Casa abierta al tiempo

PC y TI

UNIVERSIDAD AUTÓNOMA METROPOLITANA - IZTAPALAPA
DIVISIÓN DE CIENCIAS BÁSICAS E INGENIERÍA

SÍNTESIS ESTADÍSTICA PARAMÉTRICA DE VOZ

Tesis que presenta
Marvin Coto Jiménez
Para obtener el grado de
Maestro en Ciencias y Tecnologías de la Información

Asesores: Dr. John Goddard Close

M.I. Fabiola Martínez Licona

Jurado calificador:

Presidente: Dr. José Abel Herrera Camacho

Secretaria: M.I. Fabiola Martínez Licona

Vocal: M.C. Alma Martínez Licona

México, D.F. Agosto 2014

Resumen

La síntesis estadística paramétrica de voz es una técnica de producción de voces artificiales que utiliza como modelo matemático dominante los Modelos Ocultos de Markov sobre una representación paramétrica del habla. Esto permite que una voz pueda ser codificada utilizando parámetros espectrales, de frecuencia fundamental y de duración de sus unidades fonéticas, para luego entrenar los modelos matemáticos que permitan producir nuevas frases, con ventajas significativas sobre otros procedimientos de síntesis de voz, tales como su mayor flexibilidad y menor requerimiento en almacenamiento.

En este trabajo se presenta el desarrollo teórico, la adaptación a nivel lingüístico y computacional y una propuesta de extensiva de experimentación y evaluación de voces artificiales producidas a partir de síntesis estadística paramétrica de voz, en una variante de español latinoamericano. Para este fin se han definido una serie de contextos de implementación y se han adaptado y desarrollado aplicaciones computacionales como aportes a distintos niveles, desde la extracción de información hasta la evaluación de resultados. Esto ha permitido plantear una gran cantidad de experimentos para estudiar la influencia de diversos factores a la calidad de voces obtenidas.

Los principales aportes del proyecto son: La documentación de los elementos teóricos y prácticos para la creación de voces utilizando la síntesis estadística paramétrica. De acuerdo con el estudio de referencias realizado, este documento constituye el primer aporte a la documentación de ambos aspectos. En segundo lugar la creación de programas para la extracción y análisis de parámetros y para la evaluación de resultados, además de las aplicaciones desarrolladas para probar las voces en situaciones reales. En tercer lugar la incorporación de parámetros acústicos como elementos de evaluación de voces sintetizadas, así como pruebas de significancia estadísticas entre éstos y voces originales para evaluar la calidad de los resultados. Y finalmente, se han identificado áreas de potencial desarrollo a partir de la incorporación de métodos heurísticos y otros de inteligencia computacional para mejorar los procesos de creación de voces y su evaluación.

Abstract

Statistical parametric speech synthesis is a technique for producing artificial voices using Hidden Markov Models as the dominant mathematical model and a parametric representation of speech. This allows a voice to be coded using spectral parameters, fundamental frequency and duration of its phonetic units. This technique then trains the mathematical models to produce parameters of new sentences with significant advantages over other methods of speech synthesis, such as its greater flexibility and lower footprint.

Presented in this paper is the theoretical development, adaptation to linguistic and computational level, as well as a proposal for extensive experimentation and evaluation of artificial voices produced from statistical parametric speech synthesis for Latin American Spanish of Mexico. For this purpose we have defined a number of contexts of implementation and have adapted and developed computer applications as contributions at several levels, from the information extraction to the evaluation of results. This process has provided many opportunities for experimentation to study the influence of various factors on the quality of voices obtained.

The main contributions of this project are: The documentation of the theoretical and practical elements for creating voices using statistical parametric synthesis. According to the study of references carried out, this document is the first contribution to the documentation of both aspects simultaneously. The development of software for extraction and analysis of parameters, and the evaluation of results, in addition to applications developed to test the voices in realistic contexts. The incorporation of the acoustic parameters of pitch, jitter and shimmer as assessment elements of synthesized voices, and statistical significance tests to discriminate quality results. Finally,the identification of areas of potential development with the incorporation of heuristic methods or other computational intelligence methods to improve the processes of creating voices and evaluation methods.

Agradecimientos

Al **Consejo Nacional de Ciencia y Tecnología (CONACyT)** por haber otorgado el financiamiento que ha hecho posible la realización de este proyecto de investigación, así como a la **Universidad de Costa Rica** por el apoyo económico y administrativo brindado, gracias al cual ha sido posible desarrollar este grado académico en México.

A mis asesores, **Dr. John Goddard Close** y **M.I. Fabiola Martínez Licona** por su confianza depositada en mi persona, su decisivo impulso intelectual y personal, y sus sabios consejos a lo largo de todo el proceso de estudio e investigación.

A **Andrea** y **Gabriel** por todo el amor, apoyo y armonía ideales para poder luchar y entregarme a este proyecto de estudio.

A mi mamá **Mayra**, a mi hermana **Kattia** y hermano **Jorge** que han creído y apoyado este proyecto de estudios, y con su amor fraterno inspiran cada día.

Al **Dr. Alfonso Prieto Guerrero** y al **Dr. Sergio de los Cobos Silva** que ofrecieron el apoyo académico, administrativo y personal que me abrió las puertas a la Universidad y al país.

Al **Dr. Javier Trejos Zelaya** y **Dr. Jorge Romero Chacón** por haber creído en mi persona y apoyado desde la Universidad de Costa Rica este proyecto de estudios.

Al **Dr. José Abel Herrera Camacho** y a la **M.C. Alma Martínez Licona** por permitirme el honor de formar parte del jurado revisor y calificador para la presentación pública de este trabajo.

Al **Dr. Humberto Cervantes Maceda**, coordinador del posgrado, a los **profesores y compañeros de generación** por los conocimientos, orientación y ayuda brindada en tantos aspectos.

Contenido

Lista de Figuras

Lista de Tablas

Acrónimos

art	Razón de articulación
BW	Ancho de banda
CALL	Aprendizaje de idioma asistido por computadora
CD-HMM	HMM dependiente de contexto
DBN	Red de creencia profunda
DTW	Algoritmo de alineamiento temporal dinámico
dy	Cantidad de disfluencias
f0	Frecuencia fundamental
Fn	n-ésima formante
fp	Cantidad de pausas
FD-PSOLA	Algoritmo PSOLA en el dominio de la frecuencia
GV	Algoritmo de varianza global
HMM	Modelo oculto de Markov

HTK	Sistema de herramientas para HMM
HTS	Sistema de síntesis de voz basado en HMM
Hz	Hertz
LSF	Líneas de frecuencia espectral
MDL	Algoritmo de longitud mínima de descripción
MFCC	Coeficientes cepstrales en la escala de Mel
MGC	Cepstrum generalizado de Mel
MLLR	Regresión lineal de máxima probabilidad
mlp	Duración media de pausas
mlr	Duración media de frases
MLSA	Filtro aproximador de espectro logarítmico de Mel
MOS	Valor medio de opinión
MSD-HMM	HMM con distrubución de probabilidad multiestado
PCA	Análisis en componentes principales
PCM	Modulación por código de pulso
PSOLA	Algoritmo de superposición aditiva síncrona de tono
ptr	Razón de fonemas
RAPT	Algorimo Robusto para Rastreo de Tono
RBM	Máquina restringida de Boltzmann
RMSE	Error cuadrático medio

ros	Razón de habla
RP-PSOLA	Algoritmo PSOLA con referencia de tono
SAMPA	Alfabeto fonético legible por computadora
SLL	Aprendizaje de segundo idioma
TALP	Centro para el Lenguaje, Aplicaciones y Tecnologías del Habla
TD-PSOLA	Algoritmo PSOLA en el dominio del tiempo
tdp	Duración total de pausas
TTS	Sistema texto a habla
UPC	Universidad Politécnica de Cataluña
WER	Tasa de error en palabras

Capítulo 1

Introducción

Como punto de partida de la investigación realizada en el tema de síntesis estadística paramétrica, se presenta en este capítulo el contexto y antecedentes de la síntesis de voz, incluyendo aplicaciones y tendencias recientes. También se presenta la problemática, los objetivos planteados y la justificación. Adicionalmente se resume la metodología seguida para alcanzar los objetivos y se delinea la estructura del documento.

1.1. Contexto y antecedentes

La síntesis de voz es la producción de habla de forma artificial, de manera que pueda ser entendida e incoporada a procesos de comunicación con seres humanos. Al ser el habla el principal medio de comunicación entre personas, uno de los principales intereses en esta área es el desarrollo de sistemas que generen voz con suficiente claridad y naturalidad, para permitir y potenciar la interacción humano–máquina.

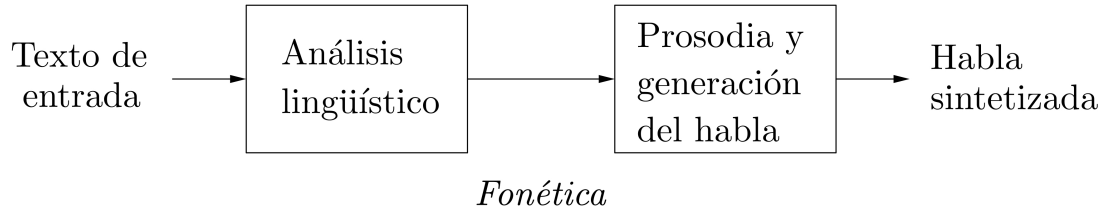


Figura 1.1: Procedimiento general de un sistema TTS [2]

1.1.1. Aspectos generales de la síntesis de voz

La producción artificial de voz se considera usualmente como la última etapa de un sistema texto a voz TTS (por las siglas en inglés de *Text to Speech*), que conlleva el procesamiento y análisis de información en formato texto, el cual es transcrito en información utilizable para generar una onda sonora de habla. El formato que se utiliza para la entrada usualmente es texto simple, pero pueden utilizarse otros, como lenguaje estructurado, para obtener mejores resultados [1]. En la Figura 1.1 se muestra el procedimiento general de un sistema TTS, donde al texto de entrada se le realiza un análisis lingüístico para tener una descripción fonética del mismo, con la cual se pueda generar la información del hablar para producir la onda de audio.

A pesar de que se los sistemas de síntesis han mejorado considerablemente las últimas décadas en cuanto a la calidad de voz resultante, y en su implementación existe en varios dispositivos y aplicaciones, se considera un área de investigación en constante crecimiento, pues se incorporan nuevas necesidades, como la producción de emociones y la generación de más voces con acentos y estilos de habla diferentes. Usualmente se desarrollan sistemas para contextos específicos, pues no es posible aún desarrollar un sistema que pueda controlar la calidad y el estilo de la voz de manera que pueda ser utilizado con todas las personas y todos

los contextos. No es suficiente la naturalidad que se logre, sino que es importante el estilo para que pueda ser mejor apreciada [3].

1.1.2. Aplicaciones

La primera aplicación comercial de la síntesis de voz fue una máquina de lectura introducida por R. Kurweil al final de la década de 1970, con la intención de ayudar a personas con discapacidad visual [2]. Consistía en un lector óptico de texto que era capaz de pronunciarlo de forma inteligible. Desde ese momento se identificó la inteligibilidad como una característica crucial en este tipo de aplicaciones, con la necesidad de mantenerla aún a distintas velocidades de habla. La naturalidad, es decir, su parecido con una forma de habla humana, también ha sido señalada como otro factor importante para hacer los sistemas de habla más aceptables al usuario.

Además de los sistemas de lectura, como el mencionado anteriormente, la síntesis de voz encuentra aplicaciones en sistemas de orientación para personas con discapacidad visual, e incluso traductores de señas a habla, para quienes se comunican utilizando este lenguaje. Existe interés por desarrollar síntesis de voz en idiomas con pocos hablantes (por ejemplo, lenguas nativas americanas), como un medio de promover el rescate y aprendizaje de éstos.

Los entornos educativos son otra área que ofrece oportunidades de desarrollo actuales, pues una computadora puede presentarse como un asistente interactivo para personas que no han aprendido a leer, o bien para el aprendizaje de otras lenguas. Presenta también utilidad potencial para personas que padecen dislexia en el proceso de lectura [4].

Como aplicaciones potenciales se señalan algunas que surgen al enfocarse en la mejora de las voces, no solamente en la conversión texto a voz, tales como [5] [6]:

- Asistentes personales incorporados en computadoras y dispositivos inteligentes.

- Clonación de voces.
- Reconstrucción de voz para personas con problemas degenerativos de habla.
- Sistemas de traducción habla a habla personalizados, es decir, traductores de voz a otro idioma que conserven las características del habla del emisor.
- Síntesis de voz adaptable al ruido.
- Análisis y reconstrucción de lenguas extintas.

En general cualquier sistema con necesidad de comunicación humano–computador es posible área de aplicación [2], incluso en videojuegos y sistemas de entretenimiento. Se puede señalar que el habla tiene el potencial para convertirse en el medio de comunicación principal entre personas y tecnología a futuro [7].

Para implementar la síntesis de voz han surgido diversos esquemas, con técnicas desde el modelado matemático del tracto articulatorio, la manipulación directa del audio o más recientemente una representación paramémtrica del mismo, tal como se presenta en la siguiente sección.

1.1.3. Tipos de síntesis

Las técnicas utilizadas en la actualidad para el proceso de síntesis de voz son presentadas por autores como [1] y [2], y se pueden clasificar en síntesis articulatoria, síntesis de formantes, síntesis concatenativa y síntesis estadística paramétrica (también llamada síntesis basada en Modelos ocultos de Markov). Esta división de técnicas obedece a la forma de analizar y producir el habla, lo cual no ha impedido la implementación híbrida de algunas de ellas. En las siguientes subsecciones se resumen las principales características de cada una.

Síntesis articulatoria

Este tipo de síntesis está basada en el modelado matemático del tracto vocal humano, de manera tiene el potencial de las voces de mayor calidad, si se cuenta con un modelo muy preciso. Sin embargo, es uno de los métodos más difíciles de implementar y tiene mayor costo computacional que otros de uso más reciente [2].

Las dificultades se deben principalmente a dos razones [1]: La primera es cómo generar los parámetros de control para el modelo del tracto vocal, y la segunda cómo hallar el balance entre un modelo que tenga alta precisión en cuanto a su aproximación a la fisiología humana, pero que sea tenga suficiente simplicidad para diseñarlo e implementarlo.

Este tipo de síntesis puede considerarse la más antigua, pues los primeros desarrollos, como la máquina de von Kempelen (primera máquina acústica-mecánica de producción, documentada en 1781), se realizaron para imitar el tracto vocal y el sistema articulatorio.

Síntesis de formantes

Las formantes son características espectrales que diferencian sonidos del habla. La síntesis de formantes crea los datos a través de reglas que correlacionan los diferentes sonidos del habla con sus espectros, sin utilizar grabaciones de voz durante la síntesis. Los resultados de esta técnica son usualmente poco naturales en comparación con sintetizadores más recientes [8]. El sintetizador Klatt, de principios de la década de 1980, es uno de los más sofisticados desarrollados a la fecha [1], y como otros sintetizadores característicos que utilizan formantes, permite flexibilidad y control sobre los parámetros acústicos.

Existen dos tipos de estructuras para incorporar la información de formantes a la onda sonora: en cascada y en paralelo, pero también son utilizadas estructuras híbridas. Se requieren al menos tres formantes del espectro del habla para producir resultados comprensi-

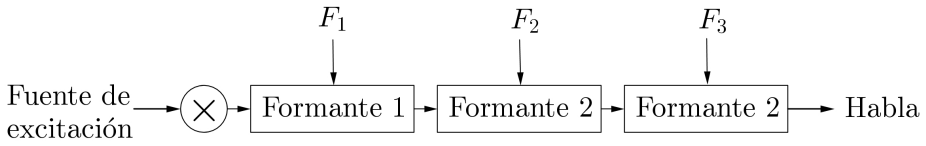


Figura 1.2: Estructura básica de un sintetizador de formantes en cascada [2]

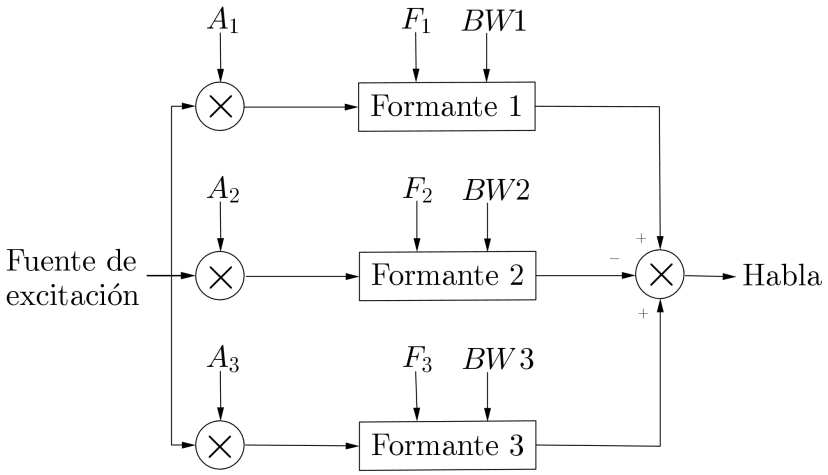


Figura 1.3: Estructura básica de un sintetizador de formantes en paralelo [2]

bles, y hasta cinco para producir habla de calidad. Cada formante es modelada mediante un resonador de dos polos, lo cual permite especificar la frecuencia del formante y el ancho de banda [2]. Un ejemplo de estructura en cascada se muestra en la Figura 1.2, y de estructura en paralelo en la Figura 1.3

Síntesis concatenativa

Este tipo de síntesis es considerada como la más sencilla de producir, más comprensible y usualmente considerada de sonido más natural [2] [8]. A grandes rasgos se basa en la utilización de grabaciones de habla real, las cuales son segmentadas y luego concatenadas

para generar nuevas frases sintetizadas.

Es común la utilización de difonos, es decir, secciones de voz que van desde la mitad de un sonido (fonema) hasta la mitad del siguiente [8]. La lista completa de difonos en un lenguaje es llamado inventario de difonos. Una vez que se determinan, es necesario encontrarlos en grabaciones reales. Para esto se siguen tres etapas [2]:

1. Grabar en lenguaje natural habla suficiente, de manera que se abarquen todos los difonos

2. Segmentar o etiquetar las unidades obtenidas en las grabaciones

3. Seleccionar las unidades más apropiadas

Los tres problemas principales que se encuentran al elaborar un sintetizador concatenativo son:

1. La distorsión producida por las discontinuidades en los puntos de concatenación

2. Altos requerimientos de memoria

3. La recolección y etiquetado de datos requiere gran cantidad de tiempo

Existe la posibilidad de seleccionar unidades de diferente tamaño, no solamente difonos, lo cual constituye la llamada síntesis de selección de unidades. En ésta, se utilizan grabaciones de larga duración (por ejemplo, varias horas) y se seleccionan unidades de diferente duración, de manera que se aproximen mejor a una pronunciación deseada, definida por ciertos parámetros [8]. Se utilizan además, algoritmos para suavizar las transiciones entre sonidos. En el Apéndice C se muestran algunos de los principales sintetizadores de voz comerciales, que utilizan en su mayoría la selección de unidades como técnica principal.

PSOLA

PSOLA no es en principio un método de síntesis de voz, sino una técnica para el suavizado en la concatenación de segmentos de audio. El nombre proviene de las siglas en inglés de superposición aditiva sincrónica de tono (*Pitch Synchronous Overlap Add*).

Existen varias implementaciones del algoritmo, que operan de forma semejante [2]. La premisa principal es que la voz humana puede caracterizarse por un tono, de manera que pueda utilizarse información sobre ese tono para sincronizar segmentos adyacentes, evitando discontinuidades de este parámetro [9].

El antecedente principal de este algoritmo es el de superposición aditiva síncrona SOLA (siglas en inglés de *Synchronous Overlap and Add*), utilizado como un recurso para alargar el tiempo de ejecución de un fragmento sonoro. En éste, el audio es dividido en bloques, en los cuales se define un área de superposición basado en la máxima similitud de los bloques. En esta área de superposición, a uno de los bloques se le aplica un efecto de intensidad ascendente, mientras que al otro se le aplica intensidad descendente. En la Figura 1.4 se muestra un esquema de aplicación del algoritmo a un fragmento de audio $x(t)$.

En el caso de PSOLA, de igual forma que el algoritmo anterior, la modificación del tono se puede realizar directamente sobre la onda. El primer paso es determinar los puntos donde se producen pulsos tonales, relacionados con la emisión de pulsos de aire a través del glotis. Estos puntos se utilizan como marcadores para generar ventanas correspondientes a cada periodo que mantiene el tono. Las ventanas se generan centradas en el punto de máxima amplitud [10].

Para aumentar el tono de la señal de voz se reduce la distancia entre los marcadores, y para reducirlo se incrementa. En el proceso de síntesis, las ventanas se unen con una distancia adecuada para el tono requerido en ese periodo. Existen efectos en el tiempo de emisión de

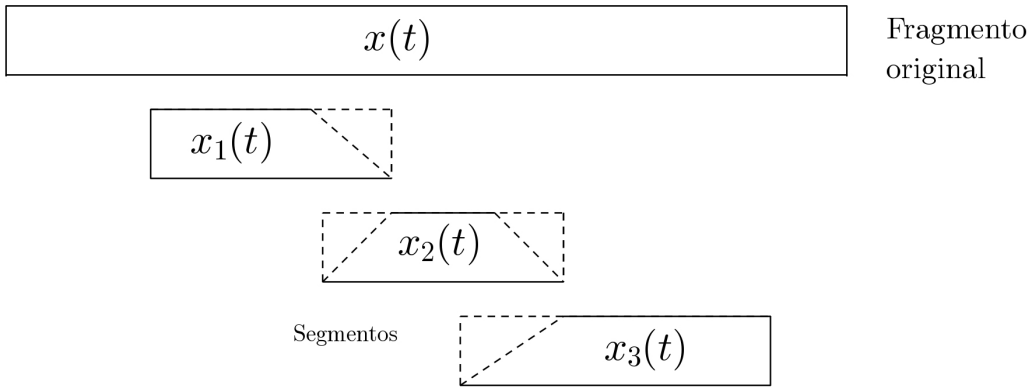


Figura 1.4: Esquema del algoritmo SOLA para manipulación del tiempo de ejecución de un fragmento de audio $x(t)$

los sonidos que deben ser corregidos, utilizando técnicas como el aumento en el tiempo de emisión de los sonidos que no tienen tono específico (como las consonantes no sonoras). Este tipo de síntesis requiere poco procesamiento de cómputo pero cantidades considerables de memoria para almacenar los segmentos de audio. En la Figura 1.5 se muestra un ejemplo de la división en ventanas y su posicionamiento en síntesis para bajar el tono de un sonido.

En conjunto con concatenación de fragmentos de habla ha sido implementada en sistemas de síntesis tales como FD-PSOLA o TD-PSOLA [11] (PSOLA en el dominio de la frecuencia o del tiempo), o RP-Psola (PSOLA con referencia de tono) [12]. El algoritmo sigue teniendo relevancia en la modificación temporal o de tono de audios de cualquier tipo.

Síntesis estadística paramétrica

En contraste con los esquemas concatenativos que han predominado en las aplicaciones de los últimos años, la síntesis estadística paramétrica se basa en la generación de parámetros que representan segmentos del habla. En una implementación típica, deben extraerse en primer lugar una representación paramétrica del habla, esto es, parámetros espectrales y de

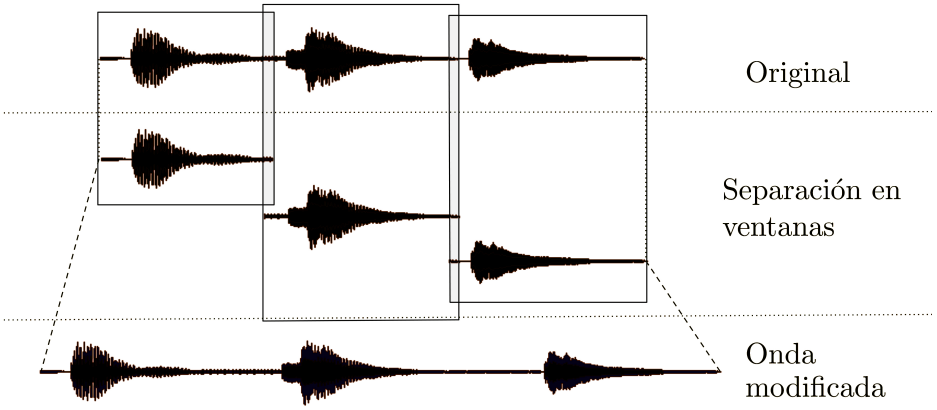


Figura 1.5: Esquema del algoritmo PSOLA para manipulación del tono de un fragmento de audio

frecuencia fundamental, para luego modelarlos usando algún modelo matemático generativo, siendo el más utilizado los Modelos ocultos de Markov HMM (por las siglas en inglés de *Hidden Markov Models*)[13].

Como se mencionó anteriormente, algunas de las aplicaciones de la síntesis de voz pueden tener como resultados aceptables aquellas voces que resulten inteligibles, aunque su estilo sea limitado en cuanto a expresividad o naturalidad. Para aplicaciones más exigentes en estos aspectos, la síntesis estadística paramétrica, ha sido objeto de intensa investigación por parte de sectores académicos y organizaciones comerciales [14], a pesar de que se considera que no alcanza la calidad de los mejores sistemas concatenativos actuales.

Este interés es debido no solamente a la ventaja de su potencial flexibilidad, o el menor tamaño de bases de datos requerida, sino a los eficiencias algoritmos de aprendizaje maquinal involucrados en el entrenamiento de los HMM. Por ejemplo, los algoritmos de Baum–Welch, Viterbi, y agrupamiento por árboles de decisión, utilizados desde hace varias décadas en el área de reconocimiento de voz han sido trasladados con éxito a la síntesis.

1.1.4. Retos actuales

Aunque ha sido señalado que las voces obtenidas mediante síntesis estadística paramétrica no alcanzan la calidad de los mejores sistemas concatenativos, es sabido que para producir las mejores voces con esta última técnica se requiere grandes cantidades de audio de alta calidad, en condiciones de ruido controladas. Aunque el contar con este tipo de grabaciones es una solución para obtener voces de calidad, el número de éstas que se pueden producir es limitado, debido a los altos costos económicos involucrados [15].

Como las aplicaciones presentes y potenciales de la síntesis de voz requieren la creación de muchos tipos de voces con cualidades distintas, es de importancia reducir los requerimientos para generar más voces y estilos de habla. La síntesis estadística paramétrica ha abierto el camino para lograr estas propiedades en la generación del habla artificial, pero los procesos involucrados encuentran numerosas dificultades y oportunidades de mejora.

Además de estos aspectos, la evaluación de la calidad resultante es tema de gran relevancia, para poder discriminar las voces más convenientes. Aún no existe consenso sobre los elementos que deban considerarse para determinar la manera en que se puedan discriminar las voces de mayor calidad en una aplicación dada.

Con el fin de promover el desarrollo de la síntesis de voz a partir de estos retos, el *Blizzard Challenge* se ha constituido como un espacio destacado de muestra e intercambio científico. Se trata de una convocatoria abierta para desarrollar voces a partir de una base de datos común [16], con métodos de evaluación también comunes a todas las voces que se desarrollan. Han sido usuales en las distintas convocatorias anuales sintetizadores que utilizan selección de unidades, así como modelos estadísticos paramétricos e híbridos de ambos.

1.2. Planteamiento del problema

En la actualidad existen más de veinte implementaciones de síntesis estadística paramétrica en diversos idiomas. Sin embargo, sobre la técnica en sí no se encuentran referencias que documenten el proceso completo de creación de voces, partiendo de la adaptación de los sistemas computacionales desarrollados para el entrenamiento de los HMM, las definiciones relacionadas con los sonidos específicos del idioma, y la teoría que sustenta esta técnica.

Por otra parte, existe un vasto campo de posibilidades para la experimentación sobre la mejora en la calidad de las voces resultantes, que pasa por el estudio de la influencia de los múltiples parámetros involucrados en todos los procesos: extracción de parámetros, entrenamiento y síntesis. Estas posibilidades han sido poco exploradas en las implementaciones realizadas hasta el momento.

En este proyecto se aborda la creación de voces utilizando esta nueva técnica de síntesis, a la vez que se pretende conjugar información teórica y práctica, pretendiendo responder los siguientes aspectos:

- ¿Cuáles son los principales fundamentos teóricos que sustentan la técnica?
- ¿Cuáles son los requerimientos computacionales para implementar nuevas voces?
- ¿Cuál es la influencia de los datos y condiciones de entrenamiento de los modelos matemáticos involucrados en el proceso, en las voces resultantes?
- ¿Cuáles métodos de evaluación aplicar para determinar la calidad en los resultados?

Para responder estas cuestiones, de importancia en la comprensión, utilización más extendida y mejora en la calidad de las voces obtenidas con síntesis estadística paramétrica, es

necesario documentar aspectos teóricos de los modelos matemáticos y lingüísticos involucrados, y los requerimientos computacionales de implementación a nivel de programas. También se requiere un marco de experimentación y evaluación extensos, con los cuales sea posible también identificar áreas donde se puedan aportar propuestas para abrir nuevos caminos a la investigación de síntesis en nuevos lenguajes, o mejorar la calidad en los existentes.

De acuerdo con el estudio de referencias realizado, este es el primer trabajo basado en esta técnica de síntesis de habla que reúne tanto el desarrollo teórico como una documentación de su implementación. También se establece como el primer trabajo que produce voces en una variante de español de América Latina sobre las que se ha realizado una evaluación extensiva de resultados, tanto objetivos como subjetivos, a partir de una experimentación en varios niveles.

1.3. Justificación

A pesar del interés generado entre investigadores y organizaciones comerciales en la síntesis estadísitca paramétrica, se encuentran hasta el momento solamente dos referencias de su implementación en el idioma español, desarrollados en 2010 en España [17] (un sistema híbrido de HMM con técnicas concatenativas), y 2013 en México [18]. Toda implementación actual tiene como núcleo el sistema HTS, el cual consiste en las herramientas informáticas necesarias para definir, entrenar y extraer parámetros de los HMM.

El HTS no cuenta hasta el día de hoy con una guía de usuario, tutorial u otras ayudas con las que su adaptación a nuevas idiomas, voces o estilos que lo haga más asequible. Por esta razón se hace necesario, junto con la comprensión teórica del proceso de producción del habla y los modelos matemáticos implicados en la síntesis estadística paramétrica, realizar aportes a la documentación de los requerimientos para la adopción del HTS, con los que se

pueda potenciar un uso más extensivo y mejoras en más contextos y aplicaciones.

Esta adaptación requiere un análisis del código desarrollado para las distintas etapas del entrenamiento, desde el análisis del texto hasta la generación de parámetros para conformar las formas de onda del habla, y una comprensión de la relación de la teoría con la implementación. Por estas razones, además de realizar pruebas en HTS con conjuntos de datos previamente no utilizados, con distintos estilos de hablantes, y para poder correlacionar la variación de los parámetros de entrenamiento con la calidad de voz resultante, es necesario contar con conocimientos de la producción humana del habla, la fonética y fonología propia del idioma a implementar y el análisis de señales de voz. Todo esto con la finalidad de dar un marco de referencia adecuado para comprender los fenómenos que se presentan en el entrenamiento y la síntesis resultante, e identificar áreas de potenciales mejoras.

Como otros métodos de síntesis, la estadística paramétrica parte de grabaciones reales de habla. En la Universidad Autónoma Metropolitana han sido desarrolladas bases de datos de alta calidad, tanto de voz de hombre como de mujer, replicando textos diseñados para obtener una cobertura adecuada de fonemas en español, para fines de estudios en reconocimiento de voz. La calidad de los datos ha sido señalada como un punto de partida importante para la calidad de cualquier método de síntesis de voz, por lo que estas bases de datos son propicias para iniciar un estudio de síntesis estadística paramétrica, a pesar de constituir un cuerpo de datos de reducido tamaño en comparación con el utilizado en otras implementaciones.

Dada la característica de tamaño de las bases de datos, es necesario delimitar los alcances de las frases que es posible pronunciar en la síntesis. Para esto se requiere definir contextos de aplicación, es decir, aplicaciones específicas que deben ser desarrolladas para limitar el rango de palabras y frases posibles de generar.

Para completar el estudio, es necesario una evaluación adecuada de los resultados, de

manera que el aporte del proyecto al desarrollo de la síntesis estadística paramétrica de voz pueda constituir una investigación replicable a nuevos estilos o lenguas, con análisis que permita cuantificar las futuras mejoras y su incorporación a aplicaciones en varios tipos de dispositivos. Por estas razones se ha formulado el presente proyecto de investigación, y así reunir los aspectos teóricos y prácticos de la creación de voces en español de México con la técnica de síntesis estadística paramétrica a partir de los datos con que se cuenta.

La información y desarrollos pueden servir como aporte a la incorporación de más voces y desarrollar sintetizadores en otras variantes de español u otras lenguas de América Latina.

1.4. Objetivos

Para lograr los propósitos de esta investigación se han planteado los siguientes objetivos:

1.4.1. Objetivo general

Desarrollar un sistema de síntesis estadística paramétrica de voz basada en HMM para el español de México.

1.4.2. Objetivos específicos

- Conocer información relevante para la síntesis de voz sobre la fonética, prosodia, y la lingüística del español
- Conocer los HMM y su la aplicación a la síntesis de voz
- Adaptar el sistema HTS a la creación de voces en español

- Estudiar la influencia de diferentes parámetros y condiciones de entrenamiento en las voces sintetizadas

- Conocer y aplicar métodos de evaluación de voces sintetizadas

1.5. Metodología

La metodología propuesta para este proyecto de investigación contempla las siguientes fases:

- Investigación. Para alcanzar los objetivos propuestos, se requiere un estudio teórico de los temas:

    - Procesamiento de la señal de voz, para una adecuada comprensión y adaptación del modelado de la voz en la extracción de parámetros propios de la síntesis estadística paramétrica.
    - Estudio básico de la fonética, prosodia y la lingüística del español, para poder definir adecuadamente las unidades del habla que se modelarán con los HMM, y establecer un lenguaje preciso con el que puedan identificarse los fenómenos que se escuchan como resultado de la síntesis.
    - Estudio de los HMM, que incluya los algoritmos de entrenamiento y el conocimiento teórico que sustenta su aplicación en la síntesis de voz.
    - Estudio de referencias sobre implementaciones de síntesis estadística paramétrica de voz, para identificar los avances recientes y los resultados esperados del proyecto en cuanto a características del audio.

- Adaptación. La implementación de síntesis estadística paramétrica está basada en el sistema HTS, los cuales permiten la definición, entrenamiento y extracción de parámetros de los HMM. De forma complementaria, se requiere un analizador de texto que permita la transcripción de una información escrita a información utilizable con los HMM. Por estas razones se requiere comprender y adaptar los sistemas:

  - HTS a partir de su código, pues no existen documentos de referencia para su utilización. El estudio del código da la posibilidad de comprender a mayor profundidad la manera en que la teoría está implementada y analizar cómo puede incorporarse futuras mejoras.

  - *Festival* como analizador de texto desarrollado para español de España, y los requerimientos para su adaptación a la nueva variante de español. Se propone la utilización de este sistema por ser un software gratuito y de código abierto, lo cual permite realizar los ajustes necesarios para la variante del español a tratar.

- Experimentación. No existen referencias sobre un estudio extensivo de la influencia en los resultados de la síntesis de los numerosos parámetros involucrados en el entrenamiento de los HMM o de las características de los datos utilizados como punto de partida. Por estas razones es necesario delimitar la experimentación, de manera que pueda significar un aporte al conocimiento de la influencia de estos parámetros a varios niveles. Para esto es necesario cubrir los siguientes aspectos:

  - Definir los parámetros de los datos o del entrenamiento que puedan afectar significativamente los resultados de la síntesis.

  - Establecer una serie de experimentos de prueba de los parámetros sobre la base de datos disponible.

  - Definir contextos de aplicación donde puedan ser implementadas las voces resultantes. Este punto involucra el desarrollo de aplicaciones computacionales.

- Reporte y evaluación de resultados. Finalmente se deben recopilar los resultados de la experimentación y realizar una evaluación adecuada de los mismos, a partir de la cual se puedan verificar los resultados esperados a partir del estudio teórico y de referencias, realizar propuestas de mejoras e indicar futuros desarrollos.

1.6. Estructura del documento

Este documento está estructuro de la siguiente manera:

- En el **Capítulo 2** se presenta el estado del arte de la síntesis estadística paramétrica, construido a partir de un estudio extensivo de las referencias. Se incorporan elementos teóricos de los Modelos Ocultos de Markov y otros modelos matemáticos utilizados en el procesamiento del habla. Se exponen las principales ventajas y desventajas de la técnica en la producción de voces artificiales.

- En el **Capítulo 3** se describe el desarrollo de la propuesta, con la adaptación del sistema HTS para producir voces en español de México, así como el planteamiento de la experimentación y métodos de evaluación de resultados.

- En el **Capítulo 4** se presentan los resultados del desarrollo de aplicaciones, la adaptación de los elementos fonéticos necesarios y la experimentación, incluyendo su evaluación con técnicas adaptadas de las referencias y otras propuestas para este trabajo.

- En el **Capítulo 5** se realiza un resumen y análisis de resultados, a partir de la información de los sistemas desarrollados para la evaluación de las voces sintetizadas, las

pruebas estadísticas y los métodos de correlación.

- En el **Capítulo 6** se presentan las conclusiones principales y las recomendaciones para futuros trabajos en el tema.

- Finalmente en el **Apéndice a** se cubre un desarrollo teórico más amplio sobre los HMM, en el **Apéndice B** se recopilan los resultados estadísticos principales utilizados en el análisis de resultados, en el **Apéndice C** se destacan algunas de los principales sintetizadores de referencia en la actualidad, y en el **Apéndice D** se recopila el estudio de los sistemas HTS y programas desarrollados como parte del proyecto como un primer intento conocido de realizar una documentación de referencia.

Capítulo 2

Estado del arte

En este capítulo se resumen el desarrollo teoríco que explica el uso de los HMM en la síntesis de voz y otros modelos matemáticos de importancia en el procesamiento del habla. Se hace una revisión del estado del arte en cuanto a la síntesis estadística paramétrica a partir de un estudio de referencias, las implementaciones y propuestas recientes para mejorar los resultados con esta técnica.

2.1. Modelos Ocultos de Markov

La síntesis estadística paramétrica es un tipo de síntesis basada en un modelo. Este modelo tiene dos características fundamentales: es paramétrico pues describe el habla usando parámetros en lugar de unidades pregrabadas, y es estadístico pues estos parámetros se describen mediante sus valores estadísticos, como medias y varianzas en distribuciones de probabilidad [19]. Los modelos matemáticos más utilizados para reproducir los parámetros involucrados en el habla son los Modelos Ocultos de Markov (HMM). Estos pueden explicarse a partir de un proceso de Markov, en el cual las transiciones entre estados están dadas por un proceso estocástico. Un segundo proceso estocástico modela la emisión de símbolos cuando se llega a cada estado. En la Figura 2.1 se muestra una representación de un HMM del tipo

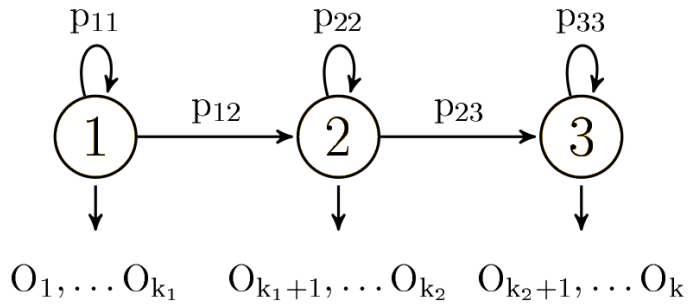


Figura 2.1: Ejemplo de un HMM tipo izquierda a derecha, con tres estados.

izquierda a derecha, llamado así pues se puede identificar un primer estado a la izquierda, a partir del cual se pueden dar transiciones hacia el mismo estado o hacia los siguientes a la derecha, pero no en sentido inverso. En esta p_{ij} representa la probabilidad de transición entre el estado i y el estado j, y O_k representa la observación emitida en el estado k.

Un HMM puede describirse mediante una tupla $\lambda = (S, \pi_i, a, b)$, donde S es el conjunto de estados, π el vector de probabilidades, en el cual la entrada i establece la probabilidad de que el estado i sea el inicial. a es la matriz de transición de probabilidad entre estados, y b las probabilidades de emisión de las observaciones en cada estado. La observación puede ser un símbolo de un alfabeto predefinido (HMM discreto), o bien un vector de coeficientes (HMM continuo). En síntesis de voz estos modelos se entrenan para emitir observaciones correspondientes a parámetros con los cuales se pueda generar audio de habla sintetizada.

En el Apéndice A se desarrolla la teoría de los HMM en cuanto a su definición formal y sus algoritmos de entrenamiento. Su primera relación con temas de habla fue como un clasificador en sistemas de reconocimiento de voz. A partir de su éxito en aplicaciones en esa área, se trasladó su utilización a la síntesis en la década de 2000. En las siguientes secciones se desarrollan los fundamentos teóricos de esta utilización de los modelos.

2.2. Descripción del proceso

En una aplicación típica de síntesis de voz, se parte de una base de datos de audio, que consiste en grabaciones de habla con su correspondiente transcripción en formato texto. De forma semejante a como se hace en reconocimiento, se debe realizar una segmentación de la base de datos, es decir, el reconocimiento de cuáles segmentos de las grabaciones corresponden a las unidades de habla que se van a modelar, típicamente fonemas. En las siguiente etapa se asigna un HMM por cada fonema, tomando en cuenta el contexto en que se encuentra. En la Figura 2.2 se muestra un esquema global del proceso, el cual se detallará en las siguientes subsecciones.

Como descripción general, se puede establecer que una vez se etiquetan los segmentos de audio, se procede a extrar representaciones paramétricas del habla, con las cuales se entrenan los HMM, utilizando un criterio de máxima probabilidad para estimar sus parámetros, tal como [13]

$$\hat{\lambda} = \arg\max_{\lambda} \{p(O|\mathcal{W}, \lambda)\}, \tag{2.1}$$

donde O es el conjunto de datos de entrenamiento y $\mathcal{W}$ es un conjunto de parámetros correspondientes a la especificación lingüística (como tono o espectro de fonemas).

En el proceso de síntesis se generan los parámetros del habla para un conjunto de unidades w, que corresponden a unidades fonética que se desean pronunciar, seleccionándolas del conjunto de modelos estimados $\hat{\lambda}$, al maximizar sus probabilidades de salida:

$$\hat{o} = \arg\max_{o} \left\{ p(o|w, \hat{\lambda}) \right\}. \tag{2.2}$$

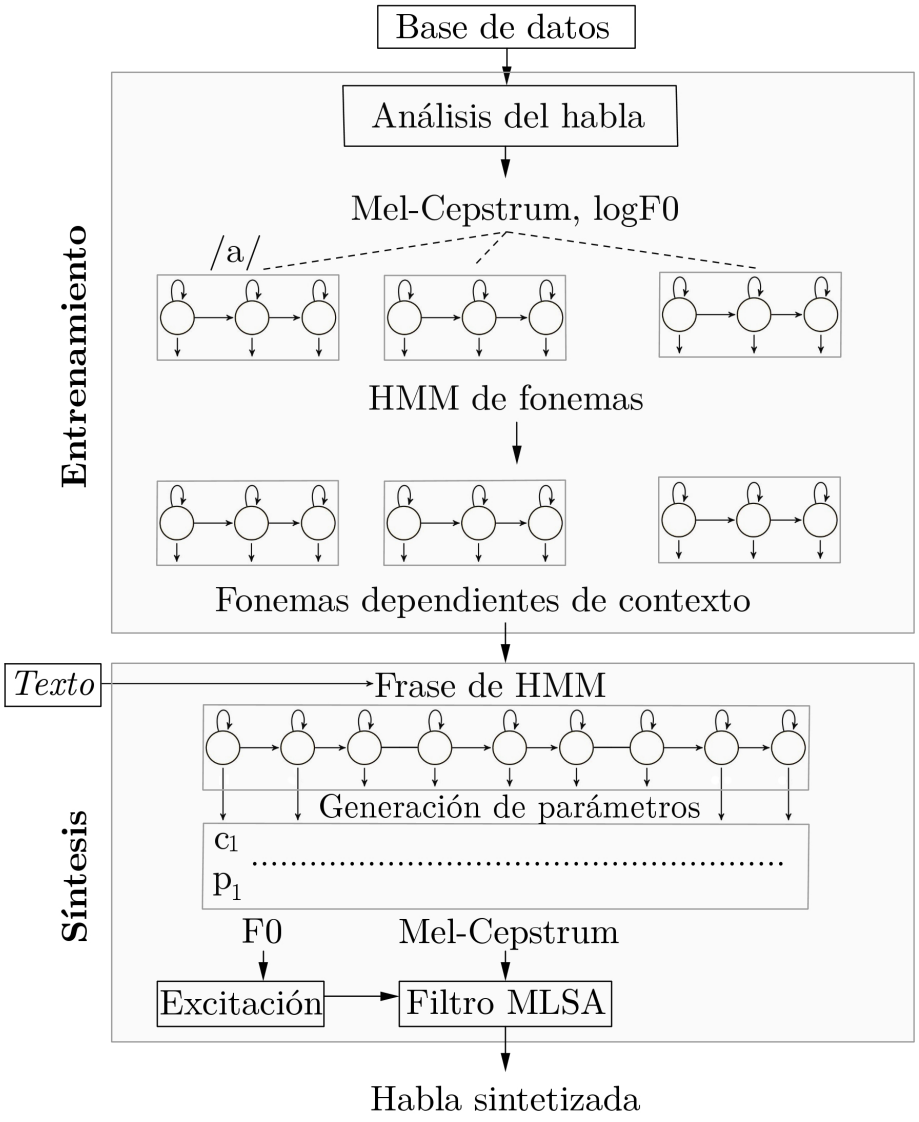


Figura 2.2: Esquema general de la síntesis estadística paramétrica de voz. Adaptado de [20]

En las siguientes subsecciones se presentan estos procedimientos con mayor detalle.

2.2.1. Análisis del habla

Este proceso consiste en la preparación de los datos del habla de la base de datos para poder utilizarla como referencia paramétrica en la síntesis. Esto requiere los siguientes pasos [17]:

1. Conversión grafema a fonema: Mediante reglas propias de cada idioma, se deben transcribir los textos presentes en la base de datos a fonemas, realizar separación silábica, de palabras y de frases, utilizando reglas específicas de cada idioma.

2. Segmentación: Consiste en identificar y establecer, en los audios de la base de datos, fronteras temporales entre todas las unidades fonéticas. Este proceso es semejante a lo que se realiza en reconocimiento de habla, donde se asume que una observación O (parámetros en el audio) es producida por una secuencia W de elementos de texto, por lo que se desea determinar la secuencia $\hat{W}$ más probable, es decir $\hat{W} = \arg\max_{W}(p(W|O))$. Luego, utilizando teorema de Bayes

$$\hat{W} = \arg\max_{W}\left(p(W|O)\right) = \arg\max_{W}\left(\frac{p(O|W)p(W)}{p(O)}\right), \tag{2.3}$$

con lo cual, dadas probabilidades previas $p(w)$ se hace depender el proceso de $p(O|w)$.

2.2.2. Extracción de parámetros

El procedimiento base para la generación de parámetros en la síntesis estadística paramétrica fue presentado en [21], y ampliado en [20]. Parte del hecho que un sintetizador necesita

información sobre espectro y tono para producir la onda de habla. Esta información debe tener como referencia parámetros de habla real, los cuales se extraen del audio de la base de datos, y deben estar ligados a información fonética del texto, obtenida a través del análisis realizado sobre el mismo.

Una vez que se tiene la correspondencia entre las fronteras temporales del audio y su correspondiente fonema, la extracción se parámetros se realiza definiendo un tipo de ventana, un ancho de ésta y su corrimiento temporal. En reconocimiento de voz, clásicamente se utilizan solamente los coeficientes espectrales, pero para el caso de síntesis, es de importancia la extracción de la frecuencia fundamental del habla.

Análisis en tiempo reducido

La señal de voz varía de forma lenta con respecto a las frecuencias de muestreo que pueden utilizarse en la actualidad, de manera que el proceso del habla puede separarse en bloques, en los cuales los propiedades de la forma de onda pueden considerarse constantes [22]. El análisis en tiempo reducido se representa mediante un vector de parámetros X en el instante $\hat{n}$, el cual se expresa mediante la ecuación

$$X_{\hat{n}} = \sum_{m=-\infty}^{\infty} T\{x[m]w[\hat{n}-m]\}, \tag{2.4}$$

donde $w[\hat{n}-m]$ representa una secuencia definida por una ventana, que selecciona un segmento de la secuencia $x[m]$. Usualmente se utiliza la ventana de Hamming, definida como

$$w_H[m] = \begin{cases} 0.54 + 0.43\cos\left(\frac{\pi m}{M}\right) & -M \leq m < M \\ 0 & \text{en otro caso} \end{cases} \tag{2.5}$$

De una señal de voz que es separada mediante ventanas, es posible obtener la información

espectral mediante la transformada de Fourier discreta en tiempo reducido, la cual es define como la transformada discreta de Fourier de la señal $x_{\hat{n}}[m] = x[m]w[\hat{n} - m]$, mediante la expresión [22]:

$$X_{\hat{n}}\left(e^{j\hat{w}}\right) = \sum_{m=-\infty}^{\infty} x[m]w[\hat{n} - m]e^{-j\hat{w}m} \tag{2.6}$$

Esta representación da información espectral puntual. Para la representación gráfica de la información espectral de la señal de voz a lo largo del tiempo, dadas las particularidades que contiene, se utiliza el espectrograma. El espectrograma es una representación tridimensional, en la cual el eje horizontal es el tiempo, el eje vertical la frecuencia, y se representa con colores (o escala de grises) la amplitud de la componente en frecuencia en cada punto. Los espectrogramas pueden ser de banda amplia o entrecha, dependiendo del tamaño de la ventana utilizado [1]. En el de banda amplia los armónicos no son distinguibles, pero es una representación adecuada de los formantes. En el caso del de banda estrecha, se tiene una representación detallada de los armónicos. En las Figuras 2.3 y 2.4, se muestran ejemplos de espectros de banda amplia y estrecha, respectivamente.

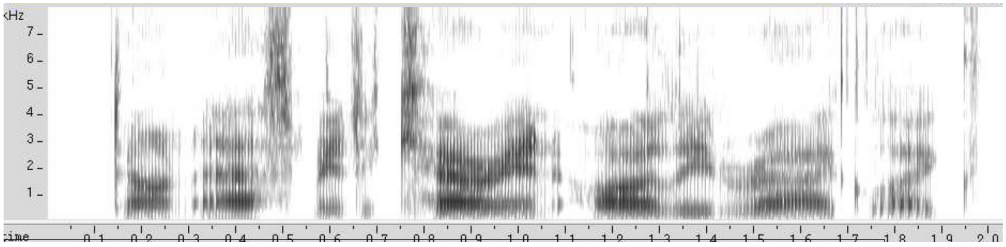


Figura 2.3: Ejemplo de espectrograma de banda amplia [1]

Otra representación espectral útil es la presentación con línea de espectros, o LSF (por las siglas en inglés de *Line-spectrum frequencies*), como el que se muestra en la Figura 2.5. Cuando las lńeas se juntan se considera evidencia de que se encuentra una formante.

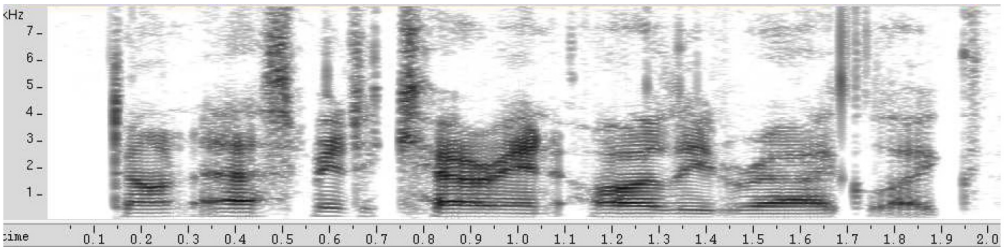


Figura 2.4: Ejemplo de espectrograma de banda estrecha [1]

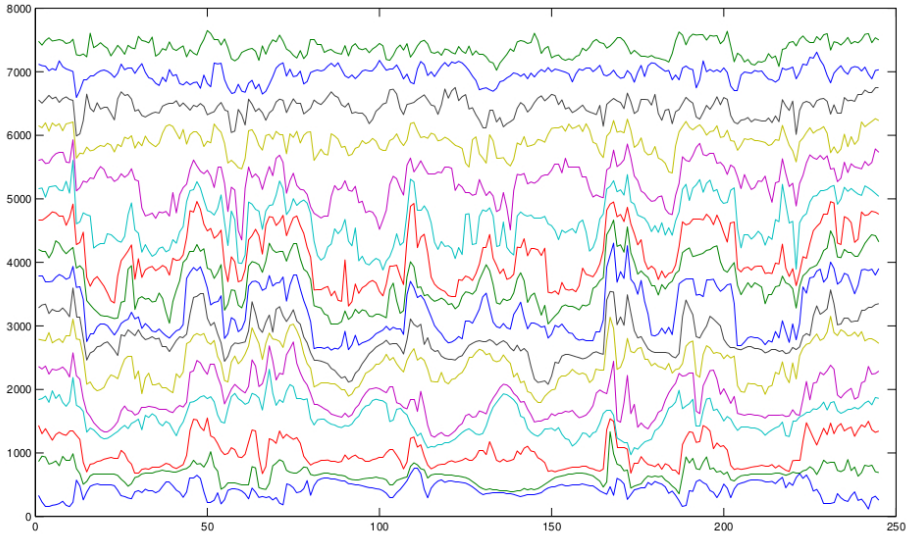


Figura 2.5: Ejemplo de representación espectral con LSF [1]

El espectrograma se utiliza para distinguir características de hablantes, y es posible utilizarlo en algunos métodos de evaluación de síntesis de voz, por ejemplo, al compararlos con espectros de habla natural. El uso de ventanas para el procesamiento de audio es de importancia en la síntesis estadística paramétrica para el establecimiento de los segmentos en los cuales se extrae la representación paramétrica, como el cepstrum, desarrollado en la siguiente subsección.

El cepstrum

Para el modelado de la señal de voz, se requiere una representación adecuada de ésta que contemple particularidades como la caracterización del eco. Para este fin, Bogert, Healy y Tukey, en 1963, definieron el cepstrum, como la inversa de la transformada de Fourier del espectro de magnitud de una señal [22]. Con este análisis de la señal, es posible detectar la presencia de resonancias propias de la voz. Pocos años después, en 1965, Oppenheim, Shafer y Stockham generalizaron el concepto dentro del área de filtrado de señales, y plantean el cepstrum de una señal discreta como

$$c[n] = \frac{1}{2\pi} \int_{-\pi}^{\pi} log \left| X(e^{jw}) \right| e^{jwn} dw, \tag{2.7}$$

donde $X(e^{jw})$ es la transformada en tiempo discreto de la señal. También definieron el cepstrum complejo como

$$\hat{x}[n] = \frac{1}{2\pi} \int_{-\pi}^{\pi} \left(\log |X(e^{jw})| + j \arg[X(e^{jw})] \right) e^{jwn} dw \tag{2.8}$$

Estas representaciones relacionan el espectro de una señal con un conjunto de coeficientes. En el análisis ceptral actual, es común la representación de la señal de voz mediante los coeficientes cepstrales en la frecuencia mel, o MFCC (por las siglas en inglés de *Mel-Frequency Cepstral Coefficient*). Estos se obtienen a partir de una combinación de la transformada discreta de Fourier y la transformada discreta de coseno [1]

$$c_i = \sqrt{\frac{2}{N}} \sum_{j=1}^{N} m_j \cos\left(\frac{i\pi}{N} (j - 0.5) \right). \tag{2.9}$$

En la Figura 2.6 se muestra un diagrama de bloques del proceso seguido para obtener los

MFCC a partir del espectro.

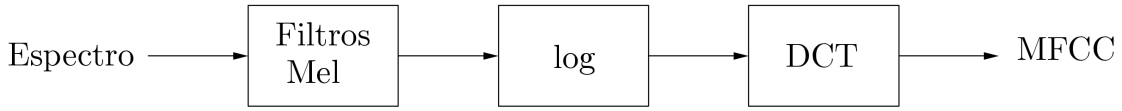


Figura 2.6: Diagrama de bloques para extracción de coeficientes MFCC

Siguiendo este análisis, la función de transferencia del tracto vocal se modela mediante los coeficientes ceptrales de orden M: $c = [c(0), c(1), \ldots . c(M)]^\top$ [20]. La frecuencia de Mel se refiere a una escala de percepción de tonos equidistantes, de manera que los coeficientes MFCC constituyen una representación del audio en una escala que es análoga a la percepción humana. La Figura 2.7 muestra un ejemplo gráfico del banco de filtros utilizado para obtener los coeficientes en la escala de Mel.

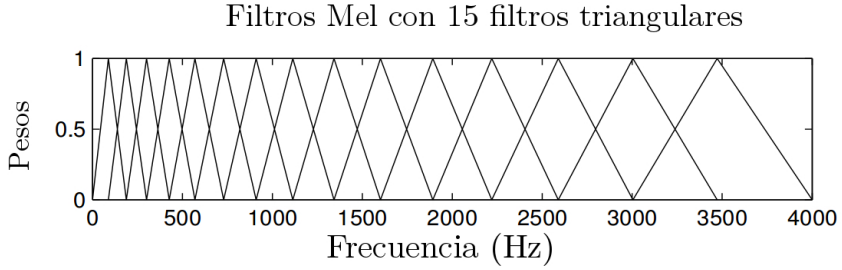


Figura 2.7: Diagrama de filtros para obtener escala Mel [23]

La representación del espectro de habla mediante MFCC tiene dos ventajas principales [1]:

- Tiene la propiedad de independencia entre coeficientes, lo que permite que sus distribuciones de probabilidad puedan modelarse con matrices de covarianza diagonal.
- La escala Mel ha mostrado tener la propiedade de discriminar mejor los fonemas.

La representación de un espectro utilizando coeficientes cepstrales es llamado MGC (por las siglas en inglés de *Mel Frequency Cepstrum*). La utilización del logaritmo permite modelar mejor la escucha humana, en la cual las intensidades del sonido no se perciben de forma lineal, además de proveer ventajas en el manejo algebraico.

Esta representación se ha usado con éxito para el reconocimiento de voz, y ha sido trasladada con éxito a la síntesis estadística paramétrica. Con los coeficientes cepstrales la función de transferencia del tracto vocal se aproxima mediante [20]

$$H(z) = \exp\left[c^\top \tilde{\boldsymbol{z}}\right] \tag{2.10}$$

$$= \exp \sum_{m=0}^{M} c(m)\tilde{z}^{-}m, \tag{2.11}$$

donde $\tilde{\boldsymbol{z}} = [1, \tilde{z}^{-1}, \ldots, \tilde{z}^{-M}]^\top$. $\tilde{z}^{-1}$ se define como un filtro de primer orden

$$\tilde{z}^{-1} = \frac{z^{-1} - \alpha}{1 - \alpha z^{-1}}, \qquad |\alpha| < 1. \tag{2.12}$$

La escala de transformación de frecuencia, para aproximarla a la escala de audición humana se representa mediante su respuesta de fase

$$\beta w = \tan^{-1} \frac{(1 - \alpha^2)\,\mathrm{sen}\, w}{(1 + \alpha^2)\cos w - 2\alpha}, \tag{2.13}$$

la cual debe ajustarse con el parámetro α, dependiente de la frecuencia de muestreo del audio. De esta manera se cuenta con un modelado del tracto vocal y los coeficientes del espectro del habla que están relacionados con la forma de percepción del sonido en los seres humanos.

Detección de f0

La detección de f0 consiste en la extracción de la frecuencia fundamental en la señal de habla. Los algoritmos para obtenerla usualmente tienen tres etapas [1]:

1. Preprocesamiento: Consiste en un filtrado y resampleado a frecuencias más bajas.
2. Determinar un límite de posibilidades de valores
3. Decidir entre las posibilidades

Existen varias técnicas que implementan el procedimiento. Una de las más utilizadas es la autocorrelación, con la ventaja sobre otras de sencillez de implementación y que se realiza en el dominio del tiempo.

La autocorrelación mide qué tan semejante es una función consigo misma cuando se desplaza, así que tendrá valores altos cuando uno tono de $f0$ está presente en un instante y se traslapa con el siguiente (como un valor sostenido a lo largo de una vocal). Al detectar estos valores se puede estimar $f0$. Programas como Praat [24] implementan esta técnica, mientras que en el análisis del habla realizado como parte del sistema HTS [25] se utiliza el Algorimo Robusto para Rastreo de Tono, RAPT [26] (por las siglas en inglés de *Robust Algorithm for Pitch Tracking*), el cual se basa en el análisis de la señal a diferentes frecuencias de muestreo, con mayor complejidad de cálculos pero generalmente mejores resultados.

HMM multiestado

La voz humana tiene la particularidad de emitir sonidos con frecuencia fundamental, representables mediante un valor numérico de frecuencia, y sonidos sin ella, representables mediante un símbolo unidimensional. Por esta razón, el modelado de $f0$ requiere la utilización

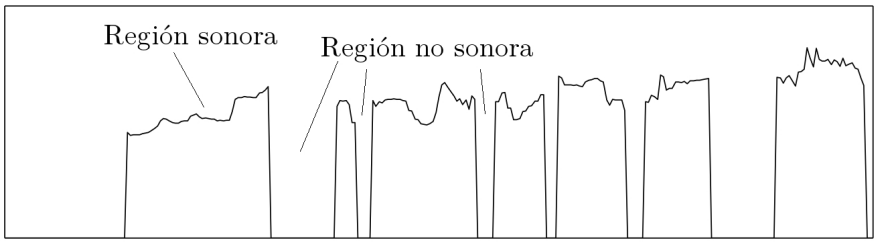


Figura 2.8: Muestra de contorno de $f0$ de un fragmento de habla

de una distribución de probabilidad multiestado MSD-HMM (por las siglas en inglés de *Multi-Space Probability Distrbution Hidden Markov Model*) [20]. Estas fueron introducidas como distribuciones para las observaciones generadas por los HMM en [27].

Los vectores que representan $f0$ tienen entonces componentes que se pueden considerar como parte de un espacio unidimensional Γ_1 con distribución de probabilidad normal, y componentes de un solo símbolo (cero dimensionales), de un espacio Γ_0. De esta manera, las observaciones generadas por los HMM pueden modelarse mediante una variable aleatoria x, e índices X que indican un sonido con $f0$ ($X = 1$), o sin él ($X = 0$), o sea

$$\boldsymbol{o} = (X, \boldsymbol{x}). \tag{2.14}$$

En la Figura 2.8 se muestra un ejemplo de contorno de $f0$ de un fragmento de audio de voz, donde se observan las regiones sonoras (con $f0$), y no sonoras (sin $f0$). La probabilidad de observación de una secuencia de $f0$ en un MSD-HMM se define como

$$b(\boldsymbol{o}) = \sum_{g \in X} w_g \mathcal{N}(\boldsymbol{x}). \tag{2.15}$$

w_g es el peso asignado a cada espacio. w_g satisface $\sum_{g=1}^{2} w_g = 1$. Tanto espectro como $f0$ pueden representarse mediante un MSD-HMM, en el cual la parte espectral se modela con

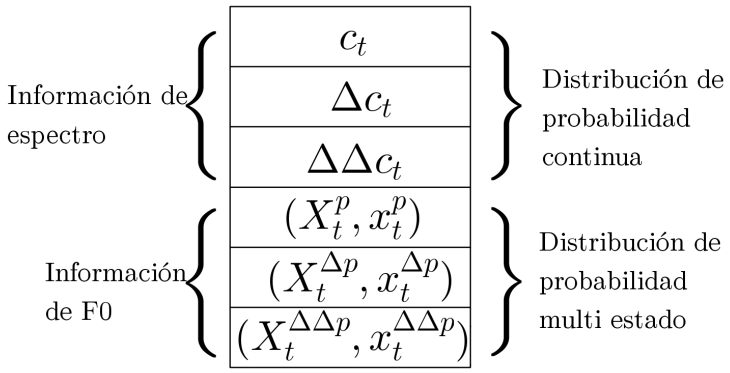


Figura 2.9: Estructura de un vector de coeficientes que representan el habla

distribución de probabilidad continua, y la parte de $f0$ como una distribución multiestado. La Figura 2.9 ilustra la estructura de un vector de parámetros que representa una ventana de habla.

Los coeficientes Δ y $\Delta\Delta$ representan aproximaciones discretas de la primera y segunda derivada de los coeficientes a través del tiempo, para incorporar elementos que miden la variabilidad en la evolución de los parámetros.

2.2.3. Entrenamiento

El proceso de entrenamiento consiste en el ajuste de parámetros de los HMM que representan cada fonema. Se desea maximizar la probabilidad de que cada HMM pueda reproducir el conjunto de vectores correspondiente al fonema en la base de datos. Como se indicó en la ecuación (2.1), esto se plantea como $\lambda_{max} = \arg\max_{\lambda} p(\boldsymbol{O}|\lambda, W)$

donde

$$p(\boldsymbol{O}|\lambda, W) = \sum_{\forall q} \pi_{q_0} \prod_{t=1}^{T} a_{q_{t-1}q_t} b_{q_t}(\boldsymbol{O}_t). \tag{2.16}$$

π_{q_0} representa la probabilidad de que q_0 sea el estado inicial, $b_{q_t}(\boldsymbol{O})$ la probabilidad de emisión de a observación $\boldsymbol{O}$ en el estado q_t y $a_{q_{t-1}q_t}$ la probabilidad de transición del estado q_{t-1} a q_t. W son los fonemas contenidos en la base de datos.

Los fonemas se presentan diferencias de acuerdo con su contexto, es decir, su posición en la frase, en la palabra, acento y otros factores. Estas diferencias se dan en cuanto a su energía y tono (información prosódica).

Para representar más adecuadamente estas variaciones, se requiere diferenciar los fonemas no solamente por su presencia en la frase, sino tomando en cuenta estos elementos. El principal problema en cuanto a estimación de los parámetros de los fonemas dependientes de contexto, que se representan mediante HMM dependientes de contexto, o CD-HMM (por las siglas en inglés de *Context-Dependent Hidden Markov Models*, es la cantidad de parámetros que se deben calcular en el entrenamiento, y la gran cantidad de información en base de datos que se requiere para hacer un ajuste adecuado.

Si la información de contexto que se utiliza es amplia, por ejemplo, considerando la posición del fonema en la sílaba, en la palabra o en la frase, el número de palabras de la frase, los fonemas precedentes y siguientes; toda aparición de un fonema en una base de datos puede ser única, lo que llevaría a representar cada aparición de fonema con un HMM individual. Para solventar este inconveniente, en [28] y [29] y se introdujo el uso de agrupamiento por árboles de decisión, donde se diferencian los HMM por su contexto, pero se agrupan los que sean semejantes para estimar sus parámetros. De esta manera, se usa la información de una manera que sea factible para el entrenamiento de HMM con bases de datos finitas. En la Figura 2.10 se muestra un ejemplo de las preguntas que se pueden utilizar para agrupar los CD-HMM.

Se utilizan árboles de decisión distintos para agrupar los parámetros de HMM de tono y de

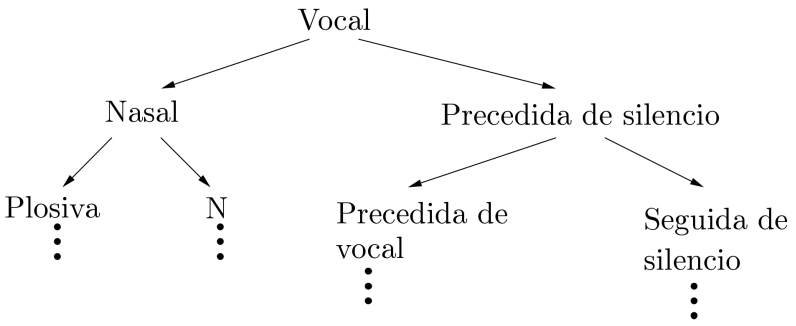


Figura 2.10: Ejemplo de árbol de decisión para agrupar CD-HMM

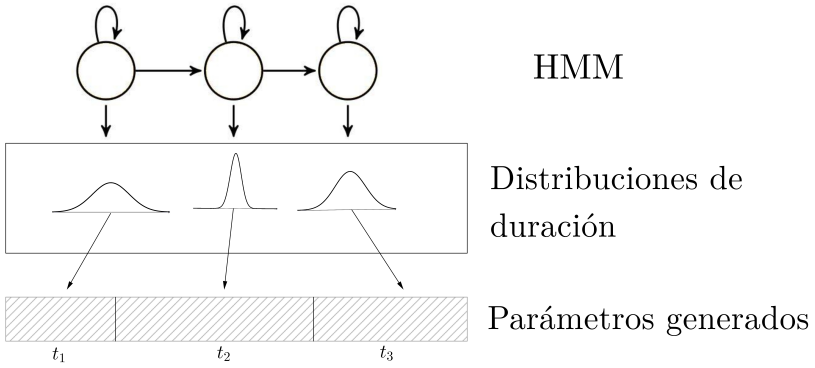


Figura 2.11: Modelado de duración de estados en HMM con distribuciones gaussianas

espectro. Adicionalmente, se requiere un modelado de la duración de ocupación de cada estado en los HMM, relacionado con la cantidad de símbolos que se emiten en cada uno. En [21] se plantea que esto puede realizarse mediante distribuciones de probabilidad gaussianas, las cuales se modelan a partir de las estadísticas de duración en cada estado, determinadas en el entrenamiento. En la Figura 2.11 se esquematiza un HMM con generación de observaciones que considera el modelado de duración de esta manera.

2.2.4. Síntesis

El proceso de síntesis parte de un texto, el cual debe ser convertido en formato identificable con los CD-HMM, de forma semejante al proceso efectuado al inicio del entrenamiento. Este proceso realiza un análisis lingüístico y genera una secuencia de fonemas con información de contexto. Se debe armar la secuencia de CD-HMM correspondiente a estas secuencias de fonemas, para generar los coeficientes cepstrales y de tono, siguiendo las duraciones establecidas por las distribuciones de probabilidad correspondientes.

Como último paso, los coeficientes generados pr los CD-HMM concatenados son procesados por un filtro MLSA (por las siglas en inglés de *Mel Log Spectral Approximation*), que genera la señal de audio correspondiente al habla del texto original. En la Figura 2.12 se esquematiza el proceso de síntesis a partir del texto.

La emisión de observaciones en los HMM está relacionada con la media y la desviación estándar de las distribuciones de probabilidad de emisión de símbolos. En la Figura 2.13 se representa la generación de parámetros para una secuencia de fonemas. En línea discontinua la media de las salidas obtenidas, sombreado la desviación de las salidas, y en línea continua los parámetros generados.

En cuanto a la reconstrucción de la forma de onda a partir de la información de $f0$ y MGC, se utiliza un filtro llamado MLSA (por las siglas en inglés de *Mel Log Spectrum Approximation*), que parte del principio de modelar la producción de voz como un sistema fuente-filtro. En la Figura 2.14 se ilustra este proceso. Los sonidos sonoros se generan a partir de un tren de pulsos con frecuencia $f0$, mientras que los no sonoros con un ruido blanco. Se utiliza un filtro digital con respuesta al impulso $h(n)$ para reconstruir la onda, y a la salida del filtro se genera ésta como convolución de la entrada y la fuente $e(n)$.

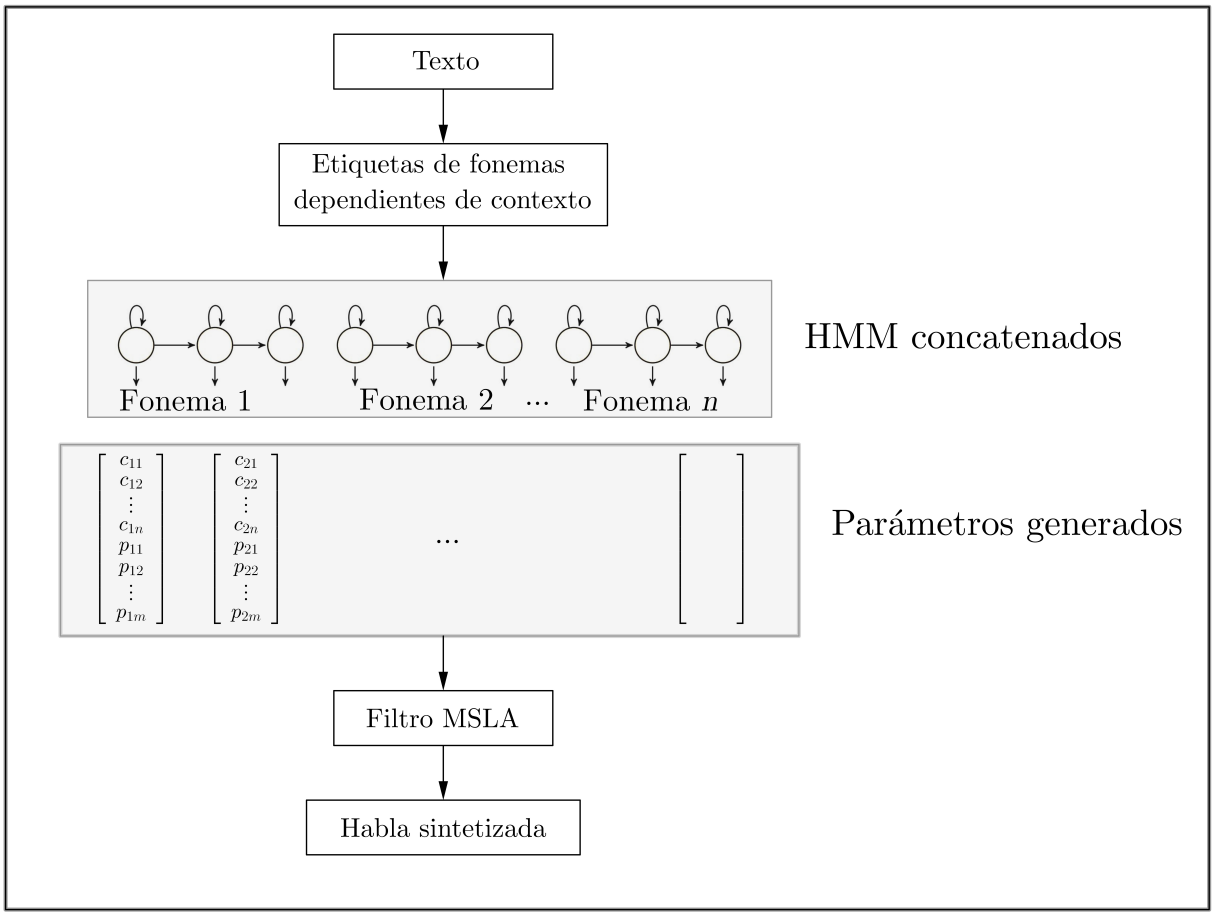


Figura 2.12: Esquema de conversión texto a habla con síntesis estadística paramétrica basada en HMM

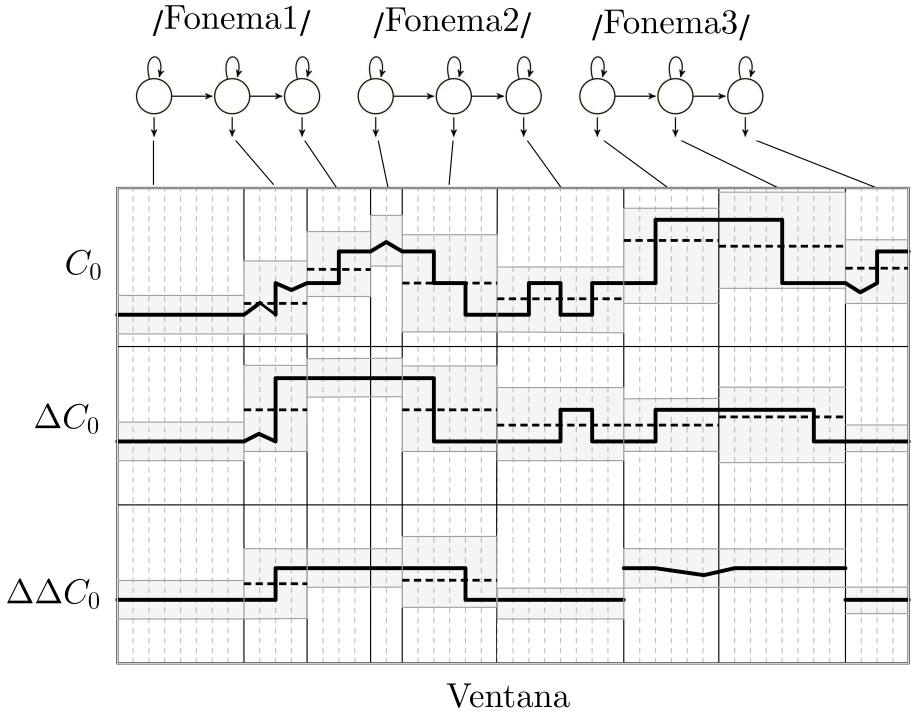


Figura 2.13: Ejemplo de generación de coeficientes a partir de HMM. Adaptada de [14]

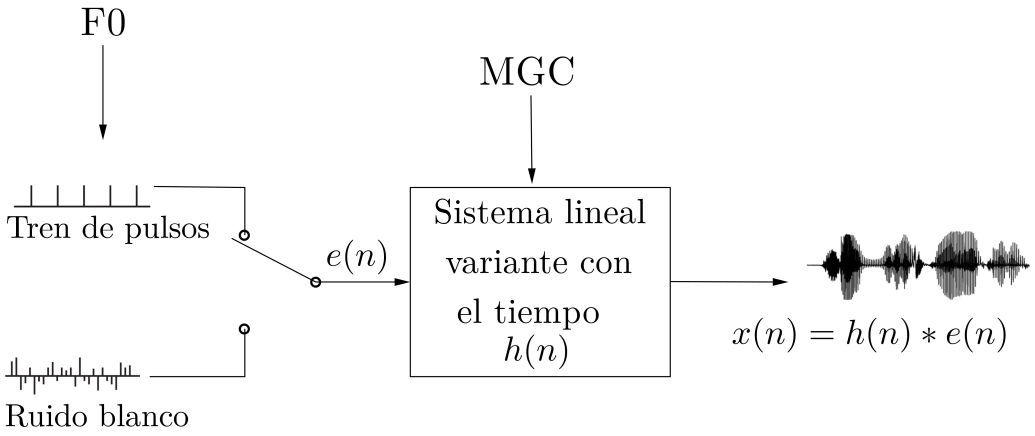


Figura 2.14: Esquema de fuente-filtro para la reconstrucción de forma de onda a partir de la información de $F0$ y MGC como salida de HMM. Adaptada de [20]

2.3. Ventajas

Las principales ventajas de la síntesis estadística paramétrica radican en su flexibilidad de cambiar estilos de habla, características de la voz y emociones [14]. Las cuatro principales técnicas existentes para lograrla [13] se detallan en las siguientes subsecciones.

2.3.1. Adaptación

Las técnicas de adaptación, originalmente desarolladas en reconocimiento de voz, se refieren a ajustar un modelo adecuadamente entrenado para equipararlo a otro producido con menos datos. El modelo adecuadamente entrenado se toma como un "modelo promedio", sobre el cual se aplican técnicas para transformarlo en el "modelo destino", a partir de las características que se pueden recabar de él usando los datos disponibles.

La principal técnica desarrollada para este fin es la Regresión Lineal de Máxima Probabilidad MLLR (por las siglas en inglés de *Maximum-Likelihood Linear Regression*). Con ésta, se estiman transformaciones lineales para mapear las distribuciones de probabilidad gaussianas del modelo promedio hacia el modelo destino. En la Figura 2.15 se ilustra este procedimiento.

Una de las principales ventajas de la creación de voces por adaptación es que se requiere solamente de unos minutos de grabación para obtener nuevas voces de calidad, con la condición de tener una base de datos adecuada como modelo promedio.

2.3.2. Interpolación

La técnica de interpolación surge del hecho de que, al utilizar representaciones paramétricas en este tipo de síntesis, los parámetros resultantes de los HMM pueden interpolarse para crear nuevas voces. A esta técnica se le llama también mezcla de voces, pues al contar con

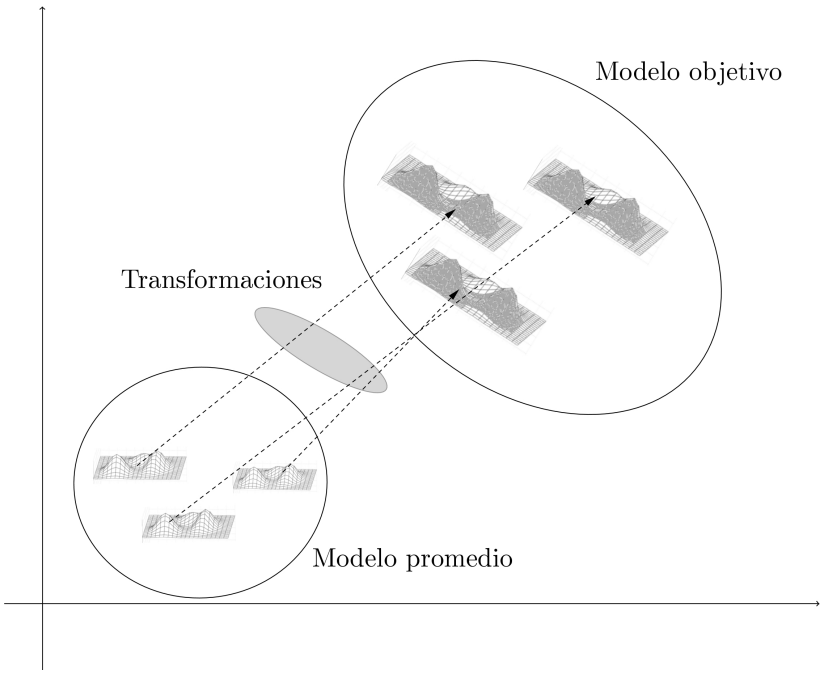


Figura 2.15: Esquema de la técnica de entrenamiento de voces por adaptación. Adaptada de [13]

dos o más voces que se pueden producir otras a partir de modelos de HMM, interpolando las ya obtenidas. En la Figura 2.16 se esquematiza la idea de este principio, donde $l(\lambda', \lambda_k)$ es la razón de interpolación, y λ_k es un HMM.

Con esta técnica se pueden producir voces con estilos de habla, características de voz y dialectos que no se encuentran en las bases de datos.

2.3.3. Producción de voces

La producción de voces por interpolación permite la creación de nuevas voces al cambiar la razón de interpolación entre los modelos. Sin embargo, si se cuenta con gran cantidad de modelos, es un problema complejo determinar la razón de interpolación requerida para lograr una voz con características deseadas [14]. Para resolver este problema, en [30] se propone

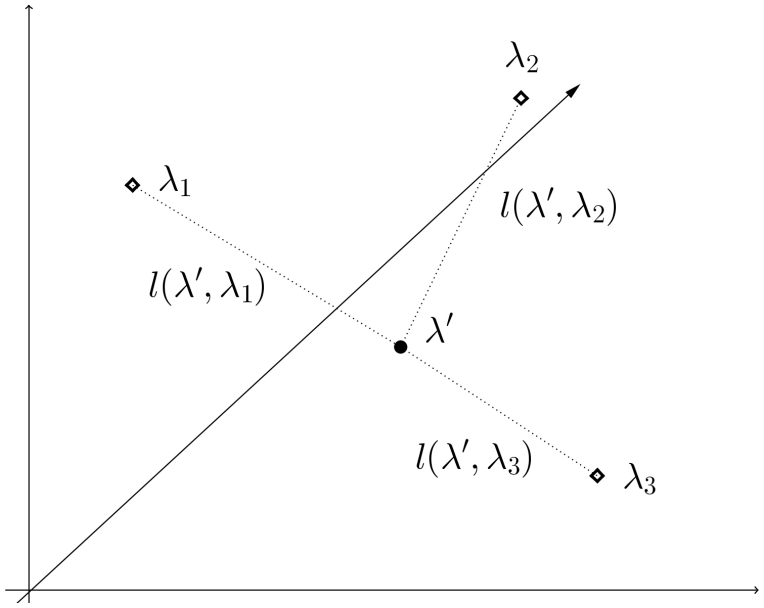


Figura 2.16: Esquema de la técnica de entrenamiento de voces por adaptación. Adaptada de [13]

utilizar análisis en componentes principales (PCA).

Al aplicar este análisis a S super-vectores (vectores con todos los parámetros de los HMM), es obtienen autovalores y autovectores. Al tomar solamente los de menor orden se reduce eficientemente la dimensionalidad del problema. En la Figura 2.17 se muestra la construcción de un super-vector utilizando la media, y un nuevo super-vector con el uso de valores y vectores propios.

2.3.4. Regresión múltiple

Una de las desventajas del enfoque de producción de voces es la dificultad de controlar las características de la voz de forma intuitiva, ya que los autovectores no están ligados a una característica física específica del habla. Para resolver el problema, en [31] se propone controlar estas características mediante un vector de control, en el cual cada elemento representa

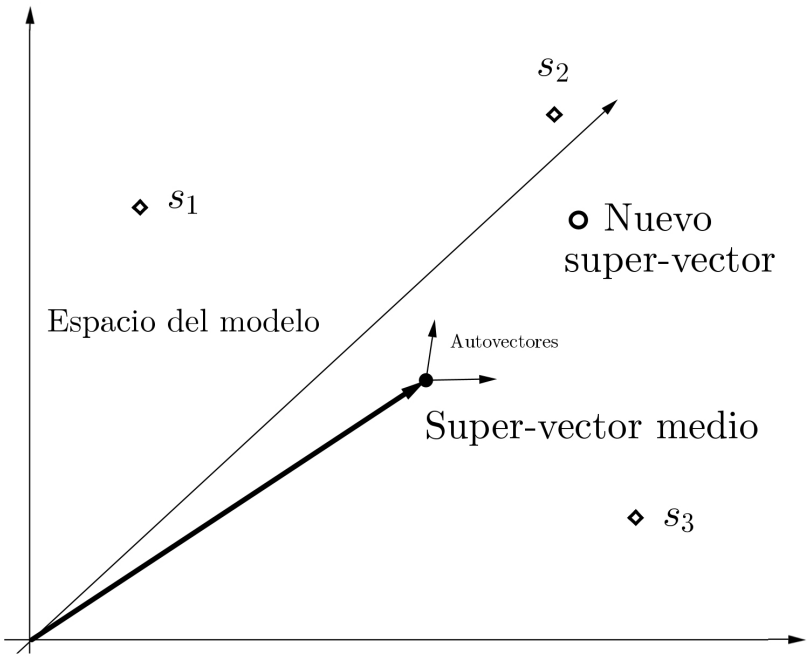


Figura 2.17: Esquema de la técnica de entrenamiento por producción de voces. Adaptada de [14]

un estilo de habla, emoción, género u otros. Al combinar estas técnicas se pueden obtener voces con características deseadas sin tener que producir nuevas bases de datos.

2.3.5. Otras ventajas

Otro atributo de la síntesis estadística paramétrica señalada en [13] [14]: el menor tamaño de almacenamiento y memoria (*footprint*) de las voces resultantes, lo cual las hace adecuadas para dispositivos portátiles. Otra ventaja es la robustez, con respecto a los errores que pueda tener la base de datos en cuanto a ruido o fluctuaciones en las condiciones de grabación [32].

2.4. Desventajas

La principal desventaja señalada en la literatura sobre síntesis estadística paramétrica es la calidad, con respecto a los mejores sintetizadores actuales que utilizan selección de unidades. Las razones principales que explican esta desventaja [13] se amplían en las siguientes subsecciones.

2.4.1. Vocoder

El vocoder se refiere al codificador de voz, que utiliza el modelo fuente-filtro. La desventaja de este modelo es su simplicidad, pues parte de un tren de pulsos y un ruido blanco.

Por otra parte, el uso de MGC con el modelo fuente-filtro para la construcción del espectro puede ser la causa de los reportes de voz sintetizada que la califican como voz con un componente de zumbido (*buzzy)*. Para resolver este problema, se han realizado propuestas, entre las que destaca STRAIGHT [33] que incorpora otros parámetros, clasificados como periódicos y aperiódicos.

En cuanto a la representación espectral, en lugar de los coeficientes MFCC se han propuesto la utilización de parámetros llamadas Par Espectral de Línea, o LSP (por las siglas en inglés de *Line Spectral Pair*), que han logrado mejoras en la percepción subjetiva con respecto a los primeros [34]. Existe una amplia línea de investigación en este sentido.

2.4.2. Modelado acústico

Como los parámetros del habla sintetizada son generados del modelo acústico (los HMM), su precisión para reproducir las características del habla afecta directamente la calidad resultante. Las principales propuestas para mejorar los HMM han sido los *trended HMM* [35], los

cuales incluyen funciones lineales en las probabilidades de emisión de los estados y los HMM de trayectoria (*trajectory HMM*) [36], los cuales han logrado mejoras bajo ciertas condiciones.

A pesar de estos logros, los HMM continúan siendo los modelos más utilizados en la actualidad.

2.4.3. Suavizado

En el proceso de modelado de los parámetros de voz en los HMM se realiza un promedio de éstos, proveniente de las diferentes apariciones de los elementos fonéticos en la base de datos. Por un lado, esto da robustez ante errores en la base de datos, pero como resultado se obtiene una voz más apagada con respecto a la natural.

Para esto se han realizado varias propuestas como el uso de Varianza Global GV (por las siglas en inglés de *Global Varianze)* [37], la cual define diferentes niveles para realizar el modelado acústico. En la Figura 2.18 se muestra un ejemplo de espectrograma de líneas de voz natural, uno generado sin GV y el otro con GV. Se aprecia un mayor dinamismo en el espectro que incluye GV, lo cual lo acerca más a la voz natural.

2.5. Revisión de literatura

En esta sección se hace una revisión de las publicaciones de trabajos relacionados con la síntesis estadística paramétrica de voz. Se divide la revisión en dos partes, la primera relacionada con la implementación en otros idiomas, y la segunda con las propuestas más recientes para el desarrollo futuro de la técnica.

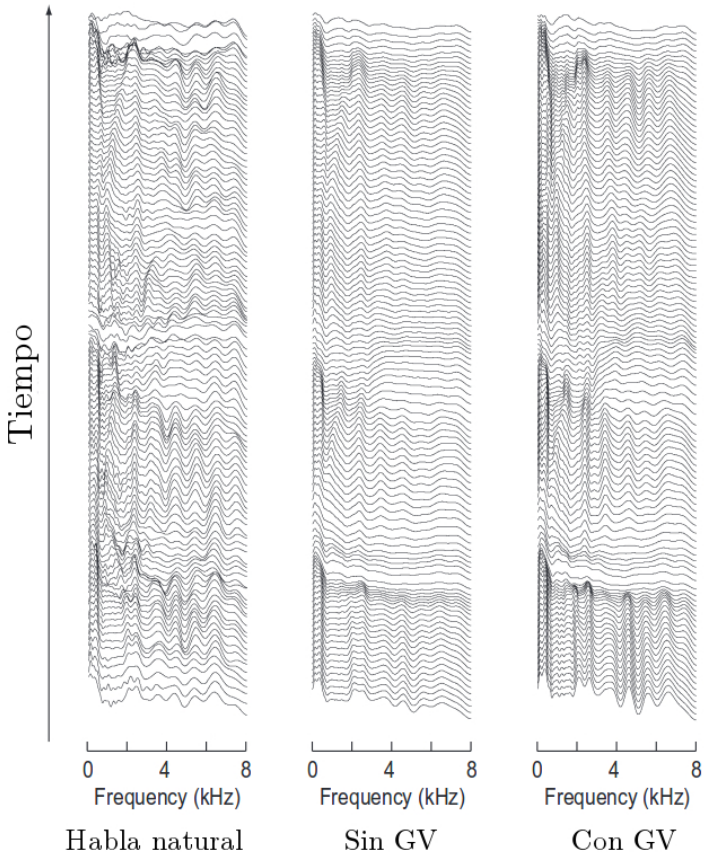


Figura 2.18: Comparación de espectros de voz natural con voz sintetizada que incluye algoritmo GV y sin él. Adaptada de [13]

2.5.1. Implementación en otros idiomas

Los primeros usos de los HMM en la síntesis de voz no utilizaron este modelo como generador de parámetros de la manera que se hace en la actualidad. Su propuesta fue utilizarlos, con salidas discretas, para mapear los fonemas. El uso de HMM para la generación de parámetros a partir de salidas continuas, nace en los trabajos de Mausko et al. [38] y Tokuda et al. [39] [17].

A partir del año 2002, con la creación del sistema HTS, las implementaciones han utilizado los HMM como modelo principal y se han incorporado mejoras para lograr las ventajas descritas en la sección anterior. En la Tabla 2.1 se hace una listado de autores, lenguaje y año de publicación.

Tabla 2.1: Desarrollo de síntesis estadística paramétrica en varios lenguajes

Año	Lenguaje	Autores
1996	Inglés	Donovan y Woodland [40]
1999	Inglés	Donovan [41]
2000	Japonés	Tokuda et al. [42]
2002	Inglés	Tokuda et al. [43]
2003	Portugués Brasileño	Maia et al. [44]
2004	Eslovaco	Vesnicer et al. [45]
2005	Persa	Hendessi et al. [46]
2005	Portugués europeo	Barros et al. [47]
2006	Mandarín	Qian et al. [48]. Zhu, Li [49]
2006	Croata	Martincic-Ipsic e Ipsic [50]

Continúa en la próxima página

Tabla 2.1 – continúa de la página anterior

Año	Lenguaje	Autores
2006	Coreano	Kim at al. [51]
2006	Árabe	Abdel-Hamid et al. [52]
2007	Castellano	Gonzalvo et al. [53]
2007	Alemán austríaco	Pucher et al. [54]
2007	Alemán	Krstulović et al. [55]
2008	Griego	Karabetsos et al. [56]
2009	Catalán	Bonafonte et al. [57]
2010	Mandarín	Li et al. [58]
2010	Francés	Lanchantin at al. [59]
2010	Castellano, Inglés	Gonzalvo [17]
2010	Vasco	Erro et al. [60]
2010	Checo	Hanzlíček [61]
2010	Xitsonga	Baloyi et al. [62]
2011	Rumano	Stan et al. [63]
2011	Lhasa	Lu et al. [64]
2012	Sueco	Bollepalli et al. [65]
2013	Tamil	Boothalingam et al. [66]
2013	4 lenguas ibéricas	Alonso et al. [67]
2013	Español mexicano	Herrera et at. [18]
2013	Vietnamita	Phan et at. [68]

A partir del año 2003, es posible observar la implementación en diferentes idiomas, incluyendo algunos con pocos hablantes nativos, como el Xitsonga (de Sudáfrica). En el caso de español, la única referenciada se da en España, en los años 2007 y 2010, y México, en 2013.

En el Cuadro 2.2 se muestra una comparativa de características importantes de las implementaciones de síntesis estadística paramétrica para aquellas referencias que las detallan: Los sistemas utilizados, el tamaño de la base de datos y el método de evaluación. Puede observarse cómo es común que las bases de datos superan las 500 frases y/o los 60 minutos por hablante.

Tabla 2.2: Comparación de sistemas y base de datos

Año, idioma	**Base de datos**	**Evaluación**
2002, Inglés	542 frases	–
2006, Croata	1111 frases, 85 minutos	–
2007, Alemán	4 hablantes, 1500 frases (aprox. 3 horas) cada uno	–
2008, Griego	1200 frases	MOS de naturalidad e inteligibilidad
2009, Catalán	10 hablantes, 9000 palabras (1 hora) cada uno	–
2010, Castellano	49 minutos	MOS de naturalidad e inteligibilidad y medidas objetivas
2010, Checo	2 hablantes, 5 horas	MOS calidad
2010, Rumano	3500 frases, 3.5 horas	MOS naturalidad e inteligibilidad

Continúa en la próxima página

Tabla 2.2 – continúa de la página anterior

Año, idioma	Base de datos	Evaluación
2013, Tamil	3732 frases, 5 horas	MOS calidad

Se destaca el hecho de que no todas las referencias disponibles utilizan sistemas de evaluación cuantitativas, o una descripción detallada de las características de sus bases de datos, parámetros de entrenamiento o la adaptación realizada a partir del sistema HTS.

2.6. Propuestas más recientes

La investigación más reciente se centra en aspectos como la mejora en los modelos de generación de parámetros [69], [70], la inclusión de nuevos parámetros [71], o el uso de modelos para dar mayor flexibilidad a la generación de nuevas voces [72]. Además de lo mencionado en la Sección 2.3 (mejoras en la creación de nuevas voces y la flexibilidad de la síntesis), y en la Sección 2.4 (para resolver las desventajas que presenta, conformando líneas de investigación abiertas, se destacan aquí las propuestas recientes que pueden dar origen a futuros desarrollos [14]:

- Inteligiblidad en ruido: En ambientes de ruido controlado, la inteligibilidad de voces sintéticas basadas en síntesis estadística paramétrica puede compararse a la voz natural. Sin embargo, en ambientes con ruido su inteligibilidad decrece considerablemente más que la de la voz natural.

  Para mejorar este aspecto, se ha planteado modificar el habla sintetizada mediante modelos estadísticos o adaptativos. Por ejemplo, el llamado efecto Lombard, en el cual

el hablante incrementa características como el tono, velocidad y duración de sonidos en presencia de ruido para hacerse entender, el cual ha sido analizado en síntesis de voz en [73].

- Uso de producto de expertos: Se utiliza esta técnica para combinar diferentes modelos acústicos, al tomar su producto y normalizar el resultado [74]. La técnica puede llevarse a muchos niveles en los múltiples factores que pueden intervenir en el entrenamiento y síntesis.

- Uso de Redes Neuronales de Aprendizaje Profundo (DBN) y Máquinas Restringidas de Boltzmann (RBM): En [75] se propone el uso de estos modelos para sustituir la parametrización de los espectros, y en su lugar trabajar con los contornos modelados con ambas técnicas. Utiliza HMM como modelo acústico, pero en las emisiones de los estados sustituye las distribuciones gaussianas por RBM o DBN. Por su parte en [75] se propone la relación entre el texto de entrada y los modelos acústicos utilizando DBN, en lugar de los árboles de decisión utilizados hasta la fecha.

Las aplicaciones de la síntesis de voz previstas, como los sistemas personalizados de habla a habla en diferentes idiomas y la clonación o reconstrucción de voces, depende en gran medida de las mejoras que puedan conseguirse a través de la integración de los avances como los mencionados.

Capítulo 3

Desarrollo de la propuesta

La implementación de voces de síntesis estadística paramétrica es posible en la actualidad tomando como base el sistema HTS [25], desarrollado en forma conjunta por un grupo de investigadores, centrados principalmente en el Nagoya Institute of Technology, de Japón. HTS se presenta como una extensión de la aplicación HTK [76], utilizada ampliamente en investigación de reconocimiento de voz.

La propuesta de trabajo de este proyecto tiene como base de implementación la adaptación de HTS, es decir, el uso de sus definiciones de modelos matemáticos y algoritmos de entrenamiento, los cuales se deben adaptar a las características de la variante del español a tratar. Con HTS como núcleo del sistema, se deben utilizar una serie de herramientas que permiten la extracción de parámetros, su uso para el entrenamiento de los HMM, la generación de parámetros y finalmente la reconstrucción de la onda de habla.

Debido a la cantidad de datos disponibles en bases de datos se debe delimitar el alcance de las frases a sintetizar dentro de ciertos contextos, de manera que sea posible utilizar mecanismos de evaluación. Estos contextos se definirán como aplicaciones informáticas potenciales que permitan mostrar los resultados de distintas pruebas, con un vocabulario y tipo de frases limitado.

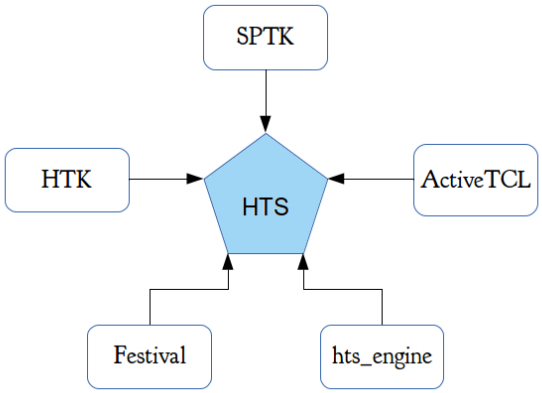


Figura 3.1: Esquema de los principales programas involucrados en la implementación

En el presente capítulo se describe la implementación de HTS y los programas necesarios para realizar la síntesis estadística paramétrica, así como las aplicaciones desarrolladas para su utilización, la propuesta de diseño de experimentos y la evaluación.

3.1. Adaptación del sistema HTS

En conjunto con las librerías de funciones de HTS, dedicadas a la definición y entrenamiento de los HMM, se utilizan un conjunto de programas para la extracción de características, el análisis de texto y la reconstrucción de señales de audio a partir de los parámetros de los HMM. La Figura 3.1 esquematiza los principales programas involucrados.

En la actualidad no se cuenta con una interfaz gráfica ni documentos de ayuda para utilizar el sistema completo en la creación de nuevas voces o la incorporación de nuevos idiomas. En el Apéndice D se realiza una descripción de los parámetros utilizados y la metodología general seguida para adaptar las implementaciones disponibles como guías en el sitio de HTS (para los idiomas inglés, japonés y portugués), a la creación de nuevas voces en español de

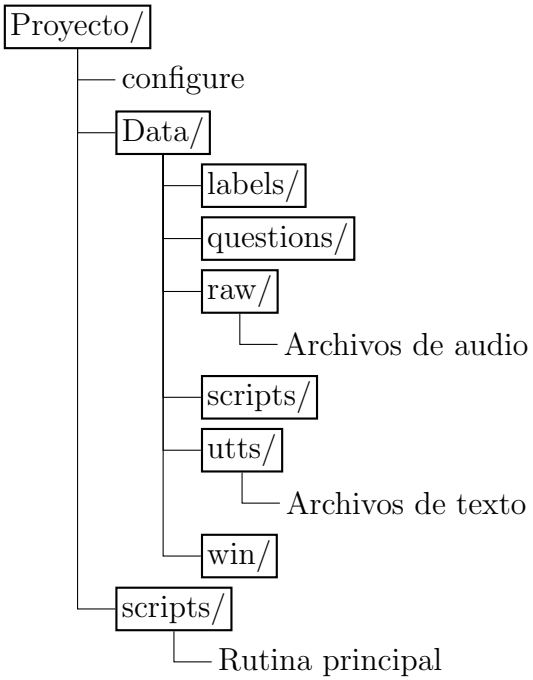


Figura 3.2: Esquema de archivos de entrada en un proyecto HTS

México. Los ajustes se realizan en diferentes archivos de configuración y rutinas de programas, inicialmente contenidos en un proyecto de implementación con la estructura de directorios y archivos resumida en la Figura 3.1. Estos corresponden a los datos iniciales de audio y su correspondiente transcripción en texto, los programas, y los archivos para la creación de árboles de decisión necesarios para el entrenamiento, en formato de HTK.

Los resultados finales de todos los procedimientos descritos en el Apéndice D son:

- Archivos de audio con nuevas frases sintetizadas, previamente definidas.
- Conjunto de archivos para utilizar en conjunto con *Festival* o bien *hts_engine* para la pronunciación de cualquier frase en estos programas.

También es de importancia la definición de las unidades de habla a tratar, y las características fonéticas del idioma particular. En este caso, la propuesta consiste en definir

como unidades fonéticas los fonemas, y establecer aquellos que mejor describan al español de México.

3.2. Descripción de los datos y tratamiento preliminar

La base de datos disponible para la creación de voces fue producida por la Universidad Autónoma Metropolitana, con dos hablantes, uno de cada género. Ambos hablantes son actores profesionales de doblaje. Las condiciones de grabación fueron controladas, pues se trabajó en un estudio de de grabación profesional. Consiste en una reproducción de la base de datos creada en *Center for Language and Speech Technologies and Applications (TALP)*, de la Universidad Politécnica de Cataluña (UPC) [77] para investigación de voz con emociones. Las frases que contiene esta base de datos consisten en frases y palabras aisladas, tal como se describen en la Tabla 3.1. En total son 11.3 minutos de grabaciones por hablante.

Inicialmente los datos se encontraban en formato PCM, codificado a 16 bit y 16kHz, sin compresión. A cada frase grabada le corresponde un archivo en formato texto con la transcripción ortográfica correspondiente a la frase pronunciada. Fue necesario utilizar programas gratuidos y de código abierto como *sox* y *Festival* para realizar la conversión a los formatos *raw* (audio sin encabezado) y *utt* (descripción fonética y silábica) de características requeridas.

Como parte de los requirimientos de ajuste de parámetros de HTS, es necesario establecer valores para la extracción de la frecuencia fundamental en cada hablante, pues se sabe que el máximo y mínimo de $f0$ emitido es dependiente del género y la emoción. Para la extracción del rango de frecuencias fundamentales se utilizó el programa Praat [24], ampliamente reconocido dentro de la investigación relacionada con procesamiento de voz, y con un flexible lenguaje de programación propio. En éste se desarrolló una rutina creada como parte de la investigación,

Tabla 3.1: Descripción de las frases en la base de datos

Identificador	Contenido
1-100	Frase afirmativa
101-134	Frase interrogativa
135-150	Párrafos
151-160	Dígitos
161-184	Palabras aisladas

en el lenguaje propio del programa, a partir del diagrama de flujo de la Figura 3.3. Se han tomado en cuenta solamente los sonidos correspondientes a vocales, debido a que se tiene la certeza de ser sonoros. Los resultados se muestran en la Tabla 3.2.

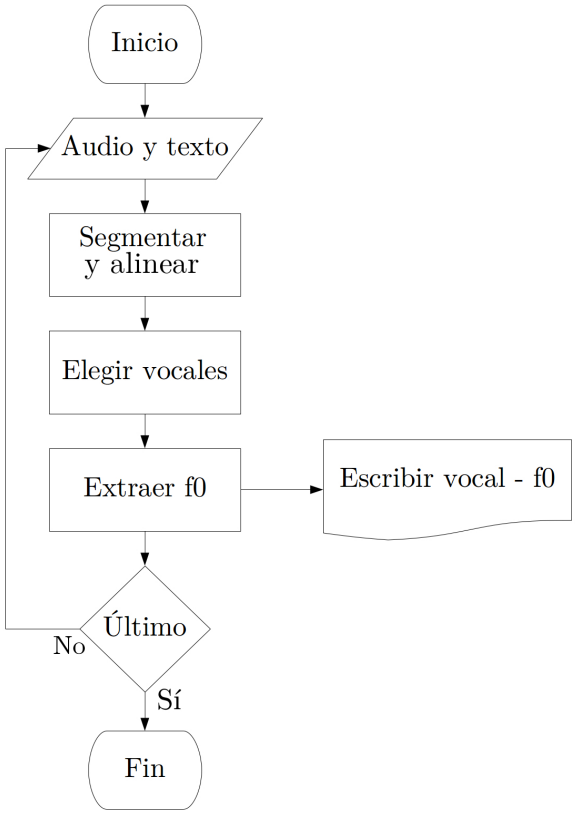


Figura 3.3: Diagrama de flujo para la extracción de la información de $f0$ en la rutina desarrollada para el programa Praat

Tabla 3.2: Rangos de frecuencia fundamental para la base de datos [78]

Hablante	Rango utilizado (Hz)
Masculino	55-255
Femenino	60-300

Además de establecer correctamente el rango de $f0$ es necesario ajustar los valores correspondientes al formato de audio (frecuencia de muestreo, longitud de ventana y desplazamiento), como se describe en el Apéndice A.

3.3. Diseño de experimentos

La experimentación que se presenta en esta sección ha tenido como propósito principal obtener nuevas voces a partir de las bases de datos en español mexicano, utilizando la síntesis estadística paramétrica. Además de su potencial utilización en diversas aplicaciones, una de las facetas más importantes de la parte experimental es verificar las ventajas reportadas en experiencias semejantes en otros idiomas, tales como el poco espacio requerido para el almacenamiento de las voces y la flexibilidad posible en los resultados.

Por otra parte, también es de importancia constatar las deficiencias que tiene la técnica de acuerdo con estas experiencias, como la menor naturalidad comparada con los mejores sistemas concatenativos, para identificar con claridad las áreas donde pueden realizarse futuras mejoras. Finalmente, dada la gran cantidad de parámetros y variables posibles en el proceso y las enormes combinaciones de éstos que pueden estudiarse, se han considerado como objeto de estudio aquellas pruebas que puedan responder algunas de las preguntas más importantes referentes a su imlementación, planteadas en [19].

En las siguientes subsecciones se detallan los experimentos planteados en cuatro grupos:

estudio de la influencia en la definición de parámetros de entrada, estudio del efecto del tamaño del conjunto de entrenamiento, estudio de la influencia de la calidad de las grabaciones, y estudio de la importancia de la información de contexto.

3.3.1. Influencia de parámetros de entrenamiento

En los procesos de entrenamiento y generación de parámetros con los HMM utilizados en la síntesis de voz con el sistema HTS, se requieren definir más de ochenta variables. Esto debido a que se deben establecer aspectos estructurales de los HMM, como la cantidad de estados, así como ajustes en los algoritmos propios del entrenamiento, como la cantidad de épocas (ciclos) a utilizar para actualizar las distribuciones de probabilidad.

Debido a la cantidad de opciones que esto presenta, tanto por el ajuste individual como por la relación entre ellos y los requierimientos computacionales para crear nuevas voces, es posible analizar solamente la influencia de una cantidad reducida de estos parámetros en los resultados de la síntesis. Para este efecto, se ha tomado un parámetro de entrenamiento, el rango de $f0$, para estudiar su relación con los resultados del proceso.

No se han encontrado referencias en la literatura con respecto a la influencia de la definición del rango de este parámetro en la calidad de voz resultante. Sin embargo, las primeras experimentaciones realizadas en el transcurso de este proyecto mostraron una dependencia significativa en la apreciación de esta calidad con respecto a su valor.

Para verificar esta dependencia, se definieron experimentos con rangos de $f0$ que pueden ser fácilmente identificados como rangos amplios y estrechos, para voces masculinas y femeninas. Un rango natural para una voz masculina puede estar entre el intervalo de $[60, 200]$ Hz, mientras que la femenina se considera usualmente en $[90, 250]$ Hz. Los rangos definidos como amplios y estrechos para ambos géneros son:

- Rango amplio de voz masculina: De 50 a 800 Hz.
- Rango amplio de voz femenina: De 50 a 800 Hz.
- Rango estrecho de voz masculina: De 60 a 90 Hz.
- Rango estrecho de voz femenina: De 80 a 110 Hz.

Para efectos de la experimentación, se han dejado todos los demás parámetros por omisión que contienen las rutinas del sistema HTS establecidos para el idioma inglés. El análisis de los resultados contempla la comparación de las voces resultantes con estos rangos, para las voces masculina y femenina, así como la obtenida con un intervalo adecuado para cada una.

Para obtener el rango preciso de $f0$, es necesario utilizar una herramienta de análisis, como el programa Praat, con el cual es posible realizar una segmentación de los fonemas de una base de datos a partir del audio y la transcripción ortográfica. Después de la segmentación, se pueden analizar los fonemas que son sonoros, es decir, en los cuales puede encontrarse una frecuencia fundamental, y de esta manera establecer estadísticas de este parámetro, como se estableció en la Sección 3.2.

Es posible entonces plantear las pruebas indicadas en la Tabla 3.3 para este apartado, a partir de las bases de datos con hablantes masculino y femenino.

Tabla 3.3: Experimentos sobre influencia de parámetros de entrenamiento

No.	Género de voz	Condición
1	Masculino	Rango amplio
2	Femenino	Rango amplio
3	Masculino	Rango estrecho
4	Femenino	Rango estrecho
5	Masculino	Rango adecuado
6	Femenino	Rango adecuado

3.3.2. Influencia del tamaño del conjunto de entrenamiento

En las referencias que incluyen información sobre el tamaño de la base de datos, se encuentran experiencias que indican uso de varias horas de grabaciones disponibles para el entrenamiento de los modelos de HMM. Por ejemplo, para sintetizadores a prueba sin contextos específicos referencias indican desde cerca de hora y media [50] [79], hasta más de diez horas de grabación [57] [66]. Para aplicaciones dentro de contextos específicos, se encuentran con cerca de dos horas de grabaciones [51]. Pocas de éstas, entre las que se encuentran la realizada para el idioma Checo [61], han experimentado con la reducción de bases de datos grandes para evaluar resultados. En la Figura 3.4 se ilustra la comparación de datos disponibles de varias experiencias recientes con los del presente proyecto.

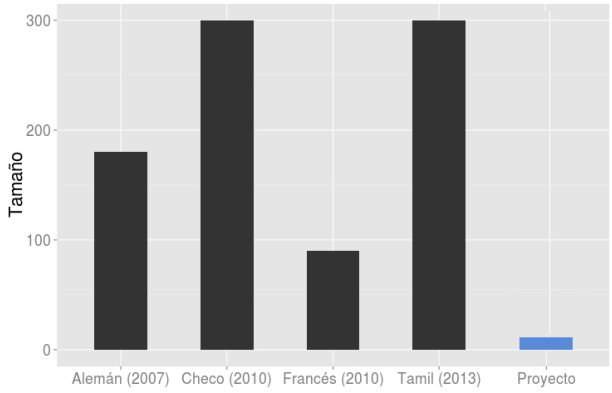


Figura 3.4: Comparación de cantidad de datos para el desarrollo de proyectos de síntesis estadística paramétrica

En el caso de las bases de datos disponibles para este proyecto, se cuenta solamente con once minutos de grabación, lo cual categoriza el proyecto como síntesis de voz a partir de pocos datos de entrenamiento. Por esta razón, para la evaluación de la influencia del tamaño del conjunto de entrenamiento, se plantean escenarios donde este tiempo de datos es reducido

y aumentado, utilizando los siguientes procedimientos:

- Para reducir el conjunto de datos: Realizar un estudio del contenido fonético de las frases que se desean sintetizar, y utilizar solamente aquellas frases de la base de datos que tienen contenido fonética que coincida con éstas, excluyendo aquellas que en principio no presentan información útil para sintetizar las frases requeridas.

- Para aumentar el conjunto: Duplicar los datos, y adicionalmente realizar cambios de velocidad sobre las grabaciones, de manera que se pueda multiplicar la cantidad de datos al presentar el conjunto de datos original con ligeras variaciones de velocidad, como nuevos datos. Las variaciones puedan afectar ligeramente las características espectrales con el mismo contenido fonético del texto, lo cual puede reforzar el entrenamiento de los HMM.

Para estudiar este aspecto, se plantean los experimentos que se muestra en la Tabla 3.4

Tabla 3.4: Experimentos sobre influencia de tamaño de conjunto de entrenamiento

No.	Género de voz	Condición
1	Masculino	Base de datos reducida
2	Femenino	Base de datos aumentada
3	Masculino	Base de datos completa
4	Femenino	Base de datos completa
5	Masculino	Base de datos duplicada
6	Femenino	Base de datos duplicada
7	Masculino	Base de datos aumentada
8	Femenino	Base de datos aumentada

Los resultados de este apartado deben compararse con los obtenidos a partir de la utilización de la base de datos completa, para ambos géneros de voces. Con las modificaciones a realizar en los datos se pretende explorar la posibilidad de utilizar este tipo de manipulaciones para bases de datos de tamaño pequeño.

3.3.3. Influencia de la calidad de grabaciones del conjunto de entrenamiento

Dado que se cuenta con las grabaciones de voz de hombre y mujer realizadas en ambiente controlado de ruido, alta calidad de equipo y hablantes profesionales, se pueden plantear la experimentación con nuevas fuentes de datos, para estudiar la influencia de la calidad del audio en los resultados. Los aspectos por analizar en la influencia de las características de las grabaciones sobre el resultado son:

- Grabaciones con menor calidad del audio en cuanto a su tasa de bits por segundo y su nivel de compresión.
- Grabaciones realizadas en ambiente con menor control de ruido, y variaciones inducidas por hablantes no profesionales.

Para el primer punto, se pueden aprovechar datos provenientes de audio libros, cuya utilización como fuente de datos para síntesis ha estado presente en investigaciones recientes, por ejemplo en [80]. Esto debido principalmente a la disponibilidad de grabaciones suficientes de audio y su transcripción textual precisa. Los datos provenientes de una fuente de este tipo, usualmente disponibles en formatos con compresión de audio, como MP3, se pueden utilizar entonces para contrastar el resultado con los obtenidos en las bases de datos disponibles de hombre y de mujer.

Por otra parte, para la obtención de voces con hablantes no profesionales, es necesario contar con voluntarios que puedan reproducir el contenido de frases de las bases de datos disponibles. Dado el tiempo necesario para crear datos adecuados, se hace factible plantear la creación de una base con distintos hablantes, de manera que también pueda verificarse la posibilidad de construir voces promedio a partir de la síntesis estadística paramétrica.

Para analizar la influencia de la calidad de grabaciones, los experimentos que se planean se presentan en la Tabla 3.5.

Tabla 3.5: Experimentos sobre la influencia de la calidad de las grabaciones

No.	Género de voz	Condición
1	Masculino	Grabación con menor calidad de audio
2	Masculino	Grabación en ambiente sin control de ruido
3	Femenino	Grabación en ambiente sin control de ruido

Las fuentes de datos para estos experimentos pueden crearse a partir de las siguientes estrategias:

1. Utilizar un audio libro con calidad de audio que utilice un formato con compresión (menor calidad con respecto a la base de datos).

2. Realizar grabaciones con voluntarios no profesionales para reproducir la base de datos disponible de voz de hombre y de mujer.

Los resultados deben compararse con los producidos con la base de datos.

3.3.4. Influencia de la información de contexto

La información de contexto determina la cantidad de HMM que se utilizan para representar las unidades fonéticas presentes en la base de datos. Esta información establece la posibilidad de discriminar entre fonemas que tienen el mismo grafema, pero que en su contexto (posición en la palabra, en la frase, en la sílaba y fonemas vecinos), son distinguibles.

Con esto se pretende incluir características prosódicas al habla sintetizada, al hacer diferencia entre el tono y energía de los fonemas a lo largo de frases sintetizadas. Los contextos se definen a partir del establecimiento de características, en un formato particular llamado

de etiquedas, las cuales pueden contener hasta 57 caracterísicas de contexto por fonema. A cada etiqueta (con descripción de un fonema particula con su contexto) le corresponde un HMM distinto en el modelo.

Para considerar distintas cantidades de información de contexto, se deben redefinir las preguntas de árboles de decisión y las reglas de conversión del texto a formato etiqueta. Aunque en el proceso de entrenamiento se realiza un agrupamiento de HMM para calcular de forma compartida sus parámetros, una menor cantidad de información de contexto puede llevar a reducir la calidad de la prosodia resultante, ya que puede, por ejemplo, no haber distinción entre fonemas ubicados al principio o al final de una frase. La conveniencia de reducir la cantidad de contexto puede estar relacionada con la cantidad de datos disponibles de entrenamiento.

Para determinar la influencia de esta información a partir de la base de datos disponible, se plantean como experimentos los indicados en la Tabla 3.6

Tabla 3.6: Experimentos sobre la influencia de la información de contexto

No.	Género de voz	Condición
1	Masculino	Información completa de contexto, fonética y prosódica
2	Masculino	Información reducida de contexto, solamente fonética
3	Femenino	Información completa de contexto, fonética y prosódica
4	Femenino	Información reducida de contexto, solamente fonética

3.4. Desarrollo de aplicaciones computacionales

Entre los primeros intereses al realizar síntesis de voz se encuentran aplicaciones para automatizar la entrega de información a usuarios de servicios, tales como la hora, saldos de cuentas u otros que pudieran consultarse y enviarse vía telefónica. En la actualidad las posibilidades de estas aplicaciones se han extendido dada la diversidad de dispositivos e información requerida por los usuarios.

Por otra parte, es conveniente limitar el alcance de frases que puedan emitirse, en cuando a su longitud y vocabulario, dadas las características reducidas en la cantidad de datos disponibles. Por ejemplo, para establecer un sistema de evaluación, donde se puedan medir sus resultados y mostrar la flexibilidad característica de esta técnica.

Se definirán entonces dos aplicaciones por desarrollar, que serán base para la evaluación de los resultados:

1. Aplicación para consultar y pronunciar la hora del sistema, a partir de una interfaz gráfica. La emisión de frases de hora se caracteriza por un uso de cantidad de palabras muy semejante en todos los casos. Las frases son de la forma

   Son las hh y mm.

   Con la variación en horas exactas, donde la frase puede ser de la forma

   Son las hh en punto.

   Debe distinguirse la hora “1”, para usar el artículo *la* en singular. Por ejemplo: “Son **la** una y treinta”.

2. Aplicación para simular la consulta, y la pronunciación de la predicción del tiempo atmosférico del día, a partir de una interfaz gráfica. El vocabulario de este tipo de aplicaciones suele ser más variado, así como la longitud de las frases. La predicción incluye palabras como "soleado", "probabilidad", "lluvias aisladas", "nubosidad variable", entre otros. Es usual incluir la predicción de temperatura mínima y máxima para el día.

Estas aplicaciones deben incorporar las voces resultantes de los experimentos definidos en la sección anterior, y las frases propias de cada una deben utilizarse para evaluación.

3.5. Evaluación del habla sintetizada

La evaluación del habla sintetizada es un aspecto fundamental para el análisis comparativo de los resultados obtenidos por medio de la experimentación propia de las diversas técnicas de síntesis. Las ventajas de su aplicación son importantes tanto para desarrolladores de sistemas como para usuarios finales [81]. Existen dos enfoques fundamentales de evaluación, el enfoque objetivo y el enfoque subjetivo, dependiendo del tipo de criterios seguidos. En las siguientes subsecciones se desarrollan los principales elementos de ambos.

3.5.1. Evaluación objetiva

La evaluación objetiva consiste en cuantificar los resultados de voces sintetizadas basándose en parámetros de la voz que sean medibles directamente sobre la onda sintetizada, sin mediar apreciaciones de escuchas humanos. Se han utilizado, por ejemplo, comparaciones de los contornos de f_0 y espectrograma entre habla natural y sintetizada [82], mediante una observación directa y establecimiento de semejanzas. En [83], se propone una comparación

de parámetros x, tomando en cuenta los valores x_N de la voz natural y los x_p predecidos en el habla sintetizada. De esta manera es posible determinar medidas como el error medio cuadrático (RMSE) y coeficientes de correlación. Esto en valores como:

- Duración, a nivel de fonema.
- Coeficientes MFCC: Considerando la alta dimensionalidad que se utiliza como conjunto de parámetros MGC, cada coeficiente se normaliza mediante

$$z_i = \frac{x_i - \mu_i}{\sigma_i}. \tag{3.1}$$

  Es posible entonces calcular el RMSE por coeficiente y uno promedio de todos.
- Valores de $f0$: De forma semejante a los coeficientes de espectro, se puede comparar los valores de $f0$ extraídos a lo largo de frases grabadas y sintetizadas, o bien utilizando comparaciones entre valores de $f0$ en fonemas sonoros.

La adopción de los métodos anteriores depende de factores como la disponibilidad de grabaciones de referencia con las mismas frases sintetizadas, y la equivalente del factor tiempo en las frases que se desea comparar. De no contarse con las condiciones para implementarlos, se puede recurrir a técnicas como DTW para establecer grados de semejanza entre conjuntos de factores que no tienen la misma longitud.

3.5.2. Evaluación utilizando técnicas de aprendizaje de segundo idioma

Existen propuestas de evaluación de habla sintetizada basadas en técnicas de reconocimiento de voz utilizadas en aprendizaje de segundo idioma SLL (por las siglas en inglés de *Second*

Language Learning). Por ejemplo, los Asistentes de aprendizaje de lenguaje computarizados CALL (por las siglas en inglés de *Computer-Assisted Language Learning*), utilizan, como un medio para evaluar la pronunciación, la calidad de los fonemas del usuario comparado con un modelo acústico previamente definido. Esto incluye retroalimentación visual, tal como formas de onda y curvas de tono para indicar diferencias prosódicas entre el usuario y el modelo [84].

La utilización de medidas cuantitativas en la evaluación de fluidez en segunda lengua, se resumen en [85] y [86], e incluyen como parámetros:

- Razón de habla *ros*: Cantidad de fonemas dividido por la duración total del habla, incluyendo pausas internas.
- Razón de fonemas *ptr*: Duración total del habla sin pausas divido por la duración total incluyendo pausas. Se da como un porcentaje.
- Razón de articulación *art*: Cantidad de fonemas dividido por la duración total del habla sin pausas.
- Pausas: Cantidad de pausas internas de las frases de menos de 0.2 s.
- Duración total de pausas *tdp*: Duración total de todas las pausas internas de menos de 0.2 s.
- Duración media de pausas *mlp*: Duración media de pausas internas de más de 0.2 ms.
- Duración media de frases *mlr*: Número promedio de fonemas que ocurren entre dos pausas de más de 0.2 ms.
- Cantidad de pausas sonoras *fp*: Número de pausas con algún sonido, como “mm”, “eh”.
- Cantidad de disfluencias *dy*: Número de repeticiones, reinicios, autocorrecciones.

Otras propuestas, como [87], consideran como medidas para retroalimentar la pronunciación dos parámetros: El primerio relacionado con la medición de la coincidencia espectral entre el hablante y los modelos estadísticos entrenados con hablantes nativos. El segundo es la razón de fonemas por segundo.

En [88] se trata el tema de detección de pronunciaciones inadecuadas, mediante reglas relacionadas con la conversión grafema a fonema, y fonema a pronunciación. Evaluaciones basadas en probabilidad es analizan en [89], donde se utilizan HMM como clasificador de fonemas, y se extraen as probabilidades dadas en el proceso de clasificación como una medida de calidad de pronunciación. Adicionalmente usa medidas de duración de elementos fonéticos.

En [90] se utiliza DTW para determinar similitud entre una pronunciación y una referencia, a partir de un esquema que utiliza $f0$ y coeficientes MFCC.

3.5.3. Evaluación subjetiva

Dado que el fin último de desarrollar sistemas que apliquen síntesis de voz es su utilización en aplicaciones reales, es de importancia evaluarlos con criterios de escuchas humanas [91], lo cual se conoce como evaluación subjetiva. En este rubro tiene destacada relevancia las categorías y escalas introducidas en el *Blizzard Challenge*, por ejemplo, las utilizadas para evaluar la similitud con el hablante original, la naturalidad y la tasa de palabras erróneas.

Las escalas que se utilizan se evalúan sobre frases elegidas al azar [82], y corresponden a silimitud, naturalidad e inteligibilidad [83]. Estas escalas pueden adaptarse para establecer los siguientes aspectos:

- Similitud (SIM):

  5 Exactamente como la persona

4 Muy parecido a la persona

3 Diferente pero se reconoce como la misma persona

2 Semejante pero se escucha como una persona diferente

1 Como una persona totalmente diferente

- Naturalidad (NAT):

  5 Completamente natural

  4 Muy natural

  3 Poco natural pero aceptable

  2 Muy poco natural

  1 Nada natural

- Inteligibilidad (INT):

  5 Completamente inteligible

  4 Muy inteligible

  3 Poco inteligible pero aceptable

  2 Muy poco inteligible

  1 No inteligible

- Tasa de palabras erróneas WER (por las siglas en inglés de *Word Error Rate*)

La adaptación de estas opciones de evaluación depende de la complejidad de su implementación, y la disponibilidad de voluntarios que puedan realizar las escuchas para la

evaluación subjetiva. De contar con muchos resultados que se desean evaluar, conviene realizar una priorización de las pruebas para realizar evaluación sobre ellas, y posteriormente profundizar en sus diferencias para dirigir potenciales mejoras.

A pesar de las numerosas experiencias con síntesis estadística paramétrica en varios idiomas, no se encuentran referencias de evaluaciones utilizando estos sistemas en variantes de español de América Latina para establecer comparaciones, por lo que se considera el planteamiento de la evaluación objetiva como exploratoria. La producción de voces para las distintas bases de datos y el sistema de evaluación propuesto se resumen en la Figura 3.5.

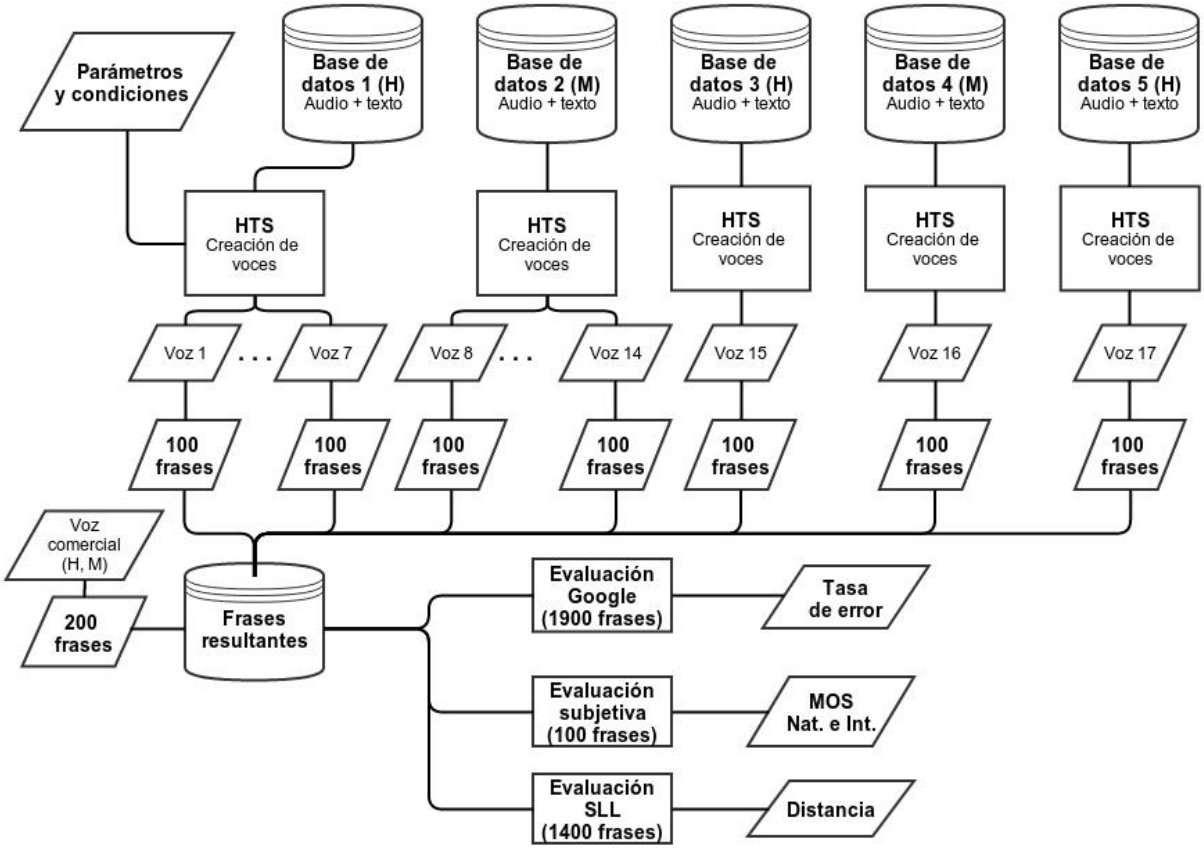


Figura 3.5: Resumen de producción de voces y sistemas de evaluación por cada aplicación

Capítulo 4

Resultados

En este capítulo se presentan los productos de las distintas pruebas realizadas en la creación de voces con síntesis estadística paramétrica. Estas pruebas se han realizado con la meta principal de comprobar la influencia de diversos parámetros y condiciones de entrenamiento en las voces resultantes, dentro de las aplicaciones desarrolladas, para determinar estrategias que puedan mejorar su calidad.

Sobre los resultados se realizan varios tipos de evaluaciones, basadas en las referencias descritas en la Sección 3.5, y se propone el análisis de parámetros acústicos previamente no considerados como criterios de evaluación (tono, *jitter* y *shimmer*). Al final del capítulo se realiza un análisis conjunto de los resultados por cada aplicación.

4.1. Definición de los HMM a partir de aspectos lingüísticos

En la creación de voces con síntesis estadística paramétrica, se utiliza un conjunto de HMM, uno por cada fonema que toma en cuenta su contexto. Los fonemas y su contexto son dependientes del idioma, por lo que es de importancia documentar los aspectos lingüísticos

que son referencia para la definión de los HMM, además de caracterizar la variante del idioma que se realiza.

4.1.1. Conjunto de fonemas

En el análisis hecho de la variante del español de los hablantes de la base de datos (de México), se ha determinado el conjunto de fonemas indicado en la Tabla 4.1. La documentación de la base de datos que se ha tomado como referencia fue realizada con la codificación SAMPA (por las siglas inglés de *Speech Assessment Methodology Phonetic Alphabet*), definida para el español de España [92], por esta razón se ha tomado como punto de partida para definir los fonemas a utilizar. Además de SAMPA, existe la codificación Mexbet [93]

Se ha estudiado, para el español de México, la importante variedad que existe en la pronunciación del fonema "x", para el cual existen cuatro sonidos diferentes [94], pero no se cuenta con reglas que permitan su transcripción fonética automática. Dado que el contenido de palabras de la base de datos no contempla estos casos exclusivos de México, no se han considerado dentro de los sḿbolos definidos para los fonemas en HTS.

Se observa que en las codificaciones SAMPA y Mexbet no se hace diferencia entre vocales acentuadas, pero en la definición de HMM es conveniente diferenciarlas para efectos de incluir elementos que den mayor semejanza con el habla natural. Además de estas diferencias, el contexto del fonema (ubicación en la sílaba, palabra o frase) marca una pauta importante en la distinción de los mismos para modelar el habla y obtener mejores resultados. En la siguiente subsección se muestra la adaptación de estos elementos.

Tabla 4.1: Fonemas definidos para la variante de español de México. A partir de [92]

Símbolo HTS	Símbolo SAMPA	Símbolo Mexbet	Clasificación	Ejemplo
a	a	a	Vocal	c**a**sa
a1	a	a	Vocal	**á**mbar
b	B	b	Fricativa	ca**b**ra
ch	tS	tS	Africada	mu**ch**o
d	d	d	Plosiva	**d**onde
e	e	e	Vocal	p**e**ro
e1	e	e	Vocal	**é**ter
f	f	f	Fricativa	f**á**cil
g	g	g	Fricativa	lue**g**o
i	i	i	Vocal	p**i**co
i0	j	i	Semivocal	p**i**e
i1	i	i	Vocal	**í**ndice
k	k	k	Plosiva	**c**asa
l	l	l	Líquida	**l**ejos
ll	L	Z	Líquida	caba**ll**o
m	m	m	Nasal	**m**ismo
n	n	n	Nasal	**n**unca
ny	J	n∼	Nasal	a**ñ**o
o	o	o	Vocal	t**o**ro
o1	o	o	Vocal	canci**ó**n
p	p	p	Plosiva	**p**adre
r	r	r(	Líquida	pu**r**o
rr	rr	r	Líquida	to**rr**e
s	s	s	Fricativa	**s**ala
t	t	t	Plosiva	**t**omo
u	u	u	Vocal	d**u**ro
u0	u	u	Semivocal	de**u**da
u1	w	u	Vocal	igl**ú**
xx	x	x	Fricativa	mu**j**er

4.1.2. Elementos de contexto

Los elementos de contexto han sido adaptados a partir de los contextos definidos en HTS para el idioma inglés. Se identifican 11 niveles, agrupados de acuerdo con lo indicado en la

Tabla 4.2. Algunos de los elementos de contexto son factores binarios, mientras que otros se refieren a cantidades de elementos (fonemas, sílabas o palabras) y otros a identidades de fonemas (cuál fonema está en determinada posición con respecto al actual).

Se puede identificar a los elementos de contexto $p1$ a $p5$ como elementos fonéticos, mientras que de $a1$ a $j3$ como elementos prosódicos. De acuerdo con los contextos establecidos, cada fonema se representa de la forma:

$$
\begin{aligned}
p1\hat{}p2 - p3 + p4 = p5@p6 - p7 \\
/A : a1 - a2 - a3 \\
/B : b1 - b2@b3 - b4 - b5 - b6\&b6 - b7\$b8 - b9; b10 \\
/C : c1 - c2 \\
/D : d1 \\
/E : e1 + e1 + e3 \\
/F : f1 \\
/G : g1 - g2 \\
/H : h1 = h2\hat{}h3 = h4 \\
/I : i1 - i2 \\
/J : j1 - j2 + j3.
\end{aligned}
$$

Esta representación permite identificar al fonema actual ($p3$) y diferenciarlo de otros por sus características en relación con su posición en la sílabra, palabra o frase. De esta manera

Tabla 4.2: Contextos definidos para los fonemas. A partir de [25]

Símbolo	Contexto representado
$p1$	fonema en la posición transanterior
$p2$	fonema en la posición anterior
$p3$	fonema en la posición actual
$p4$	fonema en la posición siguiente
$p5$	fonema en la posición dos adelante
$a1$	sílaba anterior acentuada (1) o no (0)
$a2$	sílaba anterior tildada (1) o no (0)
$a3$	número de fonemas en la sílaba anterior
$b1$	sílaba actual acentuada (1) o no (0)
$b2$	número de fonemas en la sílaba actual
$b3$	posición de la sílaba actual en la palabra (hacia adelante)
$b4$	posición de la sílaba actual en la palabra (hacia atrás)
$b5$	posición de la sílaba actual en la frase (hacia delante)
$b6$	posición de la sílaba actual en la frase (hacia atrás)
$b7$	número de sílabas acentadas anteriores a la actual en la frase
$b8$	número de sílabas acentadas posteriores a la actual en la frase
$b9$	número de sílabas entre la última acentuada y la actual
$b10$	número de sílabas entre la siguiente acentuada y la actual
$b11$	vocal de la sílaba actual
$c1$	sílaba siguiente acentuada (1) o no (0)
$c2$	número de fonemas en la sílaba siguiente
$d1$	número de sílabas en la palabra anterior
$e1$	número de sílabas en la palabra actual
$e2$	posición de la palabra actual en la frase (hacia adelante)
$e3$	posición de la palabra actual en la frase (hacia atrás)
$f1$	número de sílabas en la palabra siguiente
$g1$	número de sílabas en la frase anterior
$g2$	número de palabras en la frase anterior
$h1$	número de sílabas en la frase actual
$h2$	número de palabras en la frase actual
$h3$	posición de la frase actual en el párrafo (hacia adelante)
$h4$	posición de la frase actual en el párrafo (hacia atrás)
$i1$	número de sílabas en la frase siguiente
$i2$	número de palabras en la frase siguiente
$j1$	número de sílabas en el párrafo
$j2$	número de palabras en el párrafo
$j3$	número de frases en el párrafo

se espera que puedan sintetizar nuevas frases con los contornos de $f0$ e intensidad que lo haría el hablante de la base de datos. Como ejemplo, en la base de datos se tiene la frase "pronto", la cual es convertida en:

```
x^x-#+p=r@1_0/A:0_0_0/B:0-0@1-1-2-0\&1-0$1-1!0-0;0-0|0/C:1+4/D:0/E:0@1\&2/
F:2/G:0-0/H:0=0^1=2/I:0=0/J:2+1-1.
x^#-p+r=o1@1_4/A:0_0_0/B:1-1-4@1-2\&1-2$1-1!0-0;0-0|o1/C:0+2/D:0/E:2@1\&1/
F:0/G:0-0/H:2=1^1=1/I:0=0/J:2+1-1.
#^p-r+o1=n@2_3/A:0_0_0/B:1-1-4@1-2\&1-2$1-1!0-0;0-0|o1/C:0+0/D:0/E:2@1\&0/
F:0/G:0-0/H:2=1^1=1/I:0=0/J:2+1-1.
p^r-o1+n=t@3_2/A:0_0_0/B:1-1-4@1-2\&1-2$1-1!0-0;0-0|o1/C:0+0/D:0/E:2@1\&1/
F:0/G:0-0/H:2=1^1=1/I:0=0/J:2+1-1.
r^o1-n+t=o@4_1/A:0_0_0/B:1-1-4@1-2\&1-2$1-1!0-0;0-0|o1/C:0+0/D:0/E:2@1\&1/
F:0/G:0-0/H:2=1^1=1/I:0=0/J:2+1-1.
o1^n-t+o=#@1_2/A:1_1_4/B:0-0-2@2-1&2-1$1-1!1-0;1-0|o/C:0+0/D:0/E:2@1\&1/
F:0/G:0-0/H:2=1^1=1/I:0=0/J:2+1-1.
n^t-o+#=x@2_1/A:1_1_4/B:0-0-2@2-1\&2-1$1-1!1-0;1-0|o/C:0+0/D:0/E:2@1\&0/
F:0/G:0-0/H:2=1^1=1/I:0=0/J:2+1-1.
t^o-#+x=x@1_0/A:0_0_0/B:0-0-0@1-0\&1-1$1-1!0-0;0-0|0/C:0+0/D:0/E:0@1\&1/
F:0/G:0-0/H:0=0^1=1/I:0=0/J:2+1-1.
```

La primera línea de esta codificación se interpreta de la siguiente manera:

- $x\hat{}x$: Al fonema actual (#) no le precede ningún fonema (símbolo x), ni tiene ningún fonema en posición trasanterior.
- $-\#+$: El fonema actual es un silencio.
- $p = r$: El fonema que le sigue al actual es una p, y luego sigue una r.

- $/A$: 0_0_0: La sílaba anterior no estaba tildada, ni acentuada, y no contenía ningún fonema.
- $/B : 0-0@1-1\&1-1\#1-1\$1-1!0-0;0-0|0$: La sílaba actual no está acentuada, la sílaba actual no tiene fonemas (por ser silencio), la sílaba actual está en posición 1 en la palabra y en la frase, y en posición 2 hacia atrás. Hay una sílaba acentuada posterior al fonema, y ninguna posterior. No hay sílabas entre el fonema actual y la siguiente sílaba acentuada anterior o posterior.
- $/C : 1+4$: La sílaba siguiente esá acentuada y tiene cuatro fonemas.
- $/D$: 0: La palabra anterior no tiene sílabas.
- $/E$: 0@1&2: La palabra actual ni tiene sílabas (por ser silencio), y ocupa la posición 1 en la frase.
- $/F$: 2: La palabra siguiente tiene dos sílabas.
- $/G : 0-0$: La palabra anterior no tiene sílabas ni palabras.
- $/H : 0=0\hat{}1=1$: La frase actual no tiene sílabas ni palabras, la frase actual tiene posición 1 en la frase (hacia adelante) y 2 (hacia atrás).
- $I : 0=0$: La frase siguiente no tiene sílabas ni palabras.
- $/J : 2+1-1$: El párrafo actual tiene dos sílabas, una frase palabra y una palabra.

Además de considerar un HMM por cada fonema en su contexto, se considera uno para pausa corta (generalmente entre palabras) y una para pausa larga (entre frases).

El total de HMM para representar todos los fonemas en su contexto asciende a 13704, para la base de datos de 184 frases de cada hablante. Por esta razón es necesario realizar un

proceso de agrupamiento que haga posible el entrenamiento más adecuadado de los HMM y la gran cantidad de parámetros que esto conlleva en cada uno.

4.1.3. Agrupamiento

El agrupamiento de HMM en el proceso se realiza con dos propósitos. El primero es reducir la cantidad de parámetros que ajustar (medias y varianzas de las distribuciones de probabilidad), pues el tomar en cuenta información de contexto hace que se eleve la cantidad de HMM del modelo y por ende la cantidad de parámetros a estimar. La segunda razón es para determinar el HMM que mejor pueda representar a los fonemas de las frases que se desean sintetizar de los que se encuentran entrenados en el modelo.

Para estos propósitos se definen árboles de decisión binarios, que utilizan las preguntas establecidas para delimitar la información de contexto, y el algoritmo de Longitud Mínima de Descripción MDL (por las siglas en inglés de *Minimum Description Lenght*) [95] para establecer su tamaño. Se utilizan diferentes árboles de decisión para el tono ($f0$), la duración y el espectro (MFCC), pues estos aspectos son independientes en el habla.

4.2. Aplicaciones desarrolladas

La aplicación de la hora (llamada Reloj) tiene como objetivo pronunciar la hora el sistema. Para esto, se desarrolló una interfaz gráfica en lenguaje Python, que incluye código para leer la hora del sistema (en sistema operativo Linux), y utiliza el programa *Festival* para generar la onda de habla, de acuerdo con el diagrama de flujo presentado en la Figura 4.1.

La Figura 4.2 muestra la interfaz gráfica implementada para la aplicación Reloj. Se observan los controles de velocidad de habla y de tono, los cuales se utilizan para ajustar estos

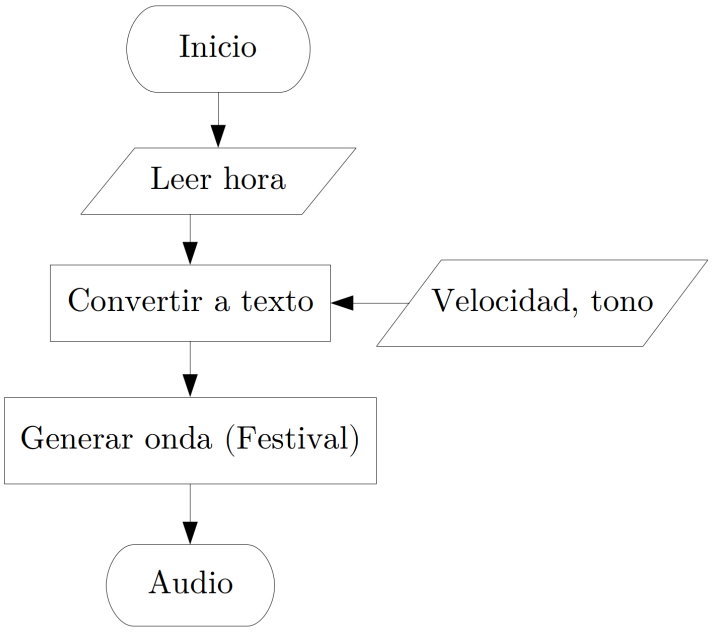


Figura 4.1: Diagrama de flujo para la aplicación de hora

parámetros en el habla sintetizada. Para esta aplicación se incluyeron voces de hombre y de mujer, así como la voz sintetizada a partir de las grabaciones de Carlos Fuentes. Esta última voz fue elegida por la disponibilidad de un audio libro leído por su autor, quien además es de nacionalidad mexicana y cuenta con características impecables de dicción y cantidad de datos.

Los controles de tono y velocidad son posibles utilizando parámetros del sistema *hts_engine* en la reconstrucción de la voz. En el siguiente código se muestran las lńeas principales que realizan la lectura de la hora del sistema y su procesamiento para establecer el texto de entrada al sintetizador:

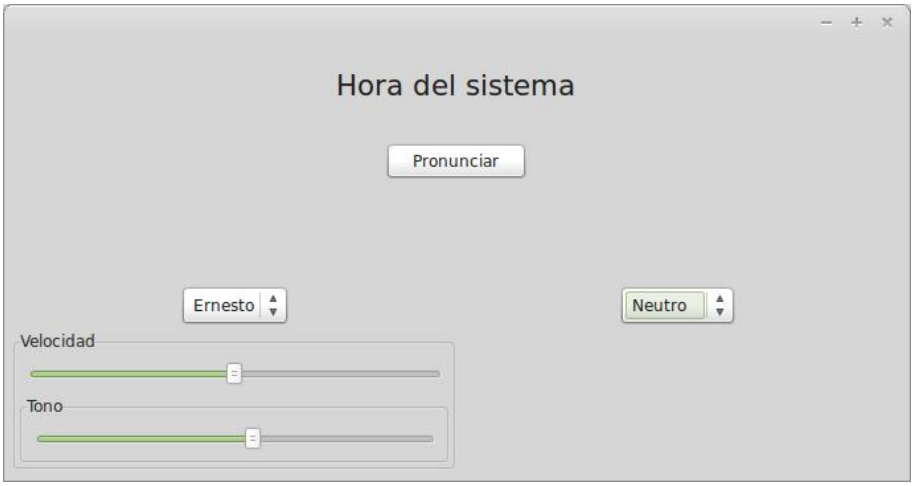


Figura 4.2: Interfaz gráfica de la aplicación Reloj

```
def m_button1OnButtonClick(self,event):
        linea1='(voice_cstr_upc_upm_spanish_hts)'
        from datetime import datetime
        tiempo=str(datetime.now())
        hora=tiempo[11:13]
        minutos=tiempo[14:16]
        linea2='(SayText "Son las '+hora+' y '+minutos+' ")'
```

Con el fin de comparar el resultado de síntesis con frases que incluyen mayor cantidad de palabras y vocabulario más diverso, se desarrolló la aplicación de simulación de predicción de tiempo atmosférico, que presupone la lectura de la predicción del tiempo en un día, incluyendo la temperatura máximo y mínima, y la pronuncia con las voces incluidas en la aplicación, tanto de hombre y de mujer producidas a partir de la base de datos creada con actores profesionales. La Figura 4.3 muestra la interfaz gráfica implementada para esta aplicación de predicción, llamada Clima.

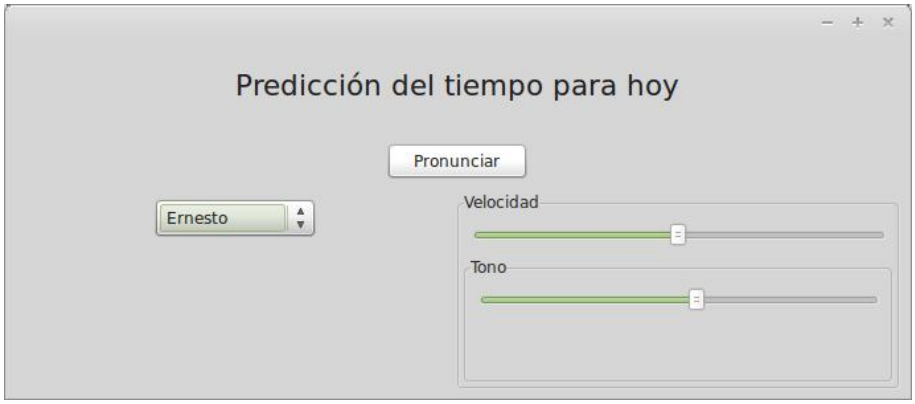


Figura 4.3: Interfaz gráfica de la aplicación Clima

Para determinar la factibilidad de los datos disponibles en la base de datos para producir las frases requeridas por ambas aplicaciones, se realizó un análisis de los fonemas y difonemas necesarios para pronunciar las frases. La Figura 4.4 muestra de forma gráfica la cantidad de fonemas de la base de datos, y los necesarios para producir todas las frases de ambas aplicaciones. Se comprueba la existencia en la base de datos de todos los fonemas necesarios para el entrenamiento de los modelos, así como los difonemas.

En cuanto a estos últimos, se ha comprobado la cobertura de la base de datos de la totalidad de difonemas requeridos en la aplicación Clima y Reloj. Para la pronunciación de frases de la hora se utilizan 70 difonemas, mientras que para las frases de tiempo atmosférico 134. La base de datos cuenta con un total de 344 difonemas.

A pesar de no ser parte de las voces evaluadas, como una manera de mostrar la flexibilidad que es posible obtener en la síntesis estadística paramétrica se incluyeron voces intermedias entre la voz masculina y femenenina. Estas voces se obtuvieron al entrenar los HMM a partir de la mezcla de ambas bases de datos con distintos grados (25 % voz de mujer - 75 % voz de hombre, 50 % voz de mujer - 50 % voz de hombre, 75 % voz de mujer - 25 % voz de hombre). Los resultados de estas mezclas de voces se han dejado para fines demostrativos, pero no se

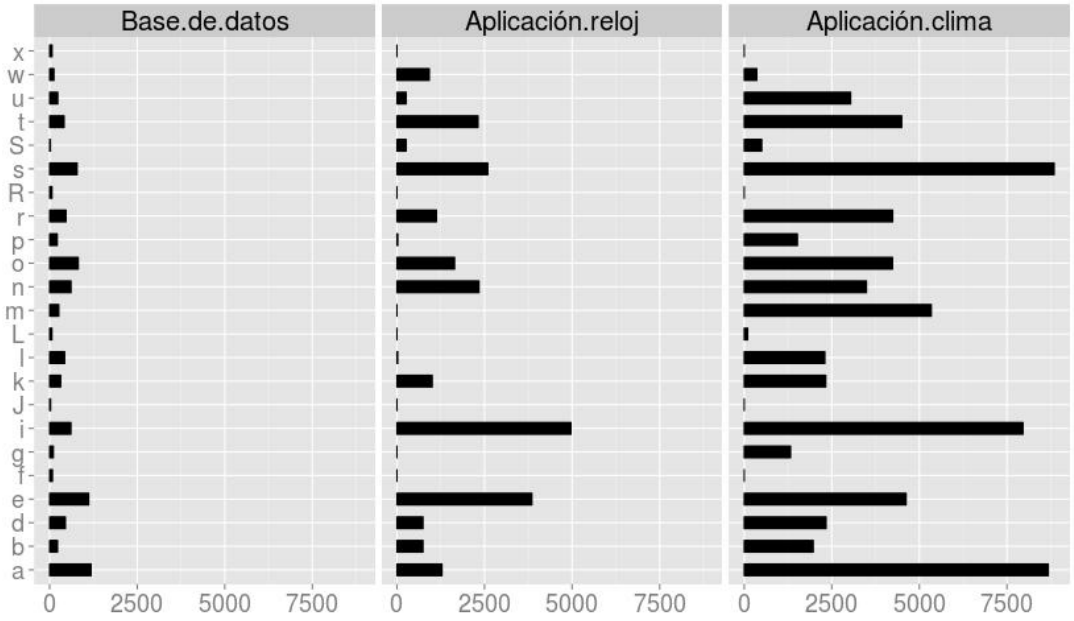


Figura 4.4: Cantidad de fonemas en la base de datos y requeridas por aplicación

incluyeron en las evaluaciones extensivas realizadas al conjunto de experimentos.

En las siguientes secciones se detallan las pruebas realizadas que fueron sujetas de evaluación, divididas de acuerdo con el tipo de entrenamiento realizado, tanto para voz de hombre como de mujer, en ambas aplicaciones.

4.3. Métodos de evaluación adoptados

Dadas las múltiples posibilidades existentes para evalur calidad de voces sintéticas, y a la falta de consenso de aquellas medidas que deben aplicarse en la creación de nuevas voces, se han adoptado las siguientes, divididas en objetivas y subjetivas:

1. Evaluación objetiva: A partir del desarrollo de rutinas en lenguaje Praat, y de proce-

samiento de la información que da este programa, se adoptaron las siguientes medidas provenientes de las técnicas SLL:

- Razón de habla *ros*: Cantidad de fonemas dividido por la duración total del habla, incluyendo pausas internas.
- Razón de fonemas *ptr*: Duración total del habla sin pausas divido por la duración total incluyendo pausas. Se da como un porcentaje.
- Razón de articulación *art*: Cantidad de fonemas dividido por la duración total del habla sin pausas.
- Duración media de pausas *mlp*: Duración media de pausas internas de más de 0.2 ms.

Para establecer un criterio de semejanza con la voz original, se utilizó distancia euclidea entre las cuatro mediciones de cada prueba y las respectivas realizadas en las bases de datos. No es posible utilizar comparaciones de otro tipo (distancias espectrales o de $f0$), pues la base de datos no contiene frases de ninguna de las dos aplicaciones.

Por otra parte, el uso de reconocimiento de palabras para evaluar voces sintetizadas ha sido propuesto en [96]. Para su incorporación en este trabajo se ha desarrollado un programa que toma 100 frases producidas en cada experimento, realiza una consulta por medio de Internet al reconocedor de voz de Google®[1], el cual devuelve la transcripción textual del audio.

Este es evaluado por el programa utilizando expresiones regulares, con las cuales se puede contar la cantidad de palabras transcritas adecuadamente y erróneamente. De

[1] `https://www.google.com/intl/es/chrome/demos/speech.html`

esta manera, se establece una tasa de error, semejante a lo establecido para clasificadores. En la Figura 4.5 se esquematiza el procedimiento programado.

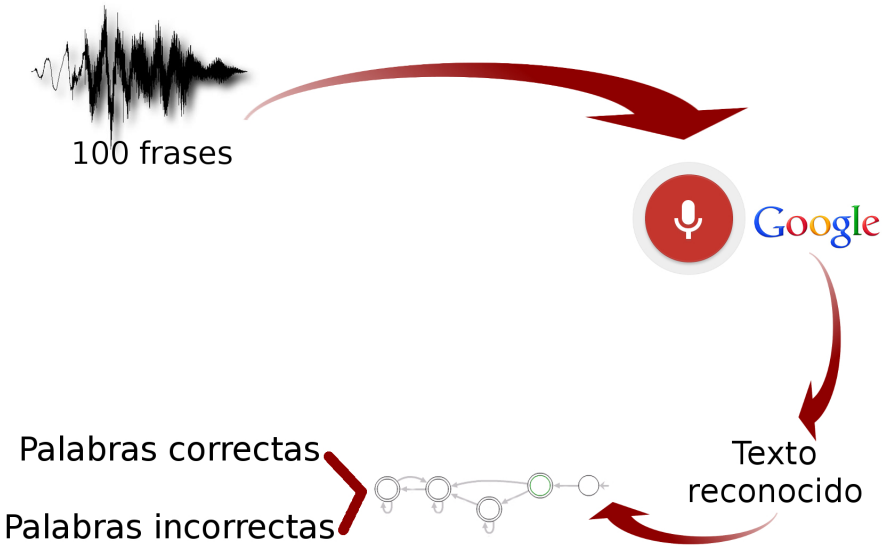


Figura 4.5: Esquema del procedimiento para establecer tasa de palabras correctas con reconocedor

Adicionalmente, se propone en este trabajo la utilización de medidas de similitud de tres parámetros acústicos de las frases sintetizadas: tono, *jitter*, *shimmer*. El tono se determinará en vocales, y se analizará de forma compartiva con el tono del hablante original. El *jitter* (medida de fluctuaciones entre periodos de la frecuencia fundamental), se define de forma local como

$$J_t = \frac{|T_i - T_{i+1}|}{\frac{1}{N}\sum_{i=1}^{N} T_i}, \tag{4.1}$$

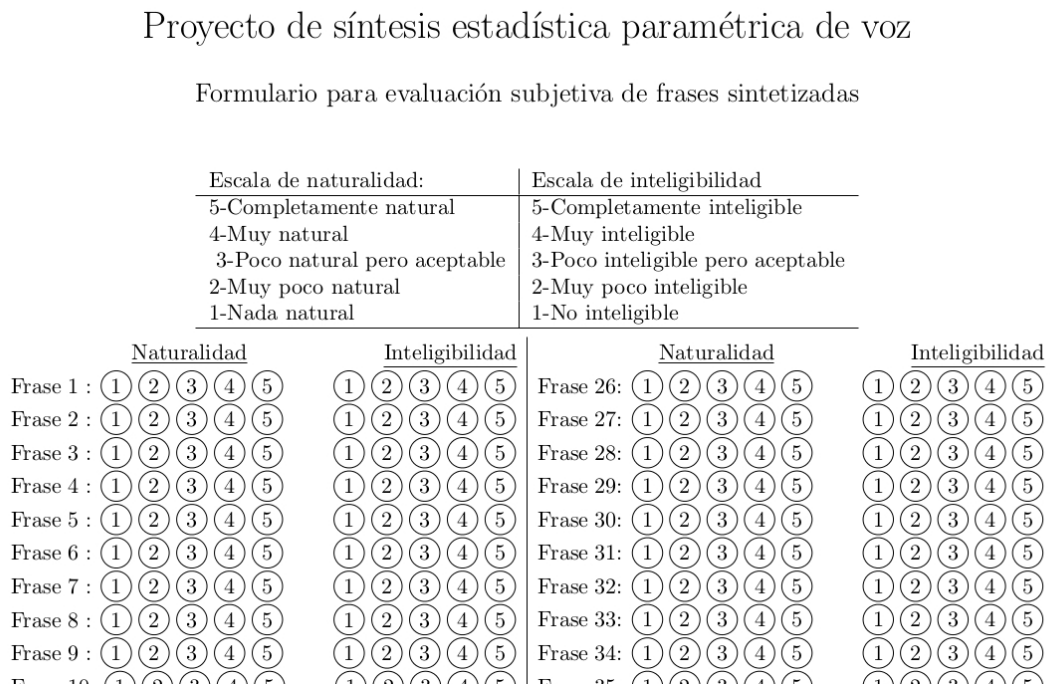
donde T_i, T_{i+1} son los periodos actual y posterior, mientras que N es el número total de intervalos. El shimmer (medida de fluctuaciones entre amplitudes de periodos), se define como

$$Shm = \frac{|A_i - A_{i+1}|}{\frac{1}{N}\sum_{i=1}^{N} A_i} \tag{4.2}$$

donde A_i, A_{i+1} son los periodos presente y posterior de amplitud, y N el número total de periodos sonoros.

Estos parámetros acústicos han sido aplicados en la distinción de voces patológicas en voces reales [97] [98], o para relacionarlas con niveles de estrés [99]. La idea de la incorporación en habla sintetizada es determinar su posible correlación con la calidad y con otros medidas subjetivas.

2. Evaluación subjetiva: Se han adoptado las dos medidas subjetivas usuales para evaluación: inteligibilidad y naturalida. Para esto se diseñó un formulario (Figura 4.6) por aplicar en grupos de escuchas humanos.


Proyecto de síntesis estadística paramétrica de voz

Formulario para evaluación subjetiva de frases sintetizadas

Escala de naturalidad:	Escala de inteligibilidad
5-Completamente natural	5-Completamente inteligible
4-Muy natural	4-Muy inteligible
3-Poco natural pero aceptable	3-Poco inteligible pero aceptable
2-Muy poco natural	2-Muy poco inteligible
1-Nada natural	1-No inteligible

	Naturalidad	Inteligibilidad		Naturalidad	Inteligibilidad
Frase 1 :	1 2 3 4 5	1 2 3 4 5	Frase 26:	1 2 3 4 5	1 2 3 4 5
Frase 2 :	1 2 3 4 5	1 2 3 4 5	Frase 27:	1 2 3 4 5	1 2 3 4 5
Frase 3 :	1 2 3 4 5	1 2 3 4 5	Frase 28:	1 2 3 4 5	1 2 3 4 5
Frase 4 :	1 2 3 4 5	1 2 3 4 5	Frase 29:	1 2 3 4 5	1 2 3 4 5
Frase 5 :	1 2 3 4 5	1 2 3 4 5	Frase 30:	1 2 3 4 5	1 2 3 4 5
Frase 6 :	1 2 3 4 5	1 2 3 4 5	Frase 31:	1 2 3 4 5	1 2 3 4 5
Frase 7 :	1 2 3 4 5	1 2 3 4 5	Frase 32:	1 2 3 4 5	1 2 3 4 5
Frase 8 :	1 2 3 4 5	1 2 3 4 5	Frase 33:	1 2 3 4 5	1 2 3 4 5
Frase 9 :	1 2 3 4 5	1 2 3 4 5	Frase 34:	1 2 3 4 5	1 2 3 4 5

Figura 4.6: Fragmento del formulario utilizado para evaluaciones subjetivas

Los resultados de frases sintetizadas se dividieron en cuatro grupos:

a) Frases de la aplicación Reloj, voces masculinas.

b) Frases de la aplicación Reloj, voces femeninas.

c) Frases de la aplicación Clima, voces masculinas.

d) Frases de la aplicación Clima, voces femeninas.

Por cada voz resultantes se seleccionaron al azar cinco frases, y se presentaron aleatoriamente en grupos de 50 frases en cada sesión de escucha. Las pruebas se aplicaron a más de ochenta personas, y después de descartar algunas completadas de forma incorrecta (incompletas), se conformó el grupo de evaluaciones con 20 por cada grupo de voces. Con las evaluaciones en la escala 1 a 5 se obtuvo la media (MOS) de naturalidad e inteligibilidad.

En evaluaciones subjetivas se incluyeron las voces:

- Voces masculinas:
  - Voz obtenida con la base de datos completa
  - Voz obtenida con rango de $f0$ estrecho
  - Voz obtenida con rango de $f0$ amplio
  - Voz obtenida con información de contexto reducida
  - Voz obtenida con base de datos reducida
  - Voz obtenida con base de datos duplicada
  - Voz obtenida con base de datos aumentada
  - Voz obtenida con la voz de Carlos Fuentes
  - Voz obtenida con la mezcla de voces de voluntarios
  - Voz obtenida con el sintetizador comercial AT& Natural Voices®

Voces femeninas:

- Voz obtenida con la base de datos completa
- Voz obtenida con rango de $f0$ estrecho
- Voz obtenida con rango de $f0$ amplio
- Voz obtenida con información de contexto reducida
- Voz obtenida con base de datos reducida
- Voz obtenida con base de datos duplicada
- Voz obtenida con base de datos aumentada
- Voz obtenida con la mezcla de la base de datos usual (emoción normal), y la misma base de datos producida por la hablante original con emoción tristeza, conformando una voz mezcla de emociones. Se ha elegido la emoción tristeza en la mezcla por no tener diferencias en variaciones de tono a lo largo de las frases.
- Voz obtenida con la mezcla de voces de voluntarios
- Voz obtenida con el sintetizador comercial AT& Natural Voices®

Las voces del sintetizador comercial, que utilizan la técnica de selección de unidades, han sido incluidas como un medio de control y comparación con las obtenidas mediante la síntesis estadística paramétrica. En las siguientes secciones se detallan las resultados por cada grupo de estudio de parámetros o condiciones de entrenamiento.

4.4. Pruebas sobre la influencia de parámetros de entrenamiento

La influencia de parámetros de entrenamiento se estudia con la variación en la definición del rango de $f0$:

- La definición de posibles valores de $f0$ dentro de un rango estrecho comparado con el rango real de un hablante masculino o femenino. Estas voces serán llamadas $f0$ estrecho, haciendo distinción entre aquellas provenientes del hablante masculino o femenino.
- La definición de posibles valores de $f0$ dentro de un rango amplio comparado con el rango real de un hablante masculino o femenino. Estas voces serán llamadas $f0$ amplio, haciendo distinción entre aquellas provenientes del hablante masculino o femenino.
- La definición de posibles valores de $f0$ dentro de un rango adecuado comparado con el rango real de un hablante masculino o femenino. Estas voces serán llamadas $f0$ ajustado, haciendo distinción entre aquellas provenientes del hablante masculino o femenino.

Se produjeron voces a partir de la condición de entrenamiento fijada en cada una de estas pruebas. En total son doce pruebas, divididas en grupos de seis por cada género, y subdividas en tres por prueba. Las voces resultantes fueron sometidas a pruebas subjetivas y objetivas, tal como se detalla en las siguientes subsecciones.

4.4.1. Evaluación objetiva

La evaluación objetiva considera parámetros utilizados en sistemas de aprendizaje de un segundo idioma SLL, parámetros acústicos y evaluación basada en un clasificador, en este caso un reconocedor de palabras.

Evaluación con parámetros SLL

El objetivo de las métricas SLL es evaluar la similitud del habla sintetizada con la voz original. En las Tablas 4.3 a 4.6 se muestran los resultados de los parámetros mpl, ros, ptr y art para las voces masculinas y femeninas. La condición de entrenamiento se refiere al ajuste del parámetro $f0$ realizado en el entrenamiento de los HMM.

Tabla 4.3: Comparación de parámetros objetivos SLL sobre pruebas de influencia de rango de $f0$ en la voz sintetizada. Voz masculina, aplicación Reloj

Voz/Condición	mlp	ros	ptr	art
Hablante	0.21	12.31	87.77 %	14.02
$f0$ ajustado	0.07	11.08	89.59 %	12.37
$f0$ amplio	0.10	10.58	86.09 %	12.29
$f0$ estrecho	0.06	11.76	92.24 %	12.75

Tabla 4.4: Comparación de parámetros objetivos SLL sobre pruebas de influencia de rango de $f0$. Voz masculina, aplicación Clima

Voz	mlp	ros	ptr	art
Hablante	0.21	12.31	87.77 %	14.02
$f0$ ajustado	0.09	12.26	84.45 %	12.98
$f0$ amplio	0.11	12.56	94.53 %	13.29
$f0$ estrecho	0.14	11.35	90.67 %	12.51

Tabla 4.5: Comparación de parámetros objetivos sobre pruebas de influencia de rango de $f0$. Voz femenina, aplicación Reloj

Voz	mlp	ros	ptr	art
Hablante	0.14	11.88	91.57 %	12.97
$f0$ ajustado	0.07	10.72	88.81 %	12.07
$f0$ amplio	0.04	11.43	94.46 %	12.10
$f0$ estrecho	0.05	11.17	92.48 %	12.08

Tabla 4.6: Comparación de parámetros objetivos sobre pruebas de influencia de rango de $f0$. Voz femenina, aplicación Clima

Voz	mlp	ros	ptr	art
Hablante	0.14	11.88	91.57 %	12.97
$f0$ ajustado	0.05	11.55	96.03 %	12.03
$f0$ amplio	0.05	12.15	96.68 %	12.57
$f0$ estrecho	0.08	11.43	94.89 %	12.05

Es notorio que la duración media de pausas sea en todos los casos inferior en las voces sintetizadas que en el habla normal. A partir de los resultados de similitud SLL en ambos hablantes y aplicaciones se puede establecer la similitud con el hablante original (base de datos) utilizando distancia euclidea de cada punto $k_i = (mlp_i, ros_i, ptr_i, art_i)$ correspondiente a las pruebas, al punto k_b característico del hablante. En las Tablas 4.7 y 4.8 se resumen los resultados.

Tabla 4.7: Similitud con hablante original masculino a partir de la distancia de parámetros objetivos en pruebas sobre la influencia del rango de $f0$

Condición de entrenamiento	$D(k_i, k_b)$
$f0$ ajustado (Reloj)	2.06
$f0$ amplio (Reloj)	2.45
$f0$ estrecho (Reloj)	1.39
$f0$ ajustado (Clima)	1.05
$f0$ amplio (Clima)	0.78
$f0$ estrecho (Clima)	1.79

Se observa que el entrenamiento con $f0$ amplio en la aplicación Clima de ambos hablantes es la que obtiene una mayor semejanza con el hablante original. No es un patrón que se mantenga para la aplicación Reloj, donde el entrenamiento con el rango de $f0$ estrecho obtiene la mayor semejanza en el caso del hablante hombre, y el de rango de $f0$ amplio tiene la menor semejanza. En el caso de mujeres, la voz resultante del entrenamiento con $f0$ amplio

Tabla 4.8: Similitud con hablante original femenino a partir de la distancia de parámetros objetivos en pruebas sobre la influencia del rango de $f0$

Condición de entrenamiento	$D(k_i, k_b)$
$f0$ ajustado (Reloj)	1.47
$f0$ amplio (Reloj)	0.98
$f0$ estrecho (Reloj)	1.14
$f0$ ajustado (Clima)	1.00
$f0$ amplio (Clima)	0.49
$f0$ estrecho (Clima)	0.99

coincide en el caso de la hablante mujer como la más semejante en medidas SLL a la voz original.

Evaluación de similitud en parámetros acústicos

La similitud de parámetros acústicos pretende realizar una comparación entre el hablante de la base de datos y las voces sintetizadas en tres medidas: tono, *jitter* (local) y *shimmer*. Dado que no todos los fonemas son sonoros, se han calculado estos parámetros solamente sobre las vocales, tanto de la base de datos como de las frases sintetizadas. Para este fin se desarrollaron rutinas en el programa Praat que extraen el conjunto de datos correspondiente a todas las frases, clasificados por vocal.

Sobre la base de datos se han utilizado todas las frases, mientras que para evaluarlos en las voces sintetizadas se han seleccionado al azar cien frases de hora o de predicción de tiempo atmosférico, según la aplicación. Dado que los parámetros presentan un rango de valores, se ha elegido como representación de resultados los digramas de caja. También se han aplicado pruebas estadísticas para determinar si las diferencias que se observan son estadísticamente significativas, dentro de un nivel de significancia. Se consideran de esta manera cuando el $p-$valor está por debajo del nivel establecido, de 0.05.

En el Apéndice B.1 se muestran los gráficos y resultados de pruebas estadísticas para la significancia en la diferencia de parámetros acústicos con la voz original. Las pruebas de Friedman que han resultado con estas diferencias significativas en parámetros acústicos se muestran en la Tabla 4.9 y 4.10.

Tabla 4.9: Diferencias estadísticamente significativas de parámetros acústicos con el hablante original de acuerdo con prueba de Friedman. Pruebas sobre la influencia de parámetros de entrenamiento. Aplicación Clima. M: Voz femenina, H: Voz masculina

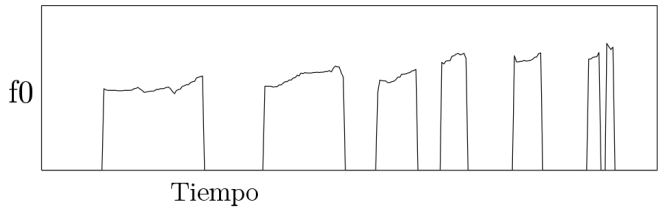
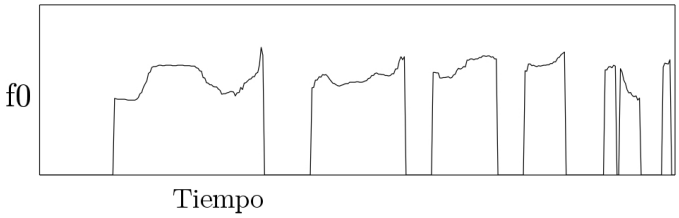
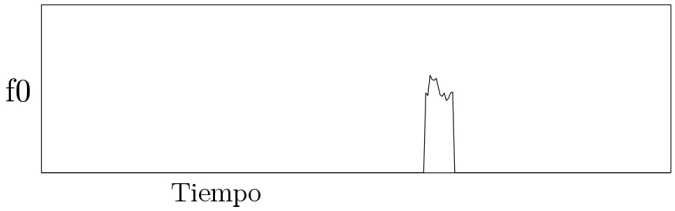
Condición de entrenamiento	Diferencia significativa		
	Tono	*Jitter*	*Shimmer*
$f0$ ajustado (H)	✓	✓	
$f0$ ajustado (M)	✓	✓	✓
$f0$ estrecho (H)	✓	✓	✓
$f0$ estrecho (M)	✓	✓	✓
$f0$ amplio (H)		✓	
$f0$ amplio (M)		✓	

Tabla 4.10: Diferencias estadísticamente significativas de parámetros acústicos con el hablante original de acuerdo con prueba de Friedman. Pruebas sobre la influencia de parámetros de entrenamiento. Aplicación Reloj. M: Voz femenina, H: Voz masculina

Condición de entrenamiento	Diferencia significativa		
	Tono	*Jitter*	*Shimmer*
$f0$ ajustado (H)		✓	
$f0$ ajustado (M)	✓		✓
$f0$ estrecho (H)	✓	✓	✓
$f0$ estrecho (M)	✓	✓	✓
$f0$ amplio (H)			
$f0$ amplio (M)	✓		

Para el caso de tono, las diferencias marcadas entre los resultados que se escuchan con las diferentes definiciones del rango de $f0$ de las vocales pueden explicarse por los datos que son presentados a los HMM en el entrenamiento, en los cuales gran cantidad de información de $f0$

Tabla 4.11: Comparación de contornos de $f0$ para la frase sintetizada de la hora: "Son las 8:45". Voz masculina

Condición de entrenamiento	Contorno $f0$
$f0$ ajustado	
$f0$ amplio	
$f0$ estrecho	

no es extraída de forma adecuada en la definición de rango estrecho, por lo que posteriormente no es generada de forma adecuada. En la Tabla 4.11 se muestran contornos de $f0$ para una frase de la hora, en los cuales pueden apreciarse diferencias considerables entre éstos, los cuales tienen gran influencia en la apreciación subjetiva, como se detalla en la siguiente sección

Evaluación con clasificador

Los resultados de esta evaluación se muestran en la Tabla 4.12, registrados como tasa de error.

En todos los casos, la menor tasa de error por hablante y aplicación se obtuvo con el rango

Tabla 4.12: Tasa de palabras correctas en clasificación de palabras con reconocedor automático. H: Voz masculina, M: voz femenina

Condición de entrenamiento	Tasa de error (%)
$f0$ ajustado. Aplicación Reloj (M)	5.80
$f0$ amplio. Aplicación Reloj (M)	15.53
$f0$ estrecho. Aplicación Reloj (M)	14.68
$f0$ ajustado. Aplicación Reloj (H)	6.14
$f0$ amplio. Aplicación Reloj (H)	17.24
$f0$ estrecho. Aplicación Reloj (H)	22.35
$f0$ ajustado. Aplicación Clima (M)	18.08
$f0$ amplio. Aplicación Clima (M)	24.31
$f0$ estrecho. Aplicación Clima (M)	33.81
$f0$ ajustado. Aplicación Clima (H)	17.88
$f0$ amplio. Aplicación Clima (H)	16.65
$f0$ estrecho. Aplicación Clima (H)	31.97

de $f0$ ajustado. La mayor tasa de error se obtuvo en el caso de la voz femenina obtenida a partir de rango estrecho de $f0$. Para contar con un valor de referencia, se utilizó este reconocedor para obtener la tasa de error de reconocimiento para las bases de datos de ambos hablantes, a partir de 50 frases. Los resultados son: Un 12.21 % de error para la voz de mujer, y un 8.14 % para la voz de hombre. A pesar de no ser el mismo tipo de palabras, se puede considerar que las voces sintetizadas a partir de una definición adecuada de rango de $f0$ en el entrenamiento tienen una tasa de error en reconocedor de palabras que no se diferencia considerablemente de la voz original.

4.4.2. Evaluación subjetiva

Para el caso de las evaluaciones subjetivas de las voces obtenidas con las diferentes opciones de rango de $f0$ consideradas, los resultados se muestran en la Tabla 4.13. El MOS se ha obtenido como la media de las calificaciones dadas por los escuchas a cada frase seleccionada,

en escala de 1 a 5.

Tabla 4.13: MOS para evaluación subjetiva de influencia del rango de $f0$ como parámetro de entrenamiento

Condición de entrenamiento	Naturalidad		Inteligibilidad	
	MOS	Desv. est.	MOS	Desv. est.
Rango $f0$ ajustado. Aplicación Reloj (M)	1.96	0.55	2.52	0.72
Rango $f0$ amplio. Aplicación Reloj (M)	2.76	0.57	3.48	0.63
Rango de $f0$ estrecho. Aplicación Reloj (M)	1.82	0.62	2.79	0.82
Rango $f0$ ajustado. Aplicación Reloj (H)	2.96	0.76	3.61	0.72
Rango $f0$ amplio. Aplicación Reloj (H)	3.10	0.53	3.83	0.65
Rango de $f0$ estrecho. Aplicación Reloj (H)	2.43	0.86	3.18	0.76
Rango $f0$ ajustado. Aplicación Clima (M)	2.40	0.59	2.80	0.57
Rango $f0$ amplio. Aplicación Clima (M)	2.37	0.58	2.84	0.56
Rango de $f0$ estrecho. Aplicación Clima (M)	2.19	0.82	2.58	0.74
Rango $f0$ ajustado. Aplicación Clima (H)	2.31	0.54	3.06	0.71
Rango $f0$ amplio. Aplicación Clima (H)	2.58	0.48	3.54	0.61
Rango de $f0$ estrecho. Aplicación Clima (H)	1.46	0.70	3.00	0.83

A partir de estos resultados subjetivos se consideró como la voz más natural a obtenida con rango de $f0$ amplio del hablante masculino en aplicación Reloj. La menos natural corresponde a la obtenida con rango estrecho de $f0$ en el entrenamiento.

En inteligibilidad la mejor evaluación la obtuvo la voz masculina producida con el rango amplio de $f0$ ajustado en la aplicación Reloj, seguido de la voz masculina obtenida a partir del rango de $f0$ ajustado. Coinciden la mejor evaluación de naturalidad e inteligibilidad, pero no la menor evaluada, pues en el caso de inteligibilidad se dio en la voz femenina obtenida con rango de $f0$ ajustado, y en masculina con rango de $f0$ estrecho.

4.5. Pruebas sobre la influencia del tamaño del conjunto de entrenamiento

La influencia del tamaño del conjunto de entrenamiento se estudia a partir de cuatro pruebas, en las que se han realizado procedimientos para modificar la base de datos, produciendo los siguientes casos:

- Entrenamiento con la base de datos completa (184 frases). Este experimento será referido como Base completa.

- Entrenamiento con la base de datos reducida: En éste se toman en cuenta únicamente aquellas frases que aportan fonemas cuya información de contexto fonético coincide con las frases que se desean sintetizar, tanto en aplicación de hora como de predicción de tiempo atmosférico. Para el caso de la hora, la base de datos reducida consta de 50 frases, con duración de 3.3 minutos. En el caso de Clima la base de datos consta de 47 frases, con duración de 4.6 minutos. Estos experimentos serán referidos como Base reducida.

- Entrenamiento con la base de datos duplicada: Las 184 frases se duplican y se considera una sola base de datos, que consiste en 368 frases. Estos experimentos serán referidos como Base duplicada.

- Entrenamiento con la base de datos aumentada: Las 184 frases son sometidas a dos procesos

    - Reducción de velocidad: Las frases son reducidas en velocidad sin alterar el tono. Se hace en dos porcentajes: 3 % y 5 %.

- Aumento de velocidad: Las frases son aumentadas en velocidad sin alterar el tono. Se hace en dos porcentajes: 3 % y 5 %.

De esta manera, se cuenta con cuatro variaciones de velocidad sobre los datos, además del conjunto original de frases, para un total de 920 frases en base de datos. La intención de realizar los cambios de velocidades es producir cambios que puedan reflejarse en el espectro, pero que no pierdan semejanza con el hablante original y puedan presentarse como nuevos datos de entrenamiento. Estos experimentos serán referidos como Base aumentada.

En todos los casos se mantuvo el rango de $f0$ ajustado, de acuerdo con la definición de la sección anterior.

4.5.1. Evaluación objetiva

Evaluación con parámetros SSL

De forma semejante a las pruebas anteriores se utilizan como parámetros mlp (duración media de pausas), ros (tasa de fonemas y pausas por segundo), ptr (Razón del total del habla sin pausas y la duración total incluyendo pausas) y art (cantidad de fonemas dividido por la duración total del habla sin pausas). En las Tablas 4.14 a 4.17 se muestran los resultados de estos parámetros para las voces de hombre y de mujer.

La tendencia de obtener un menor valor de mlp en voces sintetizadas coincide en todas las pruebas realizadas con respecto al tamaño del conjunto de entrenamiento. Esto aplica para la voz de mujer, cuyos resultados se muestran en las Tablas 4.16 y 4.17.

A partir de los resultados, y de forma semejante a lo realizado en la Sección 4.4.1 se puede establecer la similitud con el hablante original (base de datos) utilizando distancia euclidea,

Tabla 4.14: Comparación de parámetros objetivos sobre pruebas de influencia del tamaño del conjunto de entrenamiento. Voz masculina, aplicación Clima

Voz/Condición	mlp	ros	ptr	art
Hablante	0.21	12.31	87.77 %	14.02
Base completa	0.09	12.26	94.45 %	12.98
Base reducida	0.10	11.43	93.72 %	12.20
Base duplicada	0.11	12.52	93.59 %	13.38
Base aumentada	0.12	11.41	94.41 %	12.08

Tabla 4.15: Comparación de parámetros objetivos sobre pruebas de influencia del tamaño del conjunto de entrenamiento. Voz masculina, aplicación Reloj

Voz	mlp	ros	ptr	art
Hablante	0.21	12.31	87.77 %	14.02
Base completa	0.07	11.08	89.59 %	12.37
Base reducida	0.07	11.70	90.21 %	12.97
Base duplicada	0.09	11.77	88.46 %	13.31
Base aumentada	0.08	12.17	91.59 %	13.29

Tabla 4.16: Comparación de parámetros objetivos sobre pruebas de influencia del tamaño del conjunto de entrenamiento. Voz femenina, aplicación reloj

Voz	mlp	ros	ptr	art
Hablante	0.14	11.88	91.57 %	12.97
Base completa	0.07	10.72	88.81 %	12.07
Base reducida	0.04	11.98	95.90 %	12.49
Base duplicada	0.04	12.25	95.61 %	12.81
Base aumentada	0.03	11.62	96.70 %	12.02

lo cual se registra en las Tablas 4.18 y 4.19.

Se obtiene la mayor similitud con los parámetros SLL del hablante original utilizando base reducida en el caso del hablante hombre en aplicación Reloj, y en el caso de mujer con la base de datos duplicada. Las mayores diferencias se encuentran con la base de datos completa. Estos resultados distan de ser los esperados, pues si los procedimientos para aumentar el

Tabla 4.17: Comparación de parámetros objetivos sobre pruebas de influencia del tamaño del conjunto de entrenamiento. Voz femenina, aplicación predicción de tiempo atmosférico

Voz	mlp	ros	ptr	art
Hablante	0.14	11.88	91.57 %	12.97
Base completa	0.05	11.55	96.03 %	12.03
Base reducida	0.05	11.79	97.82 %	12.05
Base duplicada	0.08	11.54	95.06 %	12.14
Base aumentada	0.05	11.98	97.81 %	12.25

Tabla 4.18: Similitud con hablante original masculino a partir de la distancia de parámetros objetivos en pruebas de influencia del tamaño del conjunto de entrenamiento

Condición de entrenamiento	$D(k_i, k_b)$
Base de datos completa (Reloj)	0.97
Base de datos reducida (Reloj	0.91
Base de datos duplicada (Reloj)	1.38
Base de datos aumentada (Reloj)	0.91
Base de datos completa (Clima)	1.29
Base de datos reducida (Clima)	1.14
Base de datos duplicada (Clima)	1.40
Base de datos aumentada (Clima)	1.28

Tabla 4.19: Similitud con hablante original femenino a partir de la distancia de parámetros objetivos en pruebas de influencia del tamaño del conjunto de entrenamiento

Condición de entrenamiento	$D(k_i, k_b)$
Base de datos completa (Reloj)	1.47
Base de datos reducida (Reloj)	0.50
Base de datos duplicada (Reloj)	0.42
Base de datos aumentada (Reloj)	0.99
Base de datos completa (Clima)	1.00
Base de datos reducida (Clima)	0.93
Base de datos duplicada (Clima)	0.90
Base de datos aumentada (Clima)	0.74

tamaño de la base de datos resultan efectivos en mejorar la similitud con el hablante original, el uso de la base reducida debería desmejorar. De manera que estos resultados en similitud

SLL no muestran una consistencia con respecto a la cantidad de datos en el conjunto de entrenamiento.

Evaluación de similitud en parámetros acústicos

El desglose de resultados correspondiente se presenta en la Sección B.2. En las Tablas 4.20 y 4.21 se resumen los resultados de significancia estadística en la diferencia de los parámetros acústicos con respecto a la voz original.

Tabla 4.20: Diferencias estadísticamente significativas de parámetros acústicos con el hablante original de acuerdo con prueba de Friedman. Aplicación Clima. M: Voz femenina, H: Voz masculina

Condición de entrenamiento	Diferencia significativa		
	Tono	*Jitter*	*Shimmer*
Entrenamiento normal (H)	✓	✓	
Entrenamiento normal (M)	✓	✓	✓
Base reducida (H)		✓	
Base reducida (M)	✓	✓	✓
Base duplicada (H)			
Base duplicada (M)	✓	✓	✓
Base aumentada (H)	✓		✓
Base aumentada (M)	✓	✓	✓

Evaluación con clasificador

De forma semejante a las pruebas anteriores, se utilizaron 100 frases elegidas al azar de cada hablante y en cada aplicación para enviarlas a un reconocedor de voz y establecer una tasa de error. Los resultados se encuentran en la Tabla 4.22.

En estas pruebas objetivas con reconocedor de palabras vuelven a tener las menores tasas de error las voces sintéticas obtenidas con el entrenamiento normal, en este caso la base de datos completa. Los resultados de las voces femeninas son muy inferiores a lo obtenido en vo-

Tabla 4.21: Diferencias estadísticamente significativas de parámetros acústicos con el hablante original de acuerdo con prueba de Friedman. Aplicación Reloj. M: Voz femenina, H: Voz masculina

Condición de entrenamiento	Diferencia significativa		
	Tono	*Jitter*	*Shimmer*
Entrenamiento normal (H)		✓	
Entrenamiento normal (M)	✓		✓
Base reducida (H)		✓	
Base reducida (M)	✓		✓
Base duplicada (H)			✓
Base duplicada (M)			✓
Base aumentada (H)			✓
Base aumentada (M)		✓	✓

ces masculinas, lo cual posiblemente se deba a una condición del reconocedor de palabras. Los procedimientos realizados para aumentar el tamaño de base de datos (base duplicada y base aumentada) no parecen tener una influencia positiva en cuanto a la tasa de reconocimiento de palabras.

Para la voz en ambos géneros, las mayores tasas de error se obtienen con las voces sintéticas producidas con la base de datos reducida. Las diferencias son más notorias en la aplicación Clima en ambos casos.

4.5.2. Evaluación subjetiva

La evaluación subjetiva se realizó con escuchas voluntarios, de forma semejante a los anteriores. Los resultados se muestran en la Tabla 4.23. Se incluyó en este caso una voz adicional femenina, llamada voz mixta, producida con la mezcla de emociones.

En este caso la voz mejor evaluada en naturalidad fue la voz masculina obtenida con la base de datos completa, en la aplicación Reloj. En las voces femeninas, la mayor naturalidad la obtuvo la voz mixta.

Tabla 4.22: Tasa de palabras correctas en clasificación de palabras con reconocedor automático

Condición de entrenamiento	Tasa de error (%)
Base completa. Alicación Reloj (H)	6.14
Base reducida. Aplicación Reloj (H)	16.38
Base duplicada. Aplicación Reloj (H)	13.48
Base aumentada. Aplicación Reloj (H)	26.28
Base completa. Aplicación Clima (H)	17.88
Base reducida. Aplicación Clima (H)	40.25
Base duplicada. Aplicación Clima (H)	18.08
Base aumentada. Aplicación Clima (H)	30.44
Base completa. Aplicación Reloj (M)	5.80
Base reducida. Aplicación Reloj (M)	25.77
Base duplicada. Aplicación Reloj (M)	15.36
Base aumentada. Aplicación Reloj (M)	38.23
Base completa. Aplicación Clima (M)	18.08
Base reducida. Aplicación Clima (M)	57.30
Base duplicada. Aplicación Clima (M)	20.63
Base aumentada. Aplicación Clima (M)	23.29

Tabla 4.23: MOS para evaluación subjetiva de influencia del tamaño de la base de datos como parámetro de entrenamiento

Condición de entrenamiento	Naturalidad		Inteligibilidad	
	MOS	Desv. est.	MOS	Desv. est.
Base completa. Aplicación Reloj (H)	2.96	0.76	3.61	0.62
Base reducida. Aplicación Reloj (H)	2.74	0.65	3.32	0.45
Base duplicada. Aplicación Reloj (H)	2.68	0.62	3.37	0.57
Base aumentada. Aplicación Reloj (H)	2.87	0.58	3.31	0.48
Base completa. Aplicación Clima (H)	2.31	0.54	3.06	0.71
Base reducida. Aplicación Clima (H)	1.95	0.51	2.27	0.58
Base duplicada. Aplicación Clima (H)	2.38	0.61	3.3	0.68
Base aumentada. Aplicación Clima (H)	2.03	0.40	2.79	0.69
Base completa. Aplicación Reloj (M)	1.96	0.55	2.52	0.72
Base reducida.Aplicación Reloj (M)	2.37	0.54	2.56	0.71
Base duplicada. Aplicación Reloj (M)	2.37	0.56	2.96	0.74
Base mixta. Aplicación Reloj (M)	2.89	0.58	3.73	0.65
Base aumentada. Aplicación Reloj (M)	2.28	0.54	3.02	0.64
Base completa. Aplicación Clima (M)	2.40	0.59	2.80	0.57
Base reducida. Aplicación Clima (M)	1.95	0.61	2.26	0.55
Base duplicada. Aplicación Clima (M)	2.38	0.67	2.67	0.47
Base aumentada. Aplicación Clima (M)	2.29	0.60	2.77	0.60
Base mixta. Aplicación Clima (M)	2.32	0.68	2.85	0.65

Se destaca en inteligibilidad la voz mixta femenina en aplicación Reloj, seguida de la de hablante masculino en la misma aplicación. Las mejores apreciaciones de las voces de esta aplicación pueden explicarse por sus frases más cortas y menor vocabulario, en comparación con la aplicación Clima.

4.6. Pruebas sobre la calidad de las grabaciones

La calidad de grabaciones es analizada de forma comparativa, tomando como referencia las voces obtenidas con el entrenamiento normal, tanto para hombre como para mujer. Estas voces han sido obtenidas con condiciones controladas de ruido, y formato de audio sin pérdidas. Como experimentos se tomarán el análsis de voces obtenidas con dos procedimientos:

- Voz obtenida con un formato de audio de menor calidad: Se toma como fuente de datos un audiolibro del escritor mexicano Carlos Fuentes, leído por él mismo. El audio está en formato MP3, con una tasa de bits de 120 kpbs. Aunque la calidad es inferior, la cantidad de datos disponibles supera la base de datos normal, con aproximadamente 70 minutos de grabación.

- Voz obtenida sin condiciones controladas de ruido: Se realizaron grabaciones sin condiciones óptimas de grabación, reproduciendo el contenido de la base de datos de 184 frases. Para esto se contó con veintiséis estudiantes voluntarios, trece por cada género, los cuales grabaron partes de la base de datos. Por esta razón la voz resultante se puede considerar una voz promedio, llamada UAMI-H para el caso de voces masculinas, y UAMI-M para voces femeninas.

Para evaluación se utilizan como parámetros la tasa de error en palabras del reconocedor automático de habla, y las evaluaciones subjetivas. En este caso no se aplica el análisis de

Tabla 4.24: Tasa de error en clasificación de palabras con reconocedor automático

Condición de entrenamiento	Tasa de error (%)
Entrenamiento normal. Aplicación Reloj (H)	6.14
Voz de Carlos Fuentes. Aplicación Reloj	11.09
UAMI-H. Aplicación Reloj	28.5
Entrenamiento normal. Aplicación Clima (H)	17.88
Voz de Carlos Fuentes. Aplicación Clima	18.90
UAMI-H. Aplicación Clima	34.42
Entrenamiento normal. Aplicación Reloj (M)	5.80
UAMI-M. Aplicación Reloj	18.43
Entrenamiento normal. Aplicación Clima (M)	18.08
UAMI-M. Aplicación Clima	20.94

similitud con el hablante original, debido a que las voces tienen fuentes diferentes.

4.6.1. Evaluación objetiva

Se considera únicamente la tasa de error del clasificador de palabras, como se detalla en la siguiente subsección.

Evaluación con clasificador

El procedimiento seguido para obtener la tasa de error en clasificador de palabras es semejante a las pruebas anteriores, utilizando el reconocedor de palabras de Google. Los resultados se muestra en la Tabla 4.24.

En este caso, a pesar de tener la voz de Carlos Fuentes mayor cantidad de datos de entrenamiento, los resultados con clasificador no superan a los de la voz obtenida con entrenamiento normal, tanto masculina como femenina. Las voces obtenidas con la mezcla de trece voces obtienen resultados muy inferiores, tanto en hombres como en mujeres.

4.6.2. Evaluación subjetiva

La evaluación subjetiva se realizó de forma semejante a las pruebas anteriores, con 20 estudiantes voluntarios que escucharon 5 frases de cada voz, de forma aleatoria. Esto para cada género de voz y para cada aplicación. Los resultados se muestran en la Tabla 4.25

Tabla 4.25: MOS para evaluación subjetiva de influencia de la calidad de grabaciones

Condición de entrenamiento	Naturalidad		Inteligibilidad	
	MOS	Desv. est.	MOS	Desv. est.
Entrenamiento normal. Alicación Reloj (H)	2.96	0.76	3.61	0.62
Voz de Carlos Fuentes. Aplicación Reloj	3.59	0.51	3.16	0.43
UAMI-H. Aplicación Reloj	2.18	0.87	2.57	0.56
Entrenamiento normal. Aplicación Clima (H)	2.31	0.54	3.06	0.71
Voz de Carlos Fuentes. Aplicación Clima	4.14	0.63	4.12	0.69
UAMI-H. Aplicación Clima	1.86	0.96	1.81	0.79
Entrenamiento normal. Aplicación Reloj (M)	1.96	0.55	2.52	0.72
UAMI-M. Aplicación Reloj	2.11	0.61	2.77	0.81
Entrenamiento normal. Aplicación Clima (M)	2.40	0.59	2.80	0.57
UAMI-M. Aplicación Clima	2.34	0.64	2.93	0.70

Destaca la voz sintética obtenida a partir de la voz de Carlos Fuentes como la mejor apreciada en naturalidad e inteligibilidad, por encima en esta ocasión de las voces obtenidas con entrenamiento normal. Obtienen resultados muy bajos en ambos rubros las voces sintéticas obtenidas como mezcla de las trece voces de estudiantes, lo cual concuerda con los bajos resultados en la tasa de error en reconocimiento.

4.7. Pruebas sobre la influencia de la información de contexto

La influencia de la información de contexto se analiza comparando la voz sintética obtenida con entrenamiento normal, y una voz obtenida con la base de datos completa, pero utilizando solamente la parte de información de contexto fonética, es decir, se excluye la información prosódica. De esta manera se espera que las frases obtenidas con las voces sintéticas con contexto reducido sean menos naturales, pues no diferencian fonemas de inicio, medio o fin de frase. Sin embargo, es posible que tengan ventajas en cuanto a pronunciación, pues contarían con mayor cantidad de elementos en la base de datos por fonema, por esta característica de diferenciarlos solamente por su contenido fonético y no prosódico. Las pruebas realizadas son:

- Entrenamiento de voz masculina utilizando solamente información de contexto relacionado con la fonética.

- Entrenamiento de voz femenina utilizando solamente información de contexto relacionado con la fonética.

Estas voces son llamadas Contexto reducido. Se obtuvieron frases sintetizadas para ambos géneros, y se evaluaron en la aplicación Reloj y Clima. Los resultados de esta evaluación se muestran a continuación.

Evaluación con parámetros SLL

Los resultados de los parámetros SLL para ambas aplicaciones y las voces femeninas y masculinas se muestra en las Tablas 4.26 a 4.29. En todos los casos se observa la tendencia

a una menor duración media de silencios (mlp), con diferencias significativas con respecto a los demás parámetros.

Tabla 4.26: Comparación de parámetros objetivos sobre pruebas de influencia de información de contexto. Voz masculina, aplicación Reloj

Voz/Condición	mlp	ros	ptr	art
Hablante	0.21	12.31	87.77 %	14.02
Entrenamiento normal	0.07	11.08	89.59 %	12.37
Contexto reducido	0.10	11.19	85.30 %	13.12

Tabla 4.27: Comparación de parámetros objetivos sobre pruebas de influencia información de contexto. Voz masculina, aplicación Clima

Voz/Condición	mlp	ros	ptr	art
Hablante	0.21	12.31	87.77 %	14.02
Entrenamiento normal	0.09	12.26	84.45 %	12.98
Contexto reducido	0.09	11.65	95.18 %	12.24

Tabla 4.28: Comparación de parámetros objetivos sobre pruebas de influencia de información de contexto. Voz femenina, aplicación Reloj

Voz/Condición	mlp	ros	ptr	art
Hablante	0.14	11.88	91.57 %	12.97
Entrenamiento normal	0.07	10.72	88.81 %	12.07
Contexto reducido	0.05	10.59	92.05 %	11.50

Tabla 4.29: Comparación de parámetros objetivos sobre pruebas de influencia de información de contexto. Voz femenina, aplicación Clima

Voz/Condición	mlp	ros	ptr	art
Hablante	0.14	11.88	91.57 %	12.97
Entrenamiento normal	0.05	11.55	96.03 %	12.03
Contexto reducido	0.04	11.24	97.87 %	11.48

A partir de los resultados anteriores se puede establecer la similitud con el hablante original (base de datos) utilizando distancia euclidea. En las Tablas 4.30 y 4.31 se resumen los resultados.

Tabla 4.30: Similitud con hablante original masculino a partir de la distancia de parámetros objetivos en pruebas sobre la influencia del rango de $f0$

Condición de entrenamiento	$D(k_i, k_b)$
Entrenamiento normal (Reloj)	2.06
Contexto reducido (Reloj)	1.06
Entrenamiento normal (Clima)	0.39
Contexto reducido (Clima)	0.77

Como puede observarse en las similitudes, la mayor semejanza con el hablante original la obtiene la voz sintética con entrenamiento normal en la aplicación Clima, mientras que en la aplicación Reloj la voz sintética obtenida con contexto reducido es la más semejante.

Tabla 4.31: Similitud con hablante original femenino a partir de la distancia de parámetros objetivos en pruebas sobre la influencia del rango de $f0$

Condición de entrenamiento	$D(k_i, k_b)$
Entrenamiento normal (Reloj)	1.47
Contexto reducido (Reloj)	1.96
Entrenamiento normal (Clima)	1.00
Contexto reducido (Clima)	1.63

En el caso de la voz de mujer las voces obtenidas con contexto reducido no mejoran la semejanza con la voz original, con respecto a la voz obtenida con entrenamiento normal.

4.7.1. Evaluación con clasificador

Los resultados de la tasa de error para estas pruebas de influencia de la información de contexto obtenida con el reconocedor de palabras de Google®se muestran en la Tabla 4.32.

Tabla 4.32: Tasa de palabras correctas en clasificación de palabras con reconocedor automático

Condición de entrenamiento	Tasa de error
Entrenamiento normal. Aplicación Reloj (H)	6.14
Contexto reducido. Aplicación Reloj (H)	5.80
Entrenamiento normal. Aplicación Clima (H)	17.88
Contexto reducido. Aplicación Clima (H)	22.27
Entrenamiento normal. Aplicación Reloj (M)	5.80
Contexto reducido. Aplicación Reloj (M)	7.17
Entrenamiento normal. Aplicación Clima (M)	18.08
Contexto reducido. Aplicación Clima (M)	21.96

Esta evaluación con clasificador muestra que las voces obtenidas con contexto reducido en la aplicación Reloj, para ambos géneros de hablante, tienen resultados competitivos con respecto al entrenamiento normal, el cual obtuvo las mejores evaluaciones en todas las pruebas realizadas anteriormente. En la aplicación Clima, donde las frases son más largas, se disminuye la tasa de aciertos, al igual que en el entrenamiento normal, pero los resultados siguen siendo cercanos.

Evaluación de similitud en parámetros espectrales y de frecuencia fundamental

El desglose de resultados correspondiente se presenta en la Sección B.3. En las Tablas 4.33 y 4.34 se resumen los resultados de significancia estadística en la diferencia de los parámetros acústicos con respecto a la voz original.

4.7.2. Evaluación subjetiva

Los resultados de evaluación de MOS naturalidad e inteligibilidad se presentan en la Tabla 4.35. Como en los casos anteriores, las pruebas se realizaron con 20 estudiantes voluntarios por cada género de voz y cada aplicación.

Tabla 4.33: Diferencias estadísticamente significativas de parámetros acústicos con el hablante original de acuerdo con prueba de Friedman. Aplicación Clima. M: Voz femenina, H: Voz masculina

Condición de entrenamiento	Diferencia significativa		
	Tono	*Jitter*	*Shimmer*
Entrenamiento normal (H)	✓	✓	
Entrenamiento normal (M)	✓	✓	✓
Contexto reducido (H)		✓	✓
Contexto reducido (M)	✓	✓	✓

Tabla 4.34: Diferencias estadísticamente significativas de parámetros acústicos con el hablante original de acuerdo con prueba de Friedman. Aplicación Reloj. M: Voz femenina, H: Voz masculina

Condición de entrenamiento	Diferencia significativa		
	Tono	*Jitter*	*Shimmer*
Entrenamiento normal (H)		✓	
Entrenamiento normal (M)	✓		✓
Contexto reducido(H)		✓	✓
Contexto reducido (M)	✓	✓	✓

Se observa que la voz obtenida con entrenamiento normal, en el caso de voz masculina en aplicación Reloj obtiene la mejor evaluación en naturalidad e inteligibilidad. En la aplicación Reloj, la voz femenina con información de contexto reducida obtiene mejores resultados de naturalidad e inteligibilidad que la obtenida con entrenamiento normal. Las diferencias no son tan marcadas en el caso de voces de mujer en la aplicación Clima.

Tabla 4.35: MOS para evaluación subjetiva de influencia de información de contexto

Condición de entrenamiento	Naturalidad		Inteligibilidad	
	MOS	Desv. est.	MOS	Desv. est.
Entrenamiento normal. Aplicación Reloj (H)	2.96	0.76	3.61	0.62
Contexto reducido. Aplicación Reloj (H)	2.77	0.64	3.34	0.50
Entrenamiento normal. Aplicación Clima (H)	2.31	0.54	3.06	0.71
Contexto reducido. Aplicación Clima (H)	2.13	0.52	2.99	0.72
Entrenamiento normal. Aplicación Reloj (M)	1.96	0.55	2.52	0.72
Contexto reducido. Aplicación Reloj (M)	2.60	0.60	3.13	0.71
Entrenamiento normal. Aplicación Clima (M)	2.40	0.59	2.80	0.57
Contexto reducido. Aplicación Clima (M)	2.23	0.52	2.60	0.56

Capítulo 5

Análisis de resultados

En este capítulo se presenta un análisis de los resultados presentados en el capítulo anterior, con la finalidad de resumir y destacar aquellas pruebas que han obtenido las mejores evaluaciones y así identificar las posibles direcciones que puedan llevar a mejoras en futuras experiencias. Para tener elementos de referencia, se han incluido resultados de algunas voces previamente no consideradas, provenientes de diversas fuentes. Estas nuevas voces son:

- Una voz sintética comercial, tanto de hombre como de mujer. Se utilizaron las voces de español latinoamericano de AT&T Natural Voices (Apéndice B.5). Se incluyeron estas voces en las evaluaciones subjetivas y en las de tasa de error en reconocimiento. Estas son voces de alta calidad, que utilizan métodos concatenativos.

- Voces sintéticas obtenidas de grabaciones de actores profesionales con voz en español de España. La base de datos utilizada contiene las mismas frases de la base de datos con la cual se obtuvieron las voces masculinas y femeninas de la variante de español mexicano.

- Voz mixta masculina: De forma semejante a la voz mixta obtenida con voz femenina, la cual combina una voz con emoción neutro con una voz con emoción tristeza, se obtuvo el equivalente en voz masculina para compararla con los resultas de la voz femenina.

Tabla 5.1: Resumen de resultados, aplicación Clima

Experimento	MOS-Nat	MOS-Int	Tasa error rec.	Similitud SLL
Entrenamiento normal (H)	2.31	3.06	17.88	1.05
Entrenamiento normal (M)	2.40	2.80	18.08	1.00
f0 amplio (H)	2.58	3.54	16.65	0.78
f0 amplio (M)	2.37	2.84	24.31	0.49
f0 estrecho (H)	1.46	3.00	31.97	1.79
f0 estrecho (M)	2.19	2.58	33.81	0.99
Contexto red. (H)	2.13	2.99	22.27	0.77
Contexto red. (M)	2.23	2.60	21.96	1.63
Base aumentada (H)	2.03	2.79	30.44	1.28
Base aumentada (M)	2.29	2.77	23.29	0.74
Base duplicada (H)	2.38	3.30	18.08	1.40
Base duplicada (M)	2.38	2.67	20.63	0.90
Base reducida (H)	1.95	2.27	40.25	1.14
Base reducida (M)	1.95	2.26	57.30	0.93
Voz comercial (AT&T) H	3.92	4.12	15.83	-
Voz comercial (AT&T) M	3.69	4.05	18.08	-
Mezcla (M)	2.32	2.85	21.35	-
Mezcla (H)	-	-	21.86	-
Castellano (H)	-	-	16.45	-
Castellano (M)	-	-	15.22	-
Carlos Fuentes	4.14	4.12	18.90	-
UAMI-H	1.86	1.81	34.42	-
UAMI-M	2.34	2.93	20.94	-

5.1. Resumen de resultados

Se analizan en dos tablas el resumen de resultados, de acuerdo con la aplicación. En la Tabla 5.1 lo referente a la aplicación Clima, y en la Tabla 5.2 lo correspondiente a la aplicación Reloj. No es posible aplicar todas las evaluaciones a la totalidad de resultados, ya sea porque no aplican (similitud con el hablante original mediante parámetros SLL), o bien porque no fueron incluidas en pruebas subjetivas.

Tabla 5.2: Resumen de resultados, aplicación Reloj

Experimento	MOS-Nat	MOS-Int	Tasa error rec.	Similitud SLL
Entrenamiento normal (H)	2.96	3.61	6.14	2.06
Entrenamiento normal (M)	1.96	2.52	5.80	1.47
f0 amplio (H)	3.10	3.83	17.24	2.45
f0 amplio (M)	2.76	3.48	15.53	0.98
f0 estrecho (H)	2.43	3.18	22.35	1.39
f0 estrecho (M)	1.82	2.79	14.68	1.14
Contexto red. (H)	2.77	3.34	5.80	1.06
Contexto red. (M)	2.60	3.13	7.17	1.96
Base aumentada (H)	2.87	3.31	26.28	0.91
Base aumentada (M)	2.28	3.02	38.23	0.99
Base duplicada (H)	2.68	3.37	13.48	1.38
Base duplicada (M)	2.37	2.96	15.36	0.42
Base reducida (H)	2.74	3.32	16.38	0.91
Base reducida (M)	2.37	2.56	25.77	0.50
Voz comercial (AT&T) H	3.87	4.12	13.65	-
Voz comercial (AT&T) M	4.07	4.59	17.75	-
Mezcla (M)	2.89	3.73	16.89	-
Mezcla (H)	-	-	41.64	-
Castellano (H)	-	-	12.63	-
Castellano (M)	-	-	10.03	-
Carlos Fuentes	3.59	3.16	11.09	-
UAMI-H	2.18	2.57	28.50	-
UAMI-M	2.11	2.77	18.43	-

5.2. Análisis comparativo

En las Figuras 5.1 y 5.2 se presentan los resultados del análisis subjetivo de naturalidad para ambas aplicaciones, considerando todas las voces en las que se hizo la prueba. En ambas aplicaciones las voces de AT&T obtienen mejores resultados que las voces de hombre y mujer producidas con síntesis estadística paramétrica producidas con los diferentes experimentos. Sin embargo, la voz obtenida a partir del audiolibro de Carlos Fuentes destaca en apreciación subjetiva en algunos casos sobre estas voces comerciales.

Con esto se observa una dependencia considerable de los resultados en evaluación subjetiva con la cantidad de datos utilizada (cuando éstos datos no provienen de procedimientos como la duplicación), y menos de la calidad del audio en la base de datos. También es importante destacar que las voces obtenidas con rango amplio de $f0$ tienen también buenos resultados en estas evaluaciones, incluso sobre aquellas voces con rango de $f0$ adecuadamente ajustado de acuerdo con el análisis de las voces realizado.

Las voces producidas con rango de $f0$ estrecho producen los menores resultados en evaluación subjetiva de naturalidad en ambas aplicaciones. El ordenamiento realizado sobre estos resultados subjetivos no coincide en ambas aplicaciones, lo cual muestra una dependencia de las evaluaciones subjetivas con el tipo de frases a pronunciar, así como con el género. En cuanto al ordenamiento de los resultados subjetivos de inteligiblidad, estos se muestran en las Figuras 5.3 y 5.4.

Para la aplicación Clima los mejores resultados de inteligibilidad coinciden con los de naturalidad, al ser las voces mejor evaluadas las de AT&T y la de Carlos Fuentes. Hay diferencias considerables entre las mejores y las voces menos apreciadas de acuerdo con esta evaluación.

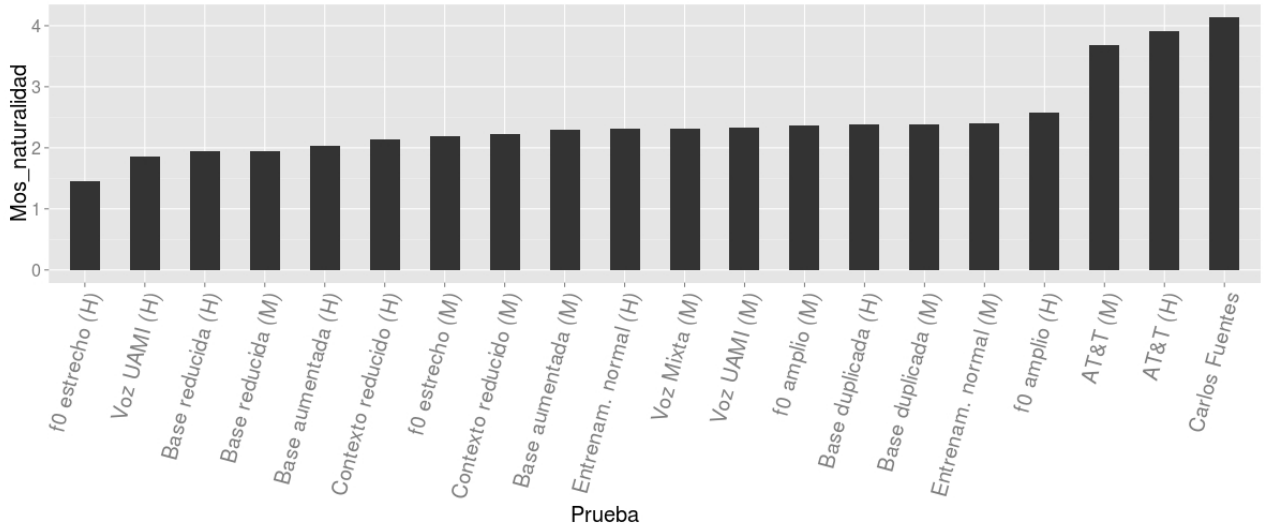


Figura 5.1: Ordenamiento de los resultados de pruebas subjetivas de naturalidad. Aplicación Clima

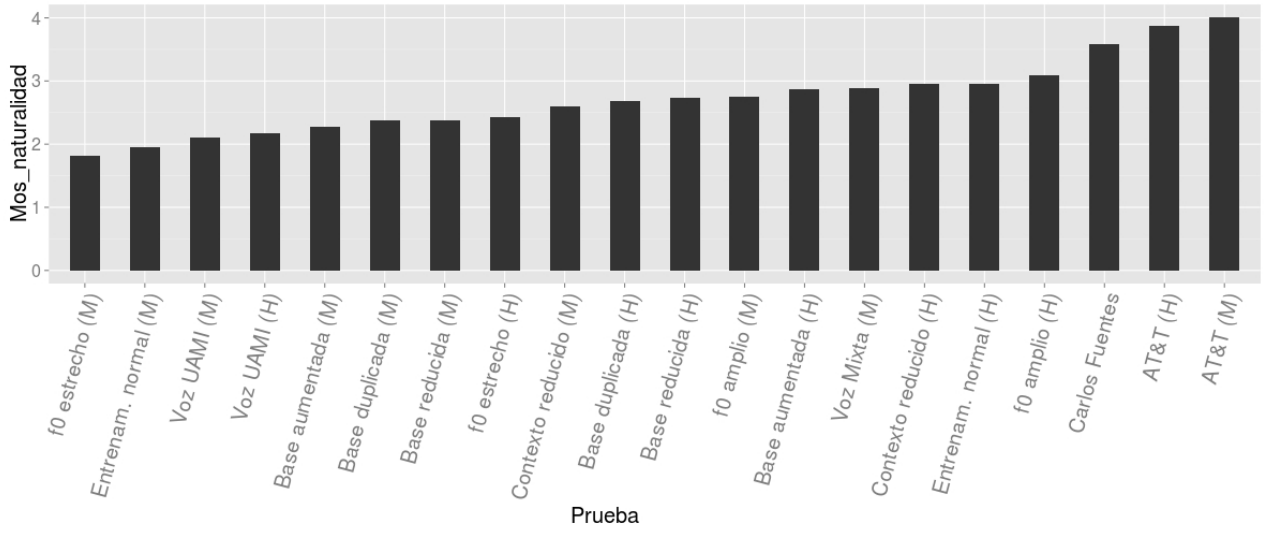


Figura 5.2: Ordenamiento de los resultados de pruebas subjetivas de naturalidad. Aplicación Reloj

El resultado que puede considerarse más dispar entre las evaluaciones subjetivas según aplicación es la voz de mujer obtenida con entrenamiento normal, la cual tiene una posición media en naturalidad e inteligibilidad en la aplicación Clima, mientras que es la menor

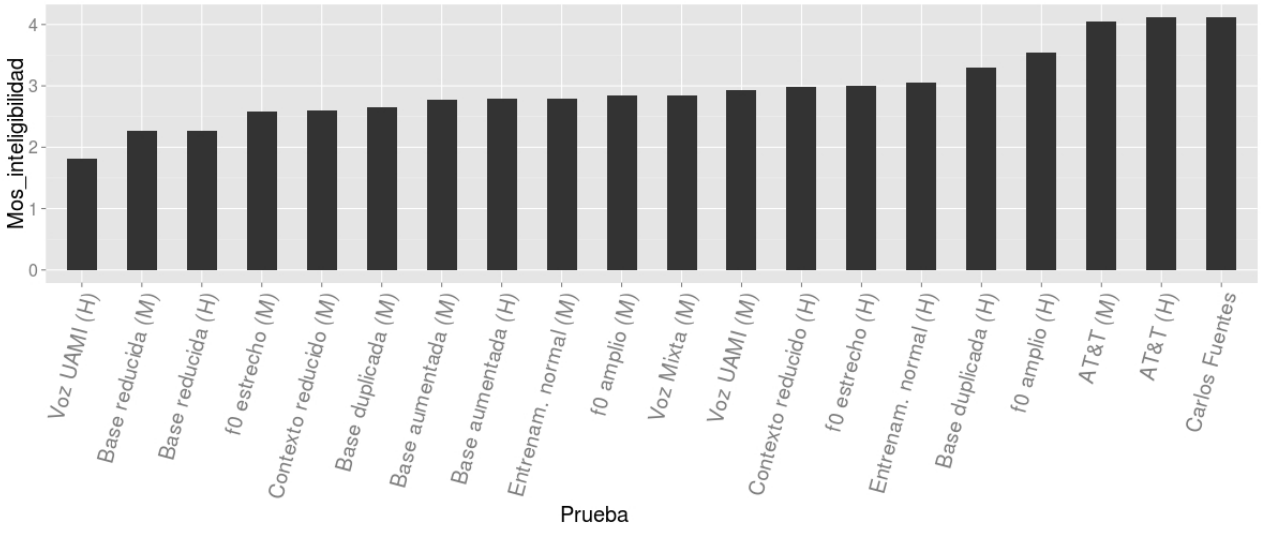


Figura 5.3: Ordenamiento de los resultados de pruebas subjetivas de inteligibilidad. Aplicación Clima

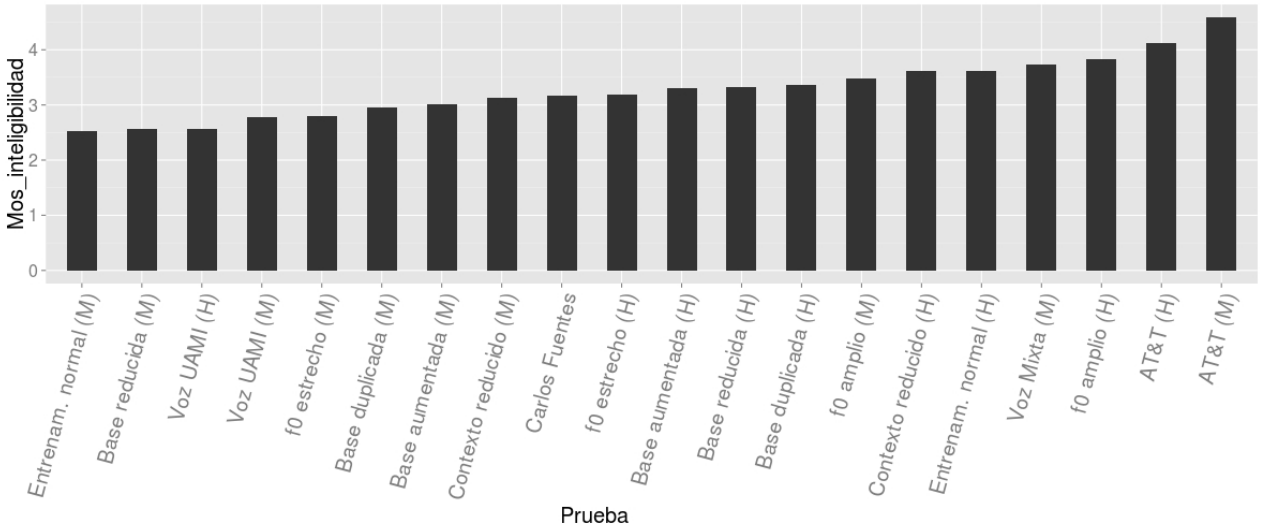


Figura 5.4: Ordenamiento de los resultados de pruebas subjetivas de inteligibilidad. Aplicación Reloj

evaluada en la aplicación Reloj. La explicación de este comportamiento en la evaluación de esta voz sintetizada puede considerarse un valor atípico entre las evaluaciones subjetivas, donde la tendencia es a obtener menores valores en MOS de las voces UAM-H, UAM-M y las obtenidas con $f0$ estrecho y base reducida.

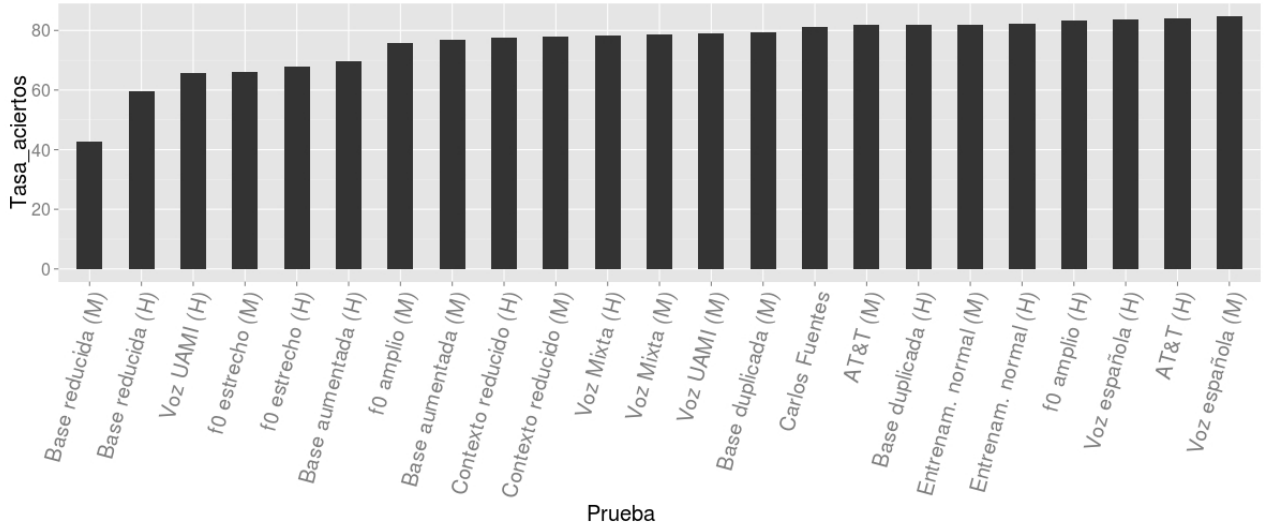


Figura 5.5: Ordenamiento de los resultados de tasa de aciertos en clasificador. Aplicación Clima

En cuanto a la tasa de aciertos de palabras reconocidas en el clasificador, en las Figuras 5.5 y 5.6 se muestran los resultados ordenados por este parámetro en ambas aplicaciones. Hay diferencias significativas entre ambas aplicaciones, al considerar las pruebas que obtienen mejores resultados.

Por ejemplo para la aplicación Clima, las voces obtenidas con las bases de datos de español de España destacan como las mejores, mientras que en Reloj son superadas por las voces obtenidas con entrenamiento normal con la base de datos completa. Las voces con mejores y peores resultados en clasificador varían de acuerdo con la aplicación y el género. Esto da un indicio también sobre las caracteristicas del reconocedor de palabras, el cual puede

tener una mayor sensibilidad ante el género de la voz, así como su acento.

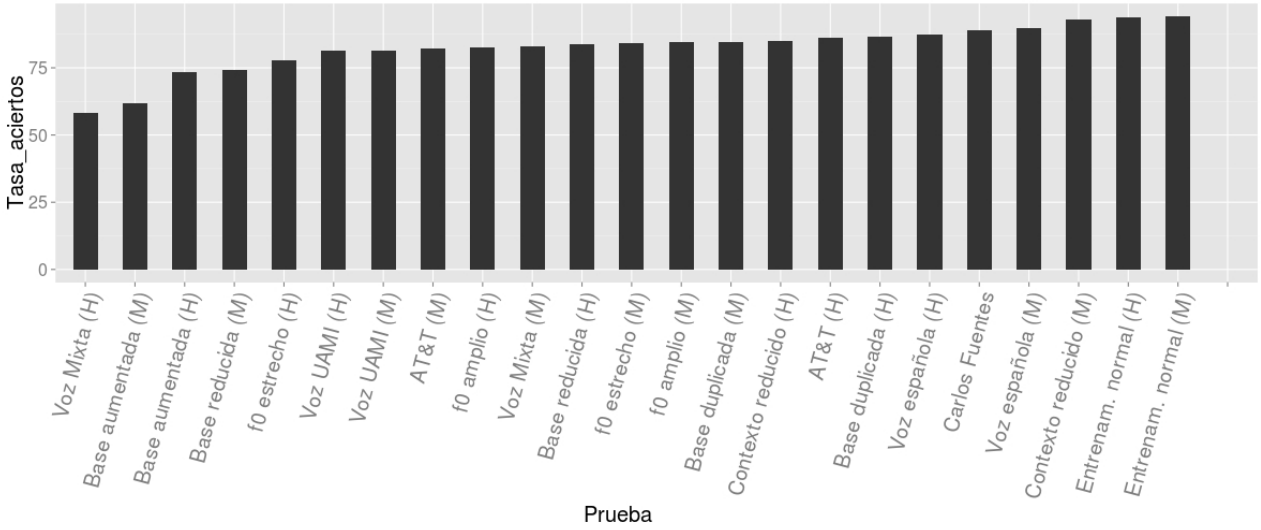


Figura 5.6: Ordenamiento de los resultados de tasa de aciertos en clasificador. Aplicación Reloj

Para determinar si las diferencias significativas en los parámetros acústicos de tono, *jitter* y *shimmer* tienen relación con las otras evaluaciones, en la Tabla 5.3 y 5.4 se indica cuáles pruebas de Friedman han tenido como resultado esa diferencia de acuerdo con la aplicación. En la Figura 5.7 y 5.8 se muestra la relación entre las evaluaciones realizadas, y se distinguen los puntos correspondientes a los experimentos cuyo resultado ha tenido diferencia significativa con el hablante original en los tres parámetros acústicos.

En el gráfico que relaciona las diferentes evaluaciones realizadas es posible observar una tendencia hacia la linealidad entre las evaluaciones subjetivas, y entre la tasa de aciertos y las evaluaciones subjetivas en ambos casos. Esto es muestra de consistencia entre estas mediciones, no así con la similitud de parámetros SLL. También se destaca que los casos cuyas evaluaciones objetivas y subjetivas tienen los valores más bajos también presentan diferencias estadísticamente significativas con el hablante original en las tres medidas acústicas de forma

Tabla 5.3: Diferencias estadísticamente significativas de parámetros acústicos con el hablante original de acuerdo con prueba de Friedman. Aplicación Clima. M: Voz femenina, H: Voz masculina

Condición de entrenamiento	Diferencia significativa		
	Tono	*Jitter*	*Shimmer*
Entrenamiento normal (H)	✓	✓	
Entrenamiento normal (M)	✓	✓	✓
$f0$ estrecho (H)	✓	✓	✓
$f0$ estrecho (M)	✓	✓	✓
$f0$ amplio (H)		✓	
$f0$ amplio (M)		✓	
Contexto reducido (H)		✓	✓
Contexto reducido (M)	✓	✓	✓
Base reducida (H)		✓	
Base reducida (M)	✓	✓	✓
Base duplicada (H)			
Base duplicada (M)	✓	✓	✓
Base aumentada (H)	✓		✓
Base aumentada (M)	✓	✓	✓

simultánea (puntos rojos), mientras que las voces con las mejores evaluaciones no presentan estas diferencias de forma simultánea.

5.3. Análisis de correlación

Finalmente, se presenta un análisis de correlación entre las distintas evaluaciones aplicadas, con la finalidad de proponer cuáles de éstas pueden ser sujeto de reemplazo por alguna de las otras en etapas de futuras experimentaciones, o bien si es factible la aplicación de técnicas de predicción de estas evaluaciones, dado que se realizó una gran cantidad de experimentos. En la Figura 5.9 y 5.10 se muestra de forma gráfica un índice de correlación de las cuatro evaluaciones realizadas sobre las voces de hombre y de mujer en la aplicación Reloj, respectivamente. En estas figuras se indica con tonos de azul una mayor correlación

Tabla 5.4: Diferencias estadísticamente significativas de parámetros acústicos con el hablante original de acuerdo con prueba de Friedman. Aplicación Reloj. M: Voz femenina, H: Voz masculina

Condición de entrenamiento	Diferencia significativa		
	Tono	*Jitter*	*Shimmer*
Entrenamiento normal (H)		✓	
Entrenamiento normal (M)	✓		✓
$f0$ estrecho (H)	✓	✓	✓
$f0$ estrecho (M)	✓	✓	✓
$f0$ amplio (H)			
$f0$ amplio (M)	✓		
Contexto reducido(H)		✓	✓
Contexto reducido (M)	✓	✓	✓
Base reducida (H)		✓	
Base reducida (M)	✓		✓
Base duplicada (H)			✓
Base duplicada (M)			✓
Base aumentada (H)			✓
Base aumentada (M)		✓	✓

entre las variables, y con rojo una correlación en sentido inverso.

Tanto en la voz masculina como en la femenina, se observa una alta correlación entre la inteligibilidad subjetiva y la tasa de aciertos del clasificador. En el caso de la voz femenina, la correlación es aún más alta con la evaluación subjetiva de naturalidad, pero disminuye en el caso de voz masculina.

No se puede identificar un patrón claro entre los resultados de similitud SLL y las evaluaciones subjetivas u objetiva de tasa de error. Con esto puede establecerse que se requiere otra combinación de parámetros o medidas de similitud que estén más acordes con las evaluaciones subjetivas.

La correlación entre evaluaciones es muy distinta en la aplicación Clima. En las Figuras 5.11 y 5.12 se muestra de forma gráfica el valor del índice de correlación. Se puede establecer que en esta aplicación sí hay una correlación entre la similitud SLL y la inteligibi-

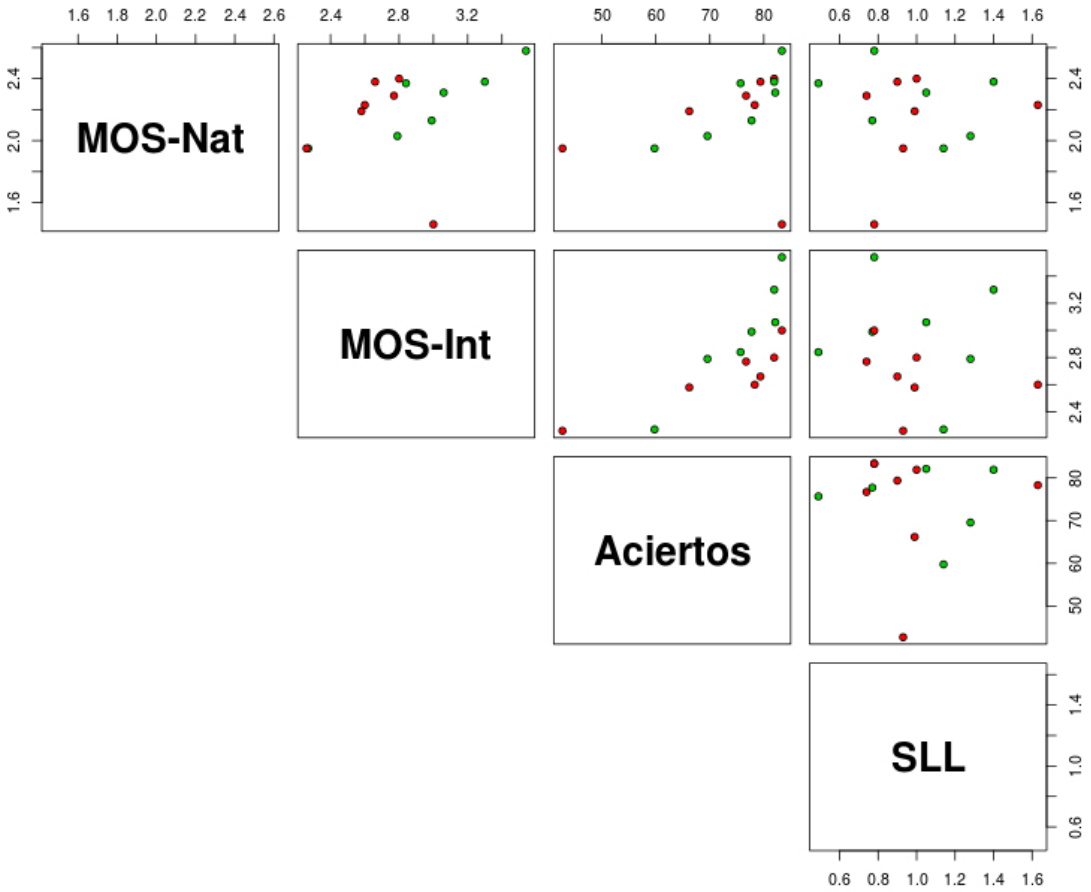


Figura 5.7: Relación entre evaluaciones subjetivas y objetivas. Aplicación Clima

lidad, especialmente en el caso de la voz masculina. Para la voz femenina los resultados son muy dispares, lo cual puede ser un indicio de que las evaluaciones subjetivas no han sido del todo fiables para este género de voz en esta aplicación.

Por último, en la Figura 5.13 se muestra de forma gráfica la correlación entre las evaluaciones aplicadas en la totalidad de experimentos, considerando ambos géneros y ambas aplicaciones. Se observa una alta correlación entre el valor de tasa de aciertos y la evaluación subjetiva de inteligibilidad, y en menor medida con la naturalidad. La similitud SLL no tiene

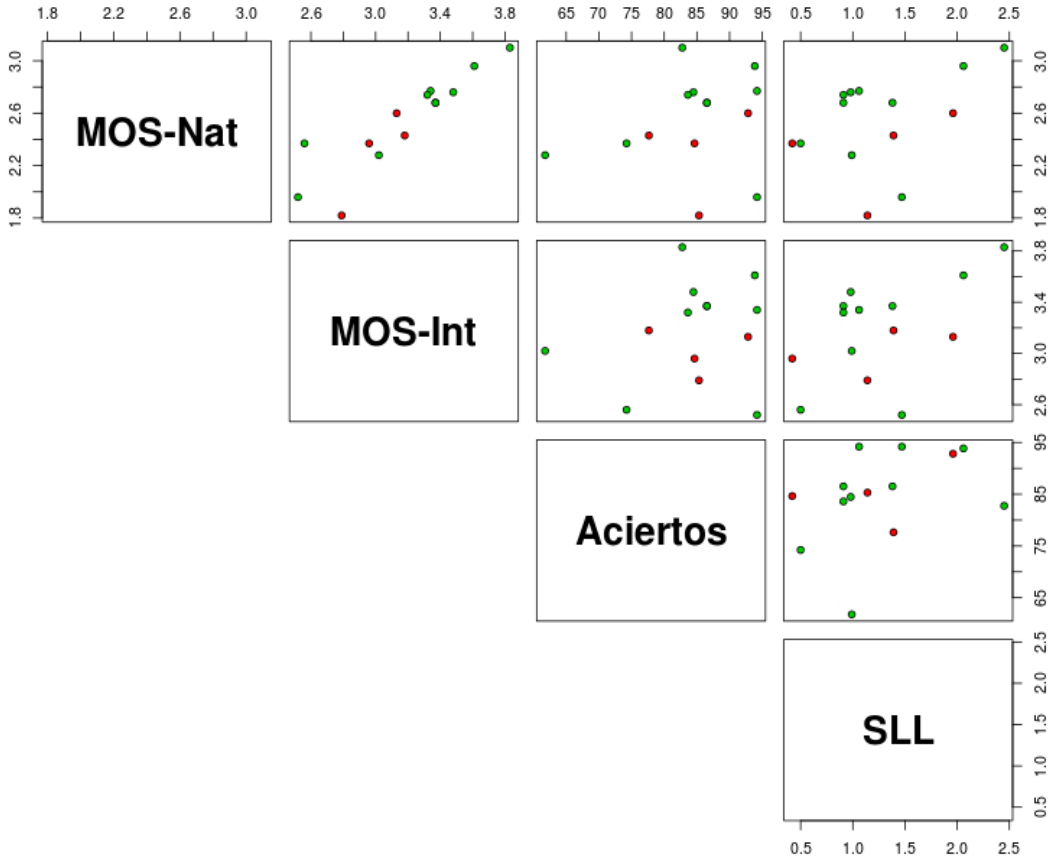


Figura 5.8: Relación entre evaluaciones subjetivas y objetivas. Aplicación Reloj

una correlación tan fuerte con ninguna de las medidas subjetivas.

5.4. Discusión

La especificación de las unidades de habla (palabras, sílabas, fonemas) a partir del texto es de importancia pues cada fonema en su contexto será representado por un HMM en el entrenamiento y posteriormente en la síntesis de nuevas frases. La implementación de

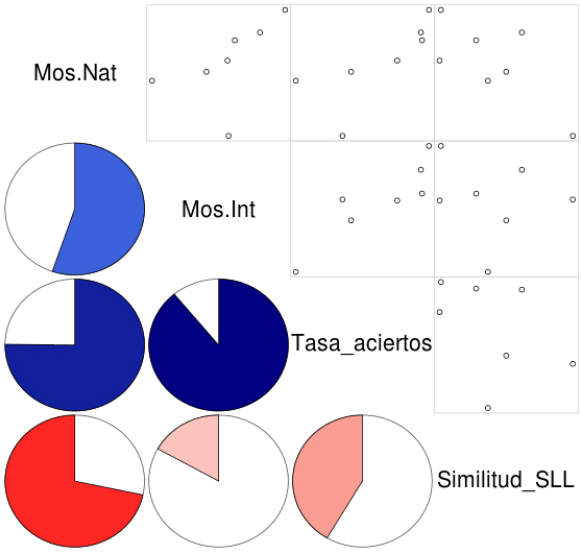


Figura 5.9: Índice de correlación entre evaluaciones de voces masculinas, aplicación Reloj

nuevos idiomas o sus variantes debe considerar este proceso para establecer con precisión la correspondencia entre grafema y fonema, la silibificación y la acentuación, con el fin de definir los contextos que mejor representen la prosodia del habla.

Estos aspectos prosódicos se derivan de la posición de los fonemas en la sílaba, la palabra y la frase, los cuales son agrupados en el entrenamiento por árboles de decisión, y establecidos en el proceso de síntesis utilizando estas mismas estructuras. Por esta razón no se especifican explícitamente en los modelos.

En este trabajo se ha propuesto una adaptación de la codificación fonética SAMPA para el español de México, a partir de la codificación del español de España, considerando los sonidos característicos de esta variante del español. La diferencia principal se da en los fonemas “s”, “z”, “c” (seguido de “e”,“i”); en los cuales no se hace distinción en el español de México. También existe diferencia en la pronunciación del grafema “x”, que tiene diversidad

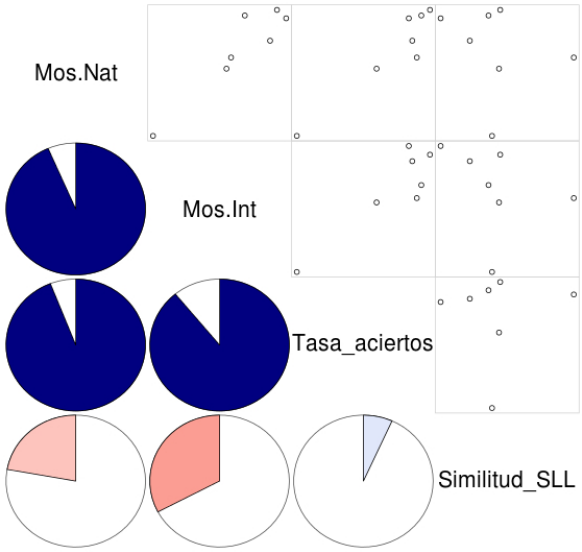


Figura 5.10: Índice de correlación entre evaluaciones de voces femeninas, aplicación Reloj

de pronunciaciones en México, pero no se ha considerado pues las bases de datos utilizadas no cuentan con este elemento. En futuras experimentaciones es conveniente crear datos que lo contemplen y establecer reglas para su pronunciación.

El el Capítulo 4 sen mostrarlo los resultados de una experimentación en varios niveles para la creación de voces con síntesis estadística paramétrica basada en HMM, la cual ha pretendido responder una serie de preguntas con respecto a la técnica en sí, y a la mejora de voces con los datos disponibles. En total se han tomado cuarenta y dos voces resultantes de los diversos experimentos para realizar evaluaciones y análisis comparativos.

A pesar de la calidad de datos con que se ha contado en bases de datos para la creación de voces, la cantidad de éstos en cuanto a duración ha llevado a la creación de aplicaciones con dominio reducido. De esta manera es posible limitar las posibilidades de palabras y longitud de frases, para evaluar de forma consistente con respecto a estas posibilidades.

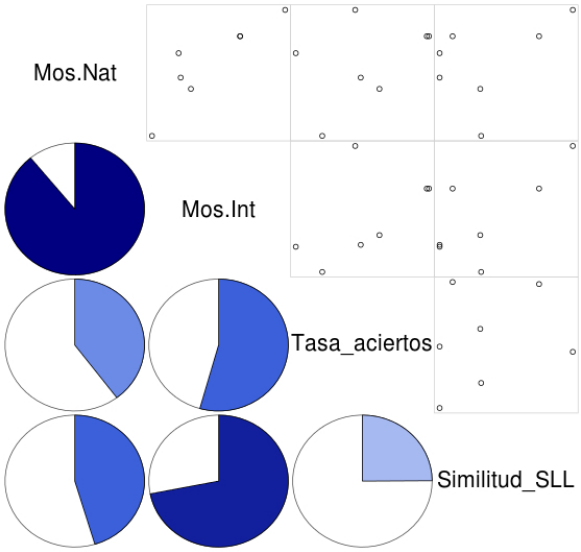


Figura 5.11: Índice de correlación entre evaluaciones de voces masculinas, aplicación Clima

Para determinar el parámetro de entrenamiento llamado rango de $f0$, se ha desarrollado una rutina en el programa Praat para determinar su valor máximo y mínimo de cada vocal en la base de datos, y con estos valores de vocales establecer uno general máximo y mínimo para cada hablante. Dado que en algunas evaluaciones la voz obtenida con un rango más amplio de $f0$ ha superado la voz con el rango adecuado, es conveniente que se trabaje con un margen mayor al extraído a partir de las vocales, o bien incluir en este cálculo los demás fonemas sonoros.

Se ha mostrado una mayor sensibilidad en la calidad de los resultados con esta técnica a la cantidad de datos disponibles, y a la definición del parámetros de entrenamiento relacionado con $f0$. Por ejemplo, la voz mejor evaluada en las pruebas subjetivas ha sido la que contó con mayor duración de su base de datos, independientemente de que la calidad de las grabaciones ha sido inferior a otras de las voces.

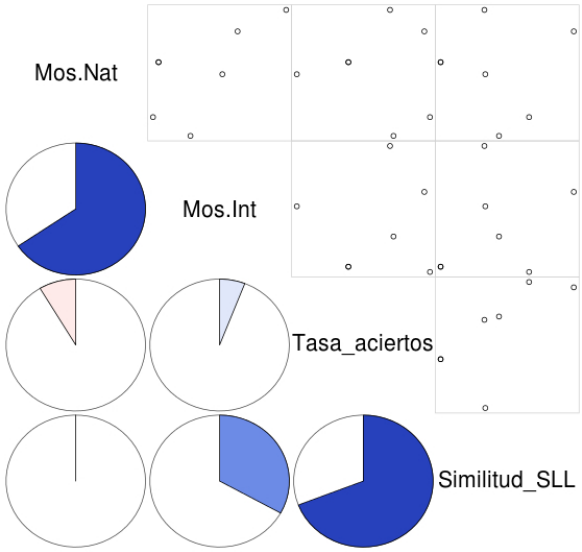


Figura 5.12: Índice de correlación entre evaluaciones de voces femeninas, aplicación Clima

Por otra parte, la voz con menor resultado en evaluación subjetiva y en general en las objetivas ha sido la producida con un ajuste muy estrecho de rango de $f0$, lo cual provoca que se pierdan valores de este parámetros en segmentos de habla, lo cual no afecta tanto su inteligibilidad como su naturalidad.

Los procedimientos realizados para aumentar el tamaño de la base de datos no han producido mejoras generales (considerando género y aplicación) en ninguna de las evaluaciones, por lo que no se consideran procedimientos adecuados para mejorar voces, con relación a aumentar la cantidad de datos con grabaciones reales. Como un elemento de comparación, la voz obtenida con una mezcla de dos grabaciones con emociones diferentes sí mejora la percepción subjetiva de calidad resultante, en comparación con la duplicación de la base de datos con una sola emoción neutral.

La identificación de las frases con contenido fonético directamente relacionado con las

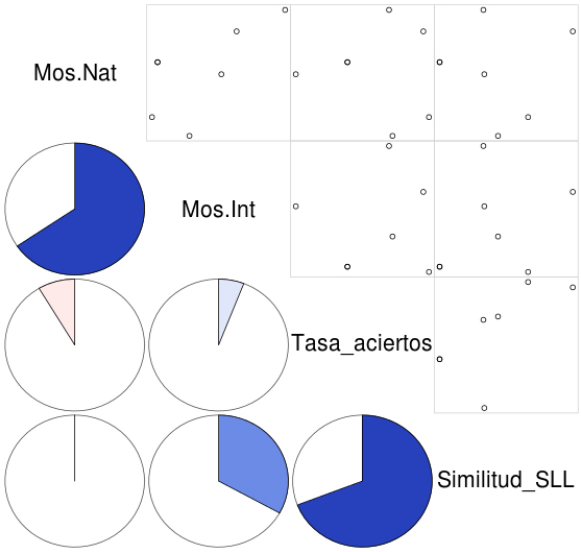


Figura 5.13: Índice de correlación entre evaluaciones de todas las pruebas realizadas

frases a sintetizar (base reducida) ha tenido resultados aceptables en la aplicación Reloj, en la cual la longitud de las frases es corta y constante. Esto a pesar de haberse reducido la base de datos a solamente tres minutos. En frases más largas, como en la aplicación Clima, el procedimiento no parece conveniente pues las evaluaciones son muy inferiores a otros resultados. A pesar de esto, la producción de una voz inteligible con tan pocos datos debe analizarse también desde el punto de vista de un requisito mínimo de contenido fonético que debe tener una base de datos para generar determinadas frases, lo cual es útil para generar voces cuando se cuenta con muy pocos datos.

La reducción del contexto ha producido buenos resultados en las distintas evaluaciones, especialmente en la aplicación Reloj. Esto puede llevar a profundizar en la definición de los contextos a utilizar de acuerdo con las frases que se desean sintetizar y las características de los datos disponibles. Además de lograr inteligibilidad en las frases a partir de pocos datos,

se reduce el tiempo de entrenamiento en la creación de voces y la complejidad de los modelos involucrados en el proceso.

La voz obtenida con la mayor cantidad de datos disponibles, aproximadamente una hora de grabaciones (Carlos Fuentes) ha tenido los mejores resultados en evaluación subjetiva, a pesar de no tener la mejor calidad en el audio original con respecto a las otras bases de datos. Esto muestra la alta sensibilidad de la técnica a la cantidad de datos, y en evaluación subjetiva una dependencia no tan marcada con respecto a la cuantificación o la compresión realidad en el audio original.

A partir del estudio de referencias, se incorporaron los métodos de evaluación subjetivos usuales para evaluar los resultados de las voces sintetizadas. Se definieron formatos y procedimientos para abarcar de forma eficiente la evaluación subjetiva de cuarenta resultados, con la participación de más de ochenta personas voluntarias.

Como parte de este trabajo se propuso el uso extensivo de un reconocedor de palabras comercial como un clasificador de las mismas, sobre el cual se definió una tasa de aciertos como elemento de comparación de las distintas voces resultantes. Para esto se definió un procedimiento y se desarrolló un programa que evalúa de forma automática una cantidad considerable de frases sintetizadas para hacer los resultados significativos.

La propuesta de incorporación de tres parámetros acústicos a la evaluación de voces sintetizadas tomando como referencia al hablante original (tono, *jitter* y *shimmer*) ha mostrado una correlación positiva de éstos con la calidad de voces a partir de evaluaciones subjetivas y objetivas.

La calidad de las voces, de acuerdo con las diversas evaluaciones, depende del hablante y de las características de las frases, especialmente su longitud y diversidad de contenido fonético, por lo que para el desarrollo de futuras aplicaciones se debe tomar en cuenta el tipo

de frases disponibles y que se desean sintetizar. Esto con la finalidad de decidir sobre los elementos de contexto y el manejo de la base de datos para mejorar los resultados.

Capítulo 6

Conclusiones y recomendaciones para trabajo futuro

Las conclusiones de este trabajo se agruparán en cinco partes, de acuerdo con las distintas áreas sobre las que se han planteado los estudios teóricos, prácticos, y los aportes realizados al desarrollo de la síntesis estadística paramétrica en su documentación, experimentación a varios niveles y evaluación.

1. Conclusiones sobre aspectos fonéticos, prosódicos y lingüísticos del español adaptados para el desarrollo de voces sintetizadas:

    - Para la implementación de voces utilizando técnicas estadísticas paramétricas es necesario un conocimiento básico de aspectos fonéticos del idioma o variante a implementar, para poder establecer una correspondencia entre los grafemas y los sonidos que representan en el habla.
    - La diferencia entre la codificación fonética para el español de México, con relación al español de España, se ha establecido en los fonemas s, z, c (seguido de e,i); en los cuales no se hace distinción en el español de México.

- La utilización de contextos y árboles de decisión definidos a partir de la posición y características de los fonemas han mostrado aspectos prosódicos en la síntesis semejantes a la forma de pronunciar del hablante original en la base de datos.
- En el caso de aplicaciones de dominio restringido, es conveniente realizar un estudio del contenido fonético de la base de datos y de las frases que correspondan en ese dominio, para establecer niveles de certeza sobre capacidad de sintetizar de las frases deseadas con los datos disponibles.

2. Conclusiones sobre los HMM y su aplicación en la síntesis de voz

   - A pesar de haber incorporado variantes de este modelo matemático para buscar mejores resultados, los HMM con distribuciones de probabilidad multi-estado, como los utilizados en este trabajo, siguen siendo el modelo dominante en la síntesis estadística paramétrica.
   - Las tres principales oportunidades de mejora de la técnica de síntesis basada en HMM son: El modelado matemático, la información utilizada para representar la voz y la reconstrucción de la señal a parte de esta representación. En este trabajo se ha hecho énfasis en el manejo del modelo matemático para mejorar la calidad de voces resultantes. Es posible que no se considere las voces producidas con síntesis estadística paramétrica de la calidad de los mejores sintetizadores comerciales hasta que se mejore la reconstrucción de la señal con más parámetros.
   - Como parte de este trabajo se ha realizado un aporte a la documentación existente sobre la técnica de síntesis estadística paramétrica basada en HMM, la cual incluye configuración y ajuste de acuerdo con el lenguaje a implementar.

3. Conclusiones referentes a la implementación de nuevas voces en el sistema HTS:

- HTS es el único sistema disponible para la creación, entrenamiento y generación de parámetros basados en HMM que se encuentra disponible para implementar un sistema de síntesis estadística paramétrica de voz.
- Debido a lo reciente de la técnica, su implementación y los diversos procesos involucrados, no se cuenta con documentación oficial del sistema HTS y el grupo de programas que lo complementan, los aspectos teóricos y los aspectos prácticos de su puesta en funcionamiento.
- En el desarrollo de este trabajo se han creado diversos programas para facilitar tareas rutinarias en el proceso de creación de nuevas voces, tales como el análisis de audio para determinar parámetros importantes en el entrenamiento, así como la evaluación usando un reconocedor de voz.
- Se ha realizado un aporte a la documentación sobre el HTS, que incluye configuración y ajustes necesario para crear nuevos lenguajes o sus variantes, lo cual complementa el desarrollo teórico presentado en este trabajo para constituir una unidad sobre el estudio de la síntesis estadística paramétrica.

4. Conclusiones referentes a la experimentación sobre diversos parámetros y condiciones de entrenamiento:

   - En total se realizaron cuarenta y dos experimentos para producir voces, los cuales han tenido diferentes niveles de calidad de acuerdo con las evaluaciones establecidas.
   - Cuando la cantidad de datos disponibles para crear voces es reducida, conviene la creación de aplicaciones con dominio reducido para enmarcar las posibilidades de frases a emitir, y permitir de esta manera su evaluación.

- Los resultados de las voces con $f0$ más amplio han mejorado algunas de las evaluaciones realizadas, por lo que se considera recomendable en futuras implementaciones considerar un margen amplio a los valores de este parámetros analizados en la base de datos.

- La voz mejor evaluada en las pruebas subjetivas ha sido la que contó con mayor duración de su base de datos, y la peor evaluada la que ha tenido un ajuste muy estrecho de rango de $f0$.

- Los procedimientos realizados para aumentar el tamaño de la base de datos no se consideran adecuados para mejorar las voces, con relación a aumentar la cantidad de datos con grabaciones reales.

- La identificación de las frases con contenido fonético directamente relacionado con las frases a sintetizar (base reducida) ha tenido resultados aceptables en frases cortas. En frases más largas el procedimiento no parece conveniente pues las evaluaciones son muy inferiores a otros resultados.

- Cuando las frases que se desean sintetizar son de longitud reducida y constante, la reducción del contexto modelado en el texto ha producido buenos resultados en las distintas evaluaciones, lo cual es conveniente también por el menor tiempo de entrenamiento en la creación de voces y la complejidad de los modelos involucrados en el proceso.

- La voz obtenida con la mayor cantidad de datos disponibles (Carlos Fuentes) ha tenido los mejores resultados en evaluación subjetiva, y la hacen una voz comparable en calidad, dentro de las aplicaciones definidas, a voces comerciales ampliamente reconocidas.

5. Conclusiones sobre los métodos de evaluación

   - El reconocedor de palabras comercial utilizado ha mostrado su conveniencia como un clasificador, sobre el cual es posible definir una tasa de aciertos para comparar las distintas voces resultantes.
   - El estudio de correlación realizado muestra que en algunos casos podría sustituirse la evaluación subjetiva de inteligibilidad por el uso del reconocedor de palabras, al menos como un medio de discriminación previo de voces de mayor o menor calidad.
   - Se ha observado una correlación de los parámetros acústicos propuestos en la evaluación objetiva con la calidad de voces a partir de evaluaciones subjetivas y objetivas.
   - La calidad de las voces resultantes ha mostrado una dependencia del hablante original y de las características de las frases, especialmente su longitud y la complejidad de su contenido fonético.

En cuanto a trabajo futuro, se tienen las siguientes recomendaciones:

- Continuar con el estudio de la incorporación de parámetros acústicos a la evaluación de voces, por ejemplo otros tipos de *jitter*, o bien incorporar la extracción e incorporación de estos parámetros en otras unidades fonéticas o fragmentaciones del audio. Una correlación segura de distintos parámetros permitiría la automatización de las pruebas de creación de voces y su evaluación, incorporando diversas heurísticas que podrían mejorar los resultados al abarcar mayor cantidad de parámetros que puedan variarse de los distintos procesos y ajustarse de acuerdo con el hablante y las frases requeridas, automatizando la experimentación extensiva en la búsqueda de nuevos resultados.

- Evaluar otras técnicas de inteligencia computacional para la agrupación de HMM y el análisis de texto, además de los árboles de decisión que se utilizan en la actualidad. Conárboles los resultados han sido satisfactorios, pero tienen participación en el suavizamiento poco conveniente que sido atribuido a la técnica.

- Construir bases de datos de mayor tamaño en español, que permitan construir voces con nuevas técnicas estadísticas paramétricas, como la adaptación, lo cual podría llevar a contar con información suficiente en modelos promedio para requerir cada vez menor cantidad de datos en la creación de nuevas voces.

- Realizar síntesis basada en HMM sobre otras unidades fonéticas, como las sílabas, o bien una combinación de unidades de distinto tamaño.

- Utilizar la gran cantidad de resultados y evaluaciones sobre los mismos para aplicar métodos de predicción de evaluaciones subjetivas que puedan llevar a una evaluación previa de voces, para discriminarlas según su calidad, de manera que se reduzca la necesidad de contar con gran cantidad de escuchas humanos cuando se realice experimentaciones extensivas.

- Comprobar la utilidad de la síntesis en aplicaciones de interés reciente, taes como el rescate de lenguas con pocos hablantes. Como se ha comprobado que es posible sintetizar frases cortas con alta calidad utilizando contextos solamente fonéticos, se reduce los requerimientos sobre el conocimiento de estas lenguas para implementar los sistemas computacionales en síntesis estadística paramétrica.

- Implementar métodos de predicción sobre el contenido de una base de datos para pronunciar una frase particular que se desea sintetizar.

- Desarrollar interfaces gráficas para facilitar la creación de nuevas voces en futuras experiencias, así como procedimientos para hacer experimentaciones extensivas.

Referencias

[1] P. Taylor, *Text-to-speech synthesis*, vol. 15. Cambridge University Press Cambridge, 2009.

[2] S. Lemmetty, "Review of speech synthesis technology," *Helsinki University of Technology*, 1999.

[3] A. Black, "Perfect synthesis for all of the people all of the time," in *Speech Synthesis, 2002. Proceedings of 2002 IEEE Workshop on*, pp. 167–170, 2002.

[4] M. Raskind and E. Higgins, "Effects of speech synthesis on the proofreading efficiency of postsecondary students with learning disabilities," *Learning Disability Quarterly*, vol. 18, no. 2, pp. 141–158, 1995.

[5] J. Yamagishi, "New and emerging applications of speech synthesis," Presented at Speech synthesis seminar series 9th February, 2011.

[6] M. Perea, "The application of speech synthesis and speech recognition techniques in dialectal studies," *Anuario del Seminario de Filología Vasca "Julio de Urquijo"*, pp. 131–150, 2013.

[7] G. Rehm and H. Uszkoreit, *META-NET Strategic Research Agenda for Multilingual Europe 2020*. Springer Publishing Company, Incorporated, 2013.

[8] M. Schröder, "Emotional speech synthesis: A review," in *Proceedings of EUROSPEECH*, vol. 1, pp. 561–564, 2001.

[9] U. Zolzer, *DAFX: Digital Audio Effects*. Wiley Publishing, 2011.

[10] J. Holmes and W. Holmes, *Speech Synthesis and Recognition 2e (HBK)*. Taylor & Francis, 2002.

[11] X. Sun, "Voice quality conversion in td-psola speech synthesis," in *Acoustics, Speech, and Signal Processing, 2000. ICASSP'00. Proceedings. 2000 IEEE International Conference on*, vol. 2, pp. II953–II956, IEEE, 2000.

[12] D. Vine and R. Sahandi, "Synthesis of emotional speech using RP-PSOLA," 2000.

[13] H. Zen, K. Tokuda, and A. Black, "Statistical parametric speech synthesis," *Speech Communication*, vol. 51, no. 11, pp. 1039–1064, 2009.

[14] K. Tokuda, Y. Nankaku, T. Toda, H. Zen, J. Yamagishi, and K. Oura, "Speech synthesis based on Hidden Markov Models," 2013.

[15] J. Yamagishi, B. Usabaev, S. King, O. Watts, J. Dines, J. Tian, Y. Guan, R. Hu, K. Oura, Y. Wu, *et al.*, "Thousands of voices for hmm-based speech synthesis–analysis and application of tts systems built on various asr corpora," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 18, no. 5, pp. 984–1004, 2010.

[16] C. Bennett, "Large scale evaluation of corpus-based synthesizers: results and lessons from the blizzard challenge 2005.," in *INTERSPEECH*, pp. 105–108, 2005.

[17] X. Gonzalvo-Fructuoso, *HMM-based speech synthesis applied to Spanish and English, its applications and a hybrid approach.* PhD, Universitat Ramon Llull, 2010.

[18] A. Herrera-Camacho and F. Del Rio-Ávila, "Development of a Mexican Spanish Synthetic Voice Using Synthesizer Modules of Festival Speech and HTS-Straight.," *International Journal of Computer & Electrical Engineering*, vol. 5, no. 1, 2013.

[19] S. King, "An introduction to statistical parametric speech synthesis," *Sadhana*, vol. 36, no. 5, pp. 837–852, 2011.

[20] J. Yamagishi, *An Introduction to HMM-Based Speech Synthesis.* Technical report, Tokyo Institute of Technology, 2006.

[21] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Simultaneous modeling of spectrum, pitch and duration in hmm-based speech synthesis," 1999.

[22] L. Rabiner and R. W. Schafer, "Introduction to digital speech processing," *Foundations and trends in signal processing*, vol. 1, no. 1, pp. 1–194, 2007.

[23] S. Panchapagesan, "Frequency warping by linear transformation of standard MFCC.," in *Interspeech*, 2006.

[24] Praat, "*Praat: Doing Phonetics by Computer.*" [En línea]. Disponible en: `http://www.fon.hum.uva.nl/praat/`. [Consultado: 20 de abril de 2014].

[25] HTS Working Group, "*HMM-based Speech Synthesis System.*" [En línea]. Disponible en: `http://hts.sp.nitech.ac.jp/`. [Consultado: 25 de abril de 2014].

[26] D. Talkin, "A robust algorithm for pitch tracking (rapt)," *Speech coding and synthesis*, vol. 495, p. 518, 1995.

[27] K. Tokuda, T. Masuko, N. Miyazaki, and T. Kobayashi, "Hidden Markov Models based on multi-space probability distribution for pitch pattern modeling," in *Acoustics, Speech, and Signal Processing, 1999. Proceedings., 1999 IEEE International Conference on*, vol. 1, pp. 229–232, IEEE, 1999.

[28] S. J. Young, J. Odell, and P. Woodland, "Tree-based state tying for high accuracy acoustic modelling," in *Proceedings of the workshop on Human Language Technology*, pp. 307–312, Association for Computational Linguistics, 1994.

[29] K. Shinoda and T. Watanabe, "MDL-based context-dependent subword modeling for speech recognition," *Journal of Acoustic Society of Japan (E)*, vol. 21, no. 2, pp. 79–86, 2000.

[30] K. Shichiri, A. Sawabe, T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Eigenvoices for HMM-based speech synthesis.," in *INTERSPEECH*, 2002.

[31] K. Fujinaga, M. Nakai, H. Shimodaira, and S. Sagayama, "Multiple-regression Hidden Markov Model," in *Acoustics, Speech, and Signal Processing, 2001. Proceedings.(ICASSP'01). 2001 IEEE International Conference on*, vol. 1, pp. 513–516, IEEE, 2001.

[32] J. Yamagishi, Z. Ling, and S. King, "Robustness of HMM-based speech synthesis," 2008.

[33] Z. Heiga, T. Tomoki, M. Nakamura, and K. Tokuda, "Details of the Nitech HMM-based speech synthesis system for the Blizzard Challenge 2005," *IEICE transactions on information and systems*, vol. 90, no. 1, pp. 325–333, 2007.

[34] M. Marume, H. Zen, Y. Nankaku, K. Tokuda, and T. Kitamura, "An investigation of spectral parameters for HMM-based speech synthesis," in *Proc. Autumn Meeting of ASJ*, pp. 185–186, 2006.

[35] J. Dines and S. Sridharan, "Trainable speech synthesis with trended Hidden Markov Models," in *Acoustics, Speech, and Signal Processing, 2001. Proceedings.(ICASSP'01). 2001 IEEE International Conference on*, vol. 2, pp. 833–836, IEEE, 2001.

[36] H. Zen, K. Tokuda, and T. Kitamura, "Reformulating the HMM as a trajectory model by imposing explicit relationships between static and dynamic feature vector sequences," *Computer Speech & Language*, vol. 21, no. 1, pp. 153–173, 2007.

[37] T. Tomoki and K. Tokuda, "A speech parameter generation algorithm considering global variance for HMM-based speech synthesis," *IEICE TRANSACTIONS on Information and Systems*, vol. 90, no. 5, pp. 816–824, 2007.

[38] T. Masuko, K. Tokuda, T. Kobayashi, and S. Imai, "Speech synthesis using HMMs with dynamic features," in *Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings., 1996 IEEE International Conference on*, vol. 1, pp. 389–392, IEEE, 1996.

[39] K. Tokuda, T. Masuko, T. Yamada, T. Kobayashi, and S. Imai, "An algorithm for speech parameter generation from continuous mixture HMMs with dynamic features," 1995.

[40] R. Donovan, "Trainable speech synthesis," *Univ. Eng. Dept*, p. 164, 1996.

[41] R. Donovan and P. Woodland, "A Hidden Markov-Model-based trainable speech synthesizer," *Computer speech & language*, vol. 13, no. 3, pp. 223–241, 1999.

[42] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, "Speech parameter generation algorithms for HMM-based speech synthesis," in *Acoustics, Speech, and Signal Processing, 2000. ICASSP'00. Proceedings. 2000 IEEE International Conference on*, vol. 3, pp. 1315–1318, IEEE, 2000.

[43] K. Tokuda, H. Zen, and A. Black, "An HMM-based speech synthesis system applied to english," in *Speech Synthesis, 2002. Proceedings of 2002 IEEE Workshop on*, pp. 227–230, 2002.

[44] R. Maia, H. Zen, K. Tokuda, T. Kitamura, and F. Resende Jr, "Towards the development of a brazilian portuguese text-to-speech system based on HMM," in *Proc. of the European Conf. on Speech Communication and Technology (EUROSPEECH)*, pp. 2465–2468, 2003.

[45] B. Vesnicer and F. Mihelič, "Evaluation of the Slovenian HMM-based speech synthesis system," in *Text, Speech and Dialogue*, pp. 513–520, Springer, 2004.

[46] F. Hendessi, A. Ghayoori, and T. A. Gulliver, "A speech synthesizer for persian text using a neural network with a smooth ergodic HMM," *ACM Transactions on Asian Language Information Processing (TALIP)*, vol. 4, no. 1, pp. 38–52, 2005.

[47] M. Barros, R. Maia, K. Tokuda, F. Resende, and D. Freitas, "HMM-based European Portuguese TTS system," in *Ninth European Conference on Speech Communication and Technology*, 2005.

[48] Y. Qian, F. Soong, Y. Chen, and M. Chu, "An HMM-based mandarin chinese text-to-speech system," in *Chinese Spoken Language Processing*, pp. 223–232, Springer, 2006.

[49] J. Zhu and J. Li, "An HMM-based approach to automatic phrasing for mandarin text-to-speech synthesis," in *Proceedings of the COLING/ACL on Main conference poster sessions*, pp. 977–982, 2006.

[50] I. Ipsic and S. Martincic-Ipsic, "Croatian HMM-based speech synthesis," *Journal of Computing and Information Technology*, vol. 14, no. 4, pp. 307–313, 2006.

[51] S. Kim, J. Kim, and M. Hahn, "HMM-based korean speech synthesis system for hand-held devices," *Consumer Electronics, IEEE Transactions on*, vol. 52, no. 4, pp. 1384–1390, 2006.

[52] O. Abdel-Hamid, S. Abdou, and M. Rashwan, "Improving Arabic HMM-based Speech Synthesis Quality.," in *INTERSPEECH*, 2006.

[53] X. Gonzalvo, J. Socoró, I. Iriondo, C. Monzo, and E. Martínez, "Linguistic and mixed excitation improvements on a HMM-based speech synthesis for castilian spanish," in *Proceedings of the 6th ISCA Workshop on Speech Synthesis (SSW-6)*, pp. 362–367, 2007.

[54] M. Pucher, D. Schabus, J. Yamagishi, F. Neubarth, and V. Strom, "Modeling and interpolation of austrian german and viennese dialect in HMM-based speech synthesis," *Speech Communication*, vol. 52, no. 2, pp. 164–179, 2010.

[55] S. Krstulovic, A. Hunecke, and M. Schröder, "An HMM-based speech synthesis system applied to german and its adaptation to a limited set of expressive football announcements," in *proc. of Interspeech*, vol. 7, 2007.

[56] S. Karabetsos, P. Tsiakoulis, A. Chalamandaris, and S. Raptis, "HMM-based speech synthesis for the greek language," in *Text, Speech and Dialogue*, pp. 349–356, 2008.

[57] A. Bonafonte-Cávez, I. Esquerra-Llucià, L. Aguilar, S. H. Oller-Martínez, M. A. Moreno-Bilbao, *et al.*, "Recent work on the FESTCAT database for speech synthesis," 2010.

[58] Y. Li, S. Pan, and J. Tao, "HMM-based speech synthesis with a flexible mandarin stress adaptation model," in *Signal Processing (ICSP), 2010 IEEE 10th International Conference on*, pp. 625–628, 2010.

[59] P. Lanchantin, G. Degottex, and X. Rodet, "A HMM-based speech synthesis system using a new glottal source and vocal-tract separation method," in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, pp. 4630–4633, 2010.

[60] D. Erro, I. Sainz, I. Luengo, I. Odriozola, J. Sánchez, I. Saratxaga, E. Navas, and I. Hernáez, "HMM-based speech synthesis in basque language using HTS," *Proc. FALA*, 2010.

[61] Z. Hanzlíček, "Czech HMM-based speech synthesis," in *Text, Speech and Dialogue*, pp. 291–298, Springer, 2010.

[62] N. Baloyi and M. J. D. Manamela, "An HMM-based Text-to-Speech Synthesis System for Xitsonga,"

[63] A. Stan, J. Yamagishi, S. King, and M. Aylett, "The romanian speech synthesis (RSS) corpus: building a high quality HMM-based speech synthesis system using a high sampling rate," *Speech Communication*, vol. 53, no. 3, pp. 442–450, 2011.

[64] G. Lu, Y. Hongzhi, Z. Jinshuang, and F. Huaping, "Research on HMM_based speech synthesis for lhasa dialect," pp. 429–433, IEEE, Oct. 2011.

[65] B. Bollepalli, J. Beskow, and J. Gustafson, "HMM based speech synthesis system for Swedish Language," in *In The Fourth Swedish Language Technology Conference*, 2012.

[66] R. Boothalingam, V. Sherlin Solomi, A. R. Gladston, S. L. Christina, P. Vijayalakshmi, N. Thangavelu, and H. A. Murthy, "Development and Evaluation of Unit Selection and HMM-Based Speech Synthesis Systems for Tamil," in *In Communications (NCC), 2013 National Conference on, IEEE*, pp. 1–5, IEEE, 2013.

[67] A. Alonso, I. Sainz, D. Erro, E. Navas, and I. Hernaez, "Sistema de Conversión Texto a Voz de Código Abierto Para Lenguas Ibéricas," *Procesamiento del Lenguaje Natural*, vol. 51, pp. 169–175, 2013.

[68] S. Phan, T. Vu, C. Duong, and M. Luong, "A study in Vietnamese statistical parametric speech synthesis base on HMM," *International Journal*, vol. 2, no. 1, 2013.

[69] M. Shannon, H. Zen, and W. Byrne, "Autoregressive models for statistical parametric speech synthesis," 2013.

[70] Z. H. Ling, L. Deng, and D. Yu, "Modeling Spectral Envelopes using Restricted Boltzmann Machines for Statistical Parametric Speech Synthesis.," 2013.

[71] R. Maia, M. Akamine, and M. Gales, "Complex cepstrum for statistical parametric speech synthesis," 2013.

[72] J. Ni, Y. Shiga, H. Kawai, and H. Kashioka, "Experiments on unsupervised statistical parametric speech synthesis," in *In Chinese Spoken Language Processing (ISCSLP), 2012 8th International Symposium on*, pp. 155–159, IEEE, 2012.

[73] T. Raitio, A. Suni, M. Vainio, and P. Alku, "Analysis of HMM-Based Lombard Speech Synthesis.," in *INTERSPEECH*, pp. 2781–2784, 2011.

[74] H. Zen, M. Gales, Y. Nankaku, and K. Tokuda, "Product of experts for statistical parametric speech synthesis," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, no. 3, pp. 794–805, 2012.

[75] H. Zen, A. Senior, and M. Schuster, "Statistical parametric speech synthesis using deep neural networks," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pp. 7962–7966, IEEE, 2013.

[76] Machine Intelligence Laboratory, Cambridge University Engineering Department, "*HTK*." [En línea]. Disponible en: `http://htk.eng.cam.ac.uk/`. [Consultado: 25 de abril de 2014].

[77] European Language Resources Association, "*ELRA catalogue. Emotional speech synthesis database, catalogue reference: ELRA-S0329*." [En línea]. Disponible en: `http://catalog.elra.info`. [Consultado: 20 de abril de 2014].

[78] F. Martínez-Licona, J. Goddard, A. Martínez-Licona, and M. Coto-Jimńez, "Acoustic analysis of spanish vowels in emotional speech," in *8th International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications, MAVEBA*, pp. 334–339, 2013.

[79] P. Lanchantin, G. Degottex, and X. Rodet, "A HMM-based speech synthesis system using a new glottal source and vocal-tract separation method," in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, pp. 4630–4633, IEEE, 2010.

[80] Y. Zhao, D. Peng, L. Wang, M. Chu, Y. Chen, P. Yu, and J. Guo, "Constructing stylistic synthesis databases from audio books.," in *INTERSPEECH*, 2006.

[81] V. Heuven, R. Bezooijen, *et al.*, "Quality evaluation of synthesized speech.," 1995.

[82] P. Thanh-Son, "Improvement of prosodic characteristic in vietnamese speech synthesis system based on HMM,"

[83] U. Remes, R. Karhila, and M. Kurimo, "Objective evaluation measures for speaker-adaptive HMM-TTS systems," *Proc. SSW, to appear.*

[84] B. Granström, "Towards a virtual language tutor," in *InSTIL/ICALL Symposium 2004*, 2004.

[85] C. Cucchiarini, H. Strik, and L. Boves, "Quantitative assessment of second language learners,Äô fluency by means of automatic speech recognition technology," *The Journal of the Acoustical Society of America*, vol. 107, no. 2, pp. 989–999, 2000.

[86] J. Kormos and M. Dénes, "Exploring measures and perceptions of fluency in the speech of second language learners," *System*, vol. 32, no. 2, pp. 145–164, 2004.

[87] K. Precoda, C. Halverson, and H. Franco, "Effects of speech recognition-based pronunciation feedback on second-language pronunciation ability," *Proceedings of InSTILL 2000*, pp. 102–105, 2000.

[88] W. Menzel, D. Herron, P. Bonaventura, and R. Morton, "Automatic detection and correction of non-native english pronunciations," *Proceedings of INSTILL*, pp. 49–56, 2000.

[89] H. Franco, L. Neumeyer, V. Digalakis, and O. Ronen, "Combination of machine scores for automatic grading of pronunciation quality," *Speech Communication*, vol. 30, no. 2, pp. 121–130, 2000.

[90] J. Arias, N. Yoma, and H. Vivanco, "Automatic intonation assessment for computer aided language learning," *Speech communication*, vol. 52, no. 3, pp. 254–267, 2010.

[91] Y. Chang, "Evaluation of tts systems in intelligibility and comprehension tasks," in *Proceedings of the 23rd Conference on Computational Linguistics and Speech Processing*, pp. 64–78, Association for Computational Linguistics, 2011.

[92] University College, London, "*SAMPA for Spanish.*" [En línea]. Disponible en: `http://www.phon.ucl.ac.uk/home/sampa/spanish.htm`. [Consultado: 20 de abril de 2014].

[93] J. Cuétara, *Fonética de la Ciudad de México. Aportaciones desde las tecnologías del habla.* Tesis de maestría inédita, México: Universidad Nacional Autónoma de México, 2005.

[94] C. D. Hernández-Mena and A. Herrera-Camacho, "Ciempiess: A new open-sourced mexican spanish radio corpus," 2014.

[95] K. Shinoda and T. Watanabe, "Acoustic modeling based on the MDL criterion for speech recognition," in *Proc. EuroSpeech-97*, no. 1, pp. 99–102, 1997.

[96] J. Bachan, T. Kuczmarski, and P. Francuzik, "Evaluation of synthetic speech using automatic speech recognition," in *XIV International PhD Workshop (OWD 2012). Conference Archives PTETiS*, vol. 30, pp. 500–505, 2012.

[97] M. Falcone, N. Yadav, C. Poellabauer, and P. Flynn, "Using isolated vowel sounds for classification of mild traumatic brain injury," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pp. 7577–7581, IEEE, 2013.

[98] H. Wertzner, S. Schreiber, and L. Amaro, "Analysis of fundamental frequency, jitter, shimmer and vocal intensity in children with phonological disorders," *Revista Brasileira de Otorrinolaringologia*, vol. 71, no. 5, pp. 582–588, 2005.

[99] F. Martínez-Licona, J. Goddard, A. Martínez-Licona, and M. Coto-Jimńez, "Assessing stress in mexican spanish from emotion speech signals," in *8th International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications, MAVEBA*, pp. 239–242, 2013.

[100] S. Terwijn, "On the learnability of Hidden Markov Models," in *Grammatical Inference: Algorithms and Applications*, pp. 261–268, Springer, 2002.

[101] G. Fink, *Markov models for pattern recognition*. Springer Heidelberg, 2008.

[102] S. Young, G. Evermann, D. Kershaw, G. Moore, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, *The HTK book*, vol. 2. Entropic Cambridge Research Laboratory Cambridge, 1997.

[103] L. Rabiner, "A tutorial on Hidden Markov Models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[104] M. Stamp, "A revealing introduction to Hidden Markov Models," *Department of Computer Science San Jose State University*, 2004.

[105] P. Dymarski, "Hidden Markov Models, Theory and Applications," *InTech Open Access Publishers*, 2011.

[106] L. Baum, T. Petrie, G. Soules, and N. Weiss, "A maximization technique occurring in the statistical analysis of probabilistic functions of markov chains," *The annals of mathematical statistics*, vol. 41, no. 1, pp. 164–171, 1970.

[107] G. Dias and S. Jayasena, "Sinhala text to speech system," 2009.

[108] Linux Sound, "*Speech Synthesis and Analysis Software.*" [En línea]. Disponible en: `http://linux-sound.org/speech.html`. [Consultado: 10 de diciembre de 2013].

[109] The Centre for Speech Technology Research, "*The Festival Speech Synthesis System.*" [En línea]. Disponible en: `http://www.cstr.ed.ac.uk/projects/festival/`. [Consultado: 1 de abril de 2014].

[110] eSpeak, "*eSpeak Texto to Speech.*" [En línea]. Disponible en: `http://espeak.sourceforge.net/`. [Consultado: 1 de abril de 2014].

[111] TCTS Lab, "*The MBROLA Project.*" [En línea]. Disponible en: `http://tcts.fpms.ac.be/synthesis/`. [Consultado: 1 de abril de 2014].

[112] AT&T, "*AT& T Natural Voices.*" [En línea]. Disponible en: `http://www.research.att.com/`. [Consultado: 25 de abril de 2014].

[113] Cepstral, "*Cepstral Speech Synthesis.*" [En línea]. Disponible en: `http://www.cepstral.com/`. [Consultado: 25 de abril de 2014].

[114] CereProc, "*CereProc Text to Speech.*" [En línea]. Disponible en: `http://www.cereproc.com/`. [Consultado: 20 de abril de 2014].

[115] Nuance, "*Loquendo.*" [En línea]. Disponible en: `http://loquendo.com/`. [Consultado: 1 de diciembre de 2013].

[116] IVONA, "*IVONA Text to Speech.*" [En línea]. Disponible en: `http://www.ivona.com/`. [Consultado: 25 de abril de 2014].

[117] SPTK, "*Speech Signal Processing Toolkit.*" [En línea]. Disponible en: `http://sp-tk.sourgeforge.net/`. [Consultado: 25 de abril de 2014].

[118] ActiveState, "*ActiveTCL.*" [En línea]. Disponible en: `http://www.activestate.com/activetcl`. [Consultado: 25 de abril de 2014].

[119] hts_engine, "*hts_engine API.*" [En línea]. Disponible en: `http://www.hts-engine.sourceforge.net/`. [Consultado: 25 de abril de 2014].

[120] T. Raitio, A. Suni, H. Pulakka, M. Vainio, and P. Alku, "Comparison of formant enhancement methods for HMM-based speech synthesis," in *Seventh ISCA Workshop on Speech Synthesis*, pp. 334–339, 2010.

Apéndice A

Modelos ocultos de Markov

Los Modelos ocultos de Markov (HMM) son procesos estocásticos de dos etapas, entrenables para clasificar y generar parámetros, tales como los que pueden describir la voz. En las siguientes secciones se describen las principales definiciones y aspectos teóricos relacionados con estos modelos.

A.1. Definición

Un HMM se pueden definir como una tupla $\lambda = (S, \pi, a, b)$ [100] donde:

- $S = 1, \ldots, m$ es un conjunto finito de estados.
- π es un vector de probabilidades iniciales.
- a es una matriz de transición de probabilidades.
- b es una matriz de probabilidades de salida.

Se pueden caracterizar como procesos estocásticos dobles: un primer proceso estocástico describe la transición entre estados, y el segundo las salidas [101]. El comportamiento del

primer proceso en el tiempo t depende solamente del estado predecesor, lo cual se puede describir como

$$p(S_t|S_1, S_1, \ldots, S_{t-1}) = p(S_t|S_{t-1}), \tag{A.1}$$

donde $p(S_t)$ indica la probabilidad del proceso de estar en el estado S en el tiempo t.

En el segundo proceso estocástico, en cada instante de tiempo t se genera una salida O_t, la cual tiene una distribución de probabilidad asociada, dependiente solamente del estado actual. Esto se puede describir como

$$p(O_t|O_1 \ldots O_{t-1}, S_1, \ldots, S_t) = p(O_t|S_t). \tag{A.2}$$

En procesos de clasificación y reconocimiento, es de interés establecer la probabiidad de que una palabra (conjunto de coeficientes o símbolos), sea emitida por un HMM. Se define la probabilidad $L_w(\lambda)$ de una palabra binaria w como la probabilidad de que λ genere w, es decir, la probabilidad de que para todo $t < |w|$, en el estado S_t el símbolo de salida O_t sea igual a el bit t–ésimo de w. La probabilidad de un conjunto M se define como

$$L_w(\lambda) = \prod_{w \in M} L_w(\lambda). \tag{A.3}$$

El tamaño de un HMM se puede definir como tamaño(π) + tamaño(a) + tamaño(b) [100]. En la Figura A.1 se muestra un ejemplo de un HMM con cinco estados, uno utilizado como estado inicial y otro final, es decir, sin emisión de observaciones en estos dos.

La razón por la que es llamado un modelo oculto, es porque lo único observable del modelo es la secuencia de salidas, pero no la secuencia de estados que son parte de la generación de éstas. Si las salidas son continuas, la distribución de probabilidad de salida deberá serlo también, y en caso de que éstas sean discretas, de igual forma debe serlo su distribución de

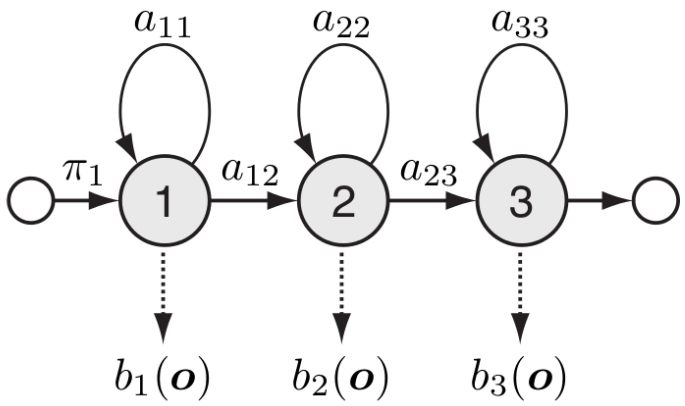


Figura A.1: Ejemplo de HMM [20]

probabilidad.

A.2. Los tres problemas en los HMM

Para aplicaciones de voz, existen tres problemas fundamentales relacionados con HMM que se desean resolver. Estos son [17] [103] [104] [105]:

1. Problema de evaluación: Dado un HMM $\lambda = (S, \pi, a, b)$ y una secuencia de observaciones O, determinar $p(O|\lambda)$, es decir, la probabilidad de que el HMM genere esa secuencia de observaciones.

2. Problema de decodificación: ¿Cuál es la secuencia de estados más probable que produce una secuencia de observaciones dada?

3. Problema de aprendizaje: ¿Cómo ajustar los parámetros del HMM (a, b, π) para maximizar $p(O|\lambda)$, dada un modelo y una secuencia de observaciones?

En la solución a los tres problemas, se sigue la formulación de [20] y [103]

1. Solución al problema 1: Dada una secuencia de estados $S = (S_1, \ldots, S_T)$ y una secuencia de observaciones $O = (O_1, \ldots, O_T)$, la probabilidad de ésta dado un HMM λ con N estados puede calcularse mediante

$$P(O|S, \lambda) = \prod_{t=1}^{T} P(O_t|S_t, \lambda), \tag{A.4}$$

es decir, con el producto de las probabilidades de salida de cada estado $\prod_{t=1}^{T} b_{q_t}(O_t)$.
La probabilidad de una secuencia de estados se puede calcular con la multiplicación de las probabilidades de transición, esto se expresa

$$P(S|\lambda) = \prod_{t=1}^{T} a_{q_{t-1}q_t}. \tag{A.5}$$

Usando el teorema de Bayes, $P(O, S|\lambda)$ se puede escribir como

$$P(O, S|\lambda) = P(O|S, \lambda)P(S|\lambda). \tag{A.6}$$

La probabilidad de la secuencia de observaciones $P(O|S, \lambda)$ se puede calcular sumando las probabilidades sobre las secuencias posibles de estados,

$$P(O|\lambda) = \sum_{q} P(O, S|\lambda) \tag{A.7}$$

$$= \sum_{q} P(O|, S, \lambda)P(S|\lambda) = \sum_{q} \prod_{t=1}^{T} a_{q_{t-1}a_{q_t}} b_{q_t}(O_t) \tag{A.8}$$

La secuencia de estados en en un HMM puede representarse como una estructura de Trellis, semejante a la Figura A.2, la probabilidad de la secuencia de observación dado el HMM λ se puede escribir como:

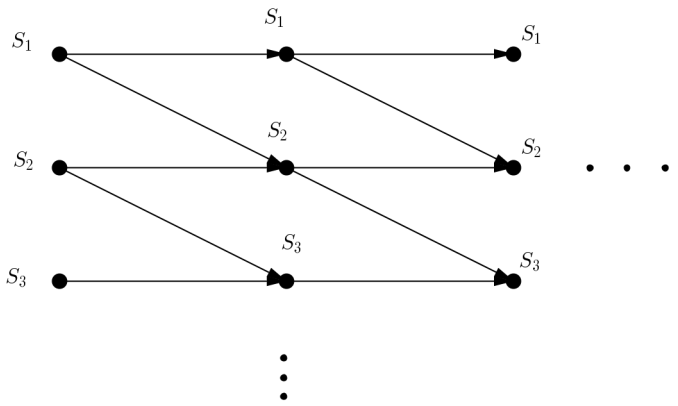


Figura A.2: Estrucura de Trellis

$$P(O|\lambda) = \sum_{i=1}^{N} P(O_1, \ldots, O_t, S_t = i|\lambda) P(O_{t+1}, \ldots, O_T|S_t = i, \lambda) \quad \forall t \in [0, T]. \tag{A.9}$$

Para el cálculo eficiente de la probabilidad se definen respectivamente las probabilidades hacia adelante y hacia atrás de la siguiente manera:

$$\alpha_T(i) = P(O_1, O_2, \ldots, O_t, q_t = i|\lambda) \tag{A.10}$$

$$\beta_T(i) = P(O_{t+1}, O_{t+2}, \ldots, O_T, q_t = i|\lambda) \tag{A.11}$$

Con estas dos probabilidades, en la Ecuación A.9 se calcula la probabilidad de la se-

cuencia observada O en el HMM λ. Existe el siguiente procedimiento recursivo para calcular $\alpha_t(i)$ y $\beta_t(i)$:

a) Inicializar, para $1 \leq i \leq N$

$$\alpha_1(i) = \pi_i b_i(O_1) \tag{A.12}$$

$$\beta_T(i) = 1 \tag{A.13}$$

b) De forma recursiva, calcular:

Para $1 \leq i \leq N$, $t = 2, \ldots, T$,

$$\alpha_{t+1}(i) = \left(\sum_{j=1}^{N} \alpha_t(j) a_{ji} \right) b_i(O_{t+1}). \tag{A.14}$$

Para $1 \leq i \leq N$, $t = T-1, \ldots, 1$,

$$\beta_t(i) = \sum_{j=1}^{N} a_{ij} b_j(O_{t+1}) \beta_{t+1}(j) \tag{A.15}$$

2. Solución al problema 2: Este problema puede formularse como el determinar $S^* = \arg\max_S P(O, S|\lambda)$. Se puede resolver mediante el algoritmo Viterbi, en el cual se define la probabilidad de la secuencia de estados más probable que termina en el estado i en el instante t como

$$\delta_t(i) = \max_{S_1, S_2, \ldots, S_{t-1}} P(O_1, \ldots, O_t, S_1, \ldots, S_{t-1}, S_t = i|\lambda). \tag{A.16}$$

En el algoritmo de Viterbi, se considera un arreglo de estados $\psi_t(i)$, en el cual se rastrea el camino de mayor probabilidad. Los pasos de este algoritmo son:

a) Inicializar, para $1 \leq i \leq N$

$$\delta_1(i) = \pi_i b_i(O_1) \tag{A.17}$$

$$\psi_1(i) = 0. \tag{A.18}$$

b) Recursión. Para $1 \leq i \leq N$, y $t = 2, \ldots, T$, calcular

$$\delta_t(j) = \max_i[\delta_t(i)a_{ij}]O_t \tag{A.19}$$

$$\psi_t(j) = \arg\max_i[\delta_t(i)a_{ij}] \tag{A.20}$$

c) Finalización.

$$P(O, S^*|\lambda) = \max_i[\delta_T(i)] \tag{A.21}$$

$$S^*(T) = \arg\max_i[\delta_T(i)] \tag{A.22}$$

d) Rastreo del recorrido. Se efectúa mediante la ecuación

$$S_t^* = \psi_{t+1}(S_{t+1}^*). \tag{A.23}$$

3. Solución al problema 3.

No existe una solución analítica al problema de entrenamiento, el cual se puede enunciar

$$\lambda^* = \arg\max_{\lambda} P(O|\lambda) \tag{A.24}$$

Es posible determinar un máximo local para este problema de optimización, a partir del algoritmo EM (por las siglas en inglés de *Expectation Maximization*), el cual fue presentado como solución a la optimización de HMM por Welch en 1970 [106], razón por la cual se le conoce como algoritmo de Baum–Welch.

El método descrito a continuación, que hace uso de una función auxiliar Q, se deriva de un HMM continuo con una distribución gaussiana para modelar la emisión de observaciones O.

La función auxiliar Q, tiene como argumento los parámetros λ' del HMM que se desea ajustar, y los nuevos parámetros λ para éste. Se define como

$$Q(\lambda', \lambda) = \sum_{S} P(S|O, \lambda') \log\left[P(O, S|\lambda).\right] \tag{A.25}$$

$P(O|\lambda)$ converge monótonamente a un cierto punto crítico, en un proceso iterativo que sustituye los parámetros de λ' por los de λ. Esto se puede demostrar a partir de los siguientes teoremas:

Teorema 1 $Q(\lambda', \lambda) \geq Q(\lambda', \lambda') \Rightarrow P(O|\lambda) \geq P(O|\lambda')$

Teorema 2 $Q(\lambda', \lambda)$ *tiene un único máximo global como función de* λ*, y este máximo es el único punto crítico.*

Teorema 3 λ *es un punto crítico de* $P(O|\lambda)$ *si y solo si es un punto crítico de* Q*.*

De manera que se busca maximizar Q para obtener un punto crítico de $P(O|\lambda)$. Como $P(O,S|\lambda) = \sum_{t=1}^{T} a_{S_{t-1}S_t}\mathcal{N}(O_t;\mu_{S_t,\Sigma_{S_t}})$, tomando logaritmo

$$\log P(O,S|\lambda) = \sum_{t=1}^{T} \log a_{S_{t-1}S_t} + \sum_{t=1}^{T} \log \mathcal{N}(O_t;\mu_{S_t},\Sigma_{S_t}). \tag{A.26}$$

Se denota $\pi_{S_1} = a_{S_0S_1}$. La función Q puede escribirse separando la suma de la ecuación A.25, al considerar la expresión A.26, de manera que

$$\begin{aligned} Q(\lambda',\lambda) &= \sum_{i=1}^{N} P(O,S_1=i|\lambda')\log\pi_i & \text{(A.27)}\\ &+ \sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{t=1}^{T-1} P(O,S_t=1,S_{t+1}=j|\lambda')\log_{a_{ij}} & \text{(A.28)}\\ &+ \sum_{i=1}^{N}\sum_{t=1}^{T} P(O,S_t=i|\lambda)\log\mathcal{N}(O_t,\mu_{S_t},\Sigma_{S_t}) & \text{(A.29)} \end{aligned}$$

Utilizando multiplicadores de Lagrange y ecuaciones en derivadas parciales de la anterior definición de Q, se obtienen los parámetros de λ que maximizan esta función, los cuales son:

$$\pi_i = \gamma_1(i) \tag{A.30}$$

$$a_{ij} = \frac{\sum_{t=1}^{T-1}\xi_t(i,j)}{\sum_{t=1}^{T-1}\gamma_t(i)} \tag{A.31}$$

$$\mu_i = \frac{\sum_{t=1}^{T} \gamma_t(i) O_t}{\sum_{t=1}^{T} \gamma_t(i)} \tag{A.32}$$

$$\Sigma_i = \frac{\sum_{t=1}^{T} \gamma_t(i)(O_t - \mu_i)(O_t - \mu_i)^\top}{\sum_{t=1}^{T} \gamma_t(i)} \tag{A.33}$$

donde $\gamma_t(i)$ y $\xi_t(i,j)$ se definen:

$$\gamma_t(i) = P(O, S_t = i|\lambda) \tag{A.34}$$

$$= \frac{\alpha_t(i)\beta_t(i)}{\sum_{j=1}^{N} \alpha_{t(j)}\beta_t(j)} \tag{A.35}$$

$$\xi_t(i,j) = P(O, S_t = i, S_{t+1} = j|\lambda) \tag{A.36}$$

$$= \frac{\alpha_t(i) a_{ij} b_j(O_{t+1}\beta_{t+1}(j))}{\sum_{l=1}^{N}\sum_{n=1}^{N} \alpha_t(l) a_{ln} b_n(O_{t+1})\beta_{t+1}(n)} \tag{A.37}$$

$\gamma_t(i)$ es la probabilidad de encontrarse en el estado i en el instante t, y $\xi_t(i,j)$ es la probabilidad de estar en el estado i en el instante t, y en el estado j en el instante $t+1$.

A.3. Tipos de HMM

La descripción dada de los HMM ha sido bastante general. De acuerdo con el contexto de aplicación, se realizan clasificaciones de éstos. Por ejemplo, en síntesis de voz se destaca la clasificación en ergódicos y no ergódicos [103], donde los primeros se refieren a aquellos en los cuales un estado puede ser alcanzado a partir de cualquier otro mediante una secuencia finita de transiciones, mientras que los no ergódicos no tienen esta propiedad.

Los HMM ergódicos en la práctica se consideran como aquellos que están completamente conectados, es decir, se puede alcanzar un estado desde cualquier otro en un solo paso. A pesar de las ventajas que esta característica pueda tener, es usual aplicar HMM no ergódicos del tipo izquierda a derecha, llamado de esta manera porque las transiciones se pueden dar solamente en ese sentido. De esta manera, puede establecerse un estado inicial y uno final, como se muestra en la Figura A.1, a diferencia del HMM ergódico de la Figura A.3.

Existen muchas otras combinaciones y variantes posibles, por ejemplo, el número de estados y las distribuciones utilizadas como salidas en cada estado.

En aplicaciones de síntesis de voz, es más utilizado un modelo de HMM que establece un modelo de duración, con lo cual deja de ser un modelo Markov. Es entonces llamado modelo semi-markoviano, o HSMM (por las siglas en inglés de *Hidden Semi–Markov Model)* [19].

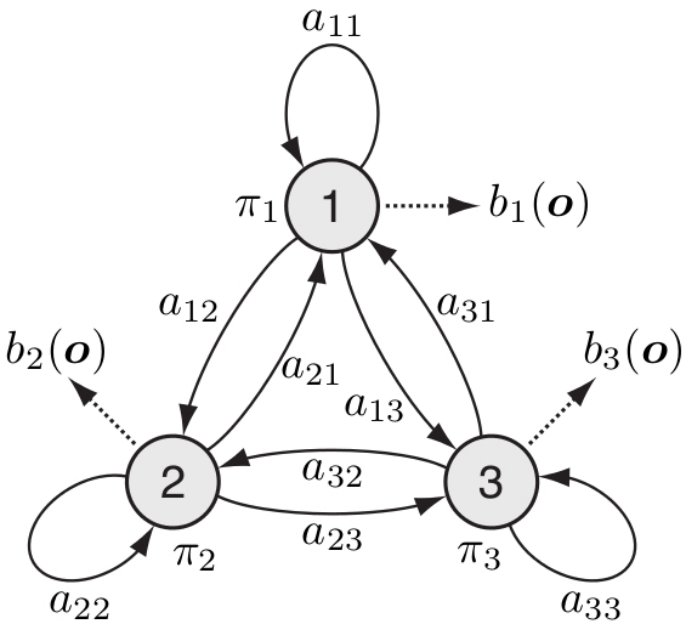


Figura A.3: HMM no ergódico tipo izquierda a derecha [17].

Apéndice B

Resultados de evaluación de parámetros acústicos

Se muestra el detalle de los valores extraídos de tono, *jitter* y *shimmer* en cada una de las pruebas, así como los resultados de las pruebas estadísticas de Friedman aplicadas para determinar diferencias estadísticamente significativas con la voz del hablante original

B.1. Pruebas sobre la influencia de parámetros de entrenamiento

En los Gráficos B.1 a B.4 se muestran los diagramas de caja de los tonos de las cinco vocales para las voces resultantes de los tres experimentos relacionados con la influencia del rango de $f0$ como parámetro de entrenamiento. Se pueden apreciar diferencias considerables entre el rango que ocupan las vocales de acuerdo con la definición del rango de $f0$ en el entrenamiento, especialmente en el caso de las vocales "i" y "u" en la voz creada utilizando un rango de $f0$ estrecho.

Para determinar si hay una diferencia estadísticamente significativa entre estos conjuntos

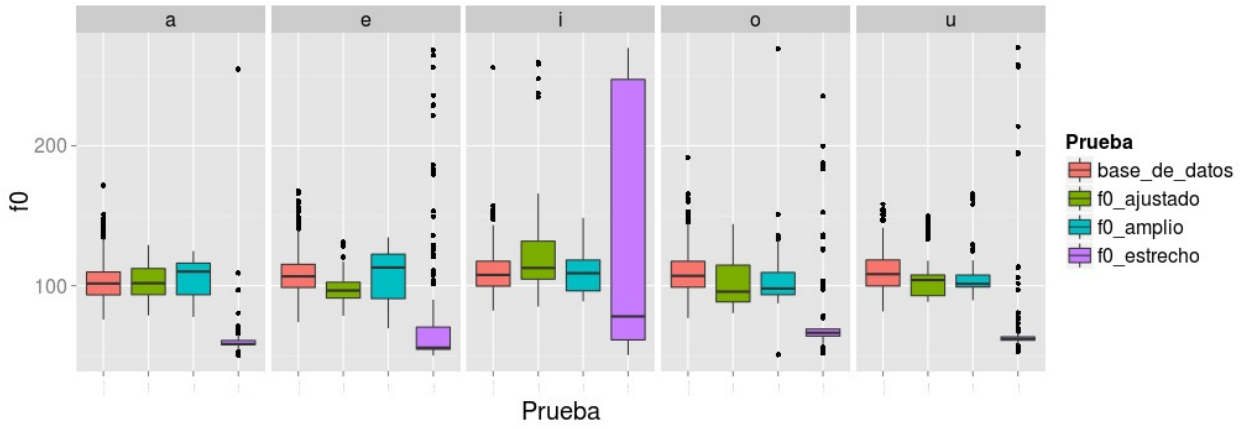


Figura B.1: Diagramas de caja para el tono de vocales según rango de $f0$ como parámetro de entrenamiento. Voz masculina, aplicación Clima

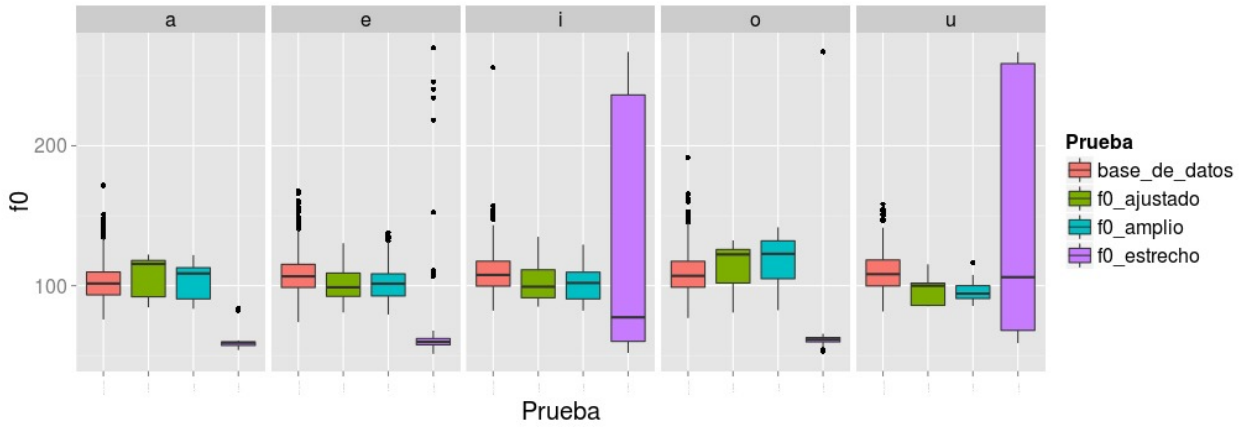


Figura B.2: Diagramas de caja para el tono de vocales según rango de $f0$ como parámetro de entrenamiento. Voz masculina, aplicación Reloj

de datos de tono en voces masculina según condición de entrenamiento, se realizó una prueba de Friedman, ya que los datos no satisfacen la condición de normalidad para aplicar un test paramétrico, como ANOVA, lo cual se determinó a partir de un análisis gráfico y una prueba Shapiro-Wilk. Las pruebas de Friedman, en todos los casos, se realizaron con un nivel de significancia de $\alpha = 0.05$. Con éste, se determina lo siguiente para cada aplicación:

- Aplicación Reloj: hay diferencia significativa entre los grupos de datos de tono de vocal en voces masculinas según parámetro de entrenamiento (p-valor $< 2.2e-16$), el cual, en análisis *Post-hoc* se determina que proviene de la diferencia entre los resultados de $f0$ estrecho con la base de datos, y de este mismo con las demás pruebas.

  De acuerdo con la prueba, no hay evidencia para rechazar la equivalencia de medias entre la base de datos, el entrenamiento con $f0$ ajustado y el entrenamiento con $f0$ amplio. En este análisis se ha excluido la vocal "a", debido a que no se cuenta con suficientes valores de tono en el caso de la voz obtenida a partir de $f0$ estrecho.

- Aplicación Clima: hay diferencia significativa en el grupo de datos de tono de vocal en voces masculinas según parámetro de entrenamiento (p-valor $< 2.2e-16$), el cual, en análisis *Post-hoc* se determina que proviene de la diferencia entre los resultados de $f0$ estrecho con la base de datos (p-valor de 0) y de $f0$ ajustado con la base de datos (p-valor de $8.78e-6$). Se presentan otras diferencias entre estos grupos de datos, pero se destaca en este análisis la diferencia significativa de las resultados con la voz original, la cual es de mayor interés para determinar la calidad de las voces sintéticas.

Se tiene entonces, que hay diferencia significativa entre los grupos de datos de tono para las cinco vocales, correspondientes a las voces resultantes de la síntesis a partir del entrenamiento con $f0$ estrecho en ambos casos de la aplicación Reloj, y en el caso de la aplicaciño Clima, también diferencia significativa con la voz resultante del entrenamiento con $f0$ ajustado.

Para el análisis estadístico de las diferencias de tono en el caso de la voz femenina, se ha considerado el análisis sin la vocal "a" en la aplicación Reloj, debido a que se cuenta con muy pocos valores extraídos en las pruebas de $f0$ estrecho. Los resultados de la prueba de Friedman indican:

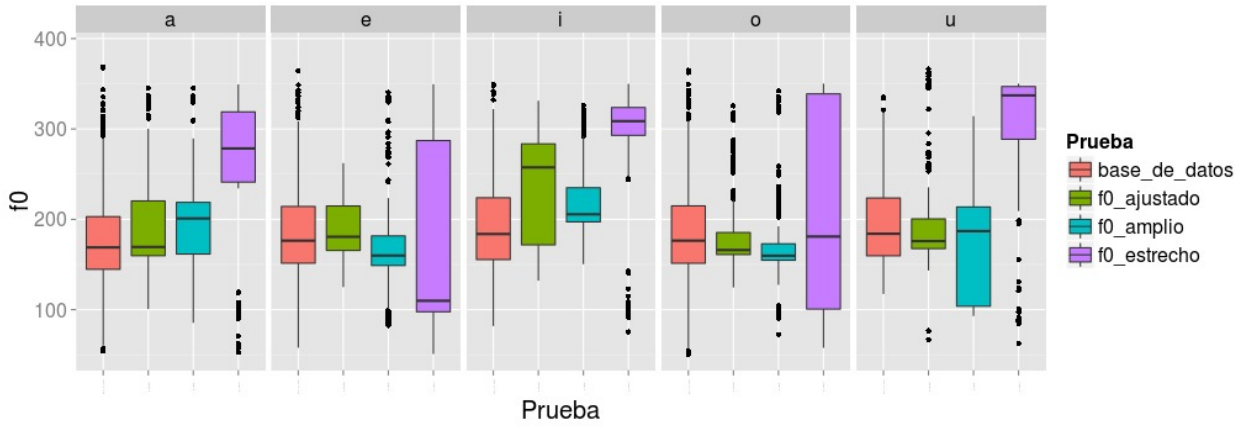


Figura B.3: Diagramas de caja para el tono de vocales según rango de $f0$ como parámetro de entrenamiento. Voz femenina, aplicación Clima

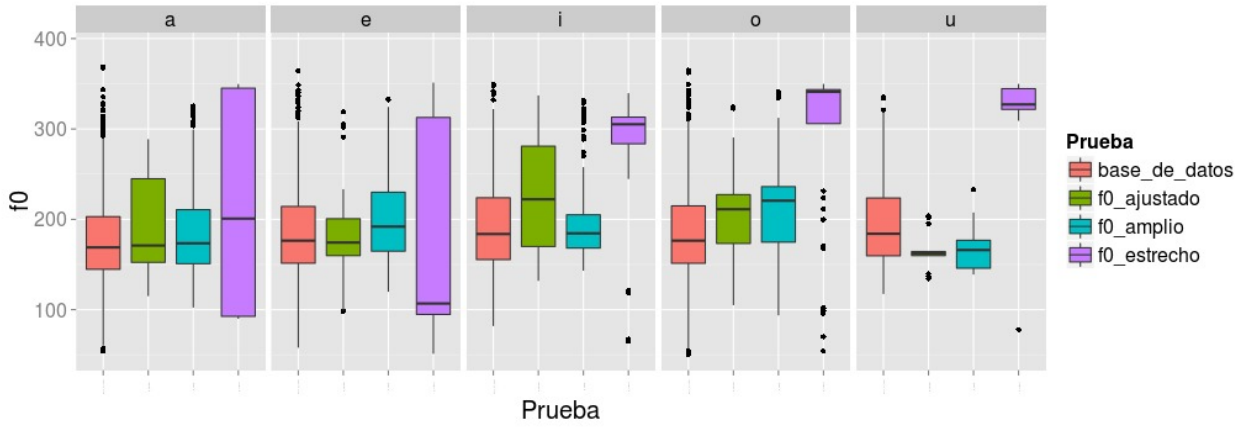


Figura B.4: Diagramas de caja para el tono de vocales según rango de $f0$ como parámetro de entrenamiento. Voz femenina, aplicación Reloj

- Aplicación Reloj: hay diferencia significativa en el grupo de datos de tono de vocal según parámetro de entrenamiento (p-valor $< 2.2e-16$), el cual, en análisis *Post-hoc* se determina que proviene de la diferencia entre los resultados de $f0$ estrecho con la base de datos (p-valor 0), de $f0$ ajustado con la base de datos (p-valor $1.3e-2$) y de $f0$ amplio con la base de datos (p-valor $1.9e-4$). Es decir, en todos las condiciones

de entrenamiento con las pruebas de diferentes rangos de $f0$ se presenta diferencia significativa con los de la base de datos.

- Aplicación Clima: hay diferencia significativa en el grupo de datos de tono de vocal según parámetro de entrenamiento (p-valor $< 2.2e-16$), el cual, en análisis *Post-hoc* se determina que proviene de la diferencia entre los resultados de $f0$ estrecho con la base de datos (p-valor de 0) y de $f0$ ajustado con la base de datos (p-valor de 0.001). De acuerdo con la prueba, no hay evidencia para rechazar la equivalencia de medias entre la base de datos y el entrenamiento con $f0$ amplio.

Se tiene entonces que para la voz femenina las diferencias de tono de vocales con la base de datos son estadísticamente significativas en todos los casos para la aplicación Reloj, mientras que en la aplicación Clima hay evidencia para considerarlo así en dos de las pruebas, lo cual coincide con lo obtenido en el caso de voz masculina.

En el caso del *jitter* obtenido en las vocales de la voz masculina, los diagramas de caja se muestran en las Figuras B.5 a B.6.

Para el estudio estadístico de diferencias significativas entre el *jitter* de vocales en voces masculinas, se ha considerado el análisis sin la vocal "a" en la aplicación Reloj, debido a que se cuenta con muy pocos valores extraídos en las pruebas de $f0$ estrecho. Los resultados de esta prueba son:

- Aplicación Reloj: hay diferencia significativa en el grupo de datos de *jitter* en vocales según condición de entrenamiento en voces masculinas (p-valor $< 2.2e-16$), el cual, en análisis *Post-hoc* se determina que proviene de la diferencia entre los resultados de $f0$ estrecho con la base de datos (p-valor de $1.57e-10$) y de $f0$ ajustado con la base de datos (p-valor de $2.36e-2$).

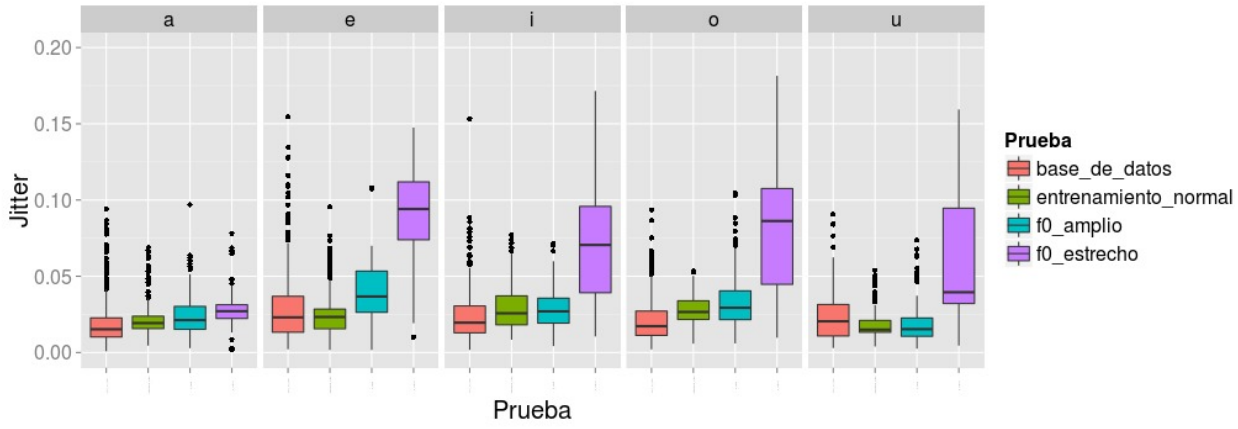


Figura B.5: Diagramas de caja para el *jitter* de vocales según rango de $f0$ como parámetro de entrenamiento. Voz masculina, aplicación Clima

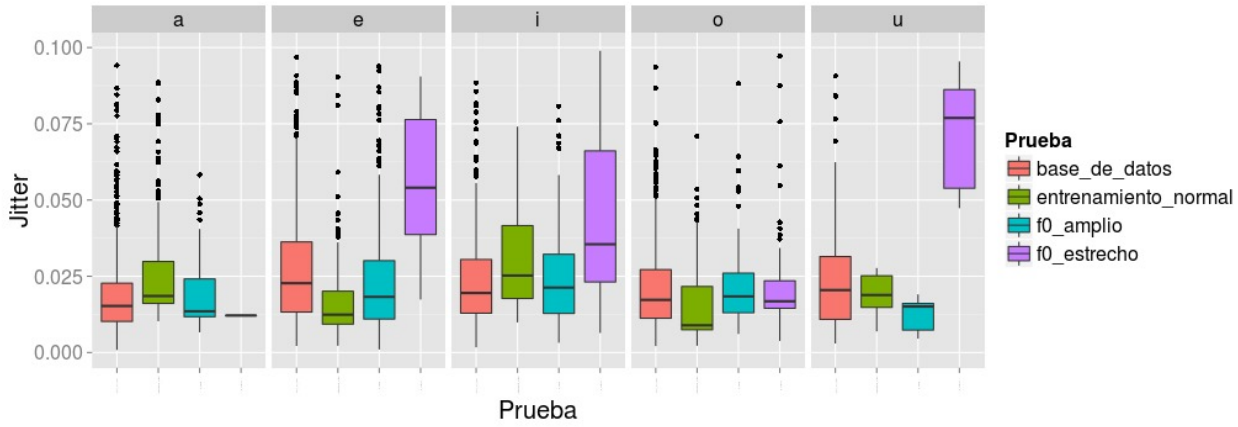


Figura B.6: Diagramas de caja para el *jitter* de vocales según rango de $f0$ como parámetro de entrenamiento. Voz masculina, aplicación Reloj

- Dentro de la aplicación Clima: hay diferencia significativa en el grupo de datos de *jitter* en vocales según condición de entrenamiento en voces masculinas (p-valor $< 2.2e-16$), el cual, en análisis *Post-hoc* se determina que proviene de la diferencia entre los resultados de $f0$ estrecho con la base de datos (p-valor 0), de $f0$ ajustado con la base de datos (p-valor $9.52e-3$) y de $f0$ amplio con la base de datos (p-valor $3.3e-13$).

Es decir, en todos las condiciones de entrenamiento con las pruebas de $f0$ se presenta diferencia significativa con la base de datos.

Para la aplicación Reloj, de forma semejante al análisis de tono de vocales la misma aplicación en voz masculina, no se encuentra evidencia para rechazar la semejanza de los resultados de *jitter* para la voz obtenida con rango amplio de $f0$. En la aplicación Clima todos los resultados tienen diferencias significativas con la voz original.

Los resultados de *jitter* para la voz de mujer en la evaluación del rango de $f0$ como parámetro de entrenamiento en voces femeninas se muestran como diagramas de caja en las Figuras B.7 y B.8.

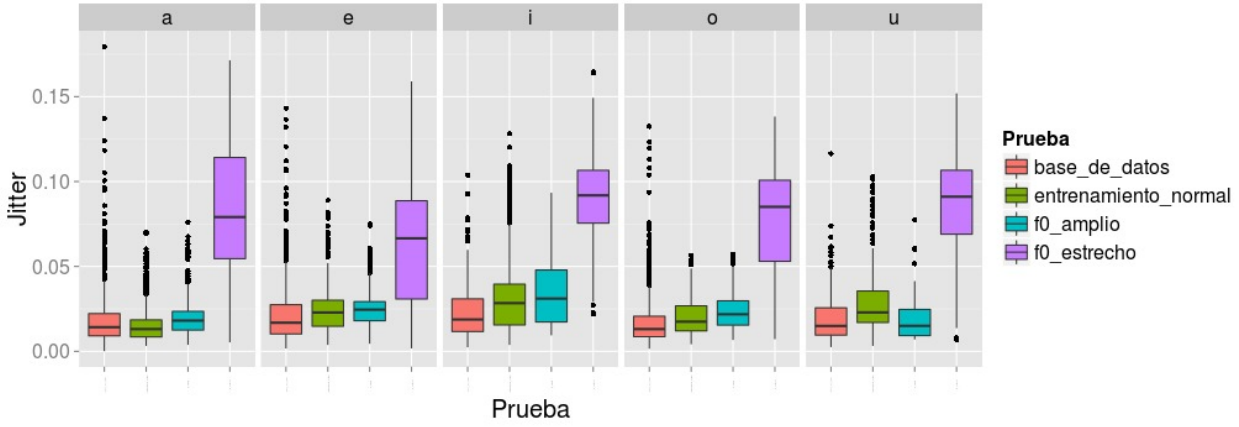


Figura B.7: Diagramas de caja para el *jitter* de vocales según rango de $f0$ como parámetro de entrenamiento. Voz femenina, aplicación Clima

El estudio de diferencias significativas realizado para *jitter* de vocales en voces femeninas con la prueba de Friedman, tiene como resultados:

- Aplicación Reloj: hay diferencia significativa en el grupo de datos de *jitter* en vocales (p-valor $< 2.2e-16$), el cual, en análisis *Post-hoc* se determina que proviene de la

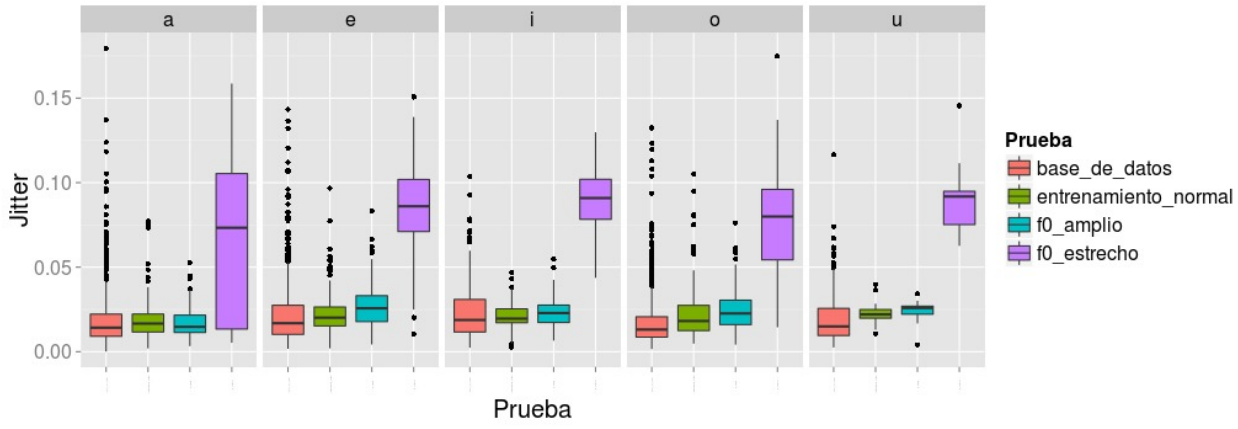


Figura B.8: Diagramas de caja para el *jitter* de vocales según rango de $f0$ como parámetro de entrenamiento. Voz femenina, aplicación Reloj

diferencia entre los resultados de $f0$ estrecho con la base de datos (p-valor de 0) y de $f0$ amplio con la base de datos (p-valor de 0.0021). No hay evidencia para rechazar la semejanza entre el resultado con $f0$ ajustado y la base de datos.

- Aplicación Clima, hay diferencia significativa en el grupo de datos de *jitter* en vocales (p-valor $< 2.2e-16$), el cual, en análisis *Post-hoc* se determina que proviene de la diferencia entre los resultados de $f0$ estrecho con la base de datos (p-valor 0), de $f0$ ajustado con la base de datos (p-valor $2.2e-4$) y de $f0$ amplio con la base de datos (p-valor $3.4e-7$). Es decir, en todos las condiciones de entrenamiento con las pruebas de $f0$ se presenta diferencia significativa con la base de datos.

En el caso de *shimmer*, los resultados para las vocales de la voz masculina según el parámetro de entrenamiento analizado se presentan en las Figuras B.9 a B.10.

La prueba de diferencias significativas para el *shimmer* de vocales según parámetro de entrenamiento en voces masculinas tiene como resultados:

- Aplicación Reloj: hay diferencia significativa en el grupo de datos de *shimmer* (p-valor

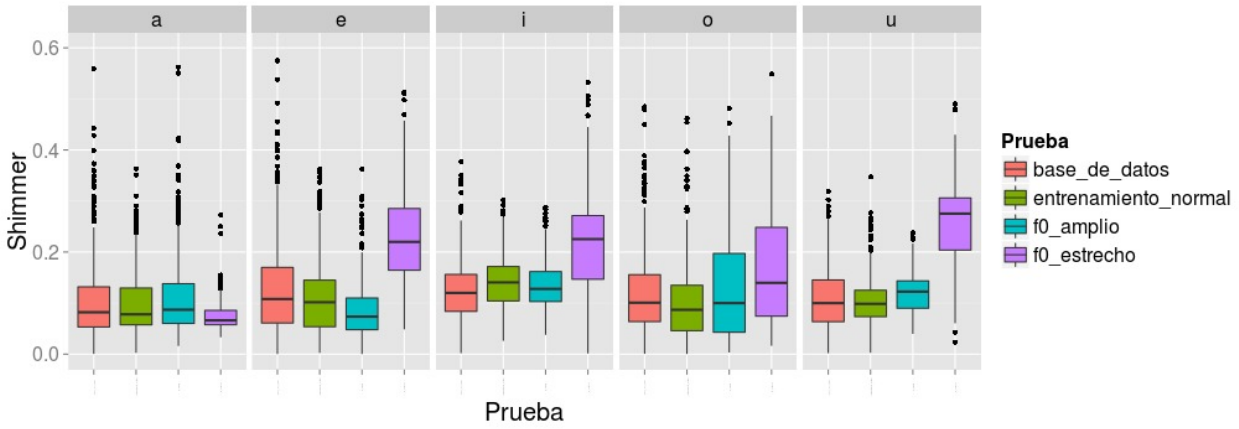


Figura B.9: Diagramas de caja para el *shimmer* de vocales según rango de $f0$ como parámetro de entrenamiento. Voz masculina, aplicación Clima

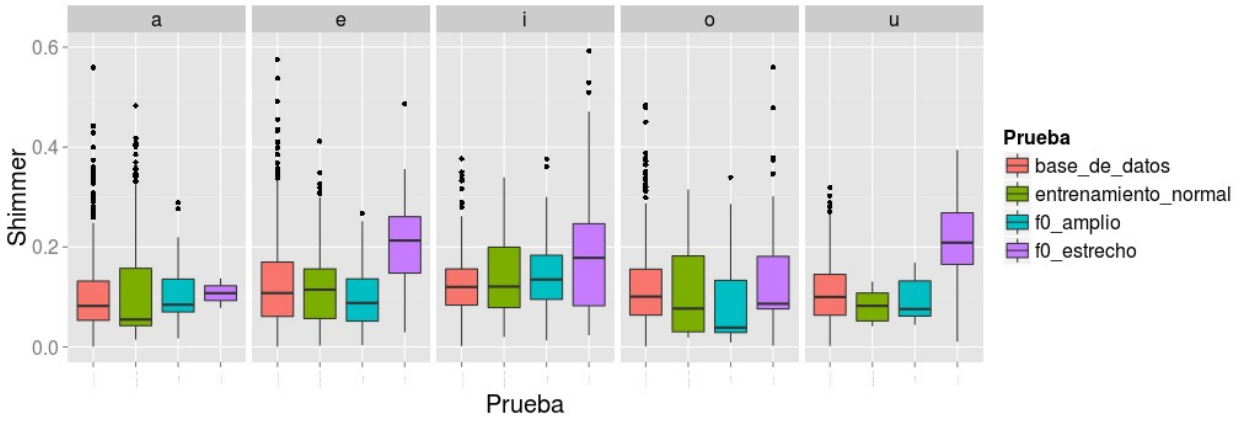


Figura B.10: Diagramas de caja para el *shimmer* de vocales según rango de $f0$ como parámetro de entrenamiento. Voz masculina, aplicación Reloj

$< 2.2e-16$), el cual, en análisis *Post-hoc* se determina que proviene de la diferencia entre los resultados de $f0$ estrecho con la base de datos (p-valor de $1.78e-5$). No hay evidencia para rechazar la semejanza entre el resultado con $f0$ ajustado y la base de datos ni de $f0$ amplio con la base de datos.

- Aplicación Clima: hay diferencia significativa en el grupo de datos de *shimmer* (p-valor $< 2.2e-16$), el cual, en análisis *Post-hoc* se determina que proviene de la diferencia entre los resultados de $f0$ estrecho con la base de datos (p-valor de 0). No hay evidencia para rechazar la semejanza entre el resultado con $f0$ ajustado y la base de datos ni de $f0$ amplio con la base de datos.

En este caso coinciden los resultados de diferencia significativa de *shimmer* en vocales de ambas aplicaciones para el hablante masculino, en cuanto a que existen estas diferencias para el caso de las voces obtenidas con rango de $f0$ estrecho. El caso de los valores obtenidos para la voz femenina se muestran en los diagramas de caja de las Figuras B.11 y B.12.

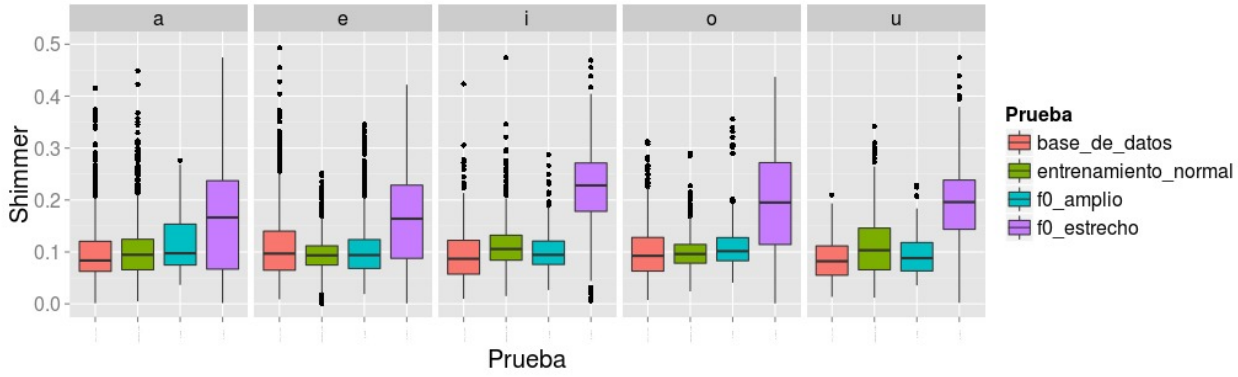


Figura B.11: Diagramas de caja para el *shimmer* de vocales según rango de $f0$ como parámetro de entrenamiento. Voz femenina, aplicación Clima

La prueba de significancia para el *shimmer* de vocales según condición de entrenamiento para voces femeninas tiene en este caso como resultados:

- Aplicación Reloj: hay diferencia significativa en el grupo de datos de *shimmer* en vocales (p-valor $< 2.2e-16$), el cual, en análisis *Post-hoc* se determina que proviene de la diferencia entre los resultados de $f0$ estrecho con la base de datos (p-valor de 0) y de

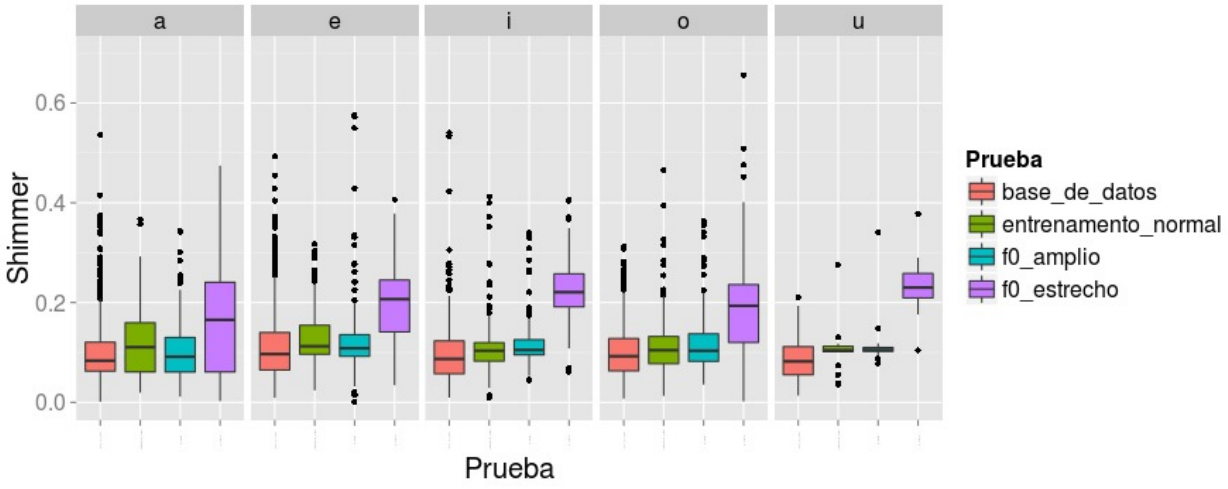


Figura B.12: Diagramas de caja para el *shimmer* de vocales según rango de $f0$ como parámetro de entrenamiento. Voz femenina, aplicación Reloj

$f0$ ajustado con la base de datos (p-valor de 0.012). No hay evidencia para rechazar la semejanza entre el resultado de $f0$ amplio con la base de datos.

- Dentro de la aplicación Clima: hay diferencia significativa en el grupo de datos de *shimmer* en vocales (p-valor $< 2.2e-16$), el cual, en análisis *Post-hoc* se determina que proviene de la diferencia entre los resultados de $f0$ estrecho con la base de datos (p-valor de $1.78e-5$) y de $f0$ ajustado con la base de datos (p-valor de 0.007). No hay evidencia para rechazar la semejanza entre el resultado de $f0$ amplio con la base de datos.

En el caso de hablante mujer, coincide la diferencia significativa de *shimmer* en vocales para ambas aplicaciones, lo cual se dio también en el caso de los hombres, aunque en este último caso se presenta el resultado de diferencia estadísticamente significativa en dos de las pruebas.

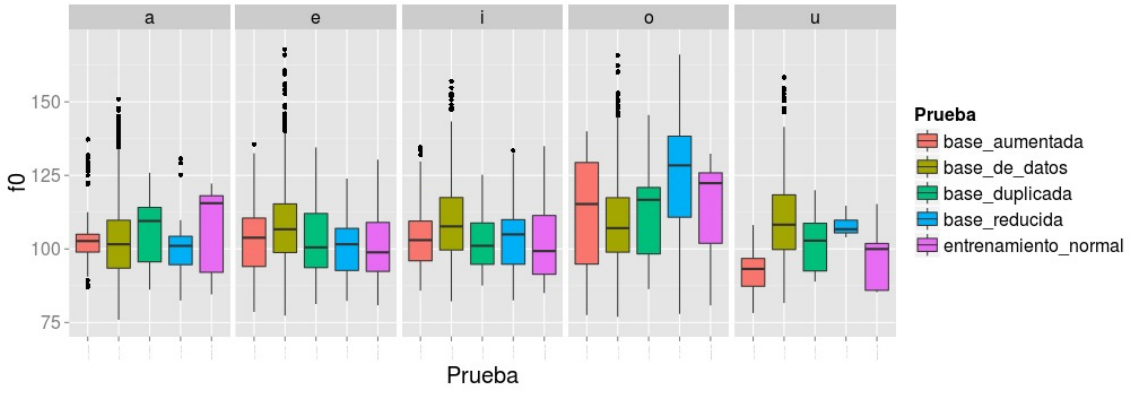


Figura B.13: Diagramas de caja para el tono de vocales según conjunto de entrenamiento. Voz masculina, aplicación Reloj

B.2. Pruebas sobre la influencia del tamaño del conjunto de entrenamiento

En las Figuras B.13 a B.14 se muestran los diagramas de caja de los tonos de las vocales para ambas aplicaciones en voces masculinas. Se pueden apreciar diferencias evidentes entre el rango que ocupan las vocales de acuerdo con la pruebas de la influencia del conjunto de entrenamiento.

Las pruebas de significancia estadística de tono de vocales para la voz masculina, en las pruebas sobre la influencia del tamaño de la base de datos tiene como resultados:

- Aplicación Reloj: no hay evidencia estadística para rechazar la hipótesis de equivalencia entre medias de todas las pruebas.

- Aplicación Clima: hay diferencia significativa en el grupo de datos de tono de vocales (p-valor $< 2.2e-16$), el cual, en análisis *Post-hoc* se determina que proviene de la diferencia entre el tono de las vocales de la voz original y la voz sintetizada a partir de

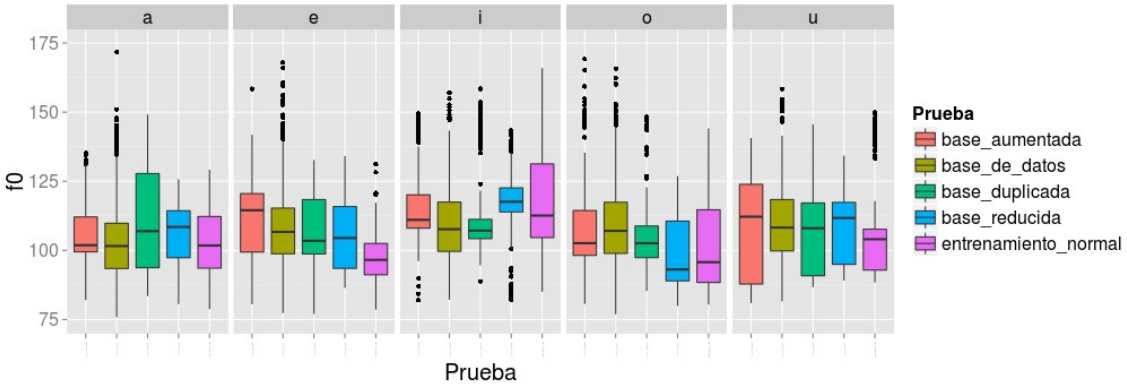


Figura B.14: Diagramas de caja para el tono de vocales según conjunto de entrenamiento. Voz masculina, aplicación Clima

la base de datos aumentada (p-valor de $1.46e-06$) y del entrenamiento normal (p-valor de $6.02e-11$). No hay evidencia para rechazar la equivalencia de medias en los demás casos.

Para las voces de mujer, los resultados de prueba estadística de significancia entre grupos de datos de tono de vocales en pruebas sobre la influencia del tamaño de la base de datos se muestran como diagramas de caja en las Figuras B.15 y B.16.

Las pruebas de significancia de diferencias en tono de vocales, para estas pruebas tienen como resultados:

- Aplicación Reloj: hay diferencia significativa en el grupo de datos de tono de vocales (p-valor $< 2.2e-16$), el cual, en análisis *Post-hoc* se determina que proviene de la diferencia entre todas las voces sintéticas con la hablante original, con excepción de la voz obtenida con la base aumentada.

- Aplicación Clima: hay diferencia significativa en el grupo de datos de tono de vocales (p-valor $< 2.2e-16$), el cual, en análisis *Post-hoc* se determina que proviene de la

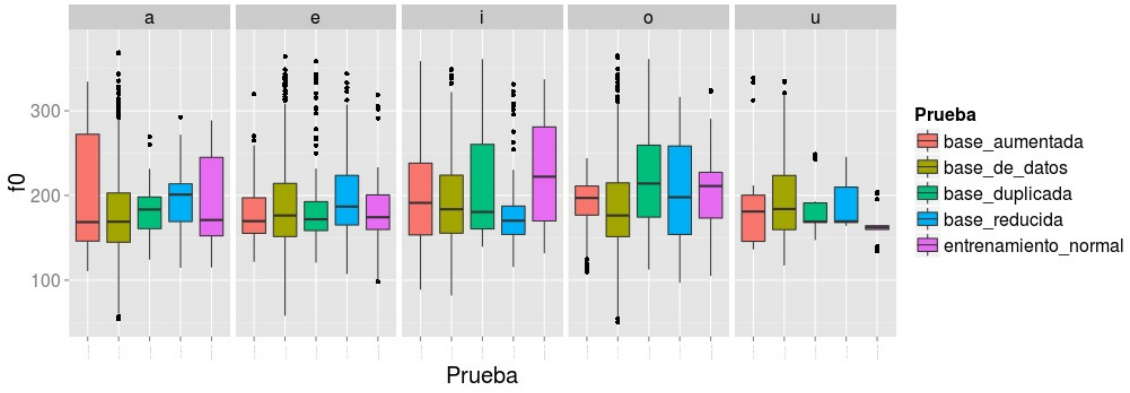


Figura B.15: Diagramas de caja para el tono de vocales según conjunto de entrenamiento. Voz femenina, aplicación Reloj

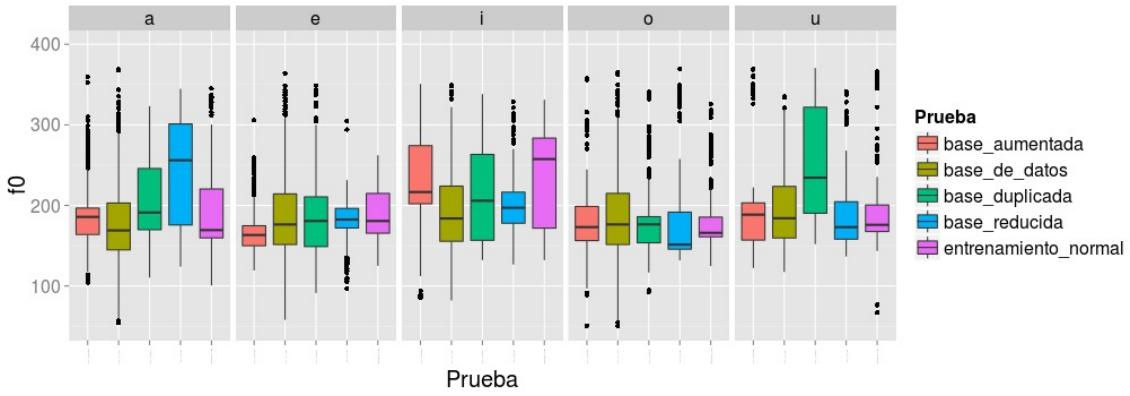


Figura B.16: Diagramas de caja para el tono de vocales según conjunto de entrenamiento. Voz femenina, aplicación Clima

diferencia de tono de vocales de todas las voces sintetizadas con la hablante original.

Se puede apreciar que los resultados de diferencias en tono de vocales varían según el género del hablante y aplicación.

Por su parte, se analiza el efecto sobre los valores de *jitter* para el hablante masculino en las pruebas sobre la influecnia de la cantidad de datos de entrenamiento en las Figuras B.17 a B.18.

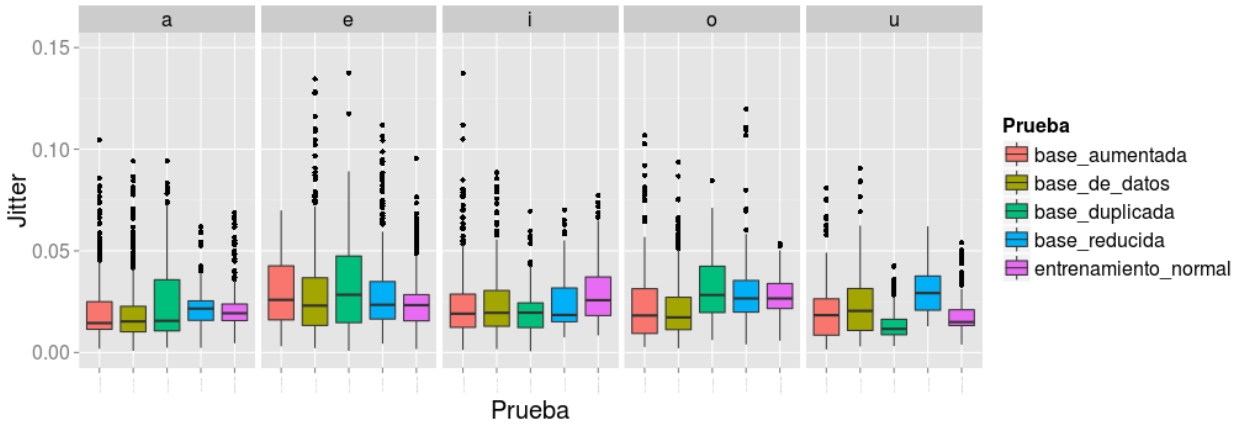


Figura B.17: Diagramas de caja para el *jitter* de vocales según condición de entrenamiento. Voz masculina, aplicación Clima

La prueba de significancia de diferencias sobre el *jitter* de vocales para el hablante masculino tiene como resultado:

- Aplicación Reloj: hay diferencia significativa en el grupo de datos de *jitter* de vocales (p-valor $< 2.2e-16$), el cual, en análisis *Post-hoc* se determina que proviene de la diferencia entre el *jitter* de las vocales de la voz sintética obtenida con la base reducida y la voz original, así como la voz obtenida con el entrenamiento normal y la voz original. En los otros dos casos no se rechaza la hipótesis de diferencia significativa.

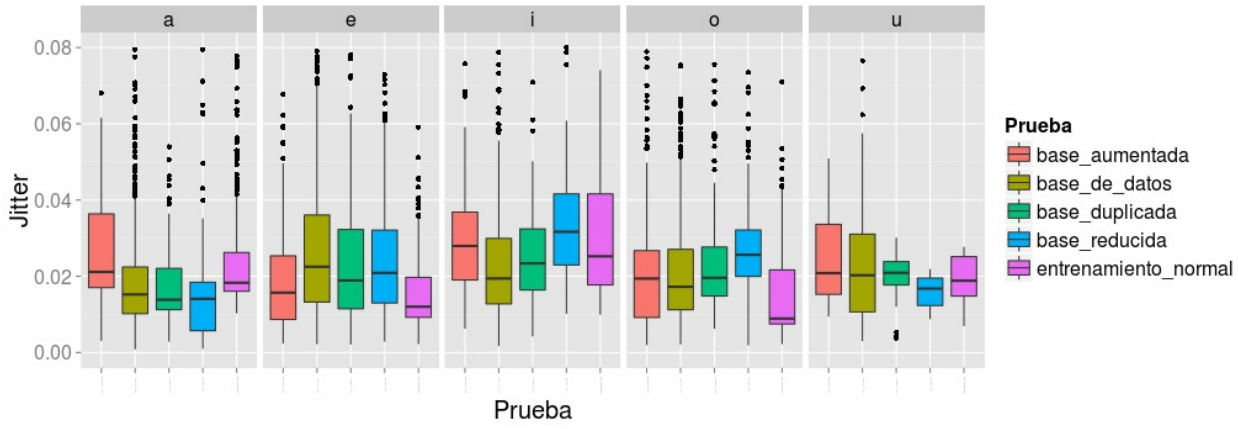


Figura B.18: Diagramas de caja para el *jitter* de vocales según condición de entrenamiento. Voz masculina, aplicación Reloj

- Aplicación Clima: hay diferencia significativa en el grupo de datos de *jitter* de vocales (p-valor $< 2.2e-16$), el cual, en análisis *Post-hoc* se determina que proviene de la diferencia entre el *jitter* de las vocales del hablante original con la voz sintética obtenida con la base de datos reducida.

El caso de la voz femenina y los diagramas de caja con las datos de *jitter* de las vocales en las diferentes pruebas realizados para determinar la influencia del tamaño del conjunto de entrenamiento se muestran en las Figuras B.19 y B.20. En este caso la prueba de significancia estadística tiene como resultados:

- Aplicación Reloj: hay diferencia significativa en el grupo de datos de *jitter* de vocales (p-valor $< 2.2e-16$), el cual, en análisis *Post-hoc* se determina que proviene de diferencia significativa entre el hablante original y la voz sintética producida con base de datos aumentada (p-valor de $1.52e-05$), el hablante original y la voz sintética producida con la base de datos duplicada (p-valor de $1.18e-02$) y la voz producida con el entrenamiento normal y el hablante original (p-valor de $2.17e-04$).

- Aplicación Clima: hay diferencia significativa en el grupo de datos de *jitter* de vocales (p-valor $< 2.2e-16$), el cual, en análisis *Post-hoc* se determina que proviene de diferencias significativas entre el hablante original y todos los resultados.

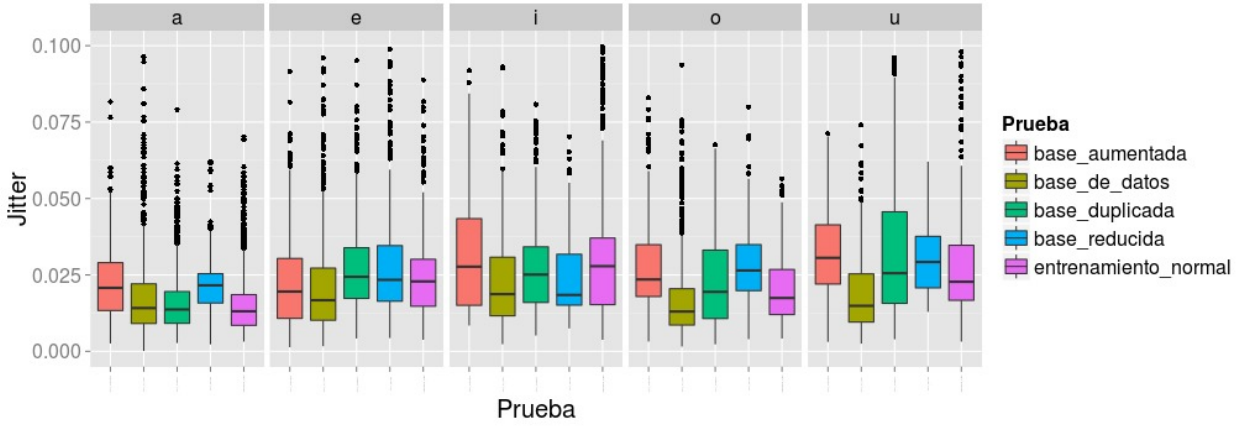


Figura B.19: Diagramas de caja para el *jitter* de vocales según condición de entrenamiento. Voz femenina, aplicación Clima

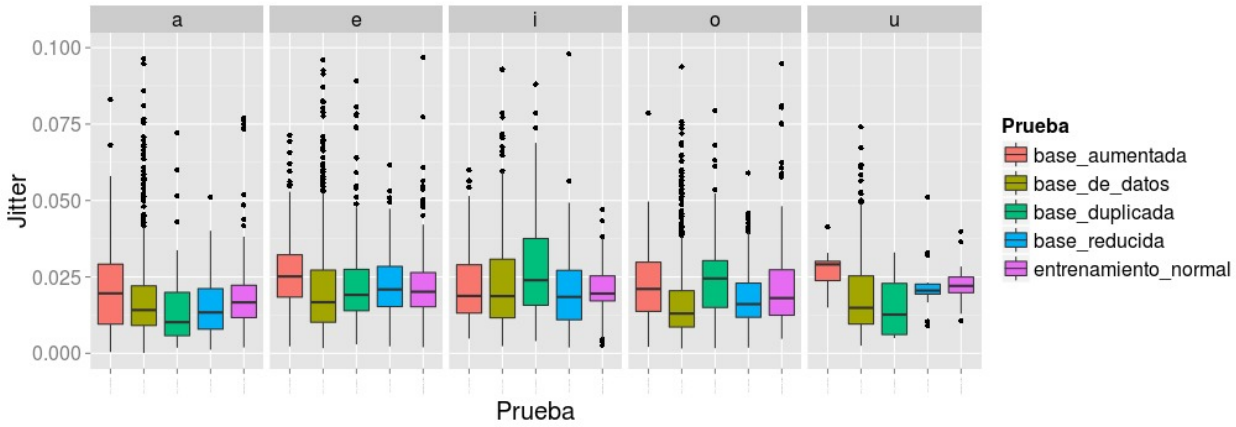


Figura B.20: Diagramas de caja para el *jitter* de vocales según condición de entrenamiento. Voz femenina, aplicación Reloj

Los datos de *shimmer* en vocales para las aplicaciones Clima y Reloj en las voces masculinas, se encuentran representados en diagramas de caja en las Figuras B.21 y B.22.

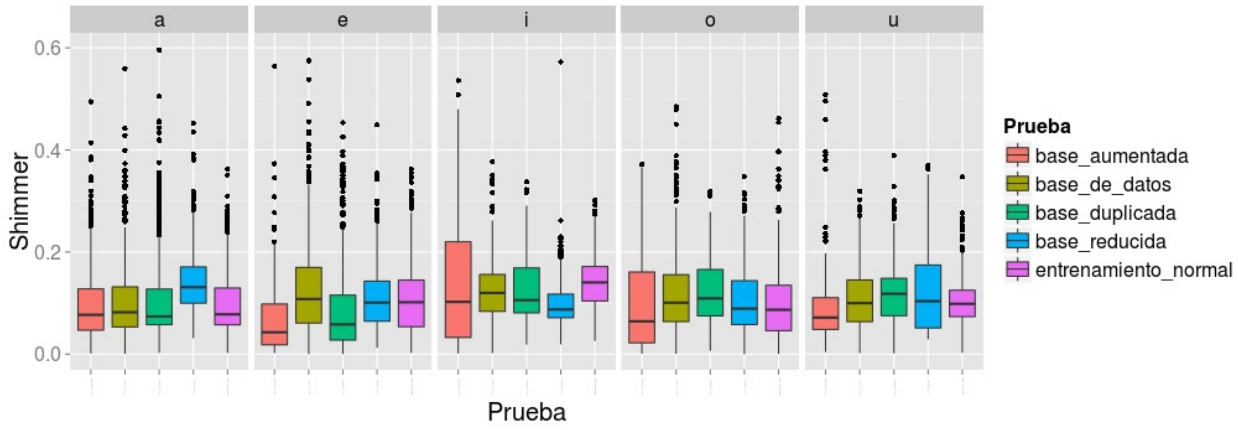


Figura B.21: Diagramas de caja para el *shimmer* de vocales según condición de entrenamiento. Voz masculina, aplicación Clima

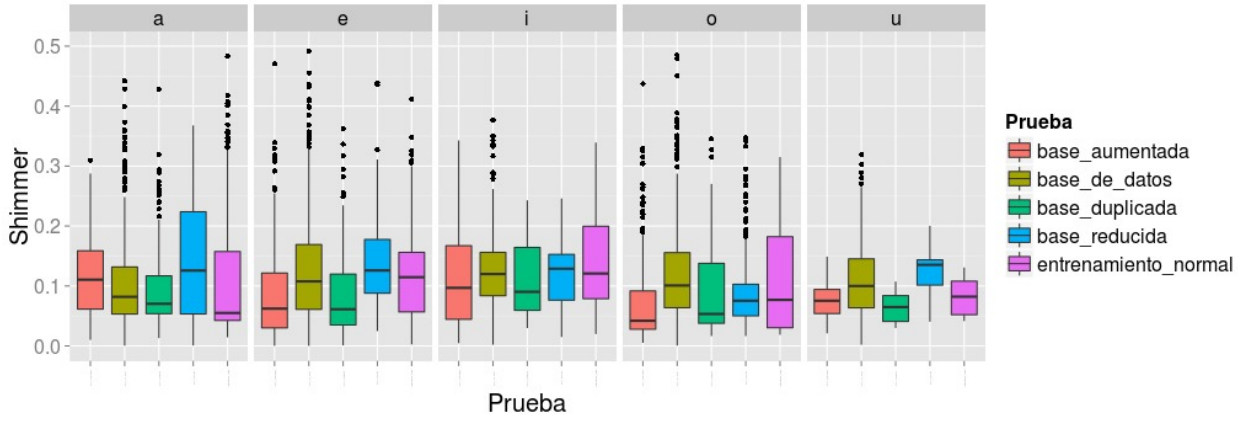


Figura B.22: Diagramas de caja para el *shimmer* de vocales según condición de entrenamiento. Voz masculina, aplicación Reloj

La prueba de Friedman para detectar diferencias estadísticamente significativas en estos conjuntos de valores indican:

- Aplicación Reloj: hay diferencia significativa en el grupo de datos de *shimmer* de vocales para las voces masculinas (p-valor $< 2.2e{-}16$), el cual, en análisis *Post-hoc* se determina que proviene de diferencia significativa entre los valores de *shimmer* de las vocales en la base de datos y la voz sintética producida con la base de datos aumentada (p-valor 2.9e-04), y la base duplicada (p-valor $2.1e-04$).

- Aplicación Clima: hay diferencia significativa en el grupo de datos de *shimmer* de vocales para las voces masculinas (p-valor $< 2.2e-16$), el cual, en análisis *Post-hoc* se determina que proviene de diferencias significativas entre el *shimmer* de las vocales en el hablante original y la voz sinética producida con la base de datos aumentada (p-valor 0) y el entrenamiento normal (p- valor $1.4e-13$).

En cuanto a los datos de *shimmer* en vocales para las aplicaciones Clima y Reloj en las voces femeninas, éstos se encuentran representados en diagramas de caja en las Figuras B.21 y B.22.

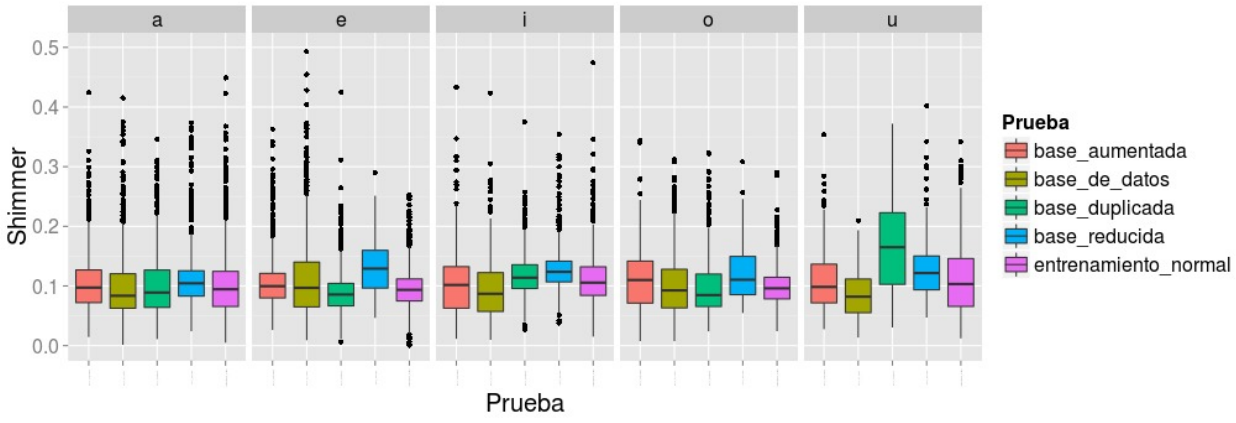


Figura B.23: Diagramas de caja para el *shimmer* de vocales según condición de entrenamiento. Voz femenina, aplicación Clima

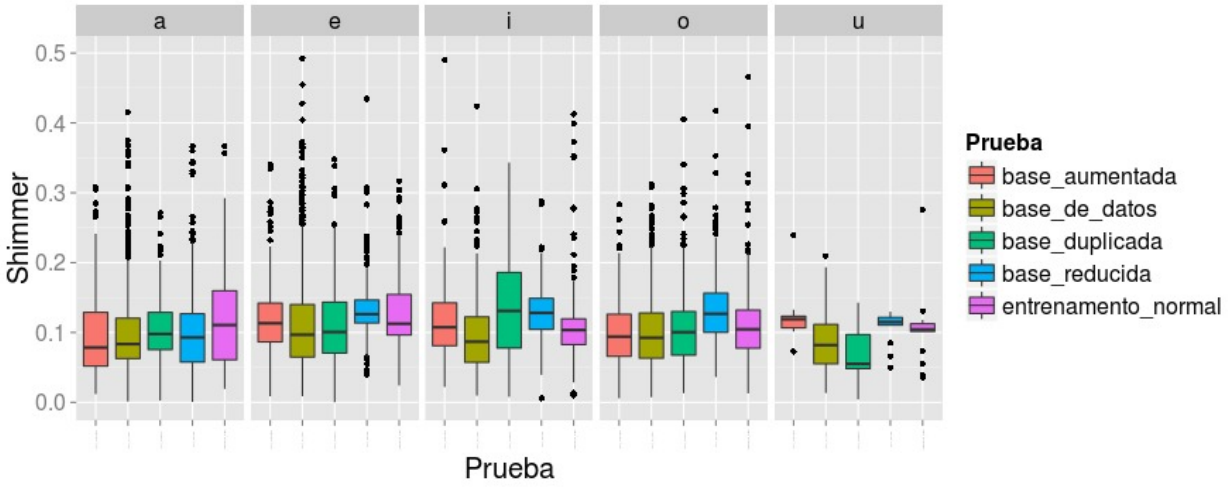


Figura B.24: Diagramas de caja para el *shimmer* de vocales según condición de entrenamiento. Voz femenina, aplicación Reloj

Las pruebas de Friedman para analizar diferencias entre los conjuntos de datos de los experimentos para analizar la influencia del tamaño del conjunto de entrenamiento dan como resultado:

- Aplicación Reloj: hay diferencia significativa en el grupo de datos de *shimmer* de vocales para las voces femeninas (p-valor $< 2.9e-06$), el cual, en análisis *Post-hoc* se determina que proviene de la diferencia entre el shimmer de las vocales del hablante original con la voz sintética producida con la base de datos aumentada (p-valor $4.6e-01$), la base duplicada (p-valor $2.8e-03$) y la base reducida (p-valor $1.7e-07$).

- Aplicación Clima: hay diferencia significativa en el grupo de datos de *shimmer* de vocales para las voces femeninas (p-valor $< 2.2e-16$), el cual, en análisis *Post-hoc* se determina que proviene de diferencias significativas del shimmer de vocales del hablante original y todas las voces sintéticas, con excepción de la obtenida con el entrenamiento normal.

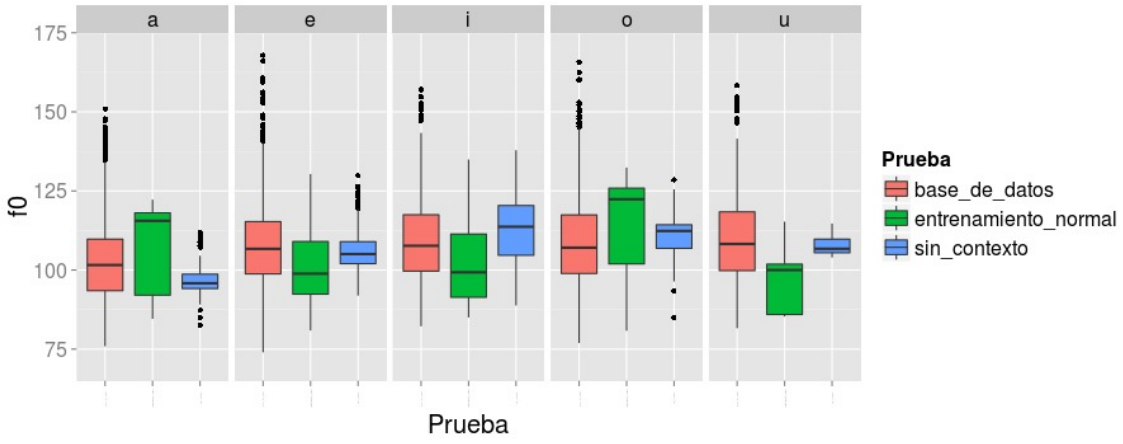


Figura B.25: Diagramas de caja para el tono de vocales según información de contexto. Voz masculina, aplicación Reloj

Las variaciones en los parámetros son dependientes del género de la voz y de la aplicación, es decir, de la longitud de frases y del vocabulario involucrado.

B.3. Pruebas sobre la similitud en parámetros espectrales y de frecuencia fundamental

En las Figuras B.25 a B.26 se presentan los diagramas de caja de los tonos de las vocales para las pruebas relacionadas con la influencia de la información de contexto, en las voces masculinas.

Las pruebas de Friedman realizadas para determinar diferencias estadísticamente significativas en estas pruebas sobre la influencia de la información de contexto, en voces masculinas, tienen como resultados:

- Aplicación Reloj: No hay diferencia significativa en el grupo de datos de tono de vocales según información de contexto (p-valor 0.32).

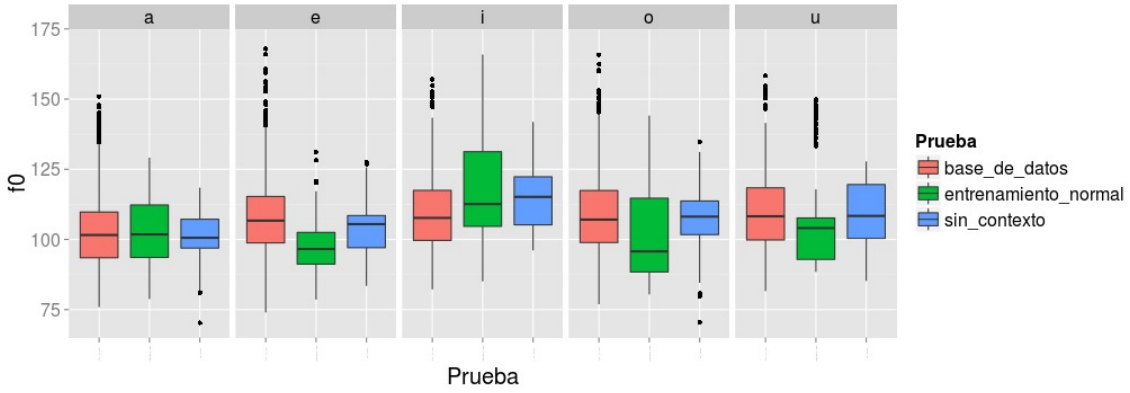


Figura B.26: Diagramas de caja para el tono de vocales según información de contexto. Voz masculina, aplicación clima

- Aplicación Clima: hay diferencia significativa en el grupo de datos de tono de vocales según información de contexto (p-valor $1.045e-10$), el cual, en análisis *Post-hoc* se determina que proviene de la diferencia significativa entre el entrenamiento normal y el hablante original (p-valor de $3.6e-11$). La evidencia no permite descartar la similitud del hablante original con la voz obtenida con contexto reducido.

Para el caso de tono de vocales en el caso de las voces femeninas, los resultados se muestran en las Figuras B.27 y B.28.

Las pruebas de significancia para el caso de tonos de vocales para hablante mujer, relacionadas con las voces obtenidas con información de contexto reducida son:

- Aplicación Reloj: hay diferencia significativa en el grupo de datos de tono de vocales (p-valor $6.14e-14$), lo cual se debe, de acuerdo con el análisis *Post-hoc*, a diferencias entre todos los grupos de datos.

- Aplicación Clima: hay diferencia significativa en el grupo de datos de tono de vocales (p-valor $< 2.2e-16$), el cual, en análisis *Post-hoc*, a diferencias entre todos los grupos

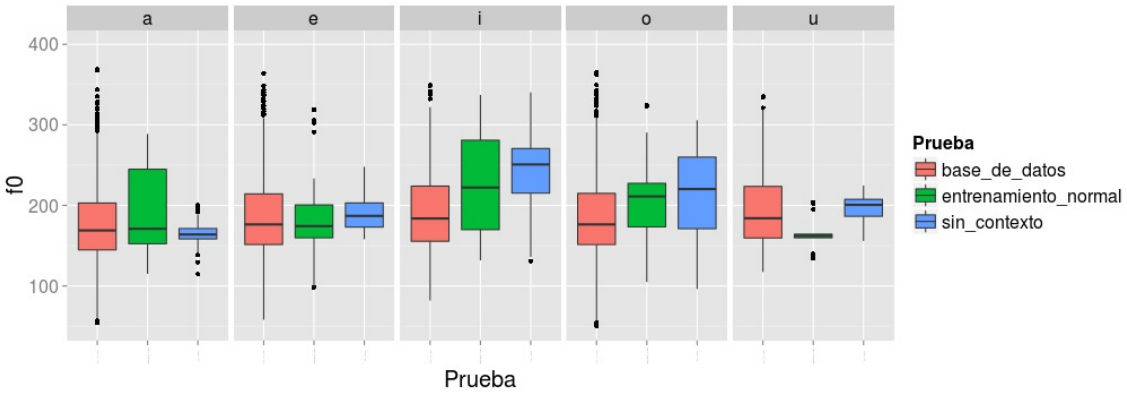


Figura B.27: Diagramas de caja para el tono de vocales según información de contexto. Voz femenina, aplicación Reloj

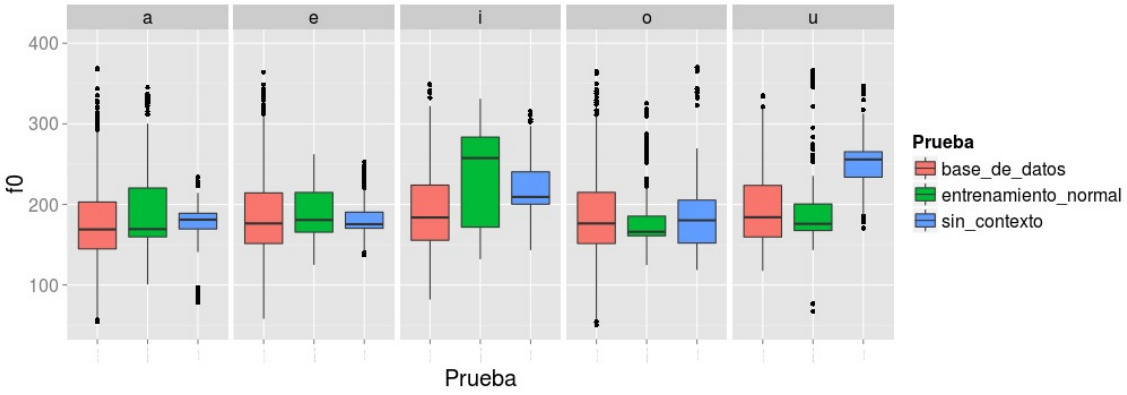


Figura B.28: Diagramas de caja para el tono de vocales según información de contexto. Voz femenina, aplicación Clima

de datos.

La prueba determina diferencias estadísticamente significativas de tono en más ocasiones con el entrenamiento normal que con el contexto reducido.

Por su parte, el *jitter* de las vocales se muestra como diagramas de caja en las Figuras B.29 y B.30 para el caso de las voces masculinas.

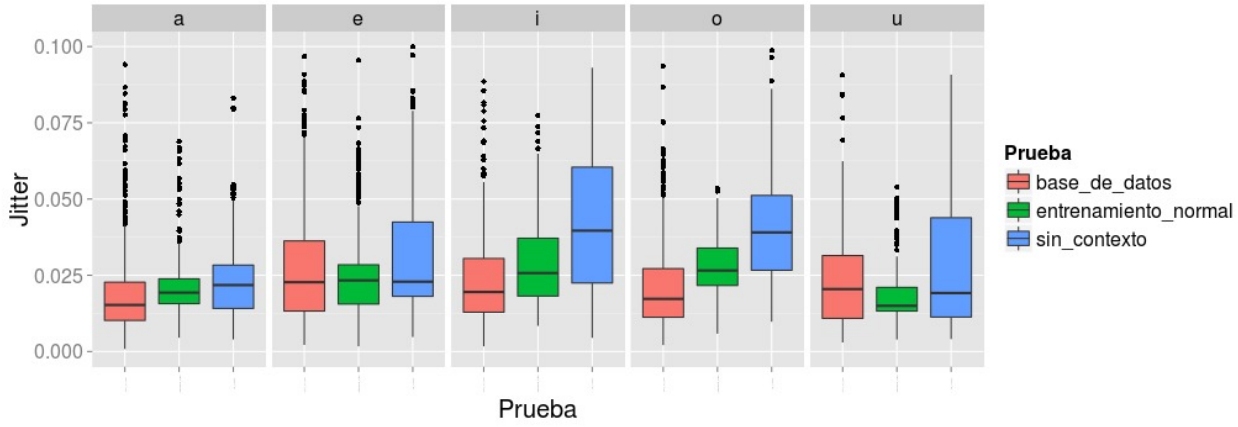


Figura B.29: Diagramas de caja para el *jitter* de vocales según condición de entrenamiento. Voz masculina, aplicación Clima

Las pruebas de significancia para el caso de *jitter* de vocales para voces de hombre, relacionadas con las voces obtenidas con información de contexto reducida son:

- Aplicación Reloj: hay diferencia significativa en el grupo de datos de *jitter* (p-valor $< 2.2e-16$), lo cual se debe, de acuerdo con el análisis *Post-hoc*, a diferencias entre el hablante original y la voz sintética obtenida con información de contexto reducida (p-valor de 0).

- Aplicación Clima: hay diferencia significativa en el grupo de datos de *jitter* (p-valor $< 2.2e-16$), el cual, en análisis *Post-hoc*, a diferencias entre todos los grupos de datos.

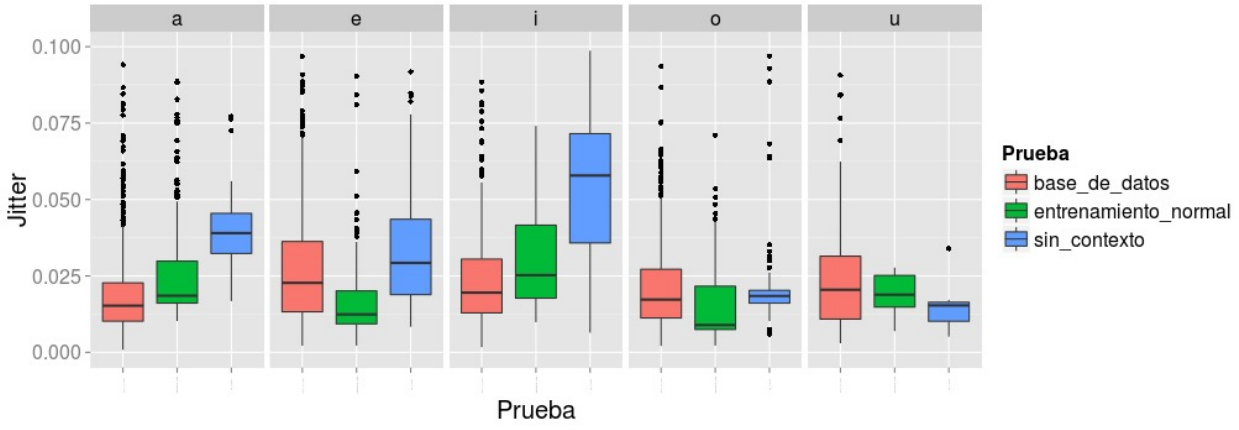


Figura B.30: Diagramas de caja para el *jitter* de vocales según condición de entrenamiento. Voz masculina, aplicación Reloj

Para el caso de voces femeninas, los resultados de *jitter* de las pruebas relacionadas con la información de contexto se presentan en las Figuras B.31 y B.32.

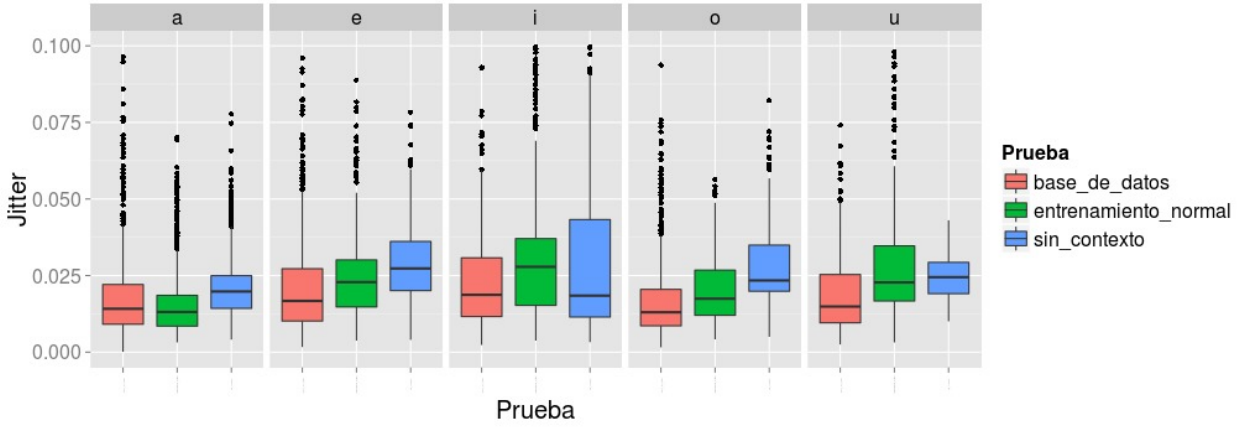


Figura B.31: Diagramas de caja para el *jitter* de vocales según condición de entrenamiento. Voz femenina, aplicación Clima

Las pruebas de diferencias significativas realizadas sobre estos datos de *jitter* en tono de vocales de voces femeninas relacionadas con la influencia de información de contexto tienen

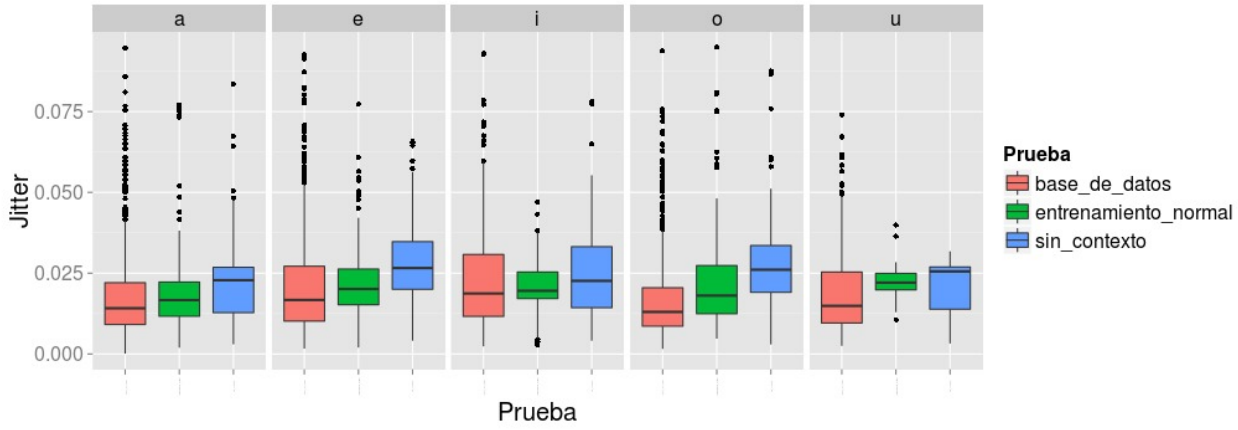


Figura B.32: Diagramas de caja para el *jitter* de vocales según condición de entrenamiento. Voz femenina, aplicación Reloj

como resultados:

- Aplicación Reloj: hay diferencia significativa en el grupo de datos de *jitter* de vocales (p-valor $< 2.2e-16$), lo cual se debe, de acuerdo con el análisis *Post-hoc*, a diferencias entre todos los grupos de datos.

- Aplicación Clima: hay diferencia significativa en el grupo de datos de *jitter* de vocales (p-valor $< 2.2e-16$), el cual, en análisis *Post-hoc*, a diferencias entre todos los grupos de datos.

En cuanto al análisis de *shimmer* en pruebas relacionadas con información de contexto, los resultados para voces masculinas se presentan en los diagramas de caja de las Figuras B.33 y B.34

Las pruebas de Friedman realizadas para determinar diferencias significativas entre estos grupos de datos de *shimmer* en vocales de voces masculinas relacionadas con información de contexto tienen como resultados:

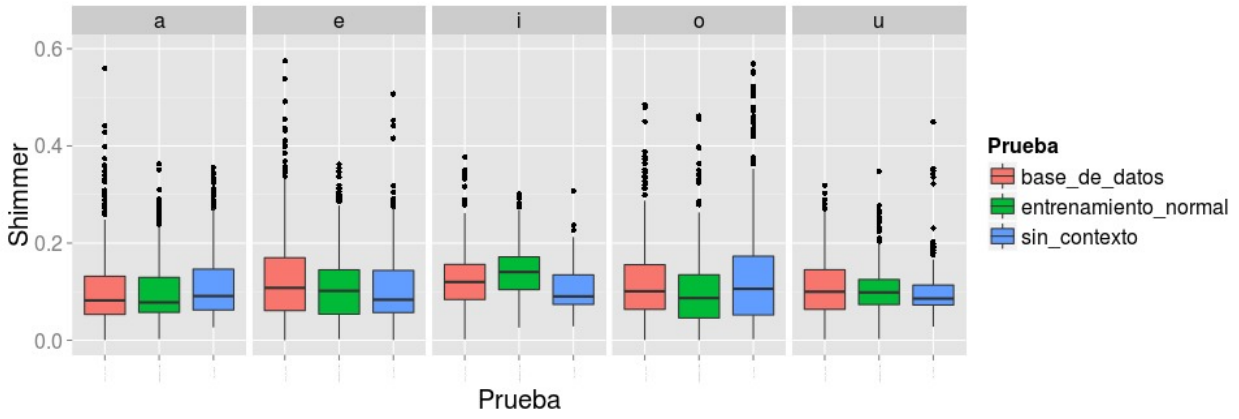


Figura B.33: Diagramas de caja para el *shimmer* de vocales según condición de entrenamiento. Voz masculina, aplicación Clima

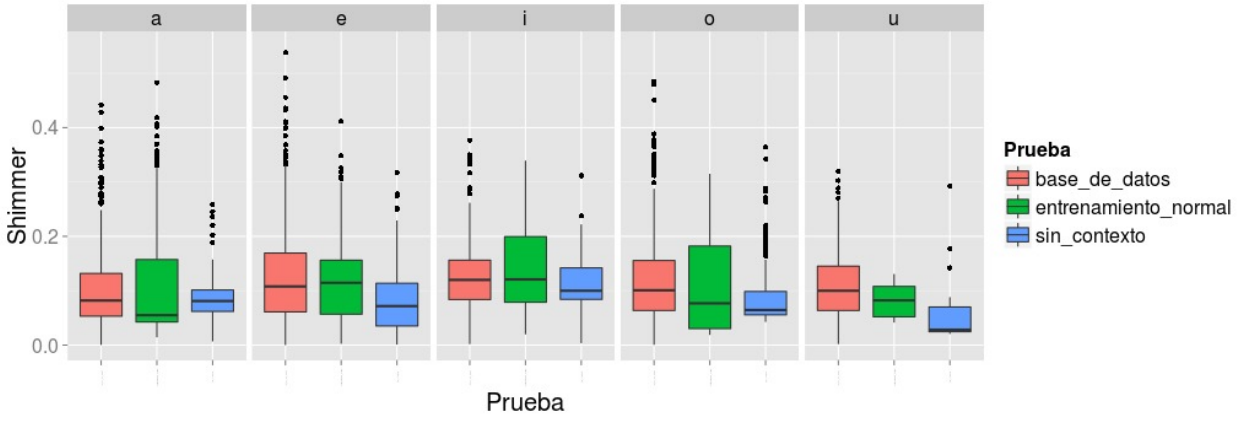


Figura B.34: Diagramas de caja para el *shimmer* de vocales según condición de entrenamiento. Voz masculina, aplicación Reloj

- Aplicación Reloj: hay diferencia significativa en el grupo de datos de *shimmer* (p-valor $< 2.2e-16$), lo cual se debe, de acuerdo con el análisis *Post-hoc*, a diferencias entre la voz original y la obtenida con información reducida de contexto (p-valor $6e-4$).

- Aplicación Clima: hay diferencia significativa en el grupo de datos de *shimmer* (p-valor

0.002), el cual, en análisis *Post-hoc*, a diferencias entre la voz original y la obtenida con información reducida de contexto (p-valor 0.002).

En el caso de las voces femeninas, los datos de shimmer en vocales, en pruebas sobre la influencia de la información de contexto se muestran en los diagramas de caja de las figuras B.35 y B.36.

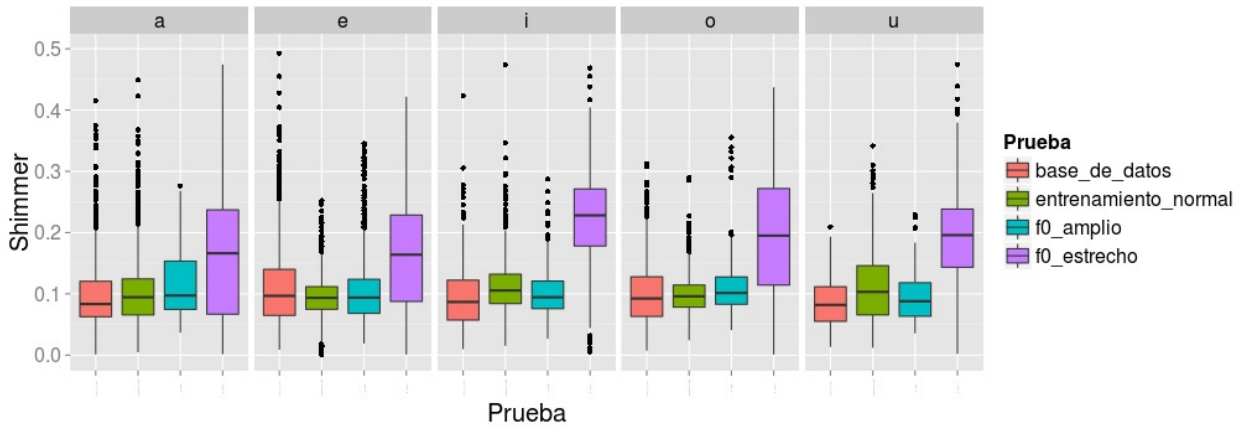


Figura B.35: Diagramas de caja para el *shimmer* de vocales según condición de entrenamiento. Voz femenina, aplicación Clima

El análisis de significancia en las diferencias de shimmer para las voces femeninas en pruebas relacionadas con la información de contexto tiene como resultados:

- Aplicación Reloj: hay diferencia significativa en el grupo de datos de *shimmer* (p-valor $< 2.2e-16$), lo cual se debe, de acuerdo con el análisis *Post-hoc*, a diferencias entre la voz original y la obtenida con información reducida de contexto (p-valor $1.6e-5$), así como con la voz obtenida con entrenamiento normal (p-valor $1.1e-6$)

- Aplicación Clima: hay diferencia significativa en el grupo de datos de *shimmer* (p-valor 0.002), el cual, en análisis *Post-hoc*, a diferencias entre la voz original y la obtenida con

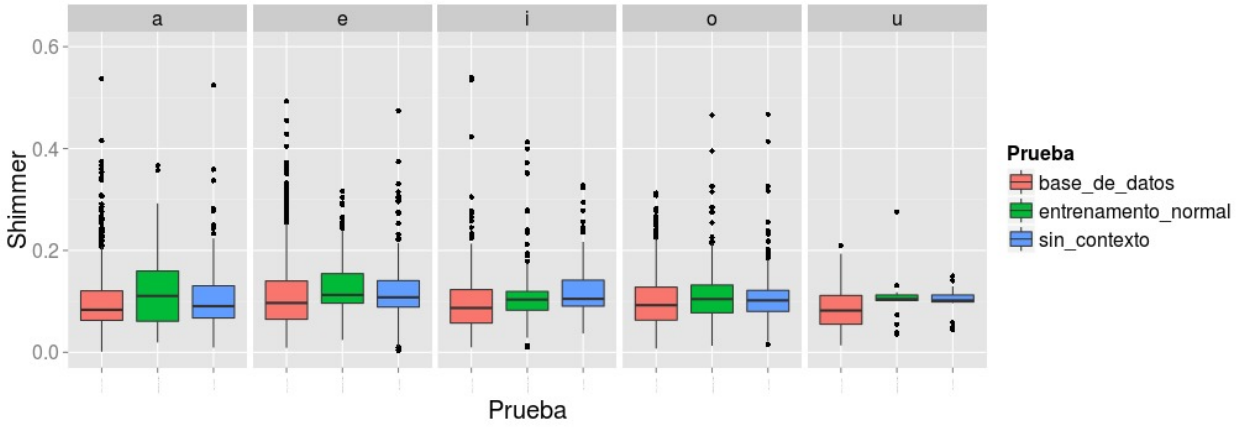


Figura B.36: Diagramas de caja para el *shimmer* de vocales según condición de entrenamiento. Voz femenina, aplicación Reloj

información reducida de contexto (p-valor $4.2e-06$) y con la voz sintética obtenida con entrenamiento normal (p-valor de $4.2e-04$).

Apéndice C

Implementaciones

Existen en la actualidad gran cantidad de implementaciones de sintetizadores en voz en software y diversos dispositivos electrónicos. Se destacan en esta sección algunos de los más utilizados, de acuerdo con los presentados por [107], [108], que adicionalmente tienen actualización a partir del año 2008. Los primeros corresponden a software de desarrollo, con licencia que permite su uso para implementar nuevas voces, y los últimos corresponden a sintetizadores comerciales que incluyen voces en español.

C.1. Festival

Desarrollado en el Centro de Investigación de Tecnología del Habla CSTR (*Centre for Speech Technology Research*) de la Universidad de Edinburgo, Inglaterra. Su licencia de uso es gratuita, y la versión más reciente incluye una implementación del sintetizador HTS, el cual está basado en HMM [109].

Sus rutinas pueden utilizarse para procesamiento del habla, extracción de características, contornos espectrales, etiquetado, entre otros. Su compatibilidad permite agregar más voces y lenguajes [107].

C.2. eSpeak

De forma semejante al anterior, su licencia permite distribución gratuita, y es de código abierto. Utiliza síntesis de formantes, y contiene implementados varios idiomas, con diverso grado de calidad [110].

C.3. MBROLA

MBROLA fue creado en Bélgica, en la *Faculté Polytechnique de Mons*, en 1996. La síntesis está basada en concatenación de difonos. Tiene como entradas una lista de éstas, junto con información prosódica, y produce habla de 16 bits [107]. La calidad de voz obtenida es más cercana al lenguaje natural de la que producen otros, como Festival. Su licencia de uso es gratuita para aplicaciones no comerciales y no militares [111].

C.4. HTS

El sistema de síntesis por HMM, HTS (por las siglas en inglés de *HMM–based Speech Synthesis System* es distribuido bajo una licencia libre. Es desarrollado por el grupo HTS (*HTS working group*), y como su nombre lo indica, está hecho para desarrollar síntesis de voz con HMM. No incluye analizador de texto, por lo que se indica que debe utilizarse en conjunto con otros programas que sí lo utilicen, como Festival o más directamente la herramienta *hts_engine* [25].

C.5. AT&T Natural Voices ®

Desarrollado por la empresa estadounidenes AT&T, utiliza una base de datos de alta calidad, la cual es etiquetada y se utiliza la técnica de concatenación para convertir nuevos textos [112]. Incluye demostración de conversión texto a voz en su sitio web, con voces en español latinoamericano.

C.6. Cepstral®

Desarrollado en Estados Unidos de América, en su sitio web [113] existe la posibilidad de realizar demostraciones de conversión texto a habla, con control del tono, velocidad y algunos efectos. Implementa dos voces en español americano.

C.7. CereProc®

Desarrollado en Escocia, utiliza síntesis concatenativa y basada en HMM para la producción de voces. En su sitio web, [114] incluye demostración de voz en español.

C.8. Loquendo®

Actualmente propiedad de la compañía Nuance, implementa gran cantidad de voces, con una codificación propia que permite incluir elementos como risas, llanto y otras expresiones de emociones en el convertir texto a habla [115]. Incluye varias voces en español, incluías de español de México, Colombia y Argentina.

C.9. IVONA®

Es propiedad de la compañía polaca IVONA Software. Utiliza una técnica basada en concatenación, con la cualimplementa voces en varios idiomas, incluyendo español castellano [116].

C.10. Verbio®

Es propiedad de la empresa Verbio Speech Technologies, y está implementado como una serie de librerías y utilidades para implementarse en diferentes aplicaciones. Incluye voces en español mexicano.

Apéndice D

Implementación de voces con HTS

D.1. Introducción

HTS es una herramienta de software desarrollada por el Grupo HTS, la cual utiliza un conjunto de rutinas y programas para la extracción de parámetros, entrenamiento de HMM y creación de archivos de configuración y registro, a partir de los cuales es posible realizar la síntesis estadística paramétrica. Sus rutinas principales se encuentran programadas en lenguaje Perl y C. Fue creado como una extensión del sistema HTK, utilizado originalmente para reconocimiento del habla. Tiene licencia libre, pero al utilizarse como extensión de HTK hereda su licencia.

Se presenta aquí la descripción y requerimientos para la creación de voces utilizando este sistema, a partir de lo definido para el idioma inglés[1]. La adaptación de se basa en las definiciones particulares del idioma y el uso de archivos de entrada específicos para crear nuevas voces.

[1] `http://hts.sp.nitech.ac.jp/archives/2.2/HTS-demo_CMU-ARCTIC-SLT.tar.bz2`

D.2. Generalidades

Un proyecto de HTS requiere una serie de archivos de entrada y de requerimientos, entre los que destacan los archivos de datos (audio y texto), en formatos específicos, de los cuales se extrae la información necesaria para la definición y entrenamiento de los HMM. El sistema tiene como salida archivos que pueden utilizarse en conjunto con el programa Festival para sintetizar nuevas frases, o bien audios que corresponden a texto definido como parte del entrenamiento.

El proceso general puede describirse como como los pasos de extracción de información, entrenamiento y síntesis de nuevas frases. El conjunto de programas involucrados en estos procedimientos en proyecto de síntesis de voz son:

- HTK [76], (siglas en inglés de *Hidden Markov Model Toolkit*, Herramienta de modelos ocultos de Markov), es desarrollada en la Universidad de Cambridge, Inglaterra. Este se utiliza para inicializar los modelos HMM y los algoritmos Baum-Welch de re–estimación de sus parámetros.

- SPTK [117] (siglas en inglés de *Speech Signal Processing Toolkit*, Herramienta de procesamiento de señales de habla), para definir ventanas y extraer parámetros espectrales.

- ActiveTcl [118], como intérprete de rutinas utilizadas en la extracción de parámetros.

- hts_engine [119], para generar ondas sonoras a partir de la información de duración, de $f0$ y los MFCC en la síntesis de nuevas frases.

- Festival [109], para analizar el texto y procesarlo, de manera que puedan identificarse los HMM que correspondan a los fonemas contenidos en las frases que se desean sintetizar.

Tabla D.1: Parámetros en el archivo de configuración config

Línea	Parámetro	Descripción
405	FRAMELEN	Longitud de ventana
409	FRAMESHIFT	Desplazamiento de ventana
425	SAMPFREQ	Frecuencia de muestreo de los audios en la base de datos
459	LOWERF0	Límite inferior de frecuencia fundamental de los audios en la base de datos
463	UPPERF0	Límite superior de frecuencia fundamental de los audios en la base de datos
545	NSTATE	Número de estados de los HMM
549	NITER	Número de iteraciones en entrenamiento

Se utiliza en conjunto con rutinas del programa *awk*, para la creación de archivos de etiquetas (.lab) que contienen la descripción de los fonemas y su contexto.

La estructura inicial de archivos, carpetas y rutinas se muestra en la Figura D.1. Los componentes de ésta son:

- Makefile.in: Intrucciones para ejecutar todo el proceso de entrenamiento y síntesis de frases.
- configure: Archivo de configuración, el cual contiene la información requerida para verificar que se cuenta con todos los programas, rutinas, parámetros y datos necesarios para ejecutar la síntesis.
- configre.ac: Archivo utilizado como entrada del configure, el cual contiene los principales parámetros involucrados en la extracción de la información y el entrenamiento de HMM, algunos de los cuales se detallan el cuadro D.1.
- Makefile.in (Data/): Contiene las instrucciones para la extracción de parámetros y la ejecución de las rutinas principales. Es llamado desde el archivo Makefile ubicado en el

directorio raíz

- labels/: Directorio que contiene inicialmente los textos que se desean sintetizar como parte del proyecto.
- questions/: Archivos donde se definen las preguntas que constituyen los árboles de decisión utilizados para agrupar los HMM, y definir los contextos que se utilizarán en el análisis del texto para sintetizar nuevas frases.
- raw/: Archivos de audio de la base de datos, en formato de audio sin encabezado
- scripts/: Archivos principales de ejecución de los algoritmos de entrenamiento, en lenguaje Perl. También contiene archivos varios en formato *awk* para la transformación de archivos de texto en archivos de etiquetas.
- utts/: Archivos de texto de la base de datos. Contienen el texto y la descripción fonética y silábica, de manera que puedan ser convertidos en archivos de etiquetas (.lab) como parte del proceso. Para producir estos archivos a partir del texto, se requiere ejecutar en Festival comandos de la forma

  ```
  (utt.save(SynthText"<texto>")"archivo_salida.utt")
  ```

  Es importante mencionar que se requiere que los archivos de audio y texto tengan un nombre raíz que coincida en ambos, seguido de un número de secuencia.
- win/: Archivos con parámetros del tipo, tamaño y desplazamiento de las ventanas a utilizar para la extracción de $f0$ y MFCC.

El primer paso es fijar las características base de todos los parámetros a modelar, agrupados de acuerdo con los tres modelos matemáticos por entrenar: de espectro (mgc), de

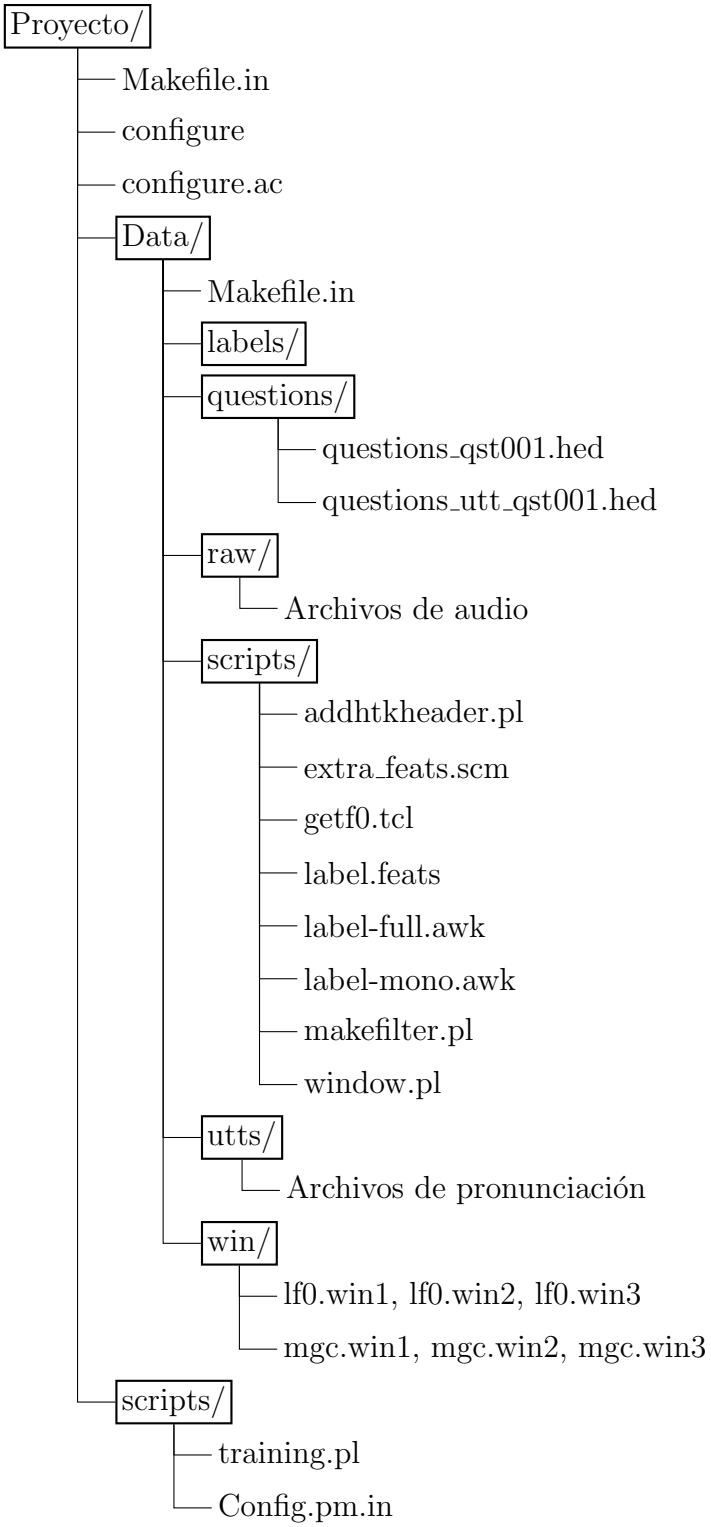


Figura D.1: Esquema de carpetas y archivos de un proyecto HTS

Tabla D.2: Ajustes de parámetros de entrenamiento

Parámetro	Valor HMM		
	mgc	lf0	dur
Límite inferior de varianza	0.01	0.01	0.01
Parámetro de control para poda de árboles de decisión con MDL	1	1	1
Inicio de trama	1	2	1
Fin de trama	1	4	5
Información MSD	0	1	0
Dimensión de características	35	1	5
Número de ventanas	3	3	0
Ganancia mínima de probabilidad para GV	0	0	–
Control de tamaño de árbol para GV	1	1	–

frecuencia fundamental (lf0, pues se trabaja con el logaritmo de $f0$) y duración (dur). Esto se realiza en el archivo *config.pm*, escrito en el lenguaje de programación Perl, y se ajustó de la manera que se describe en la Tabla D.2 partiendo de valores preestablecidos.

Los parámetros de la Tabla D.2 se justifican de la siguiente manera:

- El límite inferior de varianza se establece para impedir que los HMM ajusten dentro de límites muy estrechos los coeficientes con la representación paramétrica del habla.

- El parámetro de poda de árboles de decisión con MDL se refiere a la utilización de este algoritmo para la definición de árboles de decisión de acuerdo con los datos disponibles [95].

- El inicio y fin de trama se refiere a la identificación del bloque con la información correspondiente al espectro, frecuencia fundamental o duración. La Figura D.2 esquematiza la identificación de flujos en la salida de un estado en un HMM.

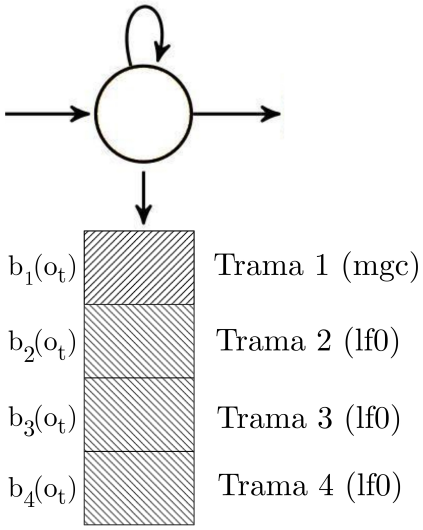


Figura D.2: Esquema de identificación de tramas a la salida de un HMM

- Información MSD es un valor binario que se refiere a la utilización de distribuciones multiestado, las cuales deben estar presentes únicamente en lo referente a $f0$.
- La dimensión de características es la cantidad de coeficientes que se manejan por modelo matemático. En este caso se utilizan 35 coeficientes MFCC, 1 coeficiente de lf0 y cinco de duración (1 por estado de HMM).
- Los parámetros de GV son valores discretos, declarados en aquellos modelos que utilizan el algoritmo GV: los referentes al espectro y a lf0.

Adicionalmente es necesario establecer cuatro grupos de parámetros relacionados con los procesos de análisis del audio, la extracción de características y la síntesis.

1. Ajustes para análisis del habla:
    - Frecuencia de muestreo *sr*: Se define de acuerdo con las características del audio presente en la base de datos.

- Periodo de análisis fs: Intervalo temporal en que se realizará el procesamiento de la señal.
- Deformación de frecuencia fw: Coeficiente de la escala de transformación de frecuencia, para ajustarla a la escala de percepción auditiva humana.
- Peso de la representación polo/cero gm: Valor discreto para indicar si se realizará análisis MGC (0) o Codificación predictiva lineal LPC (por las siglas en inglés de *Linear Predictive Coding)* (1).
- Uso de ganancia logarítmica lg: Valor discreto para indicar si se utilizará escala lineal o logarítmica de ganancias.
- Periodo de trama fr: Cociente $\frac{fs}{sr}$.

Estos parámetros de análisis no son independientes. Por ejemplo, para una base de datos en formato WAV, con frecuencia de muestreo 48000 Hz, se utiliza $sr = 48000$, $fs = 240$, $fw = 0.55$, $gm = 0$, $lg = 1$, $fr = \frac{fs}{sr} = 0.005$.

2. Ajustes para síntesis del habla:

- Factor post filtro pf: Utilización de un filtro para acentuar el modelado espectral, de manera que la dinámica entre picos y valles del espectro se incremente [120]. Este parámetro se ajustó en 1.4, valor recomendado para las características de sr del audio.
- Longitud de respuesta al impulso fl: Factor fijado para establecer un parámetro finito de respuesta al impulso. Se ajusta en 4096.
- Orden del cepstrum para aproximar el mel cepstrum generalizado co: Se ajusta el valor por omisión en 2047.

3. Ajustes del modelado: Se presentan con los valores por omisión.

    - Número de estados de los HMM: 5. Este es un valor fijo para todos los modelos.
    - Número de iteraciones para el entrenamiento incrustado: 5
    - Valor inicial, incremento y valor final del ancho de beam, para establecer un criterio dinámico de poda de árboles de decisión: 1500,100,5000
    - Máxima desviación estándar para controlar la duración máxima del HSMM: 10
    - Duración mínima de estado a ser evaluada: 5
    - Techo del peso de la mezcla: 5000
    - Uso de algoritmo DAEM para estimación de parámetros: 0 (no utilizado). Este algoritmo es una variante determinística de EM.
    - Número de iteraciones para el entrenamiento basado en algoritmo DAEM: 10
    - Calendarización de la actualización de temperatura para algoritmo DAEM: 1

4. Ajustes de la generación:

    - Cantidad máximas de iteraciones EM: 20
    - Factor de convergencia para las iteraciones EM: 0.0001
    - Uso de GV: 1 (activado)
    - Cantidad máxima de iteraciones GV: 50
    - Factor de convergencia para iteraciones GV: 0.0001
    - Mínima norma euclídea para iteraciones GV: 0.01
    - Tamaño de paso inicial: 1.0

- Factor de aceleración de paso: 1.2
- Factor de desaceleración de paso: 0.5
- Peso para la probabilidad de emisión del HMM: 1.0
- Peso para la probabilidad de emisión del GV: 1.0
- Método de optimización: Newton (Steepest, Newton o LBFGS)
- GV sin fonema de silencio y pausa: 1 (activado)
- GV dependientes de contexto: 1 (activado)

Finalmente, el archivo requiere el establecimiento de directorios de ubicación de los archivos base y la lista de comandos de los programas: HTS, SPTK, SoX.

Los parámetros definidos en *config.pl* se utilizan como entrada del programa de entrenamiento principal, llamado *training.pl*, el cual está escrito en lenguaje Perl, y es llamado desde un *Makefile* en el sistema operativo Linux. El entrenamiento puede analizarse como un proceso en dos etapas, el primero para preparación de la estructura de directorios y archivos necesarios para los múltiples pasos que requiere el ajuste, y el segundo para definir y ajustar propiamente los parámetros. Cada uno de estas etapas involucran una o más subrutinas de programa. Los pasos involucrados en ambas, hasta la creación de los archivos necesarios se presentan en la Figura D.3, y se describen:

1. Inicialización

   - Inicialización de estructura del modelo: Se establece el tamaño del vector a utilizar para representar el habla, cantidad de flujos de datos y número de distribuciones de probabilidad. Todos estos inicialmente en cero.

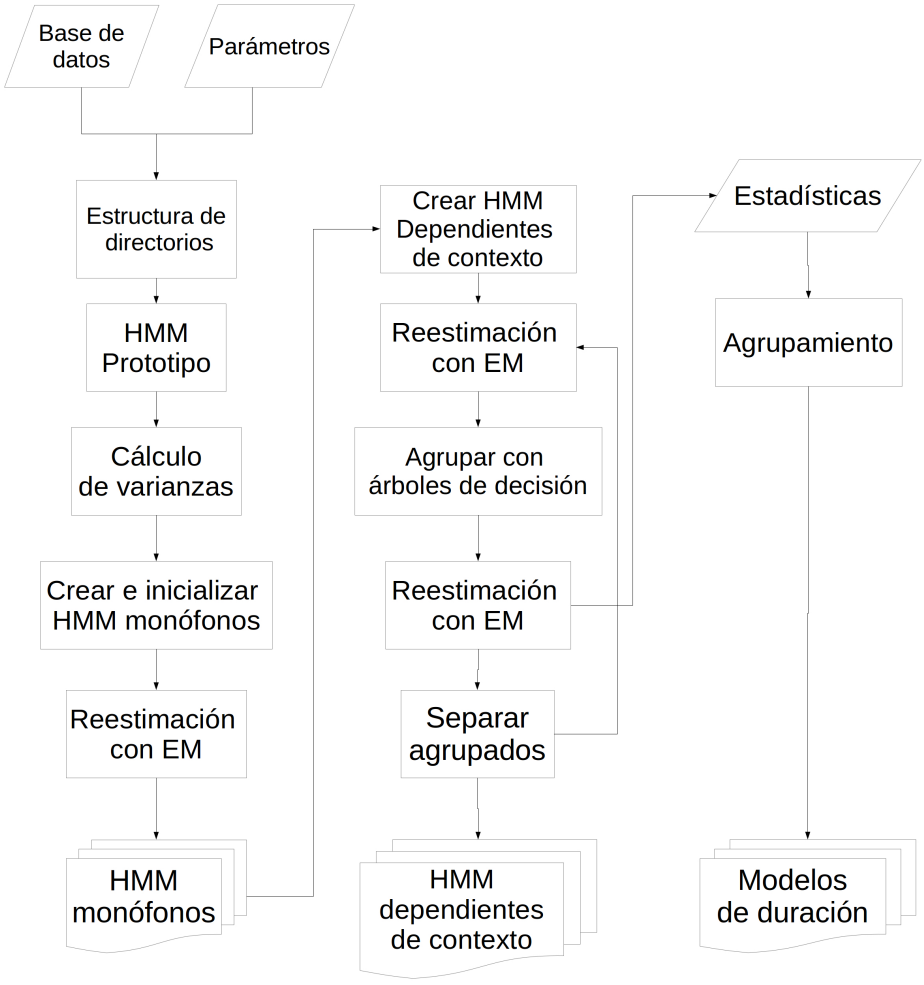


Figura D.3: Esquema del proceso de entrenamiento

- Establecer archivo de ubicaciones de datos: Son archivos de listas. Los de entrenamiento en */data/scp/train.scp* y los de síntesis en */data/scp/gen.scp*. Estos archivos, en formato texto, contienen la dirección (ubicación) de archivos con los modelos a extraer durante el entrenamiento, y a sintetizar en la parte final del proceso.

- Establecer archivos de listas de modelos: Son tres archivos en formato texto con el listado de unidades fonéticas a utilizar por cada HMM, en este caso los fonemas. El primero es */model/mono.list*, con el listado simple de fonemas (monofonemas), el $model/full.list$ con el listado de los dependientes de contexto, y $/model/all.list$ con una combinación de ambos. Para crear estos archivos se recurre a la definición de fonemas o unidades fonéticas declaradas en el programa Festival, y a los archivos de preguntas para la creación de árboles de decisión definidas en la carpeta */questions*.

- Establecer archivos maestros de etiquetas: Son dos archivos en formato texto, ubicados en $/model/mono.mlf$ y $/model/full.mlf$, los cuales contienen la información de segmentación de la base de datos, con la unidad fonética y la marca de tiempo de inicio y final en cada una de las frases.

- Establecer archivos de configuración de variables: Son archivos en formato texto con los valores de los parámetros involucrados en el entrenamiento, distribuidos entre ocho archivos, a utilizarse en diferentes momentos del proceso.

- Establecer archivo de definición de prototipo: Es un archivo en formato texto, ubicado en */proto/*, con extensión prt, el cual contiene la estructura base de los HMM: la cantidad de estados y la inicialización de vectores de media y varianza de las distribuciones de probabilidad de cada uno. La estructura de este archivo

es semejante a la utilizada en HTK para los modelos de HMM.

- Establecer archivos de modelos: Estos son los archivos que contienen la información de los HMM, en la carpeta */models/*, los cuales se van ajustando a lo largo del entrenamiento, creados a partir de la estructura prototipo. Su extensión es mmf, y son archivos binarios cuya exploración requiere la conversión a formato texto utilizando la herramienta HHEd de HTK.
- Establecer archivos de estadísticas: Son archivos en formato texto con información de las medias de duraciones en cada uno de los estados de los HMM, contenidos en la carpeta */stats*. Estos son utilizados para estimar los modelos de duración a partir de las duraciones en los estados de HMM.
- Direccionar archivos de preguntas sobre contextos: Contenidos en la carpeta */questions*, y utilizados para crear los árboles de decisión para agrupamiento de modelos.
- Establecer archivos de árboles de decisión: Son archivos en formato texto, y estructura de preguntas utilizadas por HTK para los árboles de decisión, contenidos en la carpeta */trees*.
- Establecer archivos de ventanas para la extracción de parámetros: Archivos en formato texto con coeficientes necesarios para la definición de las ventanas y su desplazamiento, tanto para los archivos de espectro como de frecuencia fundamental. Su extensión es win1; win2 y win3, y están contenidos en la carpeta */win*.
- Establecer archivos GV: Son archivos de texto con la información de segmentación, utilizados como parte del proceso de Varianza Global (GV). Están contenidos en la carpeta */gv*

2. Entrenamiento

- Crear estructura de directorios: A partir de lo indicado en la inicialización, ejecuta los comandos de sistema para creación de toda la estructura de carpetas.
- Crear archivos de configuración de variables, de acuerdo con la descripción dada anteriormente. Utiliza la subrutina *make_config()*, declarada en el mismo programa.
- Crear archivo de prototipo: Utilizando la subrutina *make_proto()*, crea el archivo descrito anteriormente con la estructura de los HMM.
- Calcular límites de varianzas: Utilizando el comando HCompV de HTK, estima un modelo promedio y determina límites inferiores de varianzas, para evitar que sean muy pequeñas.
- Crear HMM monófonos: Con el comando HHEd de HTK crea los primerios archivos con extensión mmf, correspondientes a los HMM de las unidades fonéticas, sin tomar en cuenta su contexto.
- Inicializar y reestimar: Con los comandos HInit y HRest de HTK realiza el entrenamiento inicial de los HMM.
- Crear HMM dependientes de contexto: Con el comando HHEd copia los archivos de HMM monófonos para crear nuevos archivos, de acuerdo con lo establecido en las listas de unidades fonéticas dependientes de contexto.
- Reestimar HMM dependientes de contexto: Con el comando HERest realiza entrenamiento de los HMM tomando en cuenta su contexto.
- Agrupar: Utiliza árboles de decisión para agrupar los HMM que pueden asemejarse, para estimar de forma conjunta sus parámetros.
- Reestimar HMM agrupados: Realiza ajuste de parámetros en los HMM agrupados.

- Separar estructuras agrupadas: Separa los HMM agrupados para realizar nuevamente entrenamiento de HMM.
- Reestimar nuevamente HMM dependientes de contexto.
- Agrupar nuevamente, de acuerdo con contextos.
- Alinear HSMM: Con el comando HSMMAlign, realiza alineamiento forzado de HMM del tipo semi-markoviano, lo cual es base para la estimación de los modelos de duración.
- Realizar modelos de Varianza Global (GV): Realiza un proceso de creación de prototipos, inicialización y entrenamiento de modelos para la estimación de la GV.
- Creación de modelos ocultos: Utiliza GV para crear los modelos, de los cuales se extraen parámetros y se realiza una síntesis de audio a partir de texto en formato de etiquetas, con el programa HMGenS.
- Crear archivos en formato de *hts_engine*: Estos archivos, creados a partir de los modelos mmf, corresponden al conjunto de archivos con el formato requerido para su utilización en síntesis de voz, propiamente con *hts_engine*, o bien con *Festival*.

Como se puede observar en esta descripción, el utilizar HTS a partir de su código permite el control de cada etapa y cada parámetros involucrado en los múltiples procesos. Esto permite un ajuste muy detallado de la implementación, y constituye un vasto espacio de experimentación para estudiar la influencia que los algoritos, definiciones y parámetros tiene en las voces resultantes.

Una vez definida la configuración en los archivos mencionados, para iniciar la construcción del proyecto de nuevas voces, se requiere ejecutar ./configure con las opciones adecuadas para

direccionar la ubicación de los demás programas involucrados en el proceso. Por ejemplo, se ejecuta en el directorio del proyecto:

```
./configure  --with-tcl-search-path=/opt/ActiveTcl-8.4/bin
--with-fest-search-path=/usr/share/doc/festival/examples/
--with-sptk-search-path=/home/usuario/HTS/SPTK-3.4.1/bin/*
--with-hts-search-path=/home/usuario/HTS/htk/HTKTools
--with-hts-engine-search-path=/home/usuario/HTS/hts_engine_API-1.07/bin/
```

Con esto se logra una verificación de la disponibilidad y correcto funcionamiento de los mismos. Las fases involucradas a continuación se describen en las siguiente secciones.

D.3. Entrenamiento

Si la verificación de programas tiene resultado positivo, debe terminar con la confirmación de los archivos que realizarán la extracción de información y entrenamiento, con el mensaje:

```
configure: creating ./config.status
config.status: creating data/Makefile
config.status: creating scripts/Config.pm
config.status: creating Makefile
```

La extracción de información, y entrenamiento de los HMM se realiza con la instrucción make, la cual ejecuta en secuencia los procesos de extracción de parámetros, definición de HMM, entrenamiento y síntesis de nuevas frases. Esta instrucción tendrá como resultado la siguiente secuencia de acciones:

1. Extracción de coeficientes MCG: Con llamadas al programa SPTK, y los parámetros establecidos en el archivo de configuración sobre el tipo y tamaño de ventana, extrae los coeficientes MFCC de los archivos de audio.

2. Extracción de coeficientes lf0: Con llamadas al programa SPTK y rutinas Tcl extrae el logaritmo de la frecuencia fundamental de los archivos de audio, segmentados en ventanas.

3. Creación de vectores de entrenamiento: En este paso se identifican y unen los datos para el proceso de entrenamiento de los HMM, los cuales consisten en información consolidada de MFCC y logf0 extraído de ventanas de cada uno de los archivos de audio.

4. Creación de archivos de etiquetas: Los archivos de etiquetas contienen la información fonética y de contexto de cada unas de las frases de la base de datos, lo cual se utiliza para conformar los HMM, pues a cada fonema en su contexto se le hace corresponder un HMM. Estos archivos son creados a partir de los archivos con extensión utt, utilizando las rutinas de *awk* contenidas en *scripts/*

5. Creación de listas y rutinas: Se realizan listados de los archivos de audio y etiquetas, estas últimas diferenciadas según la etapa del proceso (etiquetas de fonemas individuales o etiquetas de fonemas dependientes de contexto).

6. Ejecutar rutina principal de entrenamiento: Se ejecuta la rutina principal en lenguaje Perl que define los HMM de acuerdo con los parámetros del archivo de configuración, para proceder a realizar el entrenamiento de los mismos con el algoritmo de Baum-Welch y la información extraída de los archivos de audio. Se realiza como parte del proceso el agrupamiento de los fonemas para estimar de forma conjunta sus parámetros, y se

llevan estadísticas de duración promedio en cada estado para establecer los modelos de duración por fonema.

D.4. Resultados

Los procesos descritos en la sección anterior crean una serie de archivos de registro y carpetas donde se almacenan los productos de los pasos intermedios y finales de todo el proceso. Al finalizarlo, la estructura de carpetas y archivos mostrada en la Figura 3.2 se ve modificada ampliamente para incluir toda la información de los HMM, listas, archivos de registro, audios de las nuevas frases sintetizadas, estadísticas y árboles de decisión.

De mayor importancia para efectos de obtener nuevas frases con la voz entrenada con la base de datos se tienen dos resultados principales:

- Archivos de audio de las frases definidas previamente en la carpeta *labels/*: Se almacenan en la carpeta *gen/qst001/ver1*, la cual contiene los resultados en formato raw y wav.
- Archivos para utilizar en Festival: La voz obtenida puede ser utilizada en conjunto con el programa Festival para producir nuevas frases de audio, a partir del analizador de texto de este programa. Los resultados se almacenan en la carpeta */voices/qst001/ver1*.