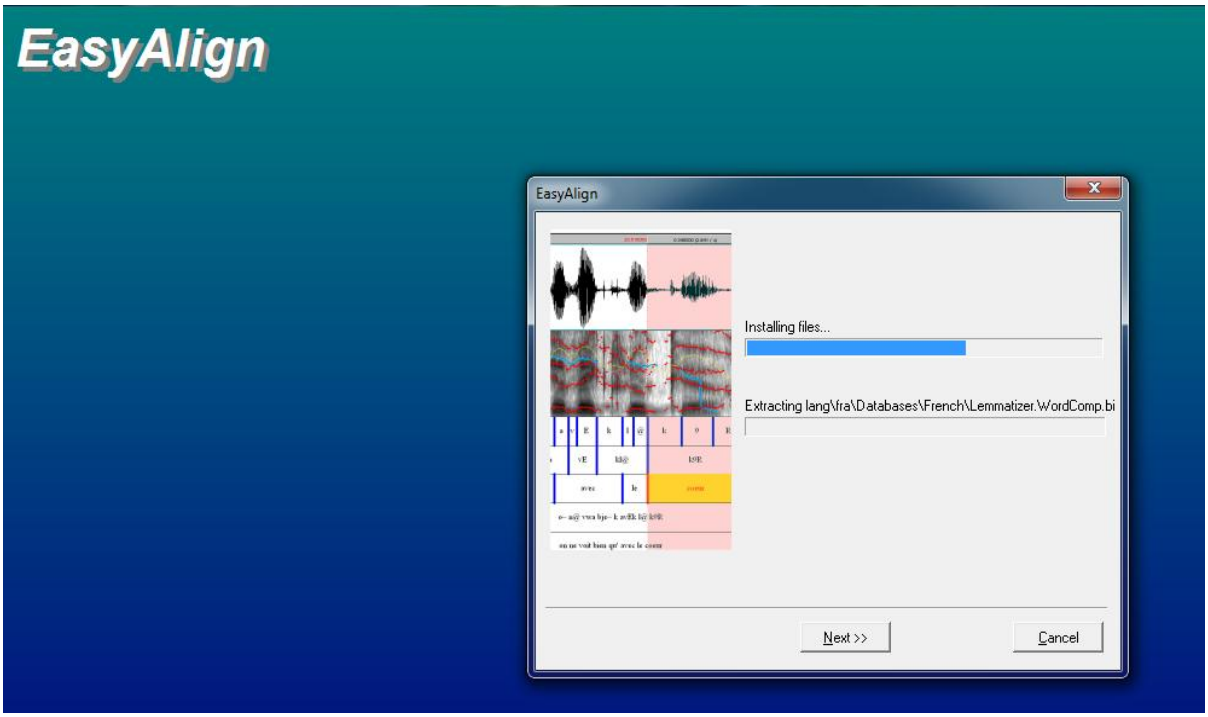


Figura 114. Ejecución de EasyAlign.

Paso 4. Se finaliza la instalación dando click en el botón *Finish* y ya se puede hacer uso de esta herramienta. Véase *Figura 15*.

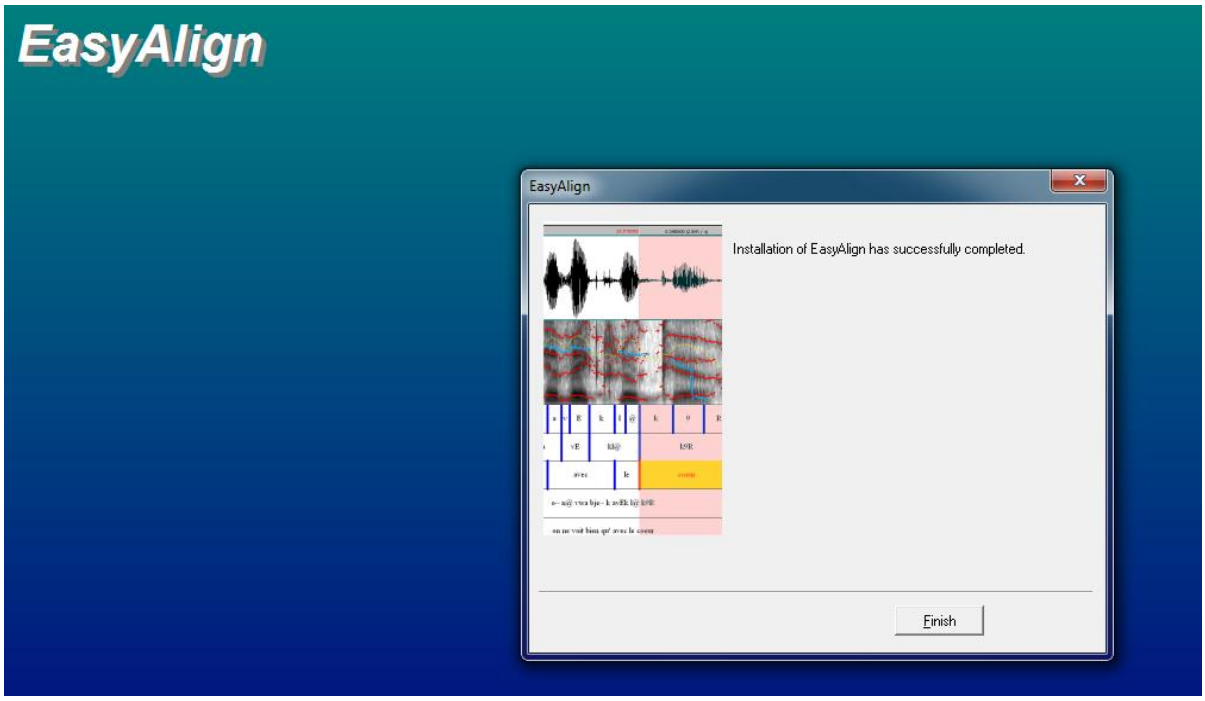


Figura 115. Instalación de EasyAlign finalizada.

Una vez instalado EasyAlign se puede observar como herramienta en el programa Praat como se muestra en la *Figura 16* encerrada en el ovalo rojo.

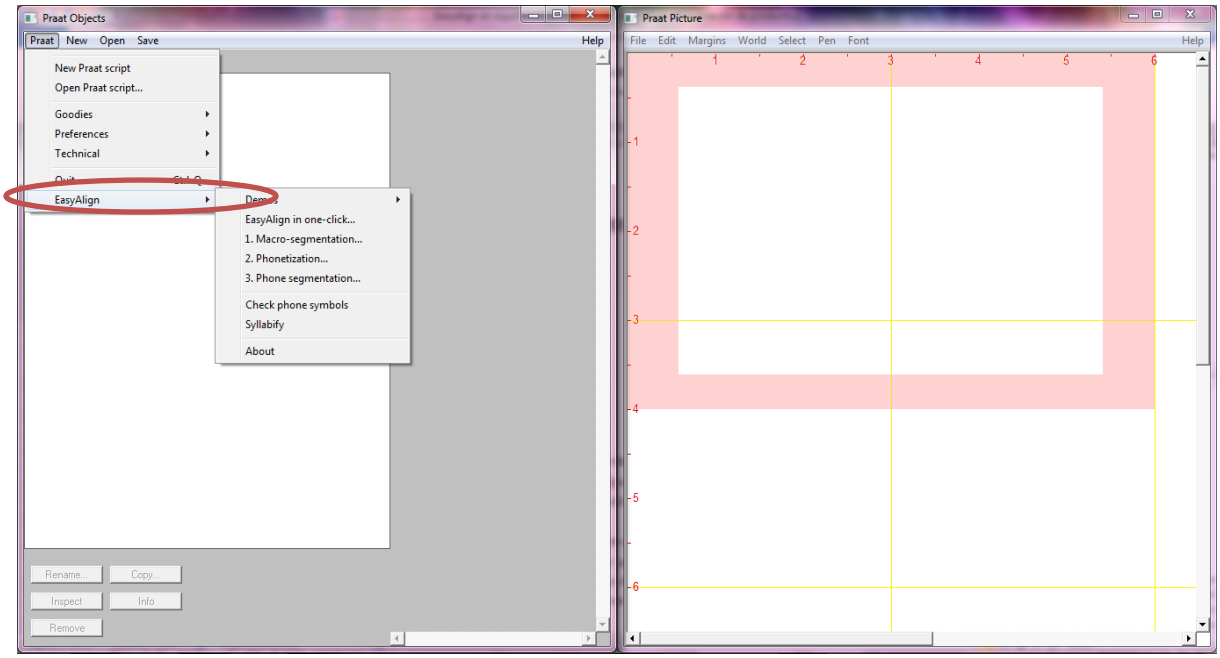


Figura 116. Instalación de EasyAlign

A4.6 Segmentación de la base de datos

Antes de la segmentación de la base de datos es necesario transformar los archivos de .l16 a .wav a través del programa en matlab ***arch2wav.m*** esto solo para la base de datos de español de España para todas las emociones, el código se encuentra en la sección de anexos.

Pasando a la segmentación se deben colocar en una carpeta todos los archivos .wav y .txt, los archivos deben llevar el mismo nombre por ejemplo AMXH001.wav y AMXH001.txt además debe de haber el mismo número de archivos, es decir, si son 184 audios .wav deben existir 184 archivos .txt los cuales contienen los enunciados, palabras o números.

Después se ejecuta el programa ***txtgrids.txt***, el cual al finalizar arroja el total de archivos .TextGrid en este caso 184, que contienen la segmentación en cinco niveles, como muestra la *Figura 17*. Como se puede observar cada uno de los audios está dividido por fonemas, sílabas de fonemas, palabras, el enunciado transcrito en fonemas y la oración original (phones, syll, words, phon, ortho) y también se puede observar la señal de audio.

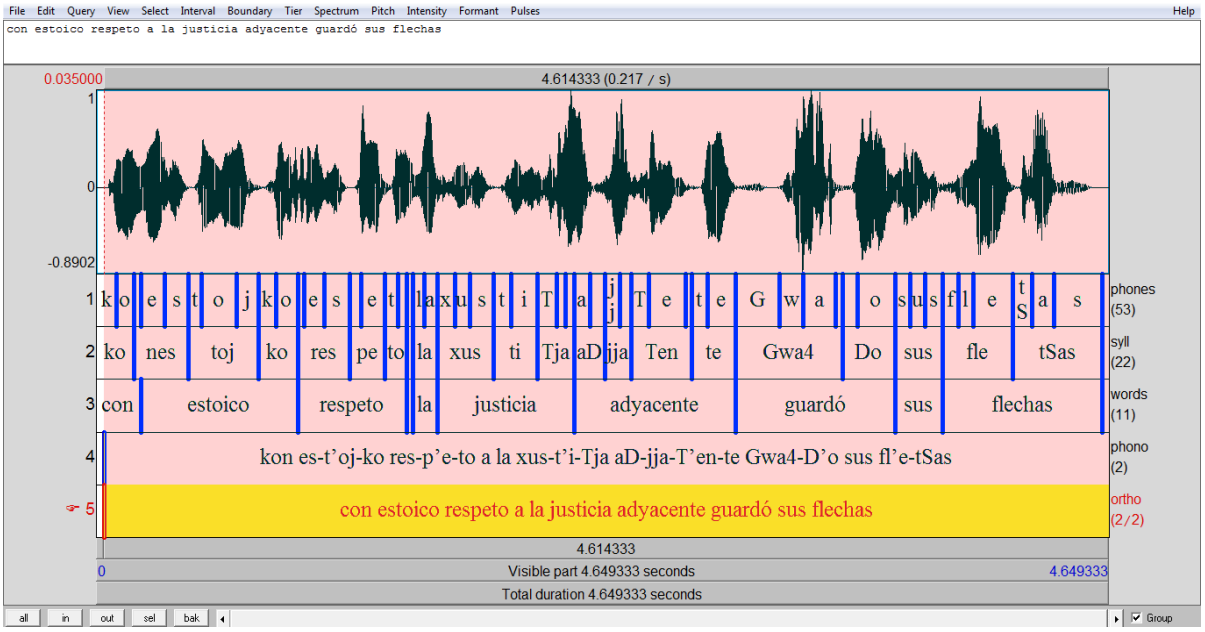


Figura 117. Segmentación de la frase "con estoico respeto a la justicia adyacente guardó sus flechas".

A continuación la oración "con estoico respeto a la justicia adyacente guardó sus flechas" desglosada en cada nivel del TextGrid resultante.

Nivel 1: Frase dividida en fonemas (phones).

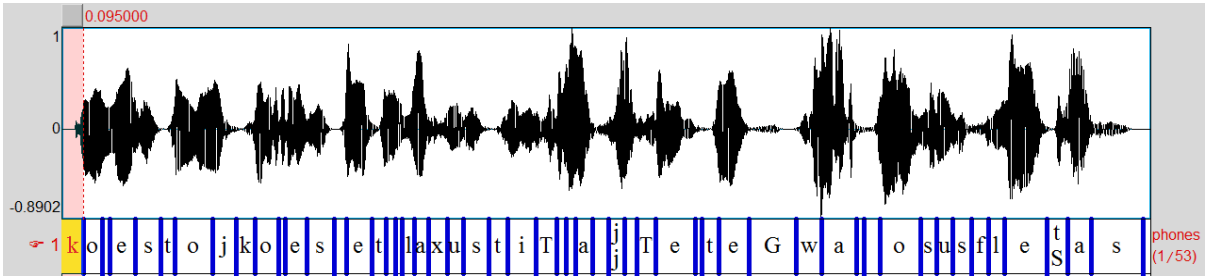


Figura 118. Frase dividida en fonemas.

Nivel 2: Frase de fonemas agrupados en sílabas (syll).

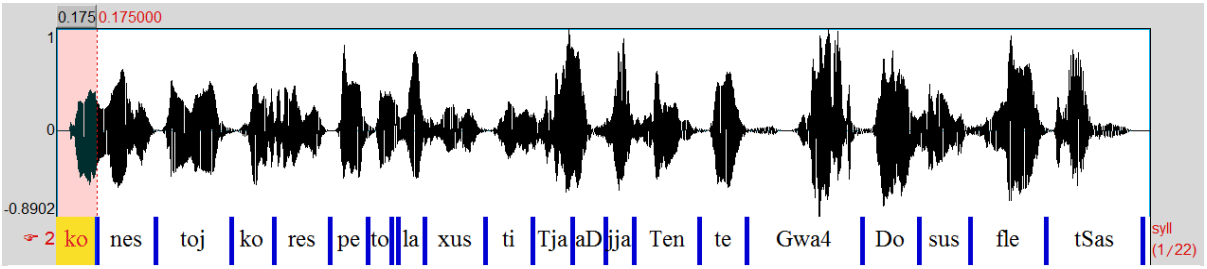


Figura 119. Frase dividida en sílabas.

Nivel 3: Frase dividida en palabras (words)

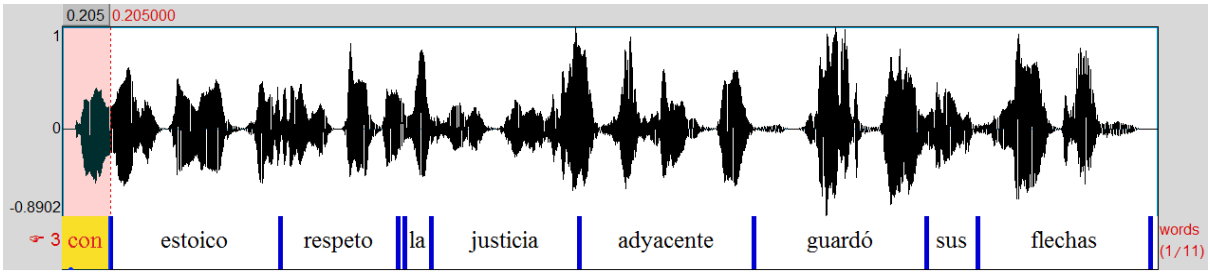


Figura 120. Frase dividida en palabras.

Niveles 4: Transcripción ortográfica-fonética de la frase (phono)

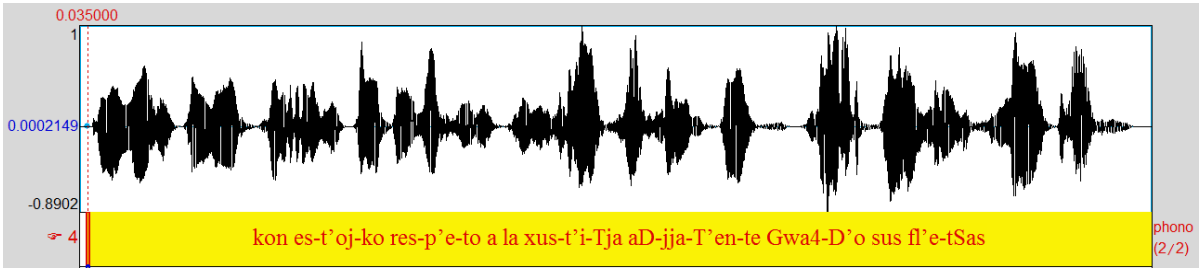


Figura 121. Frase convertida a fonemas.

Niveles 5: Oración original (ortho)

Figura 122. Oración original.

Una vez obtenidos todos los TextGrid de las bases de español España y México con las respectivas emociones, se analiza cada TextGrid junto con su audio y se escucha cada uno de los fonemas, al realizar esto se puede notar que la segmentación o alineación automática es muy buena aunque tiene algunos detalles al oído que se pueden mejorar por lo que se efectuó una segmentación manual, que fue una gran labor tratando de que cada fonema sonara lo mejor posible. Se salta entre los niveles 1, 2, y 3 para tener una mejor perspectiva del sonido. Algunos de los problemas encontrados son de fonemas donde existe un diptongo es decir un grupo de vocales donde no se distinguía claramente la separación de los fonemas, también aquellos fonemas como /f/, /k/, /s/, /r/, /n/, /m/ entre otros, que no tenían una forma definida visualmente hablando con respecto al espectrograma el cual fue de gran ayuda para esta nueva alineación, y también algunos fonemas arrastraban secciones de las pausas adyacentes por lo que se tenía que hacer la separación de estas.

A4.7 Obtención de la duración

La duración de los fonemas se obtiene al ejecutar el script calculate_segment_durations.praat, se analiza el primer nivel correspondiente al de los fonemas y se guarda con el nombre de duraciones.txt en este archivo de texto se puede observar la duración en segundos, por lo que:

1) Se debe de convertir el archivo de texto a .xlsx o archivo de Excel para trabajarlo más fácilmente.
2) Se identifican los fonemas con énfasis o acentuados.

3) Después se toma una tercia de fonemas donde exista un fonema central, un fonema anterior y un posterior, esto con la finalidad de obtener el contexto del fonema central.
4) A esas tercias, se les saca el promedio de la duración del fonema central y se convierte de segundos (s) a milisegundos (ms), lo cual servirá para alimentar al sistema de texto a habla.
5) También se toman en consideración las pausas definidas por el símbolo (_), esta se convierte en el punto central de un fonema anterior y otro posterior de la forma /o/_ /k/. Esta pausa solo tiene duración y no pitch.

A4.8 Obtención del pitch

En cuanto al pitch de los fonemas se obtiene al ejecutar el script collect_pitch_data_from_files.praat, se analiza al nivel correspondiente al de los fonemas y se guarda con el nombre de pitchresults.txt en este archivo de texto se puede observar el pitch en Hz, por lo que:

1) Se debe de convertir el archivo de texto a .xlsx o archivo Excel para trabajarlo más fácilmente.
2) Se identifican los fonemas con énfasis o acentuados
3) Después se toma una tercia de fonemas donde exista un fonema central, un fonema anterior y un posterior, esto con la finalidad de obtener el contexto del fonema central. Por ejemplo /o/ /tS/ /e/, donde el central es el fonema /tS/, el anterior /o/ y el posterior /e/.
4) De las tercias, se les saca el promedio del pitch en Hz del fonema central, lo cual servirá para alimentar al sistema de texto a habla con estos resultados.

A4.9 Qué emoción elegir

Una vez realizada la segmentación y partir de ella la obtención de duraciones y pitch, se pueden manejar estos datos para saber que emoción elegir, se tomaron como muestra las vocales para compararlas entre sí y a través de diagramas de caja realizar la elección. Las vocales se tomaron como muestra debido a la estabilidad en la señal de onda y en su duración y pitch recae la emoción.

En el habla existen elementos que permiten que el oyente comprenda con gran facilidad lo que se desea transmitir y que refuerzan el contenido del lenguaje verbal, estos

elementos son el ritmo, el tono, el volumen de voz etc., los cuales se describen en la *Figura 129*.

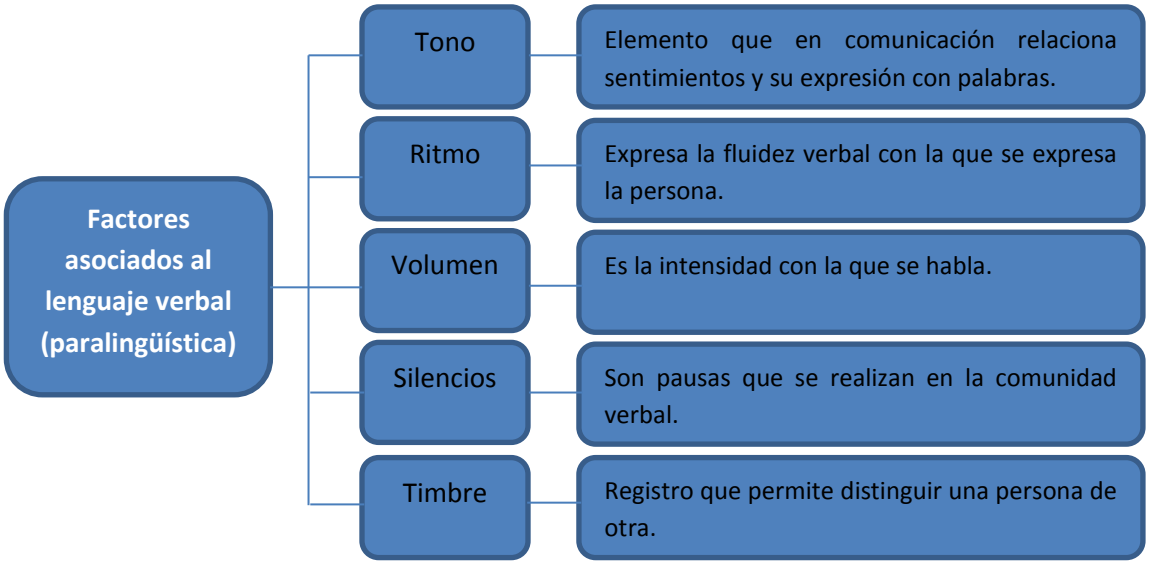


Figura 123. Paralingüística.

A4.10 Elementos paralingüísticos

F0 o pitch. Es la onda simple de frecuencia más baja entre las que forman una onda sonora compleja periódica. Es una concentración de energía acústica. Corresponde a la frecuencia de abertura y cierre de los repliegues vocales. Esta confiere a cada sonido del habla una estructura diferente o timbre.

Tono. El tono se entiende como la propiedad de los sonidos que los caracteriza como agudos o graves, en función de su frecuencia, la RAE lo define como “inflexión de la voz y modo particular de decir algo, según la actitud, la intención o el estado de ánimo de quien habla”. Sirve como regulador entre el sentimiento y la expresión, entre lo sentido y lo verbalizado.

Pausas y silencios. Las pausas funcionan como reguladores de cambio, indican cambios de unidad gramatical, tema o de turno de palabra, existe un tipo de pausa llamada reflexiva para comprender correctamente el mensaje verbal, por ejemplo la expresión “Valeria entra”, informa de una acción y “Valeria, entra”, indica un orden.

Los silencios implican comunicación, se puede invitar a hablar, a callar, se asiente o se muestra desacuerdo entre muchas otras posibilidades.

Tiempo de habla. Es la duración de las intervenciones de los interlocutores durante una conversación.

La descripción del resto de los elementos es suficiente con la definición de la figura 130, solo para dar un panorama general.

Estos aspectos son necesarios para obtener la máxima naturalidad en la voz con emoción, se busca la mayor precisión posible de los patrones que rigen la comunicación humana.

Según el estudio de PHYSTA se define el concepto de emoción con dos ideas principales. Primero, la idea básica de emoción que surge por Descartes, que dice que los humanos tienen un número concreto de emociones universales, esas emociones se pueden mezclar consiguiendo con ellos emociones complejas. Y segundo, que la biología motive las emociones, o que las emociones básicas estén motivadas por una necesidad evolutiva, las emociones se clasifican en discretas y continuas. [46]

A veces el concepto de emoción incluye todos los tipos de emociones, actitudes y convicciones y otras veces las distingue.

A4.11 Caracterización discreta de las emociones o emociones básicas.

Una cuestión actual relevante es el estudio de la emoción, de la existencia de emociones básicas o universales, según Darwin se trata de reacciones afectivas innatas, distintas entre ellas, presentes en todos los seres humanos y que se expresan en forma característica.[46]

Las emociones según Izard son placer, interés, sorpresa, tristeza, ira, asco, miedo y desprecio pero Ekman sugiere como básicas ira, alegría, asco, tristeza, sorpresa y miedo y posteriormente el desprecio.[46]

A4.12 El espacio continúo de Dominación, Dominación, Activación.

Dominación. Es la activación desde dormido a frenético.

Valencia. Se refiere a la intención ya sea positiva o negativa, es la evaluación del sentimiento.

Activación. Se manifiesta con el poder o sensación de seguridad ante una situación, sirve para distinguir entre sentimientos similares como desprecio y miedo.

El plano Activación = 0 se representa en la *Figura 130.*

Figura 124. Plano del espacio tridimensional de emociones definido por 'Activación = 0'.

A este plano le añadimos otra componente ortogonal (Activación). Este eje distinguiría, por ejemplo, miedo y espanto. El miedo tiene un nivel de control de la situación mucho mayor que espanto. Por tanto, en miedo la componente 'Activación' será menor que en espanto.

Las emociones se representarán con una cantidad entre [-100;100] para cada uno de los 3 ejes. La siguiente tabla 50 muestra las coordenadas de 5 emociones básicas[46].

Emoción	Arousal	Valence	Power
Neutro	0	0	0
Triste	-8.5	-42.9	-55.3

Enfadado	34.6	-34.9	-33.7
Asustado	31.1	-27.1	-79.4
Feliz	28.9	39.8	12.5

Tabla 43. Coordenadas de 5 emociones básicas.[46]

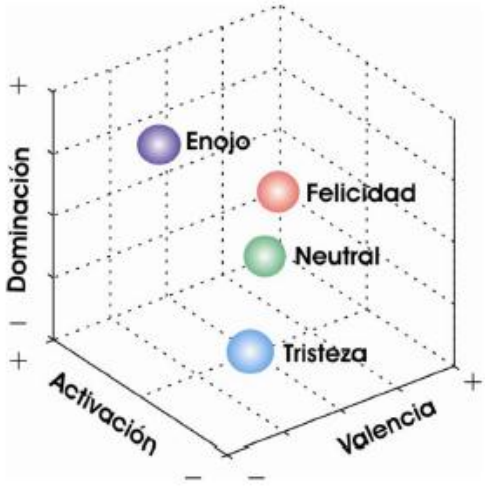


Figura 125. Emociones en las coordenadas definidas.[46]

Con este sistema se matizan las emociones. Por ejemplo, si a enfadado se aumenta la coordenada "Dominación" la emoción que obtenida es "ira".

Este tipo de sistema de categorización emocional permite trabajar con infinitas emociones.

Se puede ver gráficamente en la *Figura 131* las emociones básicas las que se toman para este proyecto son ***felicidad*** una emoción alta, ***neutral*** una emoción media y ***tristeza*** una emoción baja.

Los principales centros de investigación universitarios así como empresariales, han creado sus propias bases de datos y corpus que recogen la lengua hablada. La motivación se da principalmente por la necesidad de disponer de datos para desarrollar y evaluar las diferentes aplicaciones del procesamiento del habla, básicamente de los sistemas de reconocimiento, también la investigación básica en los más diversos aspectos de la comunicación oral.

Surge la necesidad de unificar los resultados obtenidos, de aplicar metodologías y criterios de diseños comunes y de establecer canales de distribución que permitan el acceso a un extenso abanico de usuarios.

En Europa existen diversos programas de apoyo a la investigación y desarrollo de la CEE como ESPRIT y LRE que han auxiliado en la coordinación de varios países llevando una serie de iniciativas en la constitución de corpus orales y escritos con ciertos estándares. También se perfila la constitución de corpus en otros continentes a través de organizaciones similares. Los corpus son un conjunto de datos en este caso datos el habla.

Aunque esta tendencia se ha dado de manera un tanto reducida en la lengua española, específicamente en lo referente a los corpus orales. Aun se dista de disponer de un corpus de la lengua oral para el español que permita, la investigación en fonética, fonología además del desarrollo y evaluación de sistemas de tecnología de voz.

Las aplicaciones de los corpus orales son tres:

- La investigación aplicada de la tecnología del habla
- La investigación lingüística básica
- La aplicación a la enseñanza de la lengua

La investigación aplicada de la tecnología del habla. En la aplicación de la tecnología del habla es de utilidad la obtención de modelos estadísticos de la lengua, como la probabilidad de transición entre palabras, los cuales permiten mejorar la eficacia de sistemas de conversión de habla a texto.

De igual manera para la síntesis de voz es de utilidad disponer de corpus oral, debido a tres aspectos: la información acústica de la variabilidad de las unidades, la descripción y modelado de variaciones prosódicas de la modalidad y la estructura sintáctica producida en la lectura del texto. Y finalmente con los corpus se desarrollan modelos lingüísticos utilizados para el análisis sintáctico en la calidad de la prosodia. Las aplicaciones donde existe diálogo hombre-máquina son beneficiadas del análisis de corpus que se obtiene ya sea en la interacción directa entre usuario-máquina o la simulada también llamada paradigma del Mago de Oz.

La investigación lingüística básica. La descripción fonética y fonológica del español se beneficia del corpus oral que es diseñado, recogido y estructurado de acuerdo a la variedad de realizaciones fonéticas que se encuentran en la lengua a nivel segmental y

suprasegmental. Aunque existen variaciones por la zona geográfica, social y estilística en el habla hispana que no permiten detallar la descripción en su totalidad.

La aplicación a la enseñanza de la lengua. Un vasto corpus, de extensa cobertura dialectal, sociolectal y estilística, contribuye a la creación de herramientas didácticas en el campo de enseñanza de la lengua materna y del español como lengua extranjera. Muestran una panorámica amplia sobre la variedad lingüística e entregarlos en métodos multimedia.

Las tipologías de los corpus orales son tres:

- Los que son inventarios fonéticos y fonológicos desarrollados para el estudio de los universales lingüísticos. Una compilación de corpus.
- Los diseñados para la descripción fonética y fonológica de la lengua.
- Los orientados a aplicaciones y productos en el ámbito de la tecnología de la voz.

El último punto es de gran interés para el proyecto y se describirá brevemente a continuación. Los corpus diseñados para aplicaciones tecnológicas principalmente para reconocimiento del habla suelen tener como contenido fonético palabras aisladas o frases fonéticamente equilibradas, también se incluyen textos. La base de datos TIMIT (Zue et al, 1990) es una de las más paradigmáticas en este ámbito, recoge más de 2,340 frases pronunciadas por 630 locutores.

Son herramientas útiles para alimentar con datos estadísticos el sistema de texto a habla expresivo en español, son datos sobre las diferentes realizaciones fonéticas en cuanto a su duración y también sobre los fonemas la obtención de su duración y pitch.

El corpus oral se constituye por una serie de tareas como, el tipo de información que debe recogerse, el grado de variabilidad que pretende cubrirse, el modo de obtención de los datos, su nivel de descripción y el soporte de almacenamiento y difusión, entre otros.

La *Figura 132* hace referencia de manera esquemática de las principales fases en la preparación de un corpus oral.

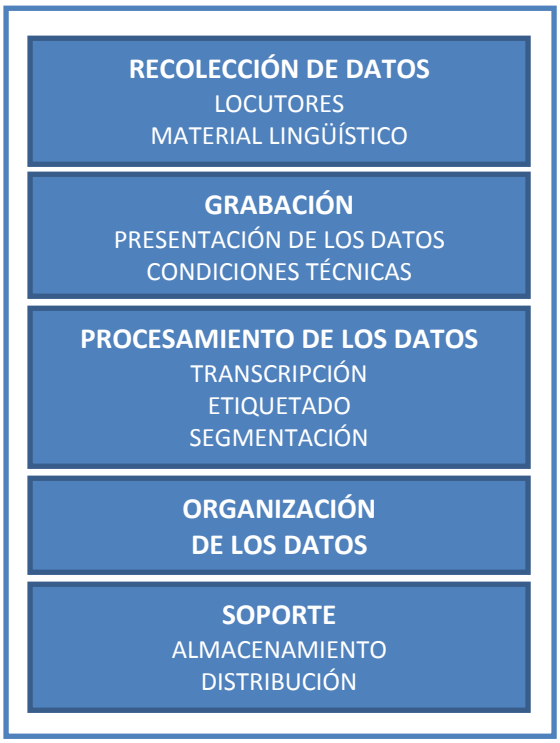


Figura 126. Fases en la preparación de un corpus oral.

10. Bibliografía

[1] J. Allen, M. S. Hunnicutt, D. Klatt, "From Text to Speech: The MITalk", Cambridge University, Press: 1987.

[2] P. Birkholz, B. Kröger. "Simulation of vocal tract growth for articulatory speech synthesis", Institute for Computer Science, University of Rostock, Rostock, Germany, 2007.

[3] A. Barbosa, "Desarrollo de una nueva voz en Español Mexicano para el Sistema de Texto a Voz Festival", Tesis de Maestría, Universidad de las Américas - Puebla, México, Otoño 1997.

[4] Kominek J., Bennett C., Black A., Evaluating and Correcting Phoneme Segmentation for Unit Selection Synthesis, Eurospeech-Geneva 2003

[5] T. Dutoit, V. Pagel, N. Pierret, F. Bataille, y O. Van der Vreken, "The mbrola project: Towards a set of high-quality speech synthesizers free of use for noncommercial purposes", Proc. ICSLP'96, Philadelphia, vol. 3, pp. 1393–1396, 1996.

[6] A. Black and K. Lenzo, "Limited Domain Synthesis", ICSLP2000, Beijin, China, 2000

[7] F. Burkhardt, Emofilt, 2012, Disponible: (http://emofilt.syntheticspeech.de/)

[8] A. W. Black, K. A. Lenzo, "Part D: Fully Automatic text-to-speech conversion 1968-1985", Carnegie Mellon University, 2013, Disponible en: (http://www.festvox.org/history/klatt.html)

[9] J. Goddard y F. Martínez-Licona, Curso de Temas Selectos de Ciencias y Tecnologías de la Información, Notas, 2012.

[10] Figueroa K., " Síntesis de Voz en español, un enfoque silábico", Morelia, Mich., Umich, 1998.

[11] Campillos L., "Tecnologías del habla y análisis de la voz. Aplicaciones en la enseñanza de la lengua", Laboratorio de Lingüística Informática - Universidad Autónoma de Madrid, España, 2010

[12] AT&T Natural Voices® Text-to-Speech Demo, AT&T Labs. Inc. Research, 2012, Disponible en: (http://www2.research.att.com/~ttsweb/tts/demo.php)

[13] Cepstral we build voices, Pittsburgh, PA, USA, 2012, Disponible en: (http://www.cepstral.com/)

[14] Naunce-Loquendo, 2013, Disponible: (http://www.sodels.com/loquendo.htm)

[15] eSpeak text to speech, Free Software Foundation, Inc. , 2012, Disponible: (http://espeak.sourceforge.net/)

[16] R. Barra-Chicote, J. Yamagishi, S. King, J. M. Montero, J. Macias-Guarasa, "Analysis of statistical parametric and unit selection speech synthesis systems applied to emotional speech", Speech Communication, Volume 52, Issue 5, 2010.

[17] F. Burkhardt, "Emofilt: the Simulation of Emotional Speech by Prosody-Transformation", Interspeech, 2005.

[18] Dutoit T., The MBROLA Project, 2012, http://tcts.fpms.ac.be/synthesis/

[19] J. Llisterri "Aplicaciones de la síntesis del habla" Departament de Filologia Espanyola, Universitat Autònoma de Barcelona [Citado de: http://liceu.uab.es/~joaquim/speech_technology/tecnol_parla/synthesis/synthesis_general/1.1.objetivos_aplicaciones.html#sintesis_aplicaciones]

[20] T. Dutoit, "A Short Introduction to Text-to-Speech Synthesis", TCTS Lab, 2013, Diponible en: (http://tcts.fpms.ac.be/synthesis/introtts_old.html)

[21] P. H. Van Santen, R. W. Sproat, J. P. Olive and J. Hirschberg, "Progress in Speech Synthesis". Springer, , 1997

[22] Castro M. J., S. España, A. Marzal, I. Salvador, "Transcriptor ortográfico-fonético para el castellano", Informe técnico, Depto. de Sistemas Informáticos y Computación Universidad Politécnica de Valencia, 2000.
[23] Well J., SAMPA - computer readable phonetic alphabe, Department of Speech, Hearing and Phonetic Sciences, University College London, 1997, (http://www.phon.ucl.ac.uk/home/sampa/)
[24] Lozano L. "Ortografía Activa", Libris Editores, 1a edición, 1997
[25] M. Kastner y B. Stangl, Exploring a text-to-speech feature by describing learning experience,enjoyment, learning styles, and values – A basis for future studies, 46th Hawaii International Conference on System Sciences
[26] J. Schroeter, "16 Text to-Speech[TTS] Synthesis", AT&T Laboratories, pág. 16-1, 2003.
[27] F. Burkhardt y J. Stegmann, "Emotional Speech Synthesis: Applications, History and Possible Future", ESSV. 2009.
[28] Chabchoub A. y Cherif A, "Estimation and optimization of prosodic to improve teh quality of the arabic synthetic speech", IJAET, pp. 632-639, Vol. 2, 2012.
[29] F. Burkhardt, "An Affective Spoken Story Teller", Interspeech, 2011.
[30] M. El Ayad, M. S. Kamel, F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases", Pattern Recognition, Vol. 44, Issue 3, pp. 572-587, 2011.
[31] Larousse Ortografía, reglas y ejercicios, Elsevier 2006.
[32] N. Audibert, V. Aubergé, A. Rilliard, "The Prosodic Dimensions of Emotion in Speech: the Relative Weights of Parameters", Eurospeech, 2005.
[33] S. Furui, "Digital Speech Processing, Synthesis and Recognition", Ed. Dekker, 1989.
[34] A. Lazaridis, "Prosody modelling using machine learning techniques for neutral and emotional speech synthesis", Ph.D. Thesis, Universidad de Patras, Grecia, 2011
[35] F. Llisterri, "Transcripción, etiquetado y codificación de corpus orales", en Etiquetación y extracción de información de grandes corpus textuales, curso impartido en Seminario de Industrias de la Lengua, Soria. 1997
[36] F. Martínez-Licona, O. Muñoz-Texzocotetla, A. Martínez-Licona and J. Goddard, "Analysis of Emotions in Mexican Spanish Speech, Proc. of the 8th WSEAS International Conference on Signal", Speech and Image Processing [SSIP '08], pp.67-71, 2008.
[37] J. M. Montero, J. Gutiérrez-Arriola, S. Palazuelos, E. Enríquez, S. Aguilera, y J. M. Pardo, Emotional Speech Synthesis: From Speech Database to TTS, ICSLP 98, Vol. 3, p. 923-926, 1998.
[38] A. Moreno y J. Mariño, "Spanish Dialects: Phonetic Transcription", Universitat Politècnica de Catalunya, Barcelona, SPAIN, pág. 2, 1998.
[39] G. A. Moreno, 2008, "Nueva Voz Concatenativa de Difonemas para el Español Mexicano en Festival", Ingeniería en Sistemas Computacionales. Escuela de Ingeniería y Ciencias, Universidad de las Américas Puebla.
[40] M. Rodríguez y E. Mora, "Conversor de texto a voz en el dialecto venezolano por medio de la concatenación de difonos", Revista Ciencia e Ingeniería. Vol. 27, No. 2, pp. 79-87, 2006.
[41] Rosenblat A., Buenas y malas palabras [en el castellano de Venezuela], vol. III, págs. 143-147 (Madrid: Mediterráneo, 1980).
[42] M. Schröder, "Emotional Speech Synthesis - A Review", Proc. Eurospeech 2001, Aalborg, Vol. 1, pp. 561-564. 2001.
[43] P. Taylor, "Text-to-Speech Synthesis", University of Cambridge, 2009.
[44] F. Tesser, P. Cosi, C. Drioli, G. Tisato, "Prosodic Data-Driven Modelling of Narrative Style in FESTIVAL TTS", Proceedings of 5th ISCA Speech Synthesis Workshop, 2004.
[45] F. Tesser, P. Cosi, C. Drioli, G. Tisato, "Emotional Festival-Mbrola TTS Synthesis", Proceedings INTERSPEECH 2005, pp. 505-508, 2005.
[46] E. Mendez, 2013 "Sistema de síntesis de texto a voz femenina en español con control prosódico basado en Mbrola", Universidad Carlos III de Madrid.

Casa abierta al tiempo
UNIVERSIDAD AUTÓNOMA METROPOLITANA

ACTA DE EXAMEN DE GRADO

No. 00075
Matrícula: 2113802076

SISTEMA DE TEXTO A HABLA EXPRESIVO EN ESPAÑOL.

En la Ciudad de México, se presentaron a las 12:00 horas del día 29 del mes de octubre del año 2018 en la Unidad Iztapalapa de la Universidad Autónoma Metropolitana, los suscritos miembros del jurado:

DRA. MARIKO NAKANO MIYATAKE
M. EN C. FABIOLA MARGARITA MARTINEZ LICONA
M. EN C. ALMA EDITH MARTINEZ LICONA

Bajo la Presidencia de la primera y con carácter de Secretaria la última, se reunieron para proceder al Examen de Grado cuya denominación aparece al margen, para la obtención del grado de:

MAESTRA EN CIENCIAS (CIENCIAS Y TECNOLOGIAS DE LA INFORMACION)

DE: NATIVIDAD FELISA NAVARRETE GOMEZ

y de acuerdo con el artículo 78 fracción III del Reglamento de Estudios Superiores de la Universidad Autónoma Metropolitana, los miembros del jurado resolvieron:

Aprobar

Acto continuo, la presidenta del jurado comunicó a la interesada el resultado de la evaluación y, en caso aprobatorio, le fue tomada la protesta.

NATIVIDAD FELISA NAVARRETE GOMEZ
ALUMNA

REVISÓ

LIC. JULIO CESAR DE LARA ISASSI
DIRECTOR DE SISTEMAS ESCOLARES

DIRECTOR DE LA DIVISION DE CBI

DR. JESUS ALBERTO OCHOA TAPIA

PRESIDENTA

DRA. MARIKO NAKANO MIYATAKE

VOCAL

M. EN C. FABIOLA MARGARITA MARTINEZ LICONA

SECRETARIA

M. EN C. ALMA EDITH MARTINEZ LICONA