UNIVERSIDAD DEL PAÍS VASCO / EUSKAL HERRIKO UNIBERTSITATEA

TÉCNICAS DE PERSONALIZACIÓN DE VOCES SINTÉTICAS PARA SU USO POR PERSONAS CON DISCAPACIDAD ORAL

Tesis doctoral presentada por Agustín Alonso Burguera

dentro del Programa de Doctorado en Tecnologías de la Información y Comunicaciones en Redes Móviles

Dirigida por la Dra. Inmaculada Hernáez Rioja

Co-dirigida por el Dr. Daniel Erro Eslava

UNIVERSIDAD DEL PAÍS VASCO / EUSKAL HERRIKO UNIBERTSITATEA

TÉCNICAS DE PERSONALIZACIÓN DE VOCES SINTÉTICAS PARA SU USO POR PERSONAS CON DISCAPACIDAD ORAL

Tesis doctoral presentada por Agustín Alonso Burguera

dentro del Programa de Doctorado en Tecnologías de la Información y Comunicaciones en Redes Móviles

Dirigida por la Dra. Inmaculada Hernáez Rioja

Co-dirigida por el Dr. Daniel Erro Eslava

El doctorando La directora El co-director

Bilbao, septiembre 2023

TÉCNICAS DE PERSONALIZACIÓN DE VOCES SINTÉTICAS PARA SU USO POR PERSONAS CON DISCAPACIDAD ORAL

Autor: Agustín Alonso Burguera
Directora: Dra Inmaculada Hernáez Rioja
Co-director: Dr Daniel Erro Eslava

Impreso en Bilbao
Primera edición, septiembre 2023

A todos los que hicieron esta tesis posible.

Resumen

La voz es un elemento fundamental en la comunicación humana. Desafortunadamente, en el mundo hay una gran cantidad de personas que sufren algún tipo de discapacidad oral que les impide emplear su voz en sus comunicaciones diarias. El uso de dispositivos de comunicación alternativa y aumentativa (AAC) que utilizan voces sintéticas puede reducir el impacto de la discapacidad. En estos dispositivos, los sistemas de conversión de texto a voz son un componente muy importante: el usuario escribe un texto arbitrario que es leído por una voz artificial o sintética. Aunque los dispositivos AAC ayudan al usuario en el proceso de comunicación, por lo general la voz artificial con la que se reproduce el mensaje está determinada por el proveedor del dispositivo. Este hecho presenta varios aspectos que pueden impactar negativamente sobre su uso, debido al desajuste de las características del usuario y de la voz en lo que se refiere al género, edad, acento y otros. Debido a esto, que los usuarios de estos sistemas tengan la capacidad de elegir una voz personalizada que se ajuste a su propia personalidad es un paso muy importante para conseguir mejorar sus comunicaciones diarias.

En esta tesis se investiga el uso de tecnologías de procesamiento del habla que permiten obtener voces personalizadas para sistemas de conversión de texto a voz. Para ello, se han investigado diferentes métodos de personalización de voces sintéticas para su uso en dispositivos AAC.

En primer lugar, se ha propuesto una estrategia novedosa que permite realizar la personalización con un número muy reducido de grabaciones. El método propuesto está basado en transformaciones

lineales de características espectrales. Al compararse con otras estrategias ha resultado ser más robusto a la escasez de datos proporcionando una mejor calidad en la voz sintética. Sin embargo, la similitud lograda en la personalización es menor.

En segundo lugar, en este trabajo se han investigado diferentes medidas para la evaluación automática de la calidad (inteligibilidad y naturalidad) de las voces sintéticas. Para ello se han estudiado medidas que se utilizan en telecomunicaciones para evaluar la calidad del canal de comunicación (que requieren el uso de una referencia) y medidas sin referencia. El método desarrollado permite estimar la calidad subjetiva de una voz sintética con objeto de eliminar las pruebas de evaluación con personas.

Finalmente, las actividades de investigación se han desarrollado en el contexto de un banco de voces, realizando importantes aportaciones al mismo. Un banco de voces puede entenderse como un gran repositorio de grabaciones de voz donde cualquier persona puede grabar su propia voz, normalmente con el objetivo de obtener una voz sintética personalizada. En esta tesis se ha contribuido al desarrollo e implantación del banco de voces ZureTTS - *ahoMyTTS*, mediante el estudio de las técnicas de adaptación más adecuadas y su implementación, el desarrollo de una metodología para la evaluación automática, el desarrollo de corpus para la síntesis y la personalización en varios idiomas. ZureTTS - *ahoMyTTS* ha tenido una gran repercusión social y actualmente tiene miles de grabaciones de usuarios.

Abstract

Voice is a fundamental element in human communication. Unfortunately, a large number of people in the world suffer from some kind of oral disability. As a consequence, they are not able to use their own voice in daily interactions. Augmentative and Alternative Communication (AAC) devices with synthetic voices can be used to assist users in their communication process by reducing the impact of voice disability. Text-to-speech systemas are part of AACs devices: a user writes an arbitrary text that is finally read by an artificial or synthetic voice.

Nevertheless, AACs vendors are used to make the decision on the voice that ends up reading users' written messages. As a consequence, there is a mismatch between user's voice and the given artificial one: gender, age, accent among other characteristics might not be neither the same nor to be similar. Daily communications can be improved by giving AACs users the ability of selecting a personalized voice that fits better to their own one.

This thesis focuses on the study of speech processing technologies that allow obtaining personalized voices for text-to-speech systems. Concretely, several methods for synthetic voices personalization in AAC systems have been investigated.

First, a novel strategy for voice personalization based on a reduced number of user recordings has been proposed. This method consists of linear transformations of spectral characteristics. Compared to other strategies, outperforms not only in terms of robustness against data scarcity but also providing a higher quality to the synthetic voice.

However, there is still room from improving voice personalization in terms of similarity.

Next, as part of present work, different measures for synthetic voices' quality (intelligibility and naturalness) automatic evaluation have also been investigated. To that end, measurements used in telecommunications to assess the quality of the communication channel (a reference is required), and measurements without reference, have been considered. A method that estimates the subjective quality of a synthetic voice to reduce human evaluation tests required has been developed.

Finally, these reseach activities have been conducted in the context of a voice bank. A voice bank can be seen as a large repository of voice recordings where a user can record his own voice, usually with the aim of obtaining a customized synthetic one. Concretely, some contrinbutions have been made to ZureTTS - *ahoMyTTS* voice bank development and set up through the study of the most appropriate adaptation techniques and their implementation, the development of a methodology for automatic evaluation, the development of a corpus for synthesis, and voice personalization in multiple languages. ZureTTS - *ahoMyTTS* has made a great social impact and currently counts on thousands of user recordings.

Laburpena

Ahotsa funtsezko elementua da giza-komunikazioan. Zoritxarrez, munduan pertsona askok dute hitz egiteko desgaitasunen bat, eguneroko komunikazioetan ahotsa erabiltzea eragozten diena. Ahots sintetikoak erabiltzen dituzten komunikazio alternatiboko eta handigarriko gailuen (AAC) erabilerak desgaitasunaren eragina murriztu dezake. Gailu horietan, testua ahots bihurtzeko sistemak oso osagai garrantzitsua dira: erabiltzaileak testu arbitrario bat idazten du, ahots artifizial edo sintetiko batek irakurtzen duena.

AAC gailuek erabiltzaileari komunikazio-prozesuan laguntzen dioten arren, oro har, mezua erreproduzitzeko erabiltzen den ahots artifiziala gailuaren hornitzaileak erabakitzen du. Horrek, erabileran eragin negatiboa izan dezaketen hainbat alderdi ditu, erabiltzailearen eta ahotsaren ezaugarrien desorekagatik, generoan, adinean, azentuan eta bestelako aspektutan. ak. Hori dela eta, sistema horien erabiltzaileentzat, beren nortasunera egokitzen den ahots pertsonalizatua aukeratzeko gaitasuna izatea oso urrats garrantzitsua da eguneroko komunikazioak hobetzeko.

Tesi honetan testu-ahotserako bihurketa-sistemetan ahots pertsonalizatuen garapena ahalbideratzen duten mintzamena prozesatzeko teknologien erabilera ikertzen da. Horretarako, ahots sintetikoak pertsonalizatzeko hainbat metodo ikertu dira, AAC gailuetan erabiltzeko..

Lehenik eta behin, pertsonalizazioa grabazio kopuru oso txikiarekin egiteko aukera ematen duen estrategia berritzaile bat proposatu da. Proposatutako metodoa espektro-ezaugarriak dituzten transformazio linealetan oinarrituta dago. Beste estrategia batzuekin alderatuz gero,

datu-urritasunarekikosendoagoa izan da, eta ahots sintetikoan kalitate hobea eman du. Hala ere, pertsonalizazioan lortutako antzekotasuna txikiagoa da.

Bigarrenik, ahots sintetikoen kalitatea (ulergarritasuna eta naturaltasuna) automatikoki ebaluatzeko hainbat neurri ikertu dira lan honetan. Horretarako, telekomunikazioetan komunikazio-kanalaren kalitatea ebaluatzeko erabiltzen diren neurriak (erreferentzia bat erabiltzea eskatzen dutenak) eta erreferentziarik gabeko neurriak aztertu dira. Garatutako metodoari esker, ahots sintetiko baten kalitate subjektiboa zenbatetsi daiteke, pertsonekin egiten diren ebaluazio-probak ekiditzeko.

Azkenik, ikerketa-jarduerak ahots-banku baten testuinguruan garatu dira, eta ekarpen garrantzitsuak egin dira bertan. Ahots-banku bat ahots-grabazioen biltegi handi bat bezala uler daiteke, non edonork grabatu dezakeen bere ahotsa, normalean ahots sintetiko pertsonalizatua lortzeko helburuarekin.

Tesi honetan ZureTTS - *ahoMyTTS* ahots-bankua garatzen eta ezartzen lagundu da, egokitzapen-teknika egokienak eta horien inplementazioak aztertuz, ebaluazio automatikorako metodologia bat garatuz, sintesirako corpusa garatuz eta hainbat hizkuntzatan pertsonalizatuz. ZureTTS - *ahoMyTTS* eragin sozial handia izan du eta gaur egun erabiltzaileek egindako milaka grabazio ditu.

Agradecimientos

Tras mucho trabajo y esfuerzo, finalmente el largo camino hasta terminar esta tesis llega a su fin. Camino que no he recorrido yo solo y que sin la ayuda y consejo de muchas personas no habría sido capaz de completar. Gente a la que quiero dedicar unas pocas palabras de agradecimiento por todo lo que han significado durante estos años.

En primer lugar, quisiera agradecer a mis directores de tesis su implicación durante todo este tiempo. A Inma, que me dio la oportunidad de empezar en Aholab siendo la que dio pie a que hoy pueda presentar este documento. Que además siempre ha sabido orientarme y ser capaz de traducir lo que decía a lo que quería decir. Y por supuesto a Daniel, que me ha transmitido todo su buen hacer investigador y tantos consejos me ha dado, no solo para el laboratorio sino también para la vida.

También al resto de gente de Aholab que durante este tiempo me ha acompañado. Eva por revisar concienzudamente todos los trabajos que he presentado. A Jon por mantener bips operativo (aunque fuera para monopolizarlo) y por enseñarme la forma más cara de transmitir un único bit de información. Ibon porque daba igual lo ocupado que estuviese, siempre sacaba tiempo y energía para explicar dudas. A Luis, David y Xabi, por todos los buenos momentos que hemos pasado juntos en el laboratorio, en los congresos y en los workshops.

Por supuesto, a mis familiares y amigos, en especial a mis padres y hermana por el apoyo que me han dado siempre. Y a mi cuadrilla de Vitoria por ser los primeros usuarios en grabar en ZureTTS y ayudarme siempre como evaluadores en mis experimentos.

Por último y más importante, a Julia. Por creer siempre en mí, darme ánimos en todo momento y ser mi revisora más crítica.

Gracias,

Agustín Alonso

septiembre 2023

Índice general

Índice de figuras

Índice de tablas

Siglas

AAC *Alternative and Augmentative Communication*. 2–5, 12, 94, 95

AS *Amplitude Scaling*. 40–42, 44, 45, 47–50, 52, 55, 61, 90, 93

ASR *Automatic Speech Recognition*. 24

BLFW *Bi-Linear Frequency Warping*. 43

CFW *Correlation Frequency Warping*. 43

CMLLR *Constrained Maximum Likelihood Linear Regression*. 31, 33, 34

CNN-LSTM *Convolutional Neural Network Long Short-Term Memory*. 74

CSMAPLR *Constrained Structural Maximum A Posteriori Linear Regression*. 31, 34, 53, 55, 56, 67

DFT *Discrete Fourier Transform*. 72

DFW *Dynamcic Frequency Warping*. 43, 45–48, 50

DML *Deep Machine Learning*. 10

DNN *Deep Neural Network*. 10, 72

DTW *Dynamic Time Warping*. 73, 76

EM *Expectation Maximization*. 32, 34

Siglas

ESTOI *Enhanced Short Time Objective Intelligibility*. 73, 77, 78, 80–84, 86, 91, 94

FW *Frequency Warping*. 40–42, 44, 45, 52, 55, 61, 90, 93

GV *Global Variance*. 27, 28

HASPI *Hearing-Aid Speech Perception Index*. 94

HMM *Hidden Markov Hodel*. 10, 23, 24, 29, 31, 119

HSMM *Hidden Semi-Markov Model*. 24–28, 30–32, 34, 44, 48–50, 59, 60

HTK *HMM ToolKit*. 21, 23, 119

HTS *HMM Speech Synthesis System*. 21–23, 52, 53, 67, 75–77

L2S *Letter-to-Sound*. 119, 120

LSF *Line Spectral Frequencies*. 15

MAP *Maximum A Posteriori*. 33, 34, 53, 56, 67

MFA *Montreal Forced Aligner*. 75

MFCC *Mel Frequency Cepstral Coefficients*. 15, 44–47, 50, 52, 53, 56, 59, 60

MGCC *Mel Generalized Cepstral Coefficients*. 15

ML *Maximum Likelihood*. 24, 31, 33

MLLR *Maximum Likelihood Linear Regression*. 31, 34

MOS *Mean Opinion Score*. 53, 56, 61, 70–72, 74, 78, 80–86, 90–92, 94

MVF *Maximum Voiced Frequency*. 14, 41, 52, 53, 56

NISQA *Non-Intrusive Speech Quality Assessment*. 74, 77, 78, 80–84, 86, 91, 94

PESQ *Perceptual Evaluation of Speech Quality*. 72

POS *Part-of-Speech*. 24, 119

RNN *Recurrent Neural Network*. 10

SAT *Speaker-Adaptive Training*. 30, 33

SIIB *Speech Intelligibility In Bits*. 73, 74, 76, 77, 81–86, 91, 94

SMAP *Structural Maximum A Posteriori*. 31, 34

SMAPLR *Structural Maximum A Posteriori Lineal Regression*. 34

STOI *Short Time Objective Intelligibility*. 72, 73, 77, 78, 80–84, 86, 91, 94

TF *Time Frequency*. 72, 73

TTS *Text-To-Speech*. 3, 4, 8, 9, 12, 16, 20–22, 36, 64–66, 72, 74, 90, 119

VTLN *Voice Track Length Normalization*. 43

CAPÍTULO

1

Introducción

1.1 Sistemas de Asistencia al Habla

La voz es el principal medio que tiene el ser humano para comunicarse. Sin embargo, existen personas que sufren algún tipo de discapacidad oral que les impide emplearla, ya sea parcial o totalmente. Esta pérdida puede ser causada por diversos factores. En algunos casos se puede sufrir desde el nacimiento. En otros casos, la pérdida puede ser debida a enfermedades degenerativas, accidentes o intervenciones quirúrgicas entre otros motivos. Independientemente de su naturaleza o severidad, este impedimento crea una barrera en las personas que afecta negativamente a la interacción con otras y a su socialización.

Para ayudar a paliar este problema existen diversos Sistemas de Asistencia al Habla o *Alternative and Augmentative Communication* (AAC) en inglés que permiten a las personas con alguna discapacidad oral comunicarse de manera más efectiva. Existen sistemas no tecnológicos como son la gesticulación, la pantomima o el lenguaje de signos con la ventaja de que siempre están disponibles para su uso. Sin embargo, es necesario que el interlocutor esté entrenado para poder interpretar correctamente el mensaje. También existen sistemas AAC tecnológicos que pueden emplearse para ayudar en la comunicación. Hay una gran variedad en función de las necesidades y capacidades del usuario. Por lo general, los AAC disponen de una entrada que permite seleccionar el mensaje a transmitir y genera una salida de audio. Hay diferentes tipos de entradas en función de las capacidades del usuario, como son:

- Cuadrículas de iconos con mensajes predefinidos, donde el usuario escoge uno usando directamente el dedo, un puntero o en algunos casos seleccionándolo mediante los ojos (*eye-tracking* en inglés).

- Codificación de mensajes, donde el usuario puede prefijar previamente mensajes comunes y activarlos con una simple combinación alfanumérica.

- Entrada de texto con predictor, donde el sistema trata de predecir, en función de las primeras letras que se escriban, la palabra o mensaje completo que se intenta introducir.

- Entrada libre de texto, donde el usuario puede escribir exactamente el mensaje a transmitir.

La salida generada es una señal de voz con el mensaje a transmitir. Aunque en algunos AAC la señal de voz de salida puede ser una frase pregrabada que se reproduce directamente al ser seleccionada, por lo general, se genera empleando un sistema texto-a-voz o *Text-To-Speech* (TTS) en inglés. Los TTS reciben un texto como entrada y como salida generan una señal de voz con el mensaje leído por una voz sintética.

En función de la tecnología que emplea el TTS, este puede contener diferentes módulos o componentes. Los módulos principales se representan en la figura 1.1 y son los siguientes:

- Análisis del texto, que se encarga de procesar el texto de entrada y transformarlo en etiquetas lingüísticas. Las etiquetas contienen la información fonética, de pausas, de sílabas y sintáctica.

- Predicción de la prosodia, que se encarga de añadir información adicional sobre la entonación, duración o intensidad.

- Síntesis, cuya función es generar la señal de voz empleando la información suministrada por los módulos previos.

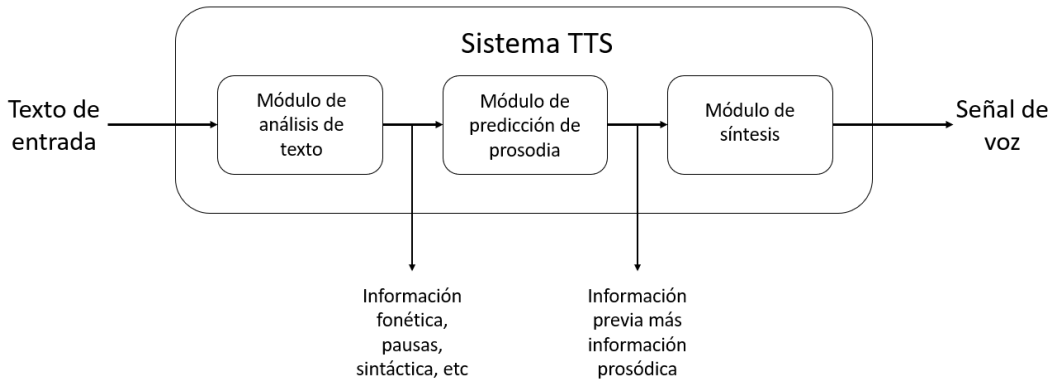


Figura 1.1: Diagrama de módulos de un sistema TTS.

Dependiendo de la tecnología empleada los límites entre un módulo y el siguiente son difusos. En algunos casos, un mismo módulo puede tener la responsabilidad de dos o más.

Actualmente, los sistemas TTS generan voz de gran inteligibilidad (cómo de claro y fácil se entiende el mensaje), naturalidad (cómo de parecida a una persona real y no a un robot suena), y calidad (cómo de bien y libre de artefactos o *artifacts* suena la voz). Además, gracias a los avances actuales en electrónica y computación, los TTS se pueden encontrar integrados de manera embebida, lo cual permite poder emplearlos sin necesidad de conectarse a servidores en la red. Esto es una gran ventaja ya que la perdida de conexión, por ejemplo, de cobertura en el interior del metro o de no disponer de un plan de *roaming* en el extranjero, no suponen una merma en la funcionalidad del sistema. A pesar de ello, suele conllevar que el software empleado suele ser cerrado y difícil de extender.

Si bien el uso de sistemas AAC ayuda a las personas con discapacidad oral a comunicarse, muchas veces están ligados y limitados a las capacidades que el proveedor del sistema haya incorporado. Esto se hace más patente en la elección de la voz que sintetiza el mensaje. Actualmente los sistemas TTS ofrecen muy buenas voces. No obstante, el usuario se puede encontrar con un sistema AAC que no incluya una voz que le represente en términos de género, edad o acento. A pesar de que el sistema pueda disponer de distintas voces a escoger, el catálogo que ofrecen suele ser limitado. Esto hace que, por ejemplo, algunas mujeres empleen voces masculinas para comunicarse o que algunos niños estén forzados a usar voces de adultos. Existen estudios como [60] que muestran cómo las personas tendemos a formarnos una impresión de la personalidad de otras por su voz, como también con el rostro o el color de la piel. Otros estudios [86] han demostrado que usar voces personalizadas puede facilitar el desarrollo intelectual de niños con falta de visión. Por lo tanto, el uso de una voz personalizada que se adapte lo máximo posible al usuario puede ayudar a reducir el impacto social de emplear un aparato electrónico para las comunicaciones diarias.

Obtener una voz sintética nueva desde cero requiere de gran cantidad de grabaciones de audio con su transcripción para poder generarla. Puede tratarse de un proceso caro, dado que las grabaciones suelen realizarlas locutores profesionales. No obstante, existen técnicas que permiten conseguir voces nuevas a

partir voces sintéticas ya generadas. Mediante unas pocas grabaciones de un nuevo locutor objetivo se modifican diversos parámetros de la voz inicial para que suene lo más parecida posible al nuevo locutor objetivo. Esto permite generar voces nuevas y, por lo tanto, aumentar el catálogo de voces disponibles de manera más barata y asequible. Sin embargo, aún es necesario que el proveedor del dispositivo AAC permita generar y acepte las voces adaptadas.

Para ayudar e incentivar a conseguir nuevas voces adaptadas han surgido varias iniciativas denominadas bancos de voces o *voice banking* en inglés. Cualquier persona pueden grabar las voces necesarias para generar una voz adaptada que suene como la suya propia. De esta manera, el catálogo de voces de un banco de voces es más amplio. El usuario de un sistema AAC compatible puede escoger entre el catálogo aquella que le resulte agradable y se adapte a sus gustos y preferencias.

La motivación de esta tesis es investigar tecnologías que permitan a las personas con discapacidad oral utilizar sistemas AAC para sus comunicaciones cotidianas disponiendo de una voz personalizada. Se tiene el convencimiento de que al proveer de una voz que se asemeje lo máximo posible a las preferencias del usuario se puede mejorar la interacción con otras personas mediante dichos sistemas electrónicos.

1.2 Modelo de Producción Fuente-Filtro

En el proceso del habla humana el modelo de producción más extendido es el llamado fuente-filtro. Como su nombre indica, en este modelo el habla se representa como una combinación de una fuente de señal y un filtro que se combinan para generar la señal de voz. La fuente se corresponde con las cuerdas vocales o la glotis, mientras que el filtro se corresponde con el tracto vocal, el cual está compuesto por la cavidad nasal, la cavidad oral, la faringe y la laringe. Una simplificación del mismo puede observarse en la figura 1.2

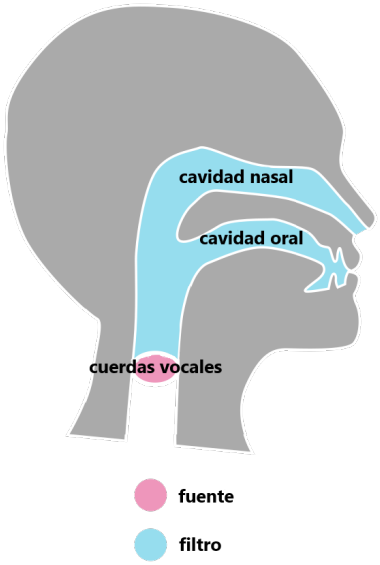


Figura 1.2: Esquema de producción de voz del modelo fuente-filtro.

En función del tipo de excitación los sonidos se pueden diferenciar entre sonoros o sordos. Es decir, si el flujo del aire que surge de los pulmones es modulado por las cuerdas vocales o la glotis, el sonido final se considera sonoro. Si, por el contrario, éstas no intervienen y el flujo de aire de los pulmones fluye directamente, se considera un sonido sordo. De la manera más simple, la excitación de los sonidos sonoros puede modelarse como un tren de deltas y la de los sonidos sordos como ruido blanco gaussiano. Modelos de excitación más complejos sustituyen el sencillo tren de deltas por pulsos glotales y combinan

ambos tipos de excitación para modelar, de manera más fiel, la generación real de la fuente de voz, dado que los sonidos sonoros no son periódicos puros y también poseen en su excitación parte de ruido.

El filtro del modelo representa el tracto vocal, la cavidad nasal, la cavidad vocal y la boca. En la aproximación más sencilla, éste se puede representar como un filtro todo polos. Para el caso de los sonidos nasales, aquellos en los que la cavidad nasal tiene relevancia para la producción del sonido, esta aproximación es bastante deficiente, siendo necesario incluir un filtro de polos y ceros para su correcta representación. Un diagrama de bloques del modelo fuente filtro se observa en la figura 1.3

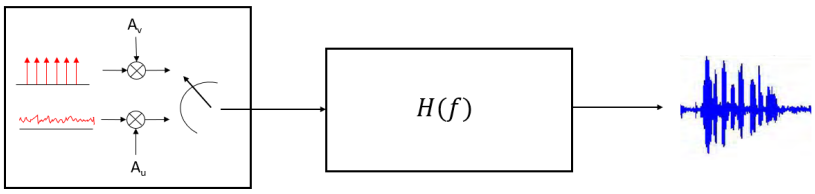


Figura 1.3: Diagrama de bloques simplificado del modelo fuente filtro de producción de voz.

1.3 Historia Síntesis de Voz

La historia de los sistemas TTS está directamente ligada a la evolución de la síntesis de voz. Sin embargo, lo contrario no se da dado que hay sistemas de síntesis que no tienen como entrada texto o parámetros lingüísticos. Por ejemplo, muchos de los primeros sistemas de síntesis tenían como entrada sistemas mecánicos que requerían interacción humana, como por ejemplo teclas, barras o pedales.

La primera máquina capaz de producir sonidos vocálicos sostenidos fue diseñada por Christian Gottlieb Kratzentein en 1780. Ésta consistía en cinco resonadores en forma de tubo de órgano que, cuando se excitaban con la vibración de una caña, producían las vocales /a, e, i, o, u/. La síntesis de frases completas llegó posteriormente en 1845 de la mano de Joseph Faber y su "maravillosa maquina parlante", a la que llamó *Euphonia*. Esta máquina disponía de un pequeño órgano conectado mecánicamente a una cabeza de madera con mandíbula, una lengua de marfil y glotis de goma. Todos estos componentes juntos eran capaces de producir voz convincente.

Síntesis articulatoria: las dos máquinas anteriores son ejemplos mecánicos de síntesis articulatoria [17, 95]. En ella se trata de producir voz sintética simulando el comportamiento del sistema articulatorio humano como son labios, lengua, glotis o tracto vocal. En teoría, la síntesis articulatoria es el método más efectivo de producir voz sintética, dado que es la manera en la que el ser humano produce voz en la vida real. Sin embargo, resulta muy difícil modelar correctamente la producción de voz empleando este método.

Síntesis por formantes: a finales del siglo XIX, estaba aceptado que el principio fundamental de la producción de voz eran las resonancias que resaltaban ciertos componentes del espectro de los sonidos, los formantes. En 1922 se presentó el primer sintetizador de voz enteramente eléctrico [101]. En él se empleaba un zumbador y dos ramas resonantes paralelas con resistencias variables que permitían modificar la posición de los formantes.

En la síntesis por formantes [5, 57, 94] la voz se genera en base a un conjunto de reglas que controlan un modelo simplificado fuente-filtro. Estas reglas, por lo general, son desarrolladas por lingüistas que tratan de imitar la estructura de los formantes y otras características espectrales de la voz de la manera más realista

posible. Sin embargo, resulta muy complejo especificar estas reglas y además la voz generada presenta una naturalidad muy baja y contiene diversos artefactos.

Síntesis por concatenación de unidades: la síntesis por concatenación de unidades fue presentada a principios de la década los años 50 del siglo XX. Ésta se basa en concatenar fragmentos de voz real que se guardan en una base de datos [45, 74, 79]. Generalmente dicha base de datos consiste en unidades de voz como sílabas, fonemas o difonemas (la parte final de un fonema junto a la parte inicial del siguiente fonema) extraídos de frases grabadas por un único locutor profesional. Durante la síntesis, el sistema TTS concatenativo busca en la base de datos las unidades que mejor se ajustan al texto de entrada y genera la forma de onda juntando dichas unidades. Por lo general este método genera una voz sintética de una alta naturalidad, inteligibilidad y calidad, dado que para ello se emplean fragmentos de voz real. Sin embargo presenta, ciertas desventajas. La cantidad de combinaciones entre fonemas que se pueden dar en un idioma es muy alta. Además, un mismo fonema puede tener variaciones en su pronunciación, por ejemplo, si se encuentra en una sílaba acentuada o no. Este hecho hace que la cantidad de audio necesario para poder escoger la unidad óptima en cada caso sea muy alto. De lo contrario, se puede dar el caso de que al juntar diferentes unidades el resultado presente saltos, discontinuidades o "clicks" que degradan la calidad. Además, la voz generada depende completamente del audio grabado y esto hace que en caso de querer modificar el timbre, emoción u otro aspecto de la prosodia, es necesario realizar nuevas grabaciones que incluyan dichos aspectos, lo cual es un proceso largo y costoso.

Síntesis estadístico paramétrica: para superar los inconvenientes de la síntesis por concatenación de unidades, en 2009 se propuso la síntesis estadístico paramétrica [115, 132, 136]. Ésta consiste en que en vez de generar directamente la señal de voz mediante concatenación, primero se generan parámetros acústicos que representan la señal y después se reconstruye generando la señal de voz. Para representar la señal se usa el modelo fuente-filtro en el que una señal de voz puede descomponerse en parámetros de excitación y parámetros espectrales. Para extraer estos parámetros de señales de voz reales se emplean *vocoders* [31, 55, 72]. Posteriormente, se entrena un modelo acústico que relaciona los parámetros extraídos por los *vocoders* y con el texto transcrito. Este modelo

acústico suele estar basado en modelos ocultos de Markov o *Hidden Markov Hodel* (HMM) en inglés [87]. El modelo ya entrenado se emplea a la hora de generar voz sintética para estimar, en función de un texto nuevo de entrada, los parámetros que mejor lo representan. Por último, los mismos *vocoders* que inicialmente han realizado la parametrización realizan el proceso contrario de reconstrucción para generar la señal de voz.

Los sistemas de síntesis estadístico paramétrica presentan varias ventajas como que el audio conseguido suena más homogéneo y libre de artefactos. Además, requiere menos cantidad de audio para generar una voz y el sistema ocupa menos espacio en memoria, dado que el modelo acústico entrenado es mucho más pequeño que una base de datos necesaria para la síntesis concatenativa. También, si se modifican correctamente los parámetros acústicos es posible modificar el resultado de la voz generada.

Síntesis por redes neuronales: las redes neuronales y el *Deep Machine Learning* (DML) han progresado rápidamente en los últimos años debido a las mejoras en *hardware* de procesamiento paralelo y la aparición de librerías de código abierto para su uso. Esto ha llevado a que también se investigue su uso en la síntesis de voz.

El primer paso para incluir redes neuronales consiste en sustituir los modelos HMM del modelo acústico por redes *Deep Neural Network* (DNN) o *Recurrent Neural Network* (RNN). En este enfoque la estructura es la misma que en la síntesis estadística anterior, cambiando un tipo de modelo acústico por otro. Algunos ejemplos de sistemas que usan esta aproximación son DeepVoice 1 [8] y DeepVoice 2 [38].

Existen otras propuestas que emplean redes neuronales con diferentes estrategias. Por ejemplo, WaveNet [80] genera directamente la forma de onda a partir de características lingüísticas, sin necesidad de usar un *vocoder* para ello. Otros, como Tacotron 1[123] y Tacotron 2 [97], FastSpeech 1 [90] y FastSpeech 2 [91] y DeepVoice 3 [83] simplifican el módulo de análisis de texto tomando como entrada directamente una secuencia de letras/fonemas, y también simplifican las características acústicas de salida generando mel-espectrogramas. Por último, existen sistemas que generan la señal de voz directamente desde el texto de entrada, como por ejemplo Clarinet [84], FastSpeech 2s [91] y EATS [27]. En estos casos

aún sigue siendo necesario utilizar un pequeño módulo de procesado previo para lidiar con la normalización del texto y en algunos casos con la desambiguación de polífonos (palabras que con una única grafía tienen más de una pronunciación).

Aunque el uso de redes neuronales ha permitido crear voces sintéticas con calidad y naturalidad mayores que los sistemas previos, aún presenta algunos inconvenientes que no la hacen apropiada para todos los casos. Para entrenar las redes es necesario disponer de una gran cantidad de material, lo cual es difícil de conseguir para muchas de las lenguas del mundo. Para las lenguas mayoritarias con gran cantidad de hablantes como el inglés o el chino mandarín es posible encontrar recursos suficientes para conseguir un entrenamiento robusto. Sin embargo, en el mundo existen más de 7000 idiomas diferentes[1] y la mayoría presentan escasez de material de entrenamiento. Existen técnicas como *cross-lingual transfer* [9, 118, 125, 137] que intentan entrenar una red usando recursos en un idioma mayoritario y después adaptarla con pocos recursos del idioma objetivo Pero la calidad final depende de la similitud fonética que exista entre ambos idiomas (por ejemplo, cuántos fonemas comparten) y el resultado final no tiene la mismas características que si se entrena usando únicamente recursos del idioma objetivo.

En el caso de entornos ruidosos existen trabajos con redes neuronales que consiguen una voz sintética con efecto Lombard [43, 81]. El enfoque que siguen estos trabajos consiste en entrenar una nueva red que sea capaz de sintetizar audio con este efecto. Esto supone una gran sobrecarga dado que es necesario entrenar o adaptar una red neuronal nueva. Empleando síntesis estadístico paramétrica, la mejora en inteligibilidad en entornos ruidosos se puede obtener con un postprocesado sin excesivo coste computacional [30] que puede ser activado o desactivado según las necesidades del momento.

En el caso de adaptación de locutor (generar una voz nueva modificando una ya entrenada con una pequeña cantidad de datos nuevos) existe un gran esfuerzo de investigación [7, 13, 14]. Cuando la red inicial no dispone de suficiente información acústica, como puede ser prosodia o timbre, la red tiende a sobre ajustar los datos de entrenamiento (*overfitting* en inglés) y la adaptación para locutores nuevos es pobre. Una forma de solucionarlo consiste en aumentar la cantidad y la diversidad de los datos de entrenamiento [20, 131]. Sin embargo, esta estrategia aumenta

[1] https://www.ethnologue.com/browse

considerablemente los costes de obtención de datos, dado que es necesario disponer de una gran cantidad de grabaciones de locutores diferentes.

Cuando se disponen de pocos datos de adaptación del locutor nuevo, existen trabajos que analizan el impacto de la escasez de datos [7], y concluyen que la calidad de la voz adaptada varía en función de la cantidad de datos disponibles para la adaptación. En estos casos se puede adaptar todo el modelo [14, 59], parte del modelo [73, 138] o únicamente el *speaker embedding* [7, 13, 14]. Un caso específico es el denominado *zero-shot adaptation* [11, 21, 51] que consiste en adaptar la red empleando unos pocos segundos de material nuevo del locutor objetivo. En este caso la calidad de la adaptación baja considerablemente si el locutor objetivo es bastante diferente del locutor con el que se ha entrenado la red. Una posible solución a este problema consiste en entrenar una red multi-locutor o *multi-speaker* en inglñes en la cual se emplea el material de varios locutores para entrenar la red de partida. De esta manera es más probable que el nuevo locutor tenga similitudes suficientes con la red entrenada para obtener una buena adaptación. No obstante, esta estrategia no garantiza que no surjan locutores objetivo lo suficientemente dispares de la red *multi-speaker* para los que se obtenga una mala adaptación. Además, para entrenar este tipo de red es necesario disponer de aún más recursos.

Además, los sistemas TTS pueden ser desplegaos en servidores en la nube o bien estar integrados de manera embebida es dispositivos electrónicos. En ambos casos es necesaria que la síntesis ser rápida para que la comunicación pueda darse de forma fluida. En el caso de servidores en la red, generalmente éstos tienen una capacidad de computación los suficientemente alta para que esto sea posible. No obstante, en el caso de que se ejecuten directamente en un dispositivo portátil, como puede ser un *smartphone*, un *tablet* o un sistema AAC dedicado, los recursos disponibles son menores. No solo es necesario disponer de una capacidad de computación lo suficientemente elevada para poder sintetizar audio, sino que además la red que se emplea puede llegar a ocupar varios centenares de MBs, lo cual dificulta su integración en dispositivos de gama baja.

Por último, la principal ventaja que presenta la síntesis por redes neuronales, la gran calidad y naturalidad que obtiene, queda enmascarada cuando se usa en entornos reales fuera del laboratorio y sin condiciones de evaluación. En estos

casos, hay diversos aspectos ajenos a la propia síntesis como los altavoces con los que se reproduce o ruidos ambientales que pueden minimizar la ventaja de las redes neuronales sobre otras tecnologías.

Debido a estos factores, el método de síntesis que se ha elegido para el trabajo realizado la presente tesis es el de síntesis estadístico paramétrica.

1.4 Vocoders

En los sistemas de síntesis estadístico paramétrica un elemento muy importante que determina la calidad de la voz obtenida son los *vocoders* (*voice coder*). Los *vocoders* son los encargados de extraer a partir de la señal de voz su representación paramétrica, que se emplea en el entrenamiento del modelo acústico, y posteriormente de reconstruir la señal de voz usando los parámetros extraídos del modelo durante la síntesis. Aunque similares, los requisitos necesarios para una correcta parametrización en los *vocoders* no son idénticos a los empleados para codificación para transmisión de voz. El objetivo principal en transmisión de voz es conseguir la máxima calidad de resintetizado empleando la menor tasa de bits posible para enviar la señal de voz. El rendimiento y eficacia en reconstrucción en tiempo real también suelen ser requisito fundamental. Aunque los *vocoders* empleados para la síntesis de voz también deben ofrecer una capacidad de reconstrucción de gran calidad de la señal, la compresión de la información no suele ser un factor fundamental. En cambio, deben proveer de parámetros capaces de modelar estadísticamente la estructura de la voz subyacente.

Los primeros *vocoders* empleados en síntesis de voz empleaban dos flujos o *streams* diferentes para representar la parametrización de la señal. La envolvente espectral para modelar el tracto vocal y la frecuencia fundamental en escala logarítmica o $log - f_0$ para modelar la excitación. Sin embargo, dada la simplicidad de este modelo, en especial en la parte de excitación, el resultado final no tenía buena calidad, dando como resultado una voz con un zumbido molesto. Versiones posteriores de *vocoders* han mejorado la representación de la excitación incluyendo información sobre la aperiodicidad de la señal para tener en cuenta la excitación mixta de la misma. Por ejemplo, STRAIGHT [55] divide la señal en diferentes bandas frecuenciales y calcula la relación de potencia entre la señal de voz y su parte aperiódica para cada banda. Ahocoder [31] por su parte emplea un único parámetro denominado *Maximum Voiced Frequency* (MVF) definido como la frecuencia en la que la energía de la parte aperiódica de la excitación iguala la energía de la parte periódica. WORLD [72] por su parte no hace uso de un componente de aperiodicidad directamente. Sin embargo, calcula la señal de

excitación a través de la forma de onda de la señal para aumentar la calidad de la síntesis.

Como representación de la envolvente espectral generalmente se emplean *Mel Frequency Cepstral Coefficients* (MFCC) [36] o *Mel Generalized Cepstral Coefficients* (MGCC) [114] aunque también se pueden emplear otras representaciones como *Line Spectral Frequencies* (LSF) [46].

1.5 Objetivos

La motivación de esta tesis es investigar tecnologías que permitan a las personas con alguna discapacidad oral que usen sistemas TTS de manera más natural. Dadas las limitaciones de estos sistemas comerciales en lo que respecta a la posibilidad de elegir la voz que se desea emplear, se pretende mejorar en este aspecto. Para ello, empleando el marco de trabajo de síntesis estadístico paramétrica, se han investigado diversas maneras de obtener nuevas voces que permitan a dichos usuarios disponer de una voz adecuada para ellos. El principal objetivo es conseguir obtener voces adaptadas que sean de una calidad lo suficientemente buena para que sean agradables de usar mientras mantiene unas características fonéticas suficientes para que puedan asociarse al usuario que las están usando.

Una primera aproximación consiste en desarrollar nuevas técnicas de adaptación capaces de obtener una nueva voz partiendo de grabaciones realizadas por personas que presenten alguna discapacidad oral. Para ello es necesario desarrollar algoritmos de adaptación con las siguientes características:

1. Robustos frente a escasez de datos. Aunque es necesario disponer de grabaciones de la voz a la que se pretende imitar, cuantos menos datos sea necesario obtener mayor aceptación por parte de los usuarios tendrá el sistema.

2. Capaces de extrapolar la información fonética disponible para simular cómo sonaría la voz en caso de que no presentara la discapacidad. Dado que se pretende obtener una voz que suene sana.

Otra aproximación consiste en usar voces adaptadas por otras personas. Para ello se puede recurrir a los bancos de voces. Éstos pueden entenderse como grandes repositorios de grabaciones de voz que se pueden emplear para obtener diferentes voces adaptadas. De esta manera, un usuario puede elegir de entre un gran catálogo la voz que más le guste o la que más se parezca a él. Dado el volumen de voces que se pueden llegar a manejar, con el fin de evitar tener que mantener de manera manual dicho catálogo es necesario diseñar y desarrollar algún método automático para facilitar dicho mantenimiento.

Además, se quiere conseguir que las voces generadas puedan ser usadas por las personas que lo necesiten sin necesidad de disponer de dispositivos electrónicos de grandes prestaciones. También es deseable que se puedan emplear sin que requieran de conexión a un servidor alojado en la nube. De esta manera se garantiza la disponibilidad del sistema en cualquier situación. Por lo tanto, es necesario que funcionen en en dispositivos de gama baja o *low-end* de manera embebida. Para ello la capacidad de computación y los requerimientos de memoria necesarios deben ser lo más bajos posibles.

1.6 Estructura del Documento

Además del presente capítulo de introducción, el documento de tesis se compone de: capítulo 2, en el que se profundiza sobre el estado del arte actual respecto a la síntesis estadístico paramétrica, la tecnología de síntesis escogida para esta investigación; capítulo 3, en el que se detalla el método de adaptación propuesto y desarrollado para adaptar nuevas voces; capítulo 4, en el que se explica qué son los bancos de voces, detallando el banco de voz desarrollado por el grupo de investigación Aholab y cómo se han empleado medidas objetivas para puntuar automáticamente la calidad de las voces personalizadas de las que dispone; capítulo 5 con la exposición de las conclusiones de esta tesis.

CAPÍTULO

2

Síntesis Estadístico Paramétrica

En el presente capítulo se describe en qué consiste la tecnología de síntesis de voz seleccionada para realizar esta investigación: la síntesis estadístico paramétrica. En primer lugar, se detalla a nivel teórico cómo se puede emplear para generar una voz sintética. A continuación, se comentan varios métodos del estado del arte que obtienen una nueva voz sintética adaptando otra voz sintética ya existente.

2.1 Introducción

La síntesis de voz estadístico paramétrica [136] ha sido muy popular desde la segunda década del siglo XXI por las diversas ventajas que presenta, tales como:

- Permite generar voces con menos material de entrenamiento que otros métodos.
- Las voces generadas presentan una gran calidad.
- Es posible adaptar voces ya existentes con una pequeña cantidad de material de entrenamiento nuevo.

En la creación de un sistema de síntesis basado en las técnicas estadístico paramétricas se distingue entre la fase de entrenamiento, cuya salida son modelos acústicos que representan una voz sintética; y la fase de síntesis, en la que se emplean dichos modelos acústicos para generar audio. La figura 2.1 ilustra ambas fases. En la fase de entrenamiento, los modelos acústicos se generan empleando una base de datos consistente en grabaciones de audio y sus transcripciones. Las grabaciones se parametrizan utilizando un *vocoder*, tal y como se detalla en la sección 1.4. De las transcripciones se extraen etiquetas lingüísticas, que contienen gran cantidad de información fonética, prosódica y sintáctica, mediante el módulo lingüístico de un TTS. Los modelos acústicos aprenden la relación entre las etiquetas lingüísticas y los parámetros acústicos a través del entrenamiento estadístico. En la fase de síntesis, dado un texto arbitrario de entrada, el módulo lingüístico empleado en la fase de entrenamiento genera las etiquetas lingüísticas de dicho texto, que se emplean para estimar los parámetros acústicos a partir de los modelos entrenados. La señal de audio de salida se reconstruye a partir

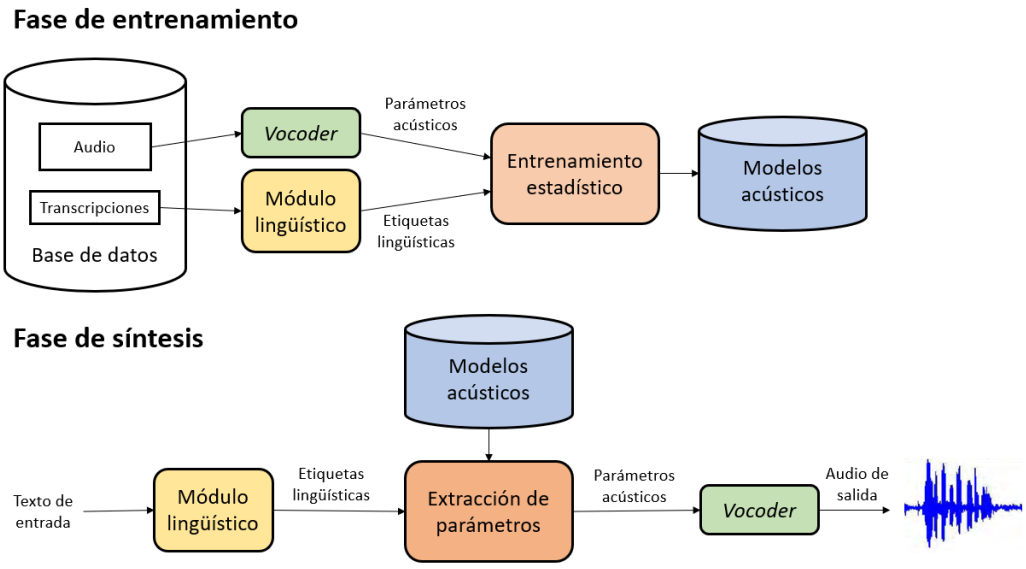


Figura 2.1: Síntesis estadístico paramétrica: fases de entrenamiento y síntesis.

de los parámetros estimados usando el mismo *vocoder* que ha parametrizado las grabaciones durante el entrenamiento.

Un factor fundamental que ha ayudado a la amplia adopción de este método de síntesis ha sido la distribución del software *HMM Speech Synthesis System* (HTS) [113]. Este *software* desarrollado por el Instituto Tecnológico de Nagoya consiste en un *framework* completo que permite generar y usar voces sintéticas en un sistema TTS. Se distribuye como un parche de *HMM ToolKit* (HTK) [134] bajo licencia de código abierto.

La facilidad que ofrece la síntesis estadístico paramétrica para adaptar una voz ya entrenada también ha ayudado a incrementar su popularidad. Empleando una pequeña cantidad de material nuevo se puede modificar una voz ya existente para cambiar diversos aspectos de la misma. Algunos usos son la adaptación de locutor [129], adaptación de estilos [69] o adaptación de emoción [76], entre otros. Esto reduce el tiempo y coste de obtención de voces nuevas. En lugar de requerir horas de audio del locutor, el estilo o la emoción que se desean imitar, se pueden obtener resultados similares con unos pocos minutos de grabaciones nuevas. El propio *software* HTS ofrece varios algoritmos que permiten la adaptación de voces, lo

que facilita aún más su obtención.

El resto del capítulo se estructura de la siguiente forma: en primer lugar se expone HTS y su método de entrenamiento estándar para obtener una voz nueva partiendo de cero. A continuación, se explica el método de generación de parámetros que se emplea para estimar los parámetros acústicos en función de las etiquetas de entrada durante la fase de síntesis. Posteriormente, se comentan diversos problemas que puede presentar la generación de parámetros y las técnicas que existen para solventarlos. Seguidamente se explican los diferentes algoritmos de adaptación HTS que ofrece para obtener nuevas voces adaptadas. Por último, se exponen las aportaciones realizadas en el ámbito de la síntesis estadístico paramétrica: un sintetizador TTS multilingüe con voces estadístico paramétricas.

2.2 HMM Speech Synthesis System

Aunque para entrenar una voz para un sistema estadístico paramétrico se pueden emplear diferentes tipos de modelos, el más extendido es el que emplea HTS. Éste utiliza los HMMs [87] de HTK [134]. En la configuración por defecto, considerada en esta tesis, los HMM constan de cinco estados, permitiendo modelar las partes transitorias entre el fonema anterior y el fonema siguiente, así como una parte central estable. Además, presentan una topología de izquierda a derecha, de manera que las matrices de transición de los estados sólo permiten permanecer en el mismo estado o pasar al estado inmediatamente siguiente. No están permitidos los saltos entre estados ni retroceder. La salida de los estados es una distribución normal gaussiana multidimensional definida como:

$$b_i(\boldsymbol{o}) = N(\boldsymbol{o}; \boldsymbol{\mu_i}, \boldsymbol{\Sigma_i}) \tag{2.1}$$

donde $\boldsymbol{\mu_i}$ y $\boldsymbol{\Sigma_i}$ son el vector media y la matriz de covarianza para el estado i, y $\boldsymbol{o}$ es el vector de observación, correspondiente a los parámetros acústicos extraídos por el *vocoder*.

Para modelar los parámetros espectrales se pueden emplear distribuciones gaussianas multidimensionales. Sin embargo, para los parámetros de excitación como la log-f_0 es difícil aplicar una única distribución. Esto se debe a que la excitación toma dos tipos de valores diferentes: un valor continuo en los sonidos sonoros, cuando las cuerdas vocales vibran; y un valor nulo en los sonidos sordos, cuando no vibran. Aunque se han investigado varios métodos para modelar la log-f_0 [50] [35] [92], el sistema HTS utiliza una distribución de probabilidad multi-espacial [116]. En este modelo la log-f_0 adopta una distribución continua en las regiones sonoras de la señal y un valor discreto en las regiones sordas.

2.2.1 Entrenamiento

Teniendo los vectores acústicos y las etiquetas lingüísticas, el problema del entrenamiento consiste en estimar los modelos λ que maximicen la verosimilitud de los parámetros acústicos extraídos por el *vocoder* $\boldsymbol{O}$ dadas las etiquetas W extraídas por el módulo lingüístico según la expresión 2.2.

$$\hat{\lambda} = \underset{\lambda}{\operatorname{argmax}}\{p(\boldsymbol{O}|W, \lambda)\} \tag{2.2}$$

Esto se realiza mediante la estimación de Máxima Verosimilitud o *Maximum Likelihood* (ML) en inglés, usando el algoritmo de Baum-Welch [25]. Este proceso es similar al empleado en el reconocimiento del habla o *Automatic Speech Recognition* (ASR) en inglés, sin embargo presenta algunas diferencias a tener en consideración:

1. Los vectores acústicos que se emplean en el entrenamiento en ASR normalmente sólo constan de parametros espectrales. Sin embargo, en síntesis estadístico paramétrica, se emplean como vectores la información espectral y la información de la excitación que proporcionan los *vocoders*.

2. La topología de izquierda a derecha modela correctamente el flujo del habla. Sin embargo, las matrices de transición hacen que la probabilidad de duración de un estado decrezca exponencialmente según aumenta la duración. Esto no modela correctamente la estructura temporal de la secuencia de los parámetros de voz. Por ello, en síntesis estadístico paramétrica, a cada estado del HMM se le asocia un modelo de duración explícito, basado en una gaussiana. Esta estructura se denomina *Hidden Semi-Markov Model* (HSMM) [135]. La distribución de salida para la duración del estado i queda definida como:

   $$p_i(d) = N(d; m_i, \sigma_i^2) \tag{2.3}$$

   donde d es la duración del estado, m_i es la media y σ_i^2 es la varianza de la distribución.

3. Además de la información fonética también se tiene en cuenta información de otros niveles lingüísticos, como la acentuación léxica, el tono o la información de *Part-of-Speech* (POS). Aunque los parámetros espectrales se ven afectados principalmente por la información fonética, la prosodia y la duración también pueden verse afectadas por información lingüística supra-segmental. Por ello se emplea toda la información proporcionada en las etiquetas lingüísticas por el módulo lingüístico.

4. En la práctica existen gran cantidad de factores contextuales en relación a la cantidad de datos de entrenamiento disponibles. Por ello, estimar los parámetros dependientes del contexto de manera robusta y precisa suele ser muy difícil. Para resolver el problema se aplican técnicas de atado o *tying* en ingles [78], para agrupar estados parecidos y unir parámetros de los modelos entre varios HSMMs dependientes del contexto. El *tying* de los estados se realiza empleando una forma de estructura de árbol jerárquico, donde cada nodo hoja corresponde a un estado HSMMs. Esto significa que siguiendo la estructura del árbol se puede llegar al mismo nodo hoja, y por lo tanto al mismo estado HSMMs, con diferentes aunque similares informaciones contextuales. Dado que el espectro, la excitación y la duración tienen dependencias contextuales diferentes, se agrupan de manera separada empleando árboles de decisión dependientes del *stream* [132].

2.2.2 Síntesis

Una vez los modelos ya han sido entrenados, el problema de la síntesis consiste en encontrar la secuencia de vectores de parámetros acústicos (los cuales incluyen tanto parámetros espectrales como parámetros de excitación) más probables dado un conjunto de HSMM $\hat{\lambda}$, y las etiquetas lingüísticas del texto de entrada w. Esto se puede determinar de la siguiente manera:

$$\hat{\boldsymbol{o}} = \underset{\boldsymbol{o}}{\operatorname{argmax}}\{p(\boldsymbol{o}|w, \hat{\lambda})\} \tag{2.4}$$

$$= \underset{\boldsymbol{o}}{\operatorname{argmax}}\{\sum_{q} p(\boldsymbol{o}, q|w, \hat{\lambda})\} \tag{2.5}$$

$$\approx \underset{\boldsymbol{o},q}{\operatorname{argmax}}\{p(\boldsymbol{o}|w, \hat{\lambda})\} \tag{2.6}$$

$$= \underset{\boldsymbol{o},q}{\operatorname{argmax}}\{P(q|w, \hat{\lambda}) \cdot p(\boldsymbol{o}|q, \hat{\lambda})\} \tag{2.7}$$

$$\approx \underset{\boldsymbol{o}}{\operatorname{argmax}}\{p(\boldsymbol{o}|\hat{q}, \hat{\lambda})\} \tag{2.8}$$

$$= \underset{\boldsymbol{o}}{\operatorname{argmax}}\{N(\boldsymbol{o}; \boldsymbol{\mu_q}, \boldsymbol{\Sigma_q})\} \tag{2.9}$$

donde $\boldsymbol{o} = [\boldsymbol{o}_1, \ldots, \boldsymbol{o}_T]$ es la secuencia de parámetros generada, $q = \{q_1, \ldots, q_T\}$ es la secuencia de estados, $\boldsymbol{\mu_q} = [\boldsymbol{\mu_{q1}}, \ldots, \boldsymbol{\mu_{qT}}]$ es el vector de medias para q, $\boldsymbol{\Sigma_q} = diag[\boldsymbol{\Sigma_{q1}}, \ldots, \boldsymbol{\Sigma_{qT}}]$ es la matriz de covarianza para q, y T es el número total de tramas en $\boldsymbol{o}$. Además, la secuencia de estados $\hat{q}$ se determina de tal manera que maximice la probabilidad de duración del estado usando las distribuciones explícitas del HSMM como sigue:

$$\hat{q} = \underset{q}{\operatorname{argmax}}\{P(q; w, \hat{\lambda})\} \tag{2.10}$$

Una vez los parámetros acústicos $\boldsymbol{o}$ han sido estimados, se emplean para reconstruir la señal de voz con *vocoders*.

2.2.3 Características Dinámicas

Durante la síntesis, debido a que las salidas de los estados de los HSMM son independientes entre sí, si se maximizara $p(\boldsymbol{o}|\hat{q}, \hat{\lambda})$, el vector de parámetros resultante es directamente el vector de medias de la distribución. Esto provocaría que la secuencia de parámetros presente la forma de función escalonada, lo cual no modela correctamente la evolución real de la voz natural, donde los parámetros acústicos varían suavemente. Como resultado, la voz sintetizada presenta saltos y discontinuidades perceptibles cuando se escucha.

Para generar una secuencia de parámetros más realista, tanto en el proceso de entrenamiento como posteriormente en la generación de parámetros, no solo se tienen en cuenta los valores estáticos, sino también el valor de los parámetros dinámicos [115]. Generalmente, se incluyen la primera y la segunda derivada junto con los parámetros estáticos. Por lo tanto, el vector de observación o_t para la trama t queda definido como:

$$\boldsymbol{o_t} = [\boldsymbol{c_t}, \Delta\boldsymbol{c_t}, \Delta^2\boldsymbol{c_t}] \tag{2.11}$$

dónde $\boldsymbol{c_t}$ es el vector de parámetros estáticos, $\Delta\boldsymbol{c_t}$ es la primera derivada y $\Delta^2\boldsymbol{c_t}$ es la segunda derivada. Estas derivadas se definen, respectivamente, como:

$$\Delta\boldsymbol{c_t} = \frac{\partial\boldsymbol{c_t}}{\partial t} \approx 0{,}5(\boldsymbol{c_{t+1}} - \boldsymbol{c_{t-1}}) \tag{2.12}$$

$$\Delta^2 \boldsymbol{c_t} = \frac{\partial^2 \boldsymbol{c_t}}{\partial t^2} \approx \boldsymbol{c_{t+1}} - 2\boldsymbol{c_t} + \boldsymbol{c_{t-1}} \tag{2.13}$$

De esta manera se pueden capturar las dependencias temporales entre los HSMM, y la secuencia de parámetros generada es más realista.

2.2.4 Oversmoothing y Global Variance

Otro factor a tener en cuenta es que la secuencia de parámetros generada se aproxima demasiado a la media de la distribución, incluso habiendo considerado las dependencias temporales entre HSMMs. Este fenómeno se conoce como *oversmoothing* o sobresuavizado y tiene como resultado una voz generada poco natural y con zumbido, al no representar correctamente las variaciones sobre la media que se dan en la voz real. Para tratar de solventarlo se han estudiado diferentes opciones, como implementar un post-filtro para enfatizar determinados aspectos relevantes del espectro [133]. Sin embargo, incluir un post-filtro puede introducir sonidos artificiales y degradar la similitud de la voz sintetizada con la del locutor original [56]. Otras soluciones serían emplear datos de entrenamiento explícitamente para generar parámetros [66] o integrar modelos estadísticos de varios niveles para generar la trayectoria de los parámetros [122]. Uno de los métodos más exitosos en esta última categoría consiste en un algoritmo de generación teniendo en cuenta la varianza global o *Global Variance (GV)* en inglés [112]. Éste trata de recuperar el rango dinámico de las trayectorias generadas acercándolas a las naturales. La GV, $\boldsymbol{v(c)}$, se define como la variación intra-oración de la trayectoria de la parte estática del vector acústico, $\boldsymbol{c}$, de la siguiente manera:

$$\boldsymbol{v(c)} = [\boldsymbol{v(1)} \dots \boldsymbol{v(M)}] \tag{2.14}$$

$$\boldsymbol{v(m)} = \frac{1}{T} \sum_{t=1}^{T} \{\boldsymbol{c_t(m)} - \boldsymbol{\mu(m)}\}^2 \tag{2.15}$$

$$\boldsymbol{\mu(m)} = \frac{1}{T} \sum_{t=1}^{T} \boldsymbol{c_t(m)} \tag{2.16}$$

siendo $\boldsymbol{c_t(m)}$ el vector de características estáticas para la trama t y la frase m, y M el número total de frases disponibles durante el entrenamiento.

Se calcula la GV para cada frase de entrenamiento y se modela usando una única distribución gaussiana multidimensional:

$$p(\boldsymbol{v(c)}|\lambda_{GV}) = N(\boldsymbol{v(c)}|\boldsymbol{\mu_{GV}}, \boldsymbol{\Sigma_{GV}}) \tag{2.17}$$

donde $\boldsymbol{\mu_{GV}}$ es el vector media y $\boldsymbol{\Sigma_{GV}}$ es la matriz de covarianza para la GV. El algoritmo de generación considerando la GV maximiza la siguiente función objetivo con respecto a c:

$$F_{GV}(\boldsymbol{c}; \lambda, \lambda_{GV}) = \omega log N(\boldsymbol{Wc}; \boldsymbol{\mu_q}, \boldsymbol{\Sigma q}) + log N(\boldsymbol{v(c)}; \boldsymbol{\mu_{GV}}, \boldsymbol{\Sigma_{GV}}) \tag{2.18}$$

donde ω es un peso que balancea las probabilidades del HSMM y GV. El segundo término de la ecuación puede entenderse como una penalización para prevenir el *oversmoothing*, dado que trata de mantener el rango dinámico de la trayectoria generada cerca de la de los datos de entrenamiento. Como resultado, la síntesis estadística genera una voz más natural.

2.3 Técnicas de adaptación

Una de las características que ha ayudado a popularizar la síntesis estadístico paramétrica es la capacidad de poder adaptar una voz ya entrenada, para obtener otra con diferente estilo, emoción o que suene como si se tratase de otro locutor. Las técnicas de adaptación permiten obtener nuevas voces sintéticas partiendo de una base de datos más reducida en comparación con la que sería necesaria para entrenar modelos acústicos de dicha voz desde cero.

Como alternativa a las técnicas de adaptación de locutor podrían emplearse técnicas de conversión de voz sobre las señales sintéticas [39, 103]. La conversión de voz consiste en modificar una señal de audio real o sintética, o sus parámetros acústicos, para conseguir una señal de salida cuyas características coincidan con las de la voz del locutor objetivo. Mientras que en la adaptación se modifican los modelos acústicos, de manera que, al extraer los parámetros éstos ya contienen las características del locutor objetivo, en la conversión son los parámetros extraídos o la señal reconstruida los que se modifican. Así, la técnica de adaptación simplifica la fase de síntesis y mejora el rendimiento del sistema, puesto que no requiere el procesamiento adicional de los parámetros acústicos necesario en la conversión.

En la literatura se pueden encontrar diversos métodos y aplicaciones de adaptación de modelos: interpolación de modelos [106], HMMs de regresión múltiple [77], modelado de estilo y emoción [128] o adaptación de estilos [107].

La figura 2.2 muestra el diagrama de la fase de adaptación, en la que se parte de una base de datos de adaptación y de unos modelos acústicos ya entrenados. La base de datos, al igual que en la fase de entrenamiento descrita en la sección 2.2.1, consiste en grabaciones de audio con sus transcripciones. En este caso, el audio de las grabaciones presenta las características de estilo, emoción o del locutor al que se desea adaptar los modelos. Tal y como sucede en la fase de entrenamiento de la figura 2.1, las grabaciones se parametrizan utilizando un *vocoder* y de las transcripciones se extraen etiquetas lingüísticas. Durante la adaptación se modifican los modelos acústicos iniciales en base a los parámetros acústicos y las etiquetas lingüísticas para que se ajusten lo máximo posible al locutor objetivo. Una vez que los modelos ya han sido adaptados se emplean en la fase de síntesis para obtener la señal de audio de salida como se describe en la sección 2.2.2.

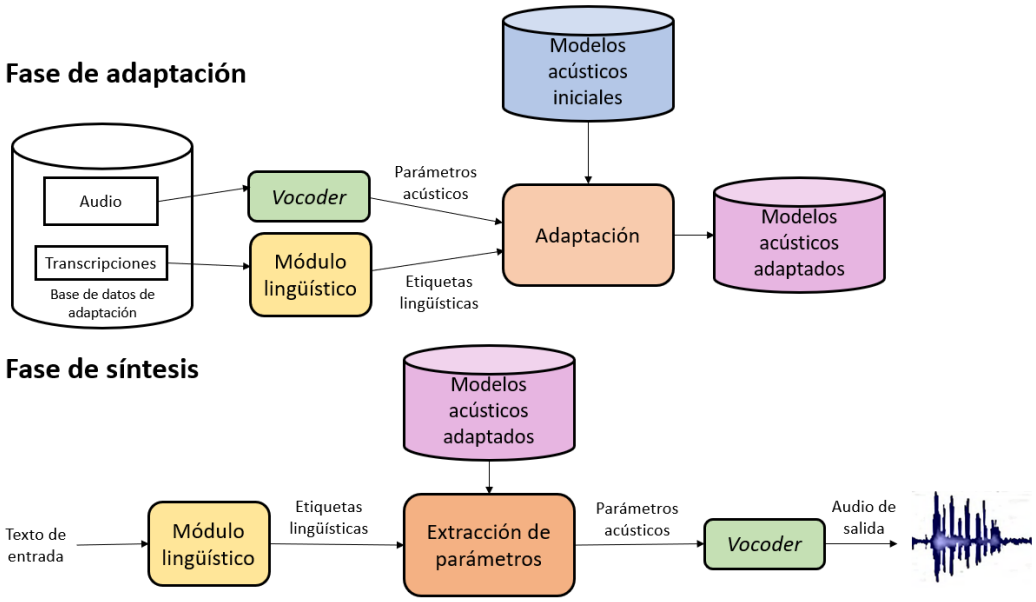


Figura 2.2: Proceso de adaptación: fases de adaptación y síntesis.

2.3.1 Adaptación de Locutor HTS

Aunque la adaptación puede aplicarse a diferentes aspectos, en esta tesis nos centramos en la adaptación de locutor con el objetivo de crear voces nuevas que suenen como locutores nuevos.

Durante la adaptación de los HSMM, como modelo inicial se puede emplear el resultado del entrenamiento descrito en la sección 2.2.1. Para la adaptación de locutor se suele emplear un modelo generado usando datos de entrenamiento consistentes en grabaciones de varios locutores. A este tipo de modelo se le denomina voz promedio. Debido a que los datos de entrenamiento incluyen gran cantidad de características dependientes del locutor, se emplea el algoritmo *Speaker-Adaptive Training* (SAT) [6] para reducir las influencias negativas de disponer de diferentes locutores [126]. En este algoritmo, los parámetros del modelo se obtienen usando un procedimiento de estimación ciega, en el que se asume que la diferencia de locutor se expresa mediante una transformación lineal del modelo de la voz promedio. El algoritmo SAT empleado en el caso de HSMM se detalla en [127].

Inicialmente únicamente se transformaba la parte de los modelos correspondiente a los parámetros espectrales [65] [108] [127] empleando técnicas desarrolladas años antes para el reconocimiento del habla como *Maximum Likelihood Linear Regression* (MLLR) [61]. Después, para adaptar simultáneamente tanto los parámetros espectrales como los de excitación se usó los HMM de distribución de probabilidad multiespacial [116] y su algoritmo particularizado de adaptación MLLR [109]. El último paso para adaptar todos los parámetros del modelo consiste en adaptar también las duraciones de los HSMM de cada estado.

Existen dos criterios que se pueden usar para estimar las transformaciones de las distribuciones de los estados (generalmente funciones de regresión lineal): el criterio ML y el *Structural Maximum A Posteriori* (SMAP) [98]. A su vez, en cada caso cada estimación puede modificar o bien solo el vector de medias del modelo o tanto el vector de medias como la matriz de covarianza. Dado que la variación que sufren los parámetros, es decir su matriz de covarianza, es un factor muy importante a la hora de adaptar un tipo de voz (especialmente en el caso de la frecuencia fundamental) es necesario que también sea adaptada al nuevo locutor. En el caso más general de MLLR el vector de medias y la matriz de covarianzas se transforman usando matrices diferentes [37]. Sin embargo, en el campo de la síntesis suele emplearse un método computacionalmente más eficiente consistente en emplear la misma matriz de transformación para adaptar tanto la media como la covarianza. A este método se le denomina *Constrained Maximum Likelihood Linear Regression* (CMLLR). El método más popular entre la comunidad científica es el denominado *Constrained Structural Maximum A Posteriori Linear Regression* (CSMAPLR) , que es el resultado de combinar SMAP con CMLLR [129].

2.3.1.1 **CMLLR**

En la adaptación CMLLR de un HMM el vector de media y la matriz de covarianza de la distribución de salida del estado son actualizados simultáneamente. Para ello, a la media se le aplica una transformación lineal multiplicando por una matriz de rotación y añadiendo un vector de traslación o *bias*. La misma matriz se emplea para actualizar la matriz de covarianza. De manera similar, en un HSMM la media

y la covarianza de la distribución de duración del estado también se modifican simultáneamente, tal como se muestra a continuación:

$$b_i(\boldsymbol{o}) = N(\boldsymbol{o}; \boldsymbol{\zeta}'\boldsymbol{\mu_i} - \boldsymbol{\epsilon}', \boldsymbol{\zeta}'\boldsymbol{\Sigma_i}\boldsymbol{\zeta}'^{\top}) \tag{2.19}$$

$$p_i(d) = N(d; \chi' m_i - \nu', \chi' \sigma_i^2 \chi') \tag{2.20}$$

La matriz $\boldsymbol{\zeta}'$ se usa para transformar tanto el vector de medias como la matriz de covarianzas de las distribuciones de salida del estado, χ' se usa para transformar las de la distribución de duración del estado. $\boldsymbol{\epsilon}'$ y ν' son los términos de *bias* de la transformación. Estas transformaciones del modelo son equivalentes a las siguientes transformaciones afines del vector de características $\boldsymbol{o}$ y la duración d del estado i:

$$b_i(\boldsymbol{o}) = N(\boldsymbol{o}; \boldsymbol{\zeta}'\boldsymbol{\mu_i} - \boldsymbol{\epsilon}', \boldsymbol{\zeta}'\boldsymbol{\Sigma_i}\boldsymbol{\zeta}'^{\top}) \tag{2.21}$$

$$= |\boldsymbol{\zeta}| N(\boldsymbol{\zeta o} + \boldsymbol{\epsilon} : \boldsymbol{\mu_i}, \boldsymbol{\Sigma_i}) \tag{2.22}$$

$$= |\boldsymbol{\zeta}| N(\boldsymbol{W\xi} : \boldsymbol{\mu_i}, \boldsymbol{\Sigma_i}) \tag{2.23}$$

$$p_i(d) = N(d; \chi' n_i - \nu', \chi' \sigma_i^2 \chi') \tag{2.24}$$

$$= |\chi| N(\chi d + \nu : m_i, \sigma_i^2) \tag{2.25}$$

$$= |\chi| N(\boldsymbol{X\phi} : m_i, \sigma_i^2) \tag{2.26}$$

donde $\boldsymbol{\zeta}' = \boldsymbol{\zeta}'^{-1}$, $\boldsymbol{\epsilon} = \boldsymbol{\zeta}'^{-1}\boldsymbol{\epsilon}'$, $\chi = \chi'^{-1}$, $\nu = \chi'^{-1}\nu'$, $\boldsymbol{\xi} = [\boldsymbol{o}^{\top}, 1]^{\top}$, y $\boldsymbol{\phi} = [d, 1]^{\top}$. $\boldsymbol{W} = [\boldsymbol{\zeta}, \boldsymbol{\epsilon}]$ y $\boldsymbol{X} = [\chi, \nu]$ son las matrices de transformación de la salida y de la duración del estado respectivamente. Se estima un conjunto de transformaciones $\Lambda = (\boldsymbol{W}, \boldsymbol{X})$ maximizando la verosimilitud de los datos de adaptación $\boldsymbol{O}$ de longitud T.

$$\tilde{\Lambda} = (\tilde{\boldsymbol{W}}; \tilde{\boldsymbol{X}}) = \underset{\Lambda}{\operatorname{argmax}}(\boldsymbol{O}|\lambda, \Lambda) \tag{2.27}$$

donde λ es el conjunto de parámetros del HSMM. Las fórmulas de reestimación que se emplean están basadas en el algoritmo *Expectation Maximization* (EM) [25]. Un método iterativo para estimar $\boldsymbol{W}$ es el descrito en [64]. Por el contrario, $\boldsymbol{X}$ tiene solución analítica que es explicada en [65].

Aunque aquí se ha expuesto el algoritmo para una única transformación global se puede extender a múltiples transformaciones utilizando regresión lineal. Para agrupar las distribuciones en el modelo y unir las transformaciones para cada grupo se emplean árboles de decisión contextuales.

Este algoritmo tiene el efecto de adaptar la información prosódica dado que las variaciones de log-f_0 y de la duración son muy importantes en la síntesis de voz. Por ejemplo, si el locutor objetivo tiene un estilo del habla caracterizado por una modulación y un ritmo, es decir variaciones de la log-f_0 más amplios que los de la voz promedio, no se puede imitar dicho estilo sin adaptar la varianza de la log-f_0.

2.3.1.2 **CSMAPLR**

El algoritmo CMLLR emplea el criterio de ML para estimar la transformación. Este criterio funciona bien en la fase de entrenamiento de la voz promedio usando el algoritmo SAT, siempre que se tengan suficientes muestras para generar dicha voz. En la fase de adaptación sin embargo la cantidad de datos disponibles para realizar la adaptación es más limitada por lo tanto se necesita un criterio más robusto, como el criterio *Maximum A Posteriori* (MAP). En la estimación MAP se estiman las transformaciones empleando la siguiente ecuación:

$$\hat{\Lambda} = (\hat{\boldsymbol{W}}; \hat{\boldsymbol{X}}) = \underset{\Lambda}{\operatorname{argmax}}(\boldsymbol{O}|\lambda, \Lambda)P(\Lambda) \tag{2.28}$$

donde $P(\Lambda)$ es la distribución *a priori* de las transformaciones $\boldsymbol{W}$ y $\boldsymbol{X}$. Para la distribución *a priori*, son convenientes las siguientes matrices combinadas de distribución normal multivariante (versiones matriciales de la distribución normal multivariante [40]):

$$\begin{aligned} P(\Lambda) \propto |\boldsymbol{\Omega}|^{-\frac{L+1}{2}}|\boldsymbol{\Psi}|^{-\frac{L}{2}}|\tau_p|^{-1}|\boldsymbol{\psi}|^{-\frac{1}{2}} \\ \times exp\{-\frac{1}{2}tr(\boldsymbol{W}-\boldsymbol{H})^\top\boldsymbol{\Omega}^{-1}(\boldsymbol{W}-\boldsymbol{H})\boldsymbol{\Psi}^{-1}\} \\ \times exp\{-\frac{1}{2}tr(\boldsymbol{X}-\boldsymbol{\eta})^\top\tau_p^{-1}(\boldsymbol{X}-\boldsymbol{\eta})\boldsymbol{\psi}^{-1}\} \end{aligned} \tag{2.29}$$

donde $\propto$ significa proporcional. $\boldsymbol{\Omega}$, $\boldsymbol{\Psi}$, $\boldsymbol{H}$, τ_p, $\boldsymbol{\psi}$ y $\boldsymbol{\eta}$ son los hiperparámetros de la distribución *a priori*.

En el criterio SMAP [98] inicialmente se realiza una estimación global empleando todos los datos de adaptación en el nodo raíz de un árbol de decisión, que posteriormente propaga a sus nodos hijos junto con los hiperparámetros $\boldsymbol{H}$ y $\boldsymbol{\eta}$. En los nodos hijos, sus transformaciones son estimadas de nuevo usando el criterio MAP con los hiperparámetros propagados. Así una estimación recursiva basada en MAP se realiza desde el nodo raíz hasta los nodos finales. La adaptación *Structural Maximum A Posteriori Lineal Regression* (SMAPLR) se desarrolló aplicando el criterio SMAP al MLLR [99]. Para el algoritmo CSMAPLR se aplica el criterio SMAP a la adaptación CMLLR y se usa el criterio MAP recursivo para estimar simultáneamente los vectores de media y las matrices de covarianza. En este caso se fijan los valores de $\boldsymbol{\Psi}$ y $\boldsymbol{\psi}$ a la matriz identidad, y $\boldsymbol{\Omega}$ a la matriz identidad escalada ($\boldsymbol{\Omega} = \tau_p \boldsymbol{I}$) de manera que el escaldado lo fija el hiperparámetro positivo τ_p, de la misma forma que en la adaptación SMAPLR. Las fórmulas de reestimación del algoritmo EM y su resolución para CSMAPLR son las detalladas en [129].

2.3.1.3 Combinación de Regresión Lineal con Adaptación MAP

En las adaptaciones anteriormente mencionadas se asume que el modelo del locutor objetivo puede representarse mediante una regresión lineal de la voz promedio. Si además se aplica una adaptación MAP al modelo transformado por la regresión lineal [15, 26], es posible modificar apropiadamente la estimación de la distribución teniendo relativamente suficiente cantidad de muestras de voz. La adaptación MAP de los vectores de media de las transformaciones realizadas por CSMAPLR se pueden estimar como sigue:

$$\hat{\boldsymbol{\mu}}_{\boldsymbol{i}} = \frac{v_b \boldsymbol{\mu_i} + \sum_{t=1}^{T} \sum_{d=1}^{t} \gamma_t^d(i) \sum_{s=t-d+1}^{t} \hat{\boldsymbol{o}}_{\boldsymbol{s}}}{v_b + \sum_{t=i}^{T} \sum_{d=1}^{t} \gamma_t^d(i)} \tag{2.30}$$

$$\hat{m}_i = \frac{v_p m_i + \sum_{t=1}^{T} \sum_{d=1}^{t} \gamma_t^d(i) \hat{d}}{v_p + \sum_{t=i}^{T} \sum_{d=1}^{t} \gamma_t^d(i)} \tag{2.31}$$

donde $\boldsymbol{\mu_i}$ y m_i son el vector de medias y la media de la salida del estado y de la distribución de la duración de la voz promedio respectivamente. $\hat{\boldsymbol{o}}_{\boldsymbol{s}} = \hat{\boldsymbol{\zeta}}_{\boldsymbol{s}} \boldsymbol{o_s} + \hat{\boldsymbol{\epsilon}}$ y $\hat{d}_s = \hat{\chi}_s d_s + \hat{\nu}$ y son el vector de observaciones y la duración linealmente transformadas usando la adaptación CSMAPLR de los HSMM respectivamente. v_b y v_p son los hiperparámetros positivos de las distribuciones de la salida del

estado y de la distribución de la duración, respectivamente. A medida que aumenta el número de datos de adaptación también, aumenta el número de distribuciones que disponen de suficientes muestras de voz, y por lo tanto, este algoritmo mejora gradualmente la calidad de la voz sintética.

2.4 Aportaciones

En el ámbito de la síntesis estadístico paramétrica, durante la realización de esta tesis se desarrolló un sistema TTS multilingüe basado en síntesis estadístico paramétrica. En el marco del proyecto *low cost telebista*[1] se identificó la necesidad de disponer de un sistema capaz de sintetizar en los diferentes idiomas oficiales de España (castellano, euskera, catalán y gallego) además de en inglés. Como punto de partida se tomó el sintetizador *AhoTTS* [42] del grupo de investigación Aholab, que dispone de módulos lingüísticos para castellano y euskera. Además, se integraron módulos lingüísticos de código abierto de otras universidades y grupos de investigación:

- Catalán: módulo festcat [33] del grupo TALP de la Universidad Politécnica de Cataluña.
- Gallego: el módulo cotovía [22] del grupo GTM de la Universidad de Vigo.
- Inglés: el modulo festival [34] del grupo CSRT de la Universidad de Edimburgo.

También se entrenaron voces para varios locutores e idiomas usando diferentes bases de datos disponibles. El tamaño de las bases de datos son los mostrados en la tabla 2.1.

Tabla 2.1: Tamaño de las bases de datos usadas para el entrenamiento de las voces del sistema TTS multilingüe.

Voz	# de frases	# de palabras	Duración aprox. (horas)
Castellano	3.995	51.380	6
Euskera	3.799	38.544	6
Catalán masculina	3.692	58.154	6
Catalán femenina	3.974	62.314	6
Gallego masculina	1.316	11.235	1
Inglés femenina	1.132	10.002	1

[1] https://aholab.ehu.eus/aholab/es/tv-social-low-cost-telebista/

El sistema se liberó con licencia de código abierto[1] y se difundió en la siguiente publicación científica:

A. Alonso, I. Sainz, D. Erro, E. Navas, I. Hernaez, "Sistema de conversión texto a voz de código abierto para lenguas ibéricas"(in Spanish), Procesamiento del Lenguaje Natural (ISSN: 1135-5948), vol. 51, pp. 169-175, 2013.

[1] https://sourceforge.net/projects/ahottsmultiling/

CAPÍTULO

3

Frequency Warping + Amplitude Scaling

En el presente capítulo se expone un tipo de transformación diferente a las vistas en el capítulo anterior, basado en proyección de frecuencia o *Frequency Warping (FW)* en inglés, y en escalado de amplitud o *Amplitude Scaling (AS)* en inglés. Aunque esta técnica suele emplearse en conversión de voz, aquí se propone su uso en adaptación de locutor en conversión de texto a voz. Se presenta un método de adaptación haciendo uso de FW+AS y los experimentos que se han llevado a cabo para comprobar su eficacia.

3.1 Introducción

Una transformación de FW es una operación que se aplica trama a trama a la componente espectral de una señal. Cambia la posición de los puntos representativos (por ejemplo, la posición o valores de frecuencia de los formantes) de un espectro fuente para mapearlos en la posición de su equivalente en el espectro objetivo. No elimina ningún detalle del espectro de la fuente, sino que simplemente mueve ciertas características relevantes de una posición a otra en el eje frecuencial, en consecuencia FW mantiene bastante bien la calidad de la señal de voz. Sin embargo, la precisión en la conversión que se consigue empleando únicamente FW resulta moderada puesto que las amplitudes relativas entre los puntos representativos del espectro no se alteran. Por ese motivo, se suele complementar con una transformación adicional de AS para compensar las diferencias en amplitud. A esta combinación de proyección de frecuencias y escalado de amplitud le llamaremos de ahora en adelante FW+AS. La figura 3.1 muestra una representación de los diferentes pasos de una transformación basada en FW+AS aplicada a la envolvente espectral de una trama de la voz fuente para transformarla a su trama equivalente de la voz objetivo.

Estas transformaciones se emplean en conversión de voz [39, 103, 119], en donde el objetivo es modificar las características de un locutor fuente para que tenga las del locutor objetivo, sin modificar el contenido del mensaje. Esta idea se puede aplicar a síntesis de voz para cambiar las características de la voz con la que se genera el mensaje leído. Con la señal de voz sintética ya generada, ésta se puede parametrizar para aplicar sobre ella las transformaciones. Después se reconstruye

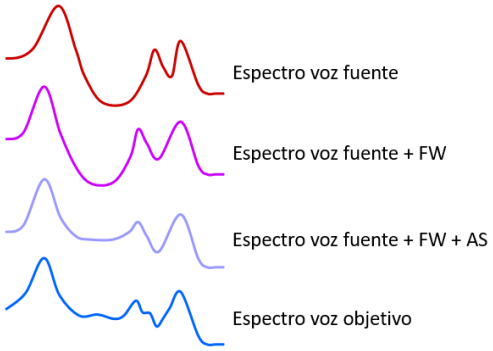


Figura 3.1: Operaciones de *Frequency Warping* y *Amplitud Scaling*.

la señal utilizando los parámetros transformados, obteniendo así el mismo mensaje con la voz objetivo.

En el caso de la síntesis estadístico paramétrica ya se dispone de los parámetros acústicos de la señal, ya que se han obtenido utilizando los modelos de síntesis. Si se aplican las transformaciones antes de reconstruir la señal con el *vocoder*, la señal generada ya presentará las características de la voz objetivo.

La idea que se presenta en este capítulo consiste en aplicar las transformaciones por FW+AS directamente sobre los modelos estadísticos, en lugar de aplicarlos sobre los parámetros ya generados por los modelos. De este modo, los parámetros que se extraen de los modelos presentarán las características de la voz objetivo y no será necesario realizar un procesamiento adicional antes de alimentar al *vocoder*.

Como se verá más adelante, al comparar las transformaciones por FW+AS con las vistas en la sección 2.3 las primeras presentan dos inconvenientes. En primer lugar, únicamente se adapta la envolvente espectral del modelo. Otros parámetros acústicos como los relacionados con la excitación (por ejemplo: pitch, log-f_0, MVF o bandas de aperiodicidad) o la duración no se modifican. En segundo lugar, ofrecen menos grados de libertad por lo tanto la capacidad de adaptación está más limitada.

No obstante, FW+AS también presenta una importante ventaja: FW+AS es más robusta en el caso de que la calidad de los datos de adaptación no sea la óptima. Es decir, si el material disponible de la voz objetivo presenta ruidos, chasquidos u otros efectos no deseados o *artifacts*, éstos no se transmiten a la voz transformada.

Esta característica hace a estas transformaciones apropiadas cuando es esperable que los datos disponibles para adaptar no sean buenos. Se garantiza que la calidad de la señal adaptada se mantenga en unos márgenes aceptables, aunque sea a costa de la capacidad de adaptación.

El resto de este capítulo se estructura de la siguiente manera: en primer lugar se explica matemáticamente una opción para calcular las transformaciones de FW+AS y aplicarlas a los modelos de una voz estadístico paramétrica. A continuación, se expone un método de adaptación empleando las transformaciones calculadas anteriormente usando únicamente fragmentos vocálicos como material de entrenamiento. Por último, se describen los experimentos llevados a cabo para comprobar la idoneidad del método propuesto y sus resultados.

3.2 Dynamic Frequency Warping

Para poder aplicar la transformación que mapea el eje frecuencial de la fuente en el eje frecuencial del objetivo, dicha transformación tiene que ser entrenada. Para ello es necesario disponer de material de adaptación, tanto de la fuente como del objetivo. Generalmente, el material de adaptación consiste en corpus paralelos, que contiene un conjunto de frases que han sido grabadas tanto por el locutor fuente como por el locutor objetivo. Hay diferentes formas de calcular la transformación como *Dynamcic Frequency Warping* (DFW) [39], *Voice Track Length Normalization* (VTLN) [103], *Bi-Linear Frequency Warping* (BLFW) [29] o *Correlation Frequency Warping* (CFW) [111].

El método descrito en [39] (DFW) es uno de los más comúnmente utilizados y es también el elegido para este trabajo. En ella se calcula la función que se debe aplicar a un conjunto de T semiespectros en amplitud logarítmica de $(N+1)$ puntos o *bins* correspondientes a la fuente $\{X_t\}$ para hacerlos lo más cercano posible a sus contra parte correspondientes al objetivo $\{Y_t\}$. Está basada en una función de coste global $D(i,j)$, que indica la distorsión log-espectral acumulada en el caso de que el *bin* i-$ésimo$ de la fuente se mapeara en el *bin* j-$ésimo$ del objetivo, siguiendo el "mejor" camino desde $D(0,0)$ hasta $D(i,j)$. Esta función se puede expresar matemáticamente como:

$$D(i,j) = \text{mín} \left\{ \begin{array}{l} D(i-1,j) + d(i,j) \\ D(i-1,j-1) + w \cdot d(i,j) \\ D(i,j-1) + d(i,j) \end{array} \right\} \tag{3.1}$$

donde $i,j = 0,1,\ldots N$ es el índice de los *bin*; $d(i,j)$ es una medida de distorsión local que tiene en cuenta únicamente el *bin* i-$ésimo$ de la fuente y el *bin* j-$ésimo$ del objetivo. w, es un parámetro que controla la penalización entre los caminos verticales y horizontales y permite evitar que se favorezcan los caminos más cortos, es decir la diagonal. Su valor varía entre $1 \leq w \leq 2$, y $w \approx 2$ significa que no hay penalización mientras que $w \approx 1$ significa una fuerte penalización. En esta tesis se propone una función de d, definida en la ecuación 3.2, que permite calcular la distorsión simultáneamente para todos los vectores de entrenamiento disponibles y así optimizar globalmente la función coste global:

$$d(i,j) = \sum_{t=1}^{T} \left(X_t[i] - Y_t[j]\right)^2 + \sum_{t=1}^{T} \alpha_t \left(X'_t[i] - Y'_t[j]\right)^2 \tag{3.2}$$

donde T es el número total de parejas de entrenamiento, $X'_t[i]$ e $Y'_t[i]$ son las derivadas respecto a la frecuencia de $X_t[i]$ e $Y_t[i]$ respectivamente, y α_t es un factor empírico que se emplea para ponderar la importancia de las derivadas respecto del espectro. La primera parte de la ecuación 3.2 captura la diferencia de valores absolutos amplitudes espectrales locales, y la segunda ayuda a alinear eventos espectrales dados por pendientes abruptas, es decir, formantes. Esta propuesta es una evolución de la reflejada en [139] donde únicamente se tiene en cuenta la primera parte de la ecuación.

El camino del *warping* P queda definido por la secuencia de puntos,

$$P = \{(0,0), (i_1, j_1), (i_2, j_2) \ldots (N,N)\} \tag{3.3}$$

en el cual la presencia del punto (i,j) indica que el bin $i\text{-}ésimo$ del espectro de la fuente se debe mapear con el bin $j\text{-}ésimo$ del espectro del objetivo. Los puntos se calculan retrocediendo desde (N,N) hasta $(0,0)$ siguiendo el camino inverso que garantiza la distorsión mínima usando la recursión expresada en la ecuación 3.1.

3.2.1 Aplicación de Frequency Warping en el Dominio Cepstral

En el dominio cepstral, las transformaciones de FW pueden expresarse como matrices multiplicativas [85] y el AS puede expresarse como un vector de bias aditivo.

De esta manera, si se aplican las transformaciones a las distribuciones de salida de un estado HSMM para adaptar los MFCC, simplemente hay que modificar la media y la varianza de su distribución de salida de los parámetros espectrales de la siguiente forma:

$$\hat{\boldsymbol{\mu}} = \check{\boldsymbol{A}}\boldsymbol{\mu} + \check{\boldsymbol{b}}, \;\; \hat{\boldsymbol{\Sigma}} = \check{\boldsymbol{A}}\boldsymbol{\Sigma}\check{\boldsymbol{A}}^{\top} \tag{3.4}$$

donde $\boldsymbol{\mu}$ y $\boldsymbol{\Sigma}$ son el vector de medias y la matriz de covarianzas de estado HSMM y $\hat{\boldsymbol{\mu}}$ y $\hat{\boldsymbol{\Sigma}}$ sus transformadas. La matriz $\check{\boldsymbol{A}}$ y el vector $\check{\boldsymbol{b}}$ se definen en la ecuación 3.5

y se emplean para modificar simultáneamente las partes estáticas y dinámicas de $\boldsymbol{\mu}$ y de $\boldsymbol{\Sigma}$.

$$\check{\boldsymbol{A}} = \begin{bmatrix} \boldsymbol{A} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{A} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{A} \end{bmatrix}, \quad \check{\boldsymbol{b}} = \begin{bmatrix} \boldsymbol{b} \\ \boldsymbol{0} \\ \boldsymbol{0} \end{bmatrix} \tag{3.5}$$

donde $\boldsymbol{A}$ es la matriz de translación que representa la operación de FW y $\boldsymbol{b}$ el vector de bias que representa la operación de AS.

Dado que DFW trabaja en el dominio log-espectral, las parejas de datos tanto de la fuente como del objetivo deben ser convertidas de vectores de orden p del dominio cepstral a $(N+1)$ puntos del semiespectro positivo. Esto se puede hacer mediante la ecuación 3.6

$$\boldsymbol{S}[n,i] = \cos(i \cdot mel(\pi n/N)), \;\; 0 \leq n \leq N, \;\; 0 \leq i \leq p \tag{3.6}$$

De manera similar, la representación en vectores MFCC de orden p del espectro discreto en log-amplitud se puede calcular mediante la técnica conocida como *regularized discrete cepstrum* [10]. En ella, se multiplican los vectores de $(N+1)$ puntos del semiespectro por un vector de la forma:

$$\boldsymbol{C} = \left(\boldsymbol{S}^T\boldsymbol{S} + \lambda\boldsymbol{R}\right)^{-1}\boldsymbol{S}^T \tag{3.7}$$

donde $\boldsymbol{S}$ está definida en la ecuación 3.6, $\boldsymbol{R}$ es una matriz de regularización que impone unas restricciones de suavizado en la envolvente cepstral definida como:

$$\boldsymbol{R} = 8\pi^2 \cdot diag\left\{0, 1^2, 2^2, \ldots p^2\right\} \tag{3.8}$$

y λ es una restricción empírica normalmente igual a $2 \cdot 10^{-4}$ [10].

En la práctica, el coeficiente de orden 0 relacionado con la energía, y el coeficiente de orden 1 relacionado con el espectro glotal, no son relevantes en términos de FW [139], por lo que se fijan a un valor 0 antes de multiplicar por $\boldsymbol{S}$. Una vez que los vectores MFCC se han trasladado al dominio log-espectral mediante la ecuación 3.6, se calcula el camino óptimo de warping P (equación 3.3) vía DFW. El camino P se puede definir de manera matricial de la siguiente forma

$$\boldsymbol{W}[j,i] = \frac{m_{i,j}}{\sum_{k=1}^{N} m_{i,k}} \tag{3.9}$$

donde $m_{i,j} = 1$ si el punto (i,j) pertenece a P, y $m_{i,j} = 0$ en caso contrario. El denominador se encarga de compensar los mapeos uno-a-varios entre los *bins* de la fuente y del objetivo, los cuales son inevitables de acuerdo a la recursión en la ecuación 3.1 y la estructura resultante de P. Una vez $\boldsymbol{W}$ se ha determinado, la matriz $\boldsymbol{A}$ que implemente la operación de DFW en el dominio MFCC queda definida como:

$$\boldsymbol{A} = \boldsymbol{C} \cdot \boldsymbol{W} \cdot \boldsymbol{S} \tag{3.10}$$

3.3 Amplitude Scaling

El término aditivo de bias que implementa la operación de AS se puede calcular como la diferencia entre los vectores fuente, tras aplicar la transformación de DFW y los vectores objetivo de la siguiente manera:

$$\boldsymbol{b} = \frac{1}{T}\sum_{t=1}^{T}(\boldsymbol{y_t} - \boldsymbol{A}\cdot\boldsymbol{x_t}) \tag{3.11}$$

donde $\boldsymbol{x_t}$ e $\boldsymbol{y_t}$ representan los vectores MFCC fuente y objetivo respectivamente, $\boldsymbol{A}$ está definida en la ecuación 3.5 y T es el número total de parejas de vectores de entrenamiento.

3.4 Método de Adaptación Propuesto

A continuación, se detalla el método propuesto para realizar la adaptación de los modelos HSMM usando DFW + AS explicado en el apartado anterior (ecuaciones 3.4). Como se ha dicho anteriormente, el objetivo principal de esta tesis consiste en proporcionar voces adaptadas para personas con alguna discapacidad oral. Así, frecuentemente, cuando existe una discapacidad oral la voz presenta alguna patología, y la pronunciación de los diferentes sonidos es dificultosa para el usuario, o incluso para algunos de ellos, será ininteligible. Por ello, al aplicar las técnicas de adaptación descritas en la sección 2.3, dicha patología quedará también reflejada en la voz adaptada, por lo que es necesario investigar otras estrategias.

De entre todos los sonidos, las vocales son a priori los sonidos más fáciles de pronunciar, ya que no exigen movimiento de los articuladores. Por ello, se espera que las vocales conserven las características del locutor. En esta sección, proponemos usar únicamente vocales para realizar la adaptación, utilizando la técnica de adaptación DFW+AS, que permite un mayor control de las características a adaptar. La información acústica extraída de las vocales se extrapolará al resto de sonidos.

3.4.1 Selección de Parejas de Datos de Entrenamiento

Para entrenar las transformaciones de DFW+AS es necesario disponer de las parejas de entrenamiento fuente-objetivo. Partiendo de grabaciones de la voz objetivo se realiza una segmentación para delimitar las partes que corresponden a vocales. Éstas se parametrizan y se selecciona la trama central de cada vocal que cumpla estas dos condiciones:

- Para asegurarnos de que el análisis espectral es preciso y existe una zona lo suficientemente libre de co-articulación, los segmentos deben tener una duración mínima, escogida de manera arbitraria, de 55ms.

- Para aliviar posibles *artifacts* debidos a una mala detección de la f_0, todas las tramas correspondientes los segmentos seleccionados deben ser sonoras.

Los vectores de la voz fuente se extraen de los modelos de la voz sintética. Con las transcripciones de las grabaciones realizadas por el locutor objetivo se generan las etiquetas lingüísticas que representan el texto. Las etiquetas se emplean para extraer del modelo el vector de parámetros correspondientes a esa información, igual que en la fase de síntesis descrita en la sección 2.2.2. Puesto que los HSMM constan de varios estados, se toma la salida del estado central puesto que representa la parte más estable del fonema y con menores efectos de coarticulación. Como vector de entrenamiento se utiliza únicamente la parte estática.

3.4.2 Cálculo y Aplicación de las Transformaciones

Teniendo un total de T parejas de vectores fuente-objetivo, mediante la ecuación 3.9 se calculan las matrices de *warping* $\boldsymbol{W_v}$, siendo v la vocal, empleando en cada caso el material correspondiente a cada vocal. Además, se calcula una matriz genérica adicional con todo el material disponible de todas las vocales. Como componente AS, se calcula un único vector empleando todas las parejas de vectores mediante la ecuación 3.11, aplicando para cada caso la matriz $\boldsymbol{W_v}$ correspondiente a la vocal.

Para adaptar el modelo usando estas transformaciones, se recorren todos los estados de todos los HSMM y se aplican las ecuaciones 3.4 para modificar los vectores de media y las matrices de covarianza. En el caso de que el estado corresponda a una vocal, se emplea la matriz $\boldsymbol{W_v}$ entrenada usando el material correspondiente a esa vocal. En otro caso, se emplea la matriz genérica entrenada usando todo el material disponible. Sin embargo, debido a que durante el entrenamiento se realiza un *tying* de los estados (ver sección 2.2.1), es difícil determinar si un HSMM corresponde a un fonema o a otro. Por ese motivo, a priori no se puede saber qué matriz $\boldsymbol{W_v}$ hay que aplicar a cada estado. Para solucionarlo se realiza una estadística que relaciona un estado con los diferentes fonemas, empleando los árboles generados durante el entrenamiento y siguiendo el siguiente proceso:

1. Por cada combinación fonema - nodo hoja del árbol se inicializa un contador a cero.

2. Se toman todas las etiquetas lingüísticas que se han usado en el entrenamiento del modelo de la voz fuente.

3. Por cada etiqueta se recorre el árbol hasta llegar al nodo hoja correspondiente.

4. Se incrementa en uno el contador del fonema correspondiente a la etiqueta lingüística en el nodo hoja.

Una vez finalizado este proceso, se conoce el nivel de pertenencia de cada nodo hoja (y por lo tanto estado HSMM) a cada fonema. Esta información se emplea para elegir la matriz correspondiente según el siguiente criterio:

- Si el HSMM corresponde a un único fonema y es una vocal, se aplica la matriz correspondiente a dicha vocal.
- Si el HSMM corresponde a varios fonemas de los cuales al menos uno es una vocal, se aplica la matriz de la vocal cuya pertenencia sea mayor.
- En cualquier otro caso, se aplica la matriz genérica.

3.4.3 Corrección de la Frecuencia Fundamental Media

Con las transformaciones basadas en DFW+AS explicadas en 3.2 únicamente se modifica la parte correspondiente a la envolvente espectral del modelo de la voz origen. No obstante, otro aspecto muy importante que ayuda a definir la identidad del locutor es la frecuencia fundamental o f_0. Para ello se propone realizar una transformación sobre la parte del modelo correspondiente a la $log\text{-}f_0$. Se emplean las mismas tramas vocálicas seleccionadas anteriormente para realizar una normalización de la media de la siguiente forma:

1. Se toma el valor de $log\text{-}f_0$ de la trama central de las vocales. Usando estos valores, se calcula la $log\text{-}f_0$ media del locutor objetivo.

2. De manera similar a los MFCC, se calcula una $log\text{-}f_0$ media del locutor fuente tomando la parte estática del vector de medias del estado central de las vocales.

3. Finalmente, la adaptación se lleva a cabo con una transformación similar a la expresada en la ecuación 3.5, donde $A = 1$ y b es igual a la diferencia entre las $log\text{-}f_0$ medias de la fuente y del objetivo.

3.5 Experimentos

A continuación, se exponen los experimentos llevados a cabo para comprobar la eficacia de este método de adaptación. Como voz objetivo se han empleado grabaciones realizadas por locutores no profesionales, en condiciones variables, ya que cada locutor ha realizado las grabaciones de forma individual y desde su propio equipamiento en el hogar. Así las condiciones de grabación no han sido controladas y la calidad que presentan las grabaciones no es tan alta como podría obtenerse con grabaciones profesionales. Como se explica más adelante, esto afecta negativamente al resultado de la adaptación. En el primer experimento, se utilizará todo el material de entrenamiento disponible y el resultado de la adaptación con la técnica FW+AS se compara con las técnicas vistas en la sección 2.3. Después, el experimento descrito en la sección 3.5.2 evalúa el comportamiento de esta estrategia ante la escasez de datos para entrenar. Finalmente, el experimento descrito en 3.5.3 evalúa el resultado en un caso de uso real, con una voz disártrica.

3.5.1 Desempeño General del Método de Adaptación

Las grabaciones empleadas en este experimento corresponden a seis locutores (3 hombres y 3 mujeres) del banco de voces ZureTTS - ahoMyTTS [3]. Cada locutor grabó 100 frases fonéticamente balanceadas. Las grabaciones se han normalizado para aumentar su rango dinámico y se han pasado por un filtro de Wiener para reducir el ruido. Se han extraído las vocales tal y como se indica en el apartado 3.4.1. La cantidad de vocales (es decir tramas) usadas por cada locutor se muestra en la tabla 3.1.

Como voz fuente se emplea una voz HTS, entrenada con 1962 frases en castellano de una locutora profesional. Todas las muestras de audio, tanto de la fuente como de los locutores objetivos, están muestreadas a 16kHz y se parametrizan empleando Ahocoder [31]. Para representar el espectro de la señal se han extraído los coeficientes MFCC de orden 39, y para representar la excitación la log-f_0y la MVF en ventanas de 25ms con un desplazamiento de 5ms.

Empleando la información de las vocales se han adaptado tanto el espectro de la señal como la log-f_0 del modelo de la voz fuente. Para evaluar el método de adaptación propuesto, se toma como referencia o *baseline* una adaptación de

Locutor	/a/	/e/	/i/	/o/	/u/	TOTAL
M1	256	154	111	153	66	740
M2	451	239	182	249	63	1184
M3	281	218	141	150	64	947
F1	317	201	169	186	74	1530
F2	453	364	244	334	135	1123
F3	368	249	196	226	84	854

Tabla 3.1: Número de vocales por locutor usadas en al adaptación.

locutor HTS del estado del arte basada en CSMAPLR+MAP descrita en la sección 2.3.1.2. Es importante recordar que este método de adaptación emplea toda la información fonética disponible, y no solo las vocales.

La comparación se ha llevado a cabo mediante una evaluación *Mean Opinion Score* (MOS). Dado que el método de referencia adapta también la MVF y la duración, además de la envolvente espectral (MFCC) y la excitación $log\text{-}f_0$, en la evaluación se ha querido restringir la comparativa únicamente a la adaptación de la envolvente espectral. Por ello, en los dos métodos de adaptación comparados, los valores de la MVF y de duración se han mantenido los del modelo de voz fuente original, sin realizar adaptación alguna. Para la adaptación de la $log\text{-}f_0$ se han utilizado únicamente las vocales, siguiendo el método 3.4.3.

Usando estos parámetros extraídos, se han sintetizado 10 frases completamente nuevas por cada locutor y método de adaptación, es decir 120 frases en total. Para cada evaluador, se han seleccionado aleatoriamente dos frases por locutor objetivo y método de adaptación, por lo que cada evaluador ha puntuado 24 frases. Por cada frase a puntuar, también se ha proporcionado una grabación de la voz original objetivo como referencia. Se ha pedido puntuar dos aspectos por cada frase, la calidad de la voz adaptada (sin tener en cuenta la calidad de la referencia original) y la similitud con la voz objetivo, ambas en una escala de 1 a 5. Las puntuaciones de ambos métodos para las seis voces se muestran en las figuras 3.2 y 3.3.

En estas figuras se observa que los resultados varían significativamente entre locutores. Estas variaciones se deben a la diferencia de calidad de las grabaciones originales. Para los locutores M1, M2 y F2, la calidad de sus grabaciones es

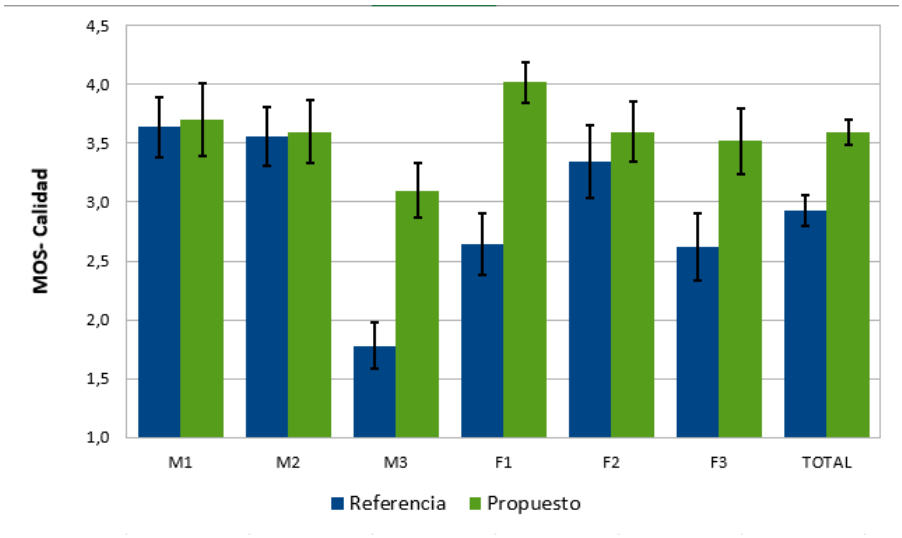


Figura 3.2: Puntuación MOS para la calidad e intervalo de confianza del 95 %.

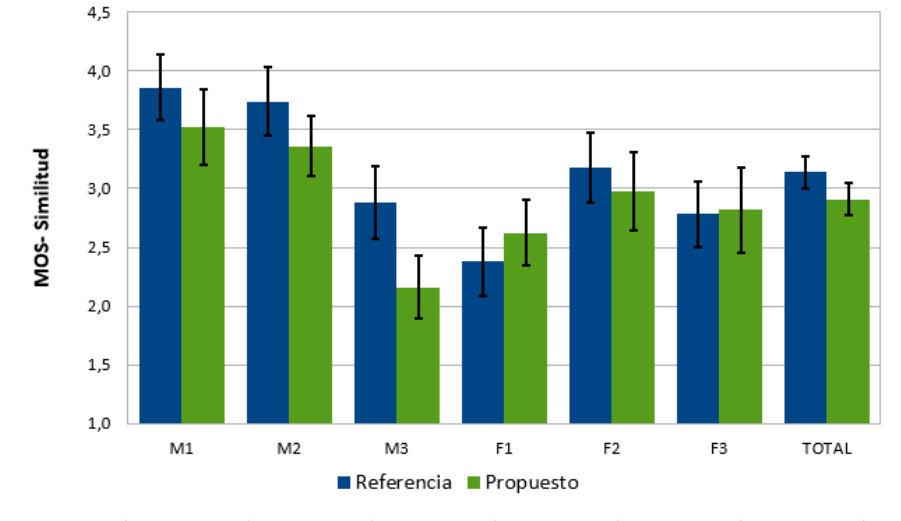


Figura 3.3: Puntuación MOS para la similitud e intervalo de confianza del 95 %.

relativamente alta y ambos métodos de adaptación obtienen buenos resultados en términos de calidad. Para M3, F1 y F3 las grabaciones presentan ruido ambiente apreciable y en el caso de M3 incluso reverberación. Este parece ser el motivo de la baja puntuación de los tres locutores en el método de referencia. Dado que el método propuesto es robusto frente a estos fenómenos porque está basado en transformaciones FW+AS, la calidad en estos casos se mantiene alta. Esto es una ventaja remarcable cuanto no se tiene control sobre el proceso de grabación, como sucede en este caso.

Respecto a la similitud, el método *baseline* se comporta ligeramente mejor. Esto es esperable dado que:

- Las transformaciones lineales CSMAPLR son más flexibles y ofrecen más grados de libertad que las basadas en FW+AS.
- El método de referencia emplea toda la información fonética de las frases de adaptación, y no solo las vocales.

No obstante, el método propuesto obtiene puntuaciones similares usando únicamente fragmentos vocálicos, lo cual en principio lo hace apropiado para algunos tipos de voces patológicas.

3.5.2 Reducción del Número de Frases de Entrenamiento

En el experimento anterior se ha mostrado que el método propuesto obtiene buenas puntuaciones en términos de calidad y similitud respecto al locutor objetivo. Sin embargo, para obtener los datos de adaptación que se han empleado es necesario que cada locutor grabe un total de 100 frases de un corpus fonéticamente balanceado. Este proceso puede ser largo y tedioso. Además, si se pretende emplear para personas con discapacidad oral, puede resultar incluso más incómodo para ellas. Por este motivo, también se ha propuesto identificar el umbral mínimo de datos de adaptación necesarios para reducir al máximo el número de grabaciones requeridas.

Dada la mala calidad de las grabaciones realizadas por los locutores M3 y F3, para el siguiente experimento se han sustituido por otros locutores del mismo género, denominados M3'y F3'. Tras extraer sus datos de entrenamiento, el número

Locutor	/a/	/e/	/i/	/o/	/u/	TOTAL
M3'	261	185	160	163	72	842
F3'	283	139	206	147	68	843

Tabla 3.2: Número de vocales por locutor nuevo usadas en al adaptación.

de vocales obtenidas para estos nuevos locutores se muestra en la tabla 3.2. Estas grabaciones se han preprocesado del mismo modo que las de los locutores sustituidos.

Para evitar tests perceptuales excesivamente largos, se ha limitado a tres el número de adaptaciones a comparar de acuerdo a test informales llevados a cabo en el laboratorio por expertos en tecnologías del habla con una cantidad variable de datos de entrenamiento:

1. Empleando todas las vocales disponibles.
2. Empleando el 10 % de las muestras disponibles de cada vocal.
3. Disponer de únicamente una muestra de cada vocal, como caso extremo.

Las muestras escogidas cuando se limita la cantidad de datos de entrenamiento se han seleccionado aleatoriamente. La estrategia usada es la misma que en el experimento anterior. Como voz fuente se toma la misma que en el experimento anterior y se emplea como referencia la adaptación del estado del arte basada en CSMAPLR+MAP. Para la parametrización se vuelve a emplear Ahocoder que extrae los coeficientes MFCC de orden 39, $log - f_o$ y MVF. De nuevo, se restringe la comparación a la adaptación del los MFCC usando la misma estrategia descrita anteriormente. De esta manera se dispone de cuatro adaptaciones para realizar una evaluación MOS: tres empleando el método propuesto y una adicional empleando el método del estado del arte.

Para evaluar las diferentes adaptaciones, se han sintetizado 10 frases nuevas por cada locutor y método de adaptación, siendo en total 240 frases. En la evaluación 11 evaluadores han tomado parte. Por cada uno de ellos se han seleccionado aleatoriamente dos frases de cada locutor y adaptación, de modo que cada evaluador ha puntuado 48 frases. Por cada frase sintética también se ha

proporcionado una señal del locutor original como referencia. A los evaluadores se les ha pedido que puntúen en una escala de 1 a 5 dos aspectos diferentes de cada frase: la calidad de la voz adaptada, sin tener en cuenta la calidad de la grabación original, y la similitud con la voz objetivo. Las puntuaciones medias así como el intervalo de confianza del 95 % se muestran en las figuras 3.4 y 3.5.

Como se aprecia en los resultados, se confirman las conclusiones del experimento anterior. El método propuesto se comporta mejor que el estado del arte en términos de calidad, mientras que no es tan bueno en términos de similitud. A este respecto, la diferencia en términos de calidad con el estado del arte se ha disminuido respecto al experimento anterior. Esto puede ser debido a que al sustituir los locutores M3 y F3 del apartado anterior, el método de referencia ha salido más beneficiado.

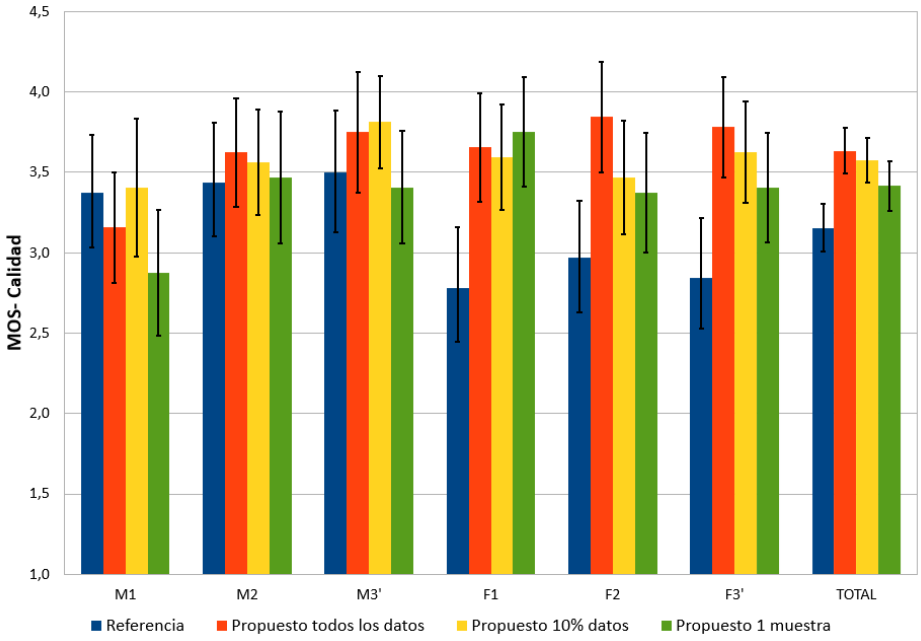


Figura 3.4: Reducción datos entrenamiento: puntuación MOS para la calidad e intervalo de confianza del 95 %.

Respecto a cómo se comporta el método propuesto cuando se reducen los datos de entrenamiento, se observa que la degradación es pequeña cuando los datos se reducen en un factor 10. Sin embargo, las diferencias son significativas en el caso

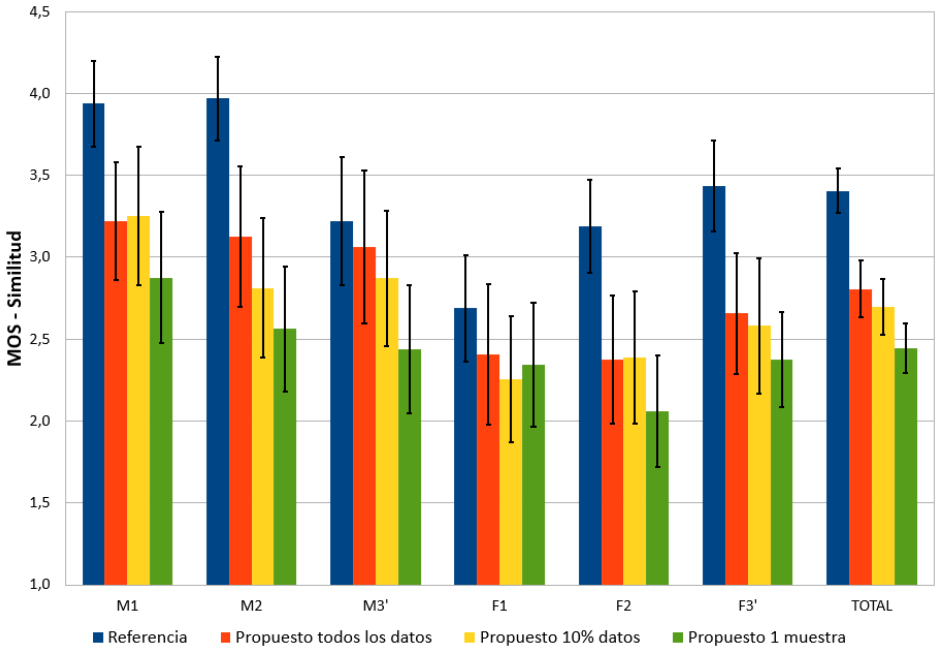


Figura 3.5: Reducción datos entrenamiento: puntuación MOS para la similitud e intervalo de confianza del 95 %.

extremo de una única muestra por vocal seleccionada aleatoriamente. Existen dos posibles interpretaciones a este fenómeno. Primero, puede ser un indicador de que una única muestra no es suficiente para entrenar el sistema. Otra posible explicación es que, al haber sido seleccionadas aleatoriamente se haya tenido la mala fortuna de escoger vocales con una fuerte co-articulación u otras características que degraden la calidad de la adaptación.

3.5.3 Adaptación con Voces Patológicas

Una aplicación práctica para este método de adaptación consiste en usarlo con personas con alguna discapacidad oral. Dado que las vocales son los sonidos más fácilmente pronunciables, es esperable que en el caso de no disponer de una voz completamente sana éstos sean los que presenten una mayor calidad. Para probar la idoneidad del método, se ha empleado una grabación de una persona que ha sufrido una operación y que presenta disartria. En dicha grabación la persona pronuncia las cinco vocales en orden, de manera sostenida, con descansos entre ellas para poder respirar. Aunque el tipo de grabación no es del mismo estilo que con los experimentos anteriores (frases con vocales co-articuladas en lugar de vocales sostenidas), se espera que el método propuesto funcione también con este tipo de voz dado que el método propuesto:

- Emplea la parte central del modelo HSMM, la parte más estable y con menos co-articulación.
- Ha demostrado ser robusto ante la escasez de datos.

El primer problema consiste en disponer de los pares de vectores de entrenamiento fuente-objetivo. En los experimentos anteriores se dispone del texto que los locutores han leído al realizar las grabaciones. Por lo tanto, se pueden generar las etiquetas lingüísticas correspondientes y extraer del modelo de la voz fuente los vectores MFCC correspondientes. Como ahora no se dispone de dicha información, no es posible conseguir un emparejamiento directo vectores fuente-objetivo.

Para conseguir las parejas de vectores para el entrenamiento se han seguido dos estrategias diferentes:

1. De la estadística calculada en la sección 3.4.2 se puede saber qué estados HSMM del modelo corresponden a una vocal o a otro fonema. Empleando esta información se seleccionan todos los HSMM que pertenecen a cada vocal y se emparejan con una trama de la grabación. Dado que las vocales pronunciadas son lo suficientemente largas, hay tramas suficientes para poder emparejar cada estado, seleccionado con una trama diferente de las vocales pronunciadas.

2. Se generan las etiquetas lingüísticas de la frase pronunciada fuera "a e i o u". Aunque estas etiquetas carecen de bastante información normalmente presente en ellas dada la peculiaridad de las mismas, es suficiente para poder extraer del modelo de la voz origen los vectores MFCC correspondientes y emparejarlos con la trama central de cada vocal pronunciada.

En ambos casos, tras aplicar la adaptación se ha conseguido una voz de calidad como en los casos anteriores. Sin embargo, tras algunas pruebas informales en el laboratorio, se ha concluido que la similitud de la voz obtenida con la de la persona antes de la operación es muy baja. Por este motivo, el objetivo de proveer de una voz sintética lo más parecida posible a como sonaría la voz sana no es factible en este caso. Dado que no se puede proveer de una voz similar, de cara al usuario final es lo mismo proveer una voz genérica o una voz adaptada usando datos de otra persona o el resultado de emplear esta adaptación, dado que el resultado final no se parece a como sonaba su voz antes.

Por ese motivo, se ha descartado realizar más experimentos en esta línea de investigación por el momento.

3.6 Aportaciones

En el presente capítulo se ha presentado un nuevo método de adaptación basado en FW y AS que emplea únicamente fragmentos vocálicos para adaptar el modelo ya entrenado de una voz fuente a una voz objetivo. Se ha demostrado que el método es útil cuando el material de adaptación presenta una calidad media-baja. Se ha comparado frente a otro método de adaptación del estado del arte en diversas evaluaciones MOS. En términos de calidad de la señal adaptada, el método propuesto supera al del estado del arte cuando las grabaciones originales tienen peor calidad inicial. En cambio, en términos de similitud, el método del estado del arte supera al método propuesto. Esto puede ser debido a las restricciones que imponen el FW que limitan los grados de libertad de las transformaciones. Además, el método del estado del arte emplea toda la información fonética y no solo las vocales. También se ha demostrado que el método propuesto es robusto en cuanto a escasez de datos de adaptación. Se ha probado reduciendo en un factor 10 el número de vocales disponibles para adaptar obteniendo puntuaciones similares. No obstante, en el caso extremo de disponer de una única muestra de cada vocal el método empieza a no funcionar tan bien.

Por último, se han realizado estudios preliminares para intentar utilizar este método para adaptar la voz de una persona con disartria. Sin embargo, tras unas pruebas iniciales se ha concluido que dicho método no es apropiado y que la similitud obtenida de la adaptación respecto al locutor objetivo es baja.

Durante el transcurso de esta investigación se han publicado los siguientes trabajos científicos:

- D. Erro, **A. Alonso**, L. Serrano, E. Navas, I. Hernaez, "Towards Physically Interpretable Parametric Voice Conversion Functions", Lecture Notes in Artificial Intelligence LNCS/LNAI (ISSN: 0302-9743), vol. 7911, pp. 75-82, 2013.

- **A. Alonso**, D. Erro, E. Navas, I. Hernaez, "Speaker Adaptation using only Vocalic Segments via Frequency Warping", Proc. Interspeech, pp. 2764-2768, Dresden, September 2015.

- D. Erro, **A. Alonso**, L. Serrano, E. Navas, I. Hernaez, Ïnterpretable Parametric Voice Conversion Functions based on Gaussian Mixture Models and Constrained Transformations", Computer Speech and Language (ISSN: 0885-2308), vol. 30, pp. 3-15, 2015.

- **A. Alonso**, D. Erro, E. Navas, I. Hernaez, "Study of the effect of reducing training data in speech synthesis adaptation based on Frequency Warping", Lecture Notes in Artificial Intelligence LNCS/LNAI (ISSN: 0302-9743), vol. 10077, pp. 3-13, 2016.

CAPÍTULO

4

Evaluación objetiva de voces sintéticas personalizadas

En el presente capítulo se explica el diseño e implementación del banco de voces del laboratorio Aholab. Este banco permite obtener una voz personalizada haciendo uso de técnicas de adaptación de locutor con unas pocas muestras de voz del locutor objetivo. Aunque este sistema se puede emplear para obtener una voz sintética para uso personal, también está habilitado su uso para realizar una donación, de forma que la voz sintética resultante puede ser utilizada por otras personas. Es decir que las personas donantes graban y generan su propia voz personalizada de manera altruista para que posteriormente la pueda emplear alguien con alguna discapacidad oral.

El componente social de este banco de voces ha hecho que sea un gran éxito, disponiendo actualmente de miles de grabaciones. Esto ha llevado a la conveniencia de disponer de un mecanismo automático de medida de la calidad de las voces personalizadas, con el fin de poder seleccionar las mejores de entre todas las disponibles. Por ese motivo, se han estudiado un conjunto de medidas de estimación de la inteligibilidad y la naturalidad de las voces sintéticas obtenidas, que pueden obtenerse de forma automática.

Después de describir en la sección 4.1 los bancos de voces en general y el banco de voces de Aholab de forma más específica, en la sección 4.2 se presentan las medidas que se han seleccionado, y a continuación, en la sección 4.3 los resultados obtenidos.

4.1 Bancos de Voces

Cuando una persona ha sufrido una pérdida total o parcial de sus facultades del habla, las tecnologías del habla pueden ayudarle en sus comunicaciones diarias. Los sistemas TTS, que toman como entrada un texto arbitrario y generan como salida una señal de voz con el texto leído por una voz sintética, son una herramienta muy útil dado que permiten una interacción natural con el resto de las personas. La facilidad de uso junto con la gran calidad de la señal de voz que generan los sistemas TTS actuales favorece la aceptación por parte de los usuarios. Además, hoy en día es posible encontrarlos embebidos en gran cantidad de aparatos electrónicos como *smartphones*, *tablets* u ordenadores. Sin embargo, por lo general no es posible escoger con qué voz se desea generar el mensaje leído, o si existe

la posibilidad de elección, la lista de voces disponibles suele ser muy limitada. Esto hace que en muchos casos la voz con la que se genera la señal de salida presente unas características de género, edad o acento con las que el usuario puede no sentirse identificado. Por ejemplo, un niño se puede ver forzado a emplear la voz de un adulto, o una mujer la voz de un hombre, y viceversa. Además, aunque la voz del sistema TTS tenga unas características que coincidan con las del usuario, dado que estas voces por defecto suelen ser genéricas e impersonales, puede no ser suficiente para que el usuario se identifique con ella. La personalización de voces, para ofrecer al usuario una voz con la que se sienta identificado es un factor muy importante a la hora de mejorar aún más la aceptación de estos sistemas por parte de los usuarios que los emplean, en sus comunicaciones diarias para interactuar con otras personas.

Cuando una persona sabe que va a perder o a sufrir una disminución de sus facultades del habla, por ejemplo, por una enfermedad degenerativa o una operación programada, puede recurrir a los bancos de voces (o *voice banking*, en inglés) para hacer un *backup* digital de su voz [2, 12, 70, 100, 110, 121]. Estos sistemas pueden entenderse como grandes repositorios de voz real, donde el usuario realiza grabaciones cuando aún dispone de su voz sana, para emplearlas en el futuro. Los mensajes estándar prefijados grabados en diversas situaciones se pueden reproducir directamente mediante un dispositivo digital. En estos casos la utilidad queda limitada a que se haya grabado un mensaje apropiado para la situación en la que se encuentre el usuario. Sin embargo, la calidad obtenida es máxima, dado que se usa una grabación con la voz real de la persona.

Muchos bancos de voces ofrecen la posibilidad de entrenar una voz sintética personalizada. Dependiendo de la tecnología que empleen, pueden ser necesarias grabaciones de duración de entre unos pocos minutos hasta varias horas. En estos casos, se obtiene una voz sintética a emplear en un dispositivo TTS integrable con el banco de voz. En algunos casos, los bancos de voces pueden ofrecer su propio sistema TTS, o también una voz compatible con sistemas de terceros. Cuando el usuario es dotado de un TTS con su propia voz personalizada, puede usarla en una variedad más amplia de situaciones: dado que puede obtener una salida de audio para cualquier texto de entrada que introduzca, no queda limitado a reproducir las frases que haya grabado con anterioridad.

Si la capacidad del habla del usuario ya se ve afectada de manera perceptible y no es posible generar una voz personalizada con la calidad deseada, algunos bancos de voces ofrecen alternativas basadas en cirugía de modelos para tratar de mejorar la voz [23, 82, 130].

Cuando el usuario ya ha perdido completamente la capacidad del habla, o está tan gravemente deteriorada que no permite su reconstrucción, existe otra opción consistente en usar grabaciones de otras personas para generar una voz personalizada. Estas personas, comúnmente denominadas donantes, suelen ser amigos o familiares a los que se les pide que realicen las grabaciones necesarias. Por ejemplo, en [100] se recomienda suministrar grabaciones de dos donantes diferentes y la voz obtenida presenta una mezcla de características de ambas. También existen bancos de voces cuyo enfoque se basa en donantes anónimos. Por ejemplo, en [121] el propio banco de voces selecciona de un catálogo de voces una voz lo más parecida posible a la del usuario objetivo usando varias muestras de su voz suministradas previamente.

Usando una voz personalizada suministrada por el banco de voces, se puede conseguir que el sistema TTS que utilice el usuario disponga de una voz con la que se sienta identificado, contribuyendo a una interacción social más satisfactoria con otras personas.

4.1.1 Banco de Voces de Aholab

El laboratorio Aholab ha desarrollado su propia solución de *voice banking* que ha ido evolucionando con el tiempo. La primera versión se denominó ZureTTS. *Zure* en euskera es el determinante posesivo "tu", que seguido de las siglas TTS puede traducirse como "Tu sistema texto a voz". Esta versión fue fruto de la colaboración de diversas universidades y centros de investigación (Universidad del País Vasco, Universidad de Vigo, Universidad Politécnica de Cataluña, Universidad Tecnológica de Nanyang, Universidad Técnica de Košice, Universidad de Ciencia y Tecnología de China, Vicomtech-IK4) en el marco del eNTERFACE'14 [28]. Entonces se desarrolló un banco de voces para siete idiomas: castellano, euskera, gallego, catalán, inglés, eslovaco y chino mandarín. Los detalles de implementación de esta versión pueden consultarse en el anexo A.

La segunda versión, disponible actualmente, se denomina *ahoMyTTS* y da soporte y mantenimiento a castellano y euskera. En ella se han optimizado esfuerzos en la renovación y el rediseño del portal web del banco de voces existente, y se ha creado una aplicación móvil basada en Android con dichos idiomas.

El banco de voces dispone de un portal web que es el punto de entrada del usuario. Una vez registrado, el usuario debe elegir en qué idioma de los disponibles desea obtener una voz personalizada, y grabar las 100 frases fonéticamente balanceadas. Este corpus se ha generado de manera automática empleando gran cantidad de texto disponible de manera libre asegurándose de que tenga la mayor cobertura fonética posible. El proceso de personalización es independiente para cada idioma, de manera que, si se desea obtener una voz para más de un idioma, deben realizarse las 100 grabaciones correspondientes para cada idioma. Las grabaciones se realizan a través del propio portal web, el usuario graba en su propio hogar, con su propio equipo, en un proceso no supervisado durante tantas sesiones como sea necesario. Una vez finalizado el proceso de grabación, el usuario puede revisar todas las grabaciones realizadas y repetir aquellas que no le gusten o que considere que no han quedado bien. Cuando esté satisfecho con las grabaciones, puede iniciar el proceso de personalización de la voz que se realiza en segundo plano, tras el cual recibe una notificación por e-mail cuando su voz personalizada ya está disponible.

En la figura 4.1 se detalla un diagrama del sistema del banco de voces de Aholab. La obtención de voces personalizadas está basada en adaptación HTS. Para cada idioma del banco de voces se dispone de una voz promedio generada con audio procedente de varios locutores profesionales. La voz promedio entrenada correspondiente a cada idioma se adapta mediante CSMAPLR+MAP, tal y como se explica en la sección 2.3.1. Como material de adaptación se emplean las grabaciones realizadas por los usuarios. Una vez el proceso de adaptación ha finalizado y la voz personalizada está lista para su uso, se puede emplear mediante el propio portal web o con la aplicación Android diseñada a tal efecto.

El banco de voces también admite donantes anónimos altruistas. En estos casos, cuando una persona quiere donar su voz para que otra pueda emplearla, el proceso de grabación es el mismo. Una vez la voz del donante está lista, puede ser seleccionada y utilizada por un usuario que requiera una voz personalizada pero

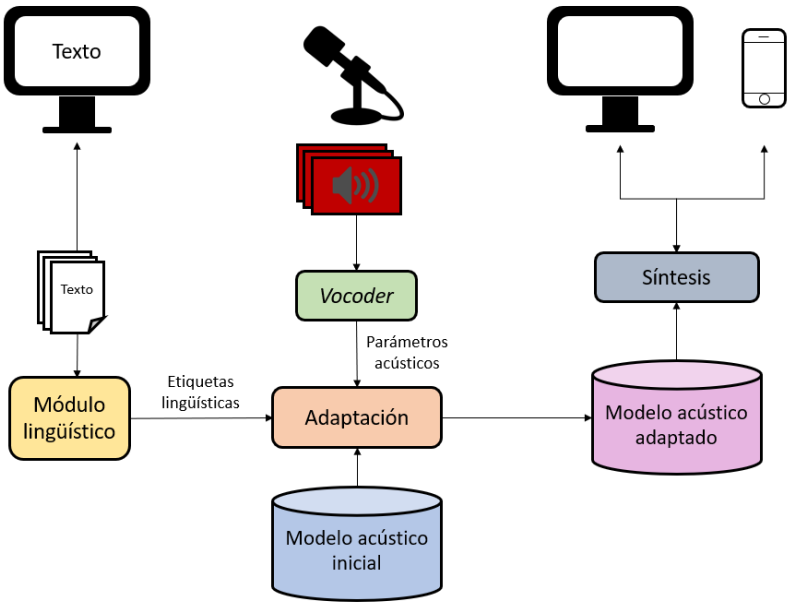


Figura 4.1: Diagrama del sistema del banco de voces de Aholab.

no pueda realizar las grabaciones por sí mismo. Actualmente, se dispone de más de mil voces de donantes. Dado que el proceso de grabación es no supervisado, con ruidos de fondo u otros efectos no deseados, el laboratorio no tiene control sobre la calidad final de las grabaciones. Aunque el portal web dispone de instrucciones con notas y consejos para que los donantes realicen las grabaciones en las mejores condiciones posibles, no se pude garantizar que sea suficiente. Como resultado, las voces disponibles de los donantes presentan una gran variabilidad en la calidad. Para ofrecer a los usuarios que necesiten la voz de un donante aquellas con mayor calidad, es necesario evaluarlas. Una opción consiste en realizar una evaluación subjetiva por cada nuevo donante, sintetizando algunas frases de ejemplo cuando el donante finalice el proceso de adaptación y evaluándolas por un grupo de personas expertas para determinar si la calidad es aceptable o no. Sin embargo, dada la cantidad de donantes existente, este proceso resulta demasiado caro y costoso, por lo que no es viable. Se ha estudiado el uso de medidas objetivas que ayuden a evaluar las voces sin necesidad de expertos humanos, pudiendo obtener de manera automática una puntuación que ayude a discriminar entre voces aptas para ser

usadas por personas con algún impedimento oral y las que no lo son.

4.2 Evaluación de Voces Sintéticas

Un aspecto muy importante a evaluar en las voces sintéticas es la calidad. Tradicionalmente se ha hecho mediante evaluaciones subjetivas MOS, en las que se pide la colaboración de varias personas para que puntúen diversos aspectos de la señal, como por ejemplo:

- Naturalidad: cómo de humana o poco robótica suena.
- Inteligibilidad: cómo de claro y fácil de entender es el mensaje.

La calidad se puntúa con una escala de valores discretos de 1 a 5, siendo 1 la puntuación más baja y 5 la más alta.

Para que la evaluación sea estadísticamente significativa es necesario que haya un mínimo de participación. Generalmente se asume que es suficiente contar con 20 evaluadores voluntarios, o recompensados económicamente por su participación. Existen diferentes enfoques respecto a cómo facilitar el acceso a la evaluación. Una opción consiste en crear una página web a la que un evaluador accede desde cualquier dispositivo para escuchar y puntuar directamente las voces. Esta estrategia tiene la ventaja de que la evaluación es fácil de distribuir y potencialmente pueden participar más evaluadores. Sin embargo, no se tiene control sobre las condiciones en las que el evaluador la realiza: entornos ruidosos, equipos (auriculares o altavoces) de mala calidad afectan en la puntuación de la calidad las voces evaluadas. Otra opción consiste en habilitar un área física donde realizar la evaluación. En este caso, se tiene control sobre el entorno donde se realiza la evaluación, garantizando que todos los evaluadores la realizan en las mismas condiciones. Sin embargo, es un enfoque más costoso puesto que hay que preparar y mantener el espacio donde se realiza, limitando el número de evaluadores y exigiendo su desplazamiento hasta el lugar habilitado.

Por cada voz, emoción, método de síntesis u otra característica de la síntesis que se desee valorar, cada evaluador debe puntuar varias señales diferentes. Si existen muchas, las evaluaciones pueden llegar a ser muy extensas, dado que deben de escucharse y puntuarse muchas señales de voz diferentes. Para evitar sesgos, las muestras de señal de voz se presentan en orden aleatorizado, de manera que en cada evaluación no se sabe a priori qué señal se va a puntuar en cada momento. En

ocasiones, además de la señal a puntuar, se suministra una señal de voz adicional como referencia.

Las evaluaciones subjetivas en general, aun siendo las más precisas dado que los resultados se obtienen preguntando a personas, presentan varios inconvenientes como:

- La evaluación tiene que estar disponible el tiempo suficiente para que los evaluadores la realicen, lo cual introduce cierto retraso en la obtención de resultados.

- Hay que considerar los costes monetarios. Si la evaluación se ha diseñado utilizando una página web, cuanto más tiempo esté disponible los costes de *hosting* aumentan. Si se ha dispuesto de un espacio específico para la evaluación también tiene un coste asociado. Además, en los gastos hay que incluir la recompensa económica por la participación de los evaluadores, caso de que así se acuerde.

- En el caso de evaluar una voz en un idioma minoritario, es difícil encontrar evaluadores suficientes para obtener una puntuación representativa. Lo ideal es que la evaluación la realicen personas cuya lengua materna sea la misma que la de la voz a evaluar, o que tengan un conocimiento alto del idioma.

Estas limitaciones hacen que diseñar una evaluación sea un aspecto crítico a tener en cuenta. Un fallo en el diseño puede provocar que sea necesario repetirla, con las complicaciones y sobrecostes que conlleva. Por todo ello, disponer de un método objetivo que permita puntuar las señales de voz sintética sin tener que recurrir a evaluaciones subjetivas resulta de gran utilidad.

Además de la síntesis de voz, otros ámbitos relacionados con las tecnologías del habla también emplean evaluaciones MOS, tales como: transmisión de señal, sistemas de codificación-decodificación, *speech enhancement*, entre otros. Actualmente se realizan grandes esfuerzos para desarrollar medidas objetivas que sustituyan las evaluaciones MOS tradicionales. Originalmente, las medidas objetivas surgieron por la necesidad de evaluar la calidad de las señales recibidas a través de un canal telefónico [1, 18, 88, 89]. En estos casos, se parte de una señal de referencia clara y limpia, que al ser transmitida sufre una distorsión

que afecta a la calidad de la señal recibida. De una medida objetiva se espera que se generalice para audio no visto durante su entrenamiento, siempre que mantenga el tipo de distorsión esperada. Dada la conveniencia de estas medidas, muchas veces se aplican en dominios para los que inicialmente no habían sido diseñadas [117], dónde la señal sufre distorsiones no previstas durante el diseño. Un ejemplo es *Perceptual Evaluation of Speech Quality* (PESQ) [88]. Esta medida fue diseñada hace más de 20 años para determinar la calidad de la señal sobre sistemas de telefonía, y hoy en día se emplea en un amplio tipo de sistemas con distorsiones diferentes (separación de voz basada en DNN, extracción de voz cantada, de-reverberación). Dado que se dispone de la señal de referencia, puede emplearse para compararla con la señal recibida. Cuando la medida hace uso de una referencia se la denomina intrusiva. Si, por el contrario, para calcularla medida no es necesario disponer de una referencia con la que comparar la señal a evaluar, la medida se denomina no intrusiva.

En esta tesis se ha estudiado el uso de cuatro medidas objetivas diferentes y cómo pueden emplearse conjuntamente para estimar una puntuación MOS en señales de voz sintéticas personalizadas. Tres de ellas pertenecen a la categoría de medidas intrusivas y una de ellas a la no intrusiva. Dado que la característica principal de los sintetizadores TTS de propósito general es poder reproducir un texto de entrada arbitrario, no se dispone de una señal de referencia para todas las posibles salidas del sistema. Por ello las medidas no intrusivas a priori pueden suponerse mejores para evaluar un sistema TTS. No obstante, existen técnicas para obtener un corpus de evaluación con señales de referencia que permitan utilizar medidas intrusivas [47].

4.2.1 STOI y ESTOI

Una medida de inteligibilidad que es ampliamente utilizada es *Short Time Objective Intelligibility* (STOI) [104, 105]. Esta medida es intrusiva, por lo que requiere una referencia alineada temporalmente con la señal a evaluar. En ella, se calcula la representación *Time Frequency* (TF) de ambas señales mediante una *Discrete Fourier Transform* (DFT) de las tramas enventanadas. Usando un análisis de un tercio de octava, agrupa los *bins* frecuenciales de la DFT en 15 bandas y calcula la

norma de cada una, la cual se denomina TF-*unit*. Emplea una medida intermedia de inteligibilidad por cada TF-*unit*, que depende de N TF-*units* consecutivas de las señales de referencia y a evaluar. Normalmente, se escoge un valor de N que haga que la medida intermedia dependa de los últimos $400ms$, aproximadamente, de la información de la señal. Para calcular la medida de inteligibilidad final, se toma la media de todas las medidas intermedias entre diferentes tramas y bandas frecuenciales. Esta operación implica que, para la medida final, la contribución de cada banda frecuencial es independiente del resto de bandas.

STOI ha demostrado que predice de manera bastante precisa la inteligibilidad en diferentes situaciones, tales como: salida de señal de teléfono [52], voz ruidosa procesada por enmascaramiento ideal de tiempo-frecuencia y algoritmos de mejora de voz de un solo canal [105] y voz procesada por implantes cocleares [32]. También es robusta ante diferentes tipos de idiomas, tales como chino mandarín [124], danés [105] u holandés [48].

Una evolución de STOI es *Enhanced Short Time Objective Intelligibility* (ESTOI) [49] en la que, al contrario que en STOI, no se asume independencia entre bandas frecuenciales, permitiendo capturar mejor los efectos del ruido de modulación en el tiempo.

Las puntuaciones obtenidas tanto de STOI como de ESTOI varían en una escala de 0 a 1. El éxito en la estimación de estas medidas ha llevado a que sean propuestas en varias áreas, por ejemplo, para evaluar la inteligibilidad de voz que presenta disartria [47]. En dicho trabajo la señal patológica se consideraba como la señal ruidosa a evaluar mientras que la señal de referencia (es decir, como sonaría la voz del paciente si no sufriera la patología) no estaba disponible. Para solventar este problema, se propuso generar una señal de referencia por cada señal a evaluar empleando varios locutores sanos mediante *Dynamic Time Warping* (DTW). Además, también se utilizó DTW para alinear las señales a evaluar con la señales de referencia generadas.

4.2.2 SIIB

Otra medida intrusiva que también trata de estimar la inteligibilidad es *Speech Intelligibility In Bits* (SIIB) [120]. En ella, la cantidad de información compartida

entre un locutor y un oyente es estimada en bits por segundo (b/s). Esta medida está motivada por la teoría de la información y sugiere que el proceso del habla se puede entender como la transmisión de un mensaje desde un locutor a un oyente. El mensaje $\{M\}$ es una secuencia de frases o de fonemas que el locutor codifica en la señal de voz $\{X\}$, la cual se envía a través de un canal de comunicación que puede distorsionar y crear una señal de voz degradada $\{Y\}$. Por lo tanto, el proceso de comunicación se puede describir como una cadena de Markov:

$$\{M\} \rightarrow \{X\} \rightarrow \{Y\} \tag{4.1}$$

donde $\{M\} \rightarrow \{X\}$ es el canal de producción de voz y $\{X\} \rightarrow \{Y\}$ es el canal ambiental.

SIIB supone que la inteligibilidad es una función de la tasa de información mutua entre el mensaje $\{M\}$ y la señal degradada $\{Y\}$. Los autores defienden que esta medida funciona mejor en condiciones generales que otras medidas diseñadas con una motivación heurística o con una distorsión o *dataset* especifico. No obstante, advierten que para que la media sea fiable, la señal a evaluar debe tener una duración de al menos $20s$. SIIB estima la inteligibilidad en una escala abierta, donde una señal óptima obtiene una puntuación de entre $150b/s$ y $180b/s$.

4.2.3 NISQA

Un aspecto muy importante que evaluar en las señales sintéticas es la naturalidad. En [68] se propone un método basado en *Non-Intrusive Speech Quality Assessment* (NISQA) [67] para medir la naturalidad de las voces sintéticas, sin la necesidad de una señal de referencia. El modelo de predicción propuesto está basado en una arquitectura de red *Convolutional Neural Network Long Short-Term Memory* (CNN-LSTM), que ha sido entrenada empleando 16 bases de datos en 12 idiomas diferentes, por lo que es independiente del idioma y puede ser empleada para evaluar cualquier sistema TTS. El modelo está publicado en [75] y puede usarse directamente para estimar la naturalidad de la señal en una escala MOS con puntuaciones que varían entre 1 y 5. En adelante, este método es denominado NISQA por simplicidad.

4.3 Experimentación

Para corroborar que las medidas objetivas expuestas anteriormente son apropiadas al determinar la calidad de las voces sintéticas personalizadas del banco de voces de Aholab, se han llevado a cabo los experimentos descritos en las siguientes secciones.

4.3.1 Obtención de Corpus de Evaluación

A continuación, se explica el proceso seguido para obtener las señales a evaluar por las diferentes medidas objetivas, así como sus respectivas señales de referencia, en el caso de que se trate de una medida intrusiva.

Dado que el único material de referencia del que se dispone son las grabaciones realizadas por el locutor cuya voz se pretende obtener, éstas son las que se toman como señales de referencia. Para casos de que se trate de un entrenamiento HTS estándar (es decir, entrenando la voz desde cero sin recurrir a técnicas de adaptación), se puede llegar a disponer de miles de muestras de voz real. Mientras que, para la adaptación de una voz personalizada del banco de voces únicamente se dispone de las 100 grabaciones del corpus de adaptación.

El siguiente paso es conseguir la señal sintética a evaluar. Dado que se dispone del texto que el locutor ha leído para hacer las grabaciones, una posibilidad es sintetizar dicho texto directamente usando la voz sintética. Sin embargo, si se sintetiza directamente, las frases obtenidas tendrán la duración que el modelo entrenado considere oportuno. Por tanto, dado que ambas señales no están alineadas temporalmente no se pueden comparar directamente. Para que las señales de voz sintética y real estén perfectamente alineadas podría realizarse una segmentación fonética de las grabaciones de voz real y forzar las duraciones en la síntesis. Sin embargo, el modelo de duración de la voz obtenido durante el entrenamiento no se tendría en cuenta, por lo que no se llegaría a evaluar el resultado real del entrenamiento.

Para que el modelo de duración se tenga en cuenta, es necesario sintetizar primero y alinear posteriormente. Para ello, mediante *Montreal Forced Aligner* (MFA) se obtienen las posiciones reales de las pausas hechas por el locutor durante la grabación, que corresponden con la omisión de algún signo de pausa del texto

de entrenamiento, o por el contrario, con nuevas pausas introducidas de manera espontánea. Empleando la información de estas pausas se sintetizan los textos leídos por el locutor mediante los modelos de su voz sintética. En este caso, no se indica cuánto tienen que durar los fonemas durante la síntesis. Solo se indica dónde existe una pausa realizada, que quizás no tiene representación gráfica en el corpus escrito. Con las frases sintetizadas usando estas pausas, sigue siendo necesario asegurarse de que cada pareja real-sintética está alineada correctamente. Aunque pueden alinearse a nivel de frase, una opción más precisa consiste en el alineamiento basado en DTW a nivel de fonema, calculando la distancia cepstral entre las señales real y sintética, donde:

- Los coeficientes cepstrales de la señal real son calculados parametrizándola usando Ahocoder [31].
- Los coeficientes cepstrales de la señal sintética se obtienen directamente del modelo de la voz.

Una vez alineadas las parejas real-sintética, se puede proceder al cálculo de las medidas intrusivas. En el caso especial del SIIB, la medida obtenida no es fiable si las duraciones de las frases a evaluar son inferiores a $20s$. Por este motivo, previamente se concatenan varias frases hasta asegurar que la entrada tiene la duración mínima recomendada.

4.3.2 Evaluación de Voces HTS Estándar

Para comprobar cuán adecuado resulta el uso de las medidas objetivas estudiadas, se han considerado voces HTS de gran calidad y calculado su puntuación. Las grabaciones empleadas durante el entrenamiento fueron realizadas por tres locutores expertos, dos mujeres (F1 y F2) y un hombre (M1), en un entorno controlado con equipamiento profesional. Cada locutor realizó grabaciones en castellano (ES) y euskera (EU), según los corpus de entrenamiento detallados en la tabla 4.1. Dado que para cada voz se disponía de material suficiente, las voces fueron entrenadas empleando el procedimiento estándar descrito en la sección 2.2, sin necesidad de recurrir a adaptación.

Tabla 4.1: Corpus de entrenamiento de voces HTS estándar.

Voz	Locutor	Idioma	Género	Número de frases
F1-ES	F1	ES	F	3.994
F1-EU	F1	EU	F	3.797
F2-ES	F2	ES	F	3.712
F2-EU	F2	EU	F	3.831
M1-ES	M1	ES	M	3.995
M1-EU	M1	EU	M	3.799

En la figura 4.2 se detallan media y desviación estándar de las medidas objetivas de inteligibilidad (STOI, ESTOI, SIIB) y naturalidad (NISQA) para las seis voces HTS estándar. De las puntuaciones obtenidas cabe destacar: para STOI y ESTOI, la voz M1-ES obtiene las puntuaciones más altas; para NISQA, la voz F2-ES tiene una puntuación significativamente más alta que las demás voces, y M1-ES tiene la puntuación más baja; para SIIB, no hay diferencias significativas. Desde la perspectiva del idioma, la puntuación NISQA de la voz M1-ES resulta inferior a la de la voz M1-EU, mientras que ocurre lo contrario para las medidas STOI y ESTOI. En cualquier caso puede concluirse que voces de gran calidad tienen asociadas puntuaciones objetivas elevadas.

4.3.3 Evaluación de Voces HTS Personalizadas

Para evaluar la idoneidad de las medidas objetivas estudiadas cuando determinan la calidad de voces personalizadas, se han empleado 1.090 voces en castellano disponibles en el banco de voces de Aholab [3] y se ha estudiado como correlan las puntuaciones con la opinión de varios evaluadores.

Inicialmente se han calculado las medidas STOI, ESTOI y NISQA para todas las voces disponibles. A continuación, las voces se han agrupado automáticamente en *clusters* con *k-means*, en función de la puntuación obtenida. Dado que cada medida tiene una escala de valores diferente, se ha realizado una normalización *z-score* de cada una de ellas, restando la media y dividiendo entre la varianza, para que todas las medidas tengan la misma importancia en el *clustering*. Como

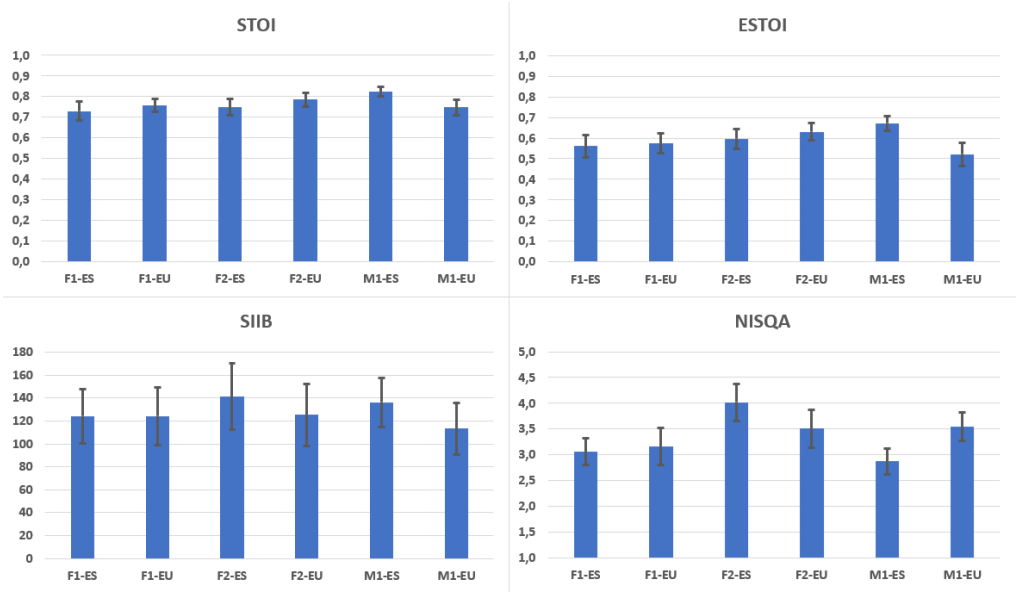


Figura 4.2: Media y desviación estándar de las medidas objetivas obtenidas para las voces HTS estándar.

la evaluación de la calidad de las voces suele realizarse mediante escala MOS de valores discretos de 1 a 5, se han empleado 5 *clusters* en los siguientes *clustering*:

A) Teniendo en cuenta solo las medidas de inteligibilidad (STOI, ESTOI).

B) Teniendo en cuenta las medidas de *A* y la medida de naturalidad (NISQA).

En la figura 4.3 se representa el resultado del *clustering A*. Se aprecia que ambas medidas están muy correladas ($\rho = 0{,}912$), pudiendo deberse a que STOI y ESTOI tienen una naturaleza similar. Además, se observa que las fronteras entre los cinco *clusters* están perfectamente delimitadas.

En las figuras 4.4a, 4.4b y 4.4c se representa el resultado del *clustering B*. En este caso, se observa que los límites entre *clusters* no están tan definidos, así como que la medida NISQA no está tan correlada con STOI y ESTOI.

Para comprobar si el resultado del *clustering* coincide con las preferencias de los evaluadores se ha realizado una evaluación MOS. Dado que realizar una evaluación con todas las voces no es viable, se ha seleccionado el donante centroide de cada *cluster* como voz representativa (figuras 4.3 y 4.4, respectivamente). Al contar con dos *clustering*, *A* y *B*, las voces a evaluar serían 10. Sin embargo,

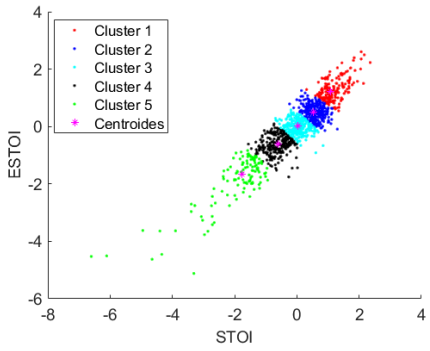
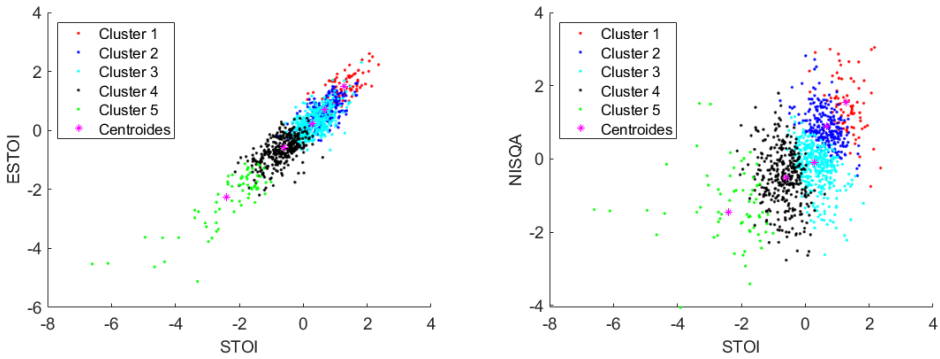
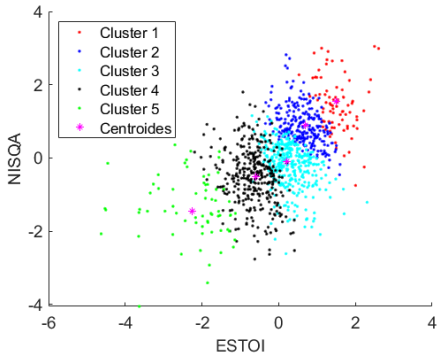


Figura 4.3: *Clustering A* de medidas objetivas STOI y ESTOI.



(a) Proyección del *clustering B* en el eje STOI-ESTOI.

(b) Proyección del *clustering B* en el eje STOI-NISQA.



(c) Proyección del *clustering B* en el eje ESTOI-NISQA.

Figura 4.4: *Clustering B* de medidas objetivas STOI, ESTOI y NISQA.

el centroide del *cluster* 2 es el mismo locutor para ambos *clustering*, por lo que finalmente se evalúan 9 voces. Las puntaciones objetivas para dichas voces son las recopiladas en la tabla 4.2.

Tabla 4.2: Puntuaciones de las medidas objetivas para las voces representativas.

Donante	STOI	ESTOI	NISQA
SPK1	0,6832	0,5045	-
SPK2	0,6689	0,4689	-
SPK3	0,6155	0,4220	-
SPK4	0,5747	0,3790	-
SPK5	0,5015	0,3093	-
SPK6	0,6895	0,5122	3,0036
SPK2	0,6689	0,4689	2,8210
SPK7	0,6343	0,4368	2,5261
SPK8	0,5691	0,3732	2,4926
SPK9	0,4601	0,2679	1,1778

Para la evaluación, se han sintetizado 10 frases completamente nuevas para cada una de las 9 voces evaluadas. En la evaluación han participado 15 evaluadores, y cada uno ha evaluado 5 frases de cada voz seleccionadas aleatoriamente, es decir, 45 frases en total. Las señales a evaluar se han facilitado vía interfaz web, donde se han recopilado puntuaciones de cada evaluador a la pregunta: "¿Usarías esta voz en un sistema TTS?", siendo 1: "No me gusta esta voz y no la usaría en un sintetizador" y 5: "Sí me agrada esta voz y la usaría en un sintetizador". Los resultados de la evaluación MOS son los mostrados en la figura 4.5, donde queda patente que mayores puntuaciones objetivas obtienen mejores resultados en la evaluación subjetiva. Además, las tres medidas objetivas consideradas por separado correlan con las puntuaciones subjetivas. La tabla 4.3 recopila los coeficientes de correlación entre las puntuaciones de las medidas objetivas y la puntuación MOS. Se observa que existe mayor correlación entre las medidas de inteligibilidad (STOI y ESTOI) con la puntuación MOS en comparación con la medida de naturalidad (NISQA).

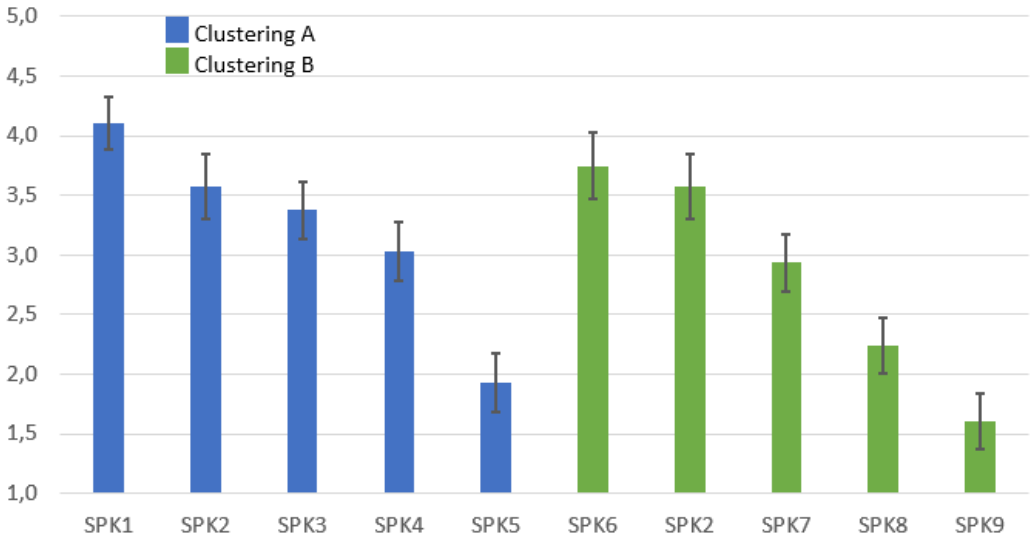


Figura 4.5: Resultados MOS con intervalo de confianza del 95 % para *clusterings A* y *B*.

Tabla 4.3: Coeficiente de correlación entre puntuaciones objetivas y MOS para cada una de las tres medidas.

STOI	ESTOI	NISQA
0,9484	0,9500	0,7858

4.3.4 Modelo de Regresión Para Predecir la Puntuación MOS

En el experimento anterior se ha comprobado que la puntuación objetiva dada por las medidas de inteligibilidad STOI y ESTOI presenta una gran correlación con la puntuación MOS. Por ello, para el siguiente experimento también se considera SIIB, que tiene un enfoque completamente diferente con respecto a STOI y ESTOI en el cálculo de las puntuaciones de inteligibilidad.

La medida SIIB se ha obtenido para todos los donantes del banco de voces para los que ya se disponía de las puntuaciones de STOI, ESTOI y NISQA. En la tabla 4.4 se detallan las cinco voces personalizadas para las que se obtiene la mayor puntuación STOI, ESTOI, SIIB y NISQA, respectivamente. De entre las veinte voces puntuadas, cinco de ellas se encuentran entre las de mayor puntuación en más de una medida objetiva, por lo que son solamente quince las

Tabla 4.4: Top 5 de voces personalizadas por cada medida objetiva.

Posición	STOI	ESTOI	SIIB	NISQA
1º	SPK01	SPK02	SPK03	SPK04
2º	SPK04	SPK04	SPK05	SPK06
3º	SPK07	SPK08	SPK09	SPK02
4º	SPK02	SPK10	SPK11	SPK12
5º	SPK10	SPK13	SPK14	SPK15

Tabla 4.5: Coeficiente de correlación entre medidas objetivas.

-	STOI	ESTOI	SIIB
ESTOI	0,912	-	-
SIIB	0,363	0,292	-
NISQA	0,449	0,476	0,258

voces diferentes con mayor puntuación asociada. Las voces personalizadas mejor puntuadas por la medida SIIB no han sido puntuadas como las cinco mejores por STOI, ESTOI y NISQA. Nótese que las voces personalizadas del presente experimento corresponden a donantes diferentes en relación con el experimento anterior (el donante SPK01 no es el mismo donante que SPK1).

Dado que algunas voces personalizadas han sido puntuadas en el top cinco por varias de las medidas objetivas estudiadas, se analiza cuán correladas están las medidas entre sí calculando el coeficiente de correlación ρ entre sus puntuaciones, cuyo resultado es el detallado en la tabla 4.5. Se aprecia que, aunque STOI y ESTOI están muy correladas como concluido en el experimento anterior, no sucede lo mismo para las demás medidas. Dado que NISQA mide la naturalidad y no la inteligibilidad, se puede esperar que correle poco con el resto de las medidas objetivas. El enfoque de SIIB es completamente diferente al de STOI y ESTOI, lo cual explica su baja correlación con éstas aunque todas midan la inteligibilidad, siendo las mejores voces para SIIB completamente diferentes a las mejores para que para STOI y ESTOI, como quedó patente en la tabla 4.4.

A continuación, se lleva a cabo una evaluación MOS considerando las quince

voces personalizadas de la tabla 4.4 (top cinco puntuadas por las medidas objetivas estudiadas). Para cada voz a evaluar se han seleccionado 10 frases sintetizadas del corpus empleado para calcular las medidas objetivas. En la evaluación han tomado parte 25 personas. Cada evaluador ha puntuado en escala MOS 5 frases sintetizadas seleccionadas aleatoriamente de entre las diez del corpus de evaluación, esto es, 75 frases sintetizadas evaluadas en total, Con puntuación discreta de 1 a 5, en función de cómo de apropiada encontraban cada voz para ser usada por personas con alguna discapacidad oral en un dispositivo de comunicación.

La figura 4.6 muestra las puntuaciones MOS obtenidas por cada voz, junto con el intervalo de confianza del 95 %. Las voces están ordenadas, de izquierda a derecha, de mayor a menor puntuación MOS. Para cada voz, se ha colocado una etiqueta indicando la medida (STOI, ESTOI, SIIB o NISQA) en la que ha sido seleccionada entre las cinco con mayor puntuación. Por ejemplo, SPK02 se encuentra en el top 5 de STOI, ESTOI y NISQA o SPK08 entre las cinco mejores puntuaciones de ESTOI.

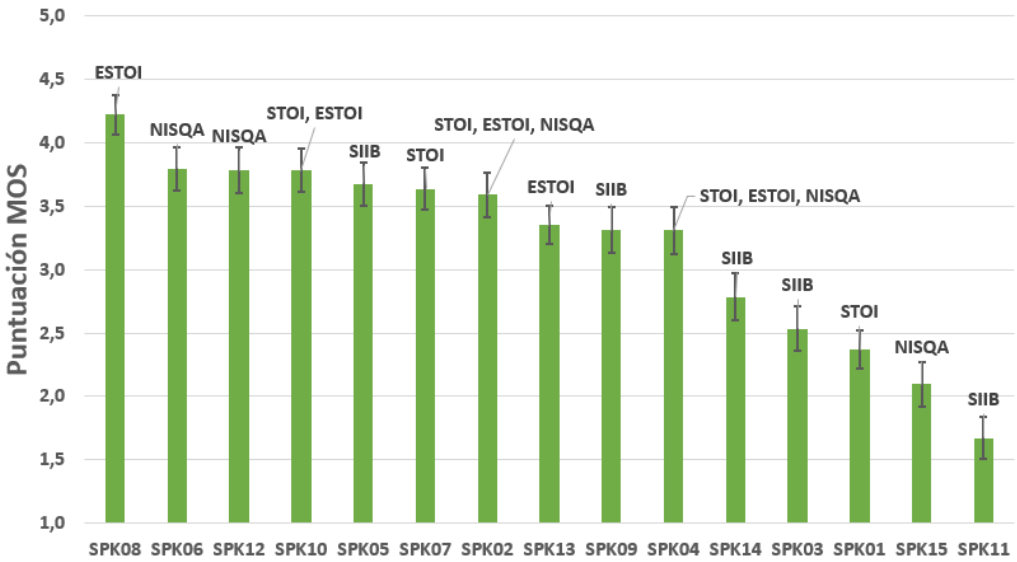


Figura 4.6: Resultado MOS con intervalo de confianza del 95 % para las cuatro medidas objetivas.

Las voces personalizadas con mayor puntuación MOS no corresponden exactamente con el top cinco de voces personalizadas seleccionadas por una única

Tabla 4.6: Coeficiente de correlación entre puntuaciones objetivas y MOS para cada una de las cuatro medidas.

STOI	ESTOI	SIIB	NISQA
0.452	0.584	-0.199	0.356

medida objetiva. La tabla 4.6 muestra cuán correladas están las puntuaciones de las medidas objetivas con las puntuaciones de la evaluación MOS. De mayor a menor coeficiente de correlación resultan: ESTOI, cuyas voces personalizadas con mayor puntuación objetiva se encuentran entre las 7 mejores de la evaluación subjetiva; STOI, cuyas mejores voces personalizadas se encuentran entre las 13 mejor puntuadas de la evaluación MOS; NISQA, cuyas cinco mejores voces personalizadas están entre el top 14 de la evaluación subjetiva; y SIIB, cuyas cinco mejores voces personalizadas corresponden a la quinta, novena, undécima, décimo segunda y décimo quinta puntuaciones de la evaluación MOS.

Dado que cada medida objetiva emplea criterios diferentes y que ninguna ha escogido individualmente las voces personalizadas con mayor puntuación MOS, es interesante investigar si una combinación de medidas objetivas podría usarse para seleccionar las mejores voces personalizadas según la evaluación MOS. Empleando las medidas objetivas disponibles y las puntuaciones MOS obtenidas en la evaluación subjetiva, se ha desarrollado un estimador del resultado MOS basado en regresión lineal. Este estimador se puede aplicar a las miles de voces sintéticas disponibles en el banco de voces para seleccionar, de manera automática, aquéllas con mayor puntuación MOS estimada.

Se ha usado la técnica de *leave-one-out* para medir la eficacia del predictor lineal, y se han generado 15 polinomios de regresión diferentes, usando en cada caso datos de 14 voces dejando la voz restante a efectos de test. La medida SIIB muestra una dificultad adicional para considerarla en la regresión lineal: como la puntuación no es fiable si la frase a evaluar tiene una duración menor de $20s$, es necesario concatenar varias frases, y no se puede obtener una puntuación por frase. Además, dado que la puntuación SIIB correla negativamente con la puntuación MOS obtenida en la evaluación, según la tabla 4.6, y que del top cinco de voces personalizadas de la medida SIIB tres se encuentran entre las cinco con menor

puntuación MOS, según la figura 4.6, finalmente se decide no considerar la medida SIIB en la regresión lineal.

Cada uno de los quince polinomios generados se ha evaluado usando las 10 frases de la voz personalizada no incluidas en la generación del polinomio. El valor medio de la puntuación MOS predicha junto con su intervalo de confianza del 95 % se muestran en la figura 4.7, donde se observa que no hay diferencias significativas para 8 (SPK06, SPK12, SPK10, SPK02, SPK13, SPK09, SPK03, y SPK11) de las 15 voces evaluadas; para las voces SPK08, SPK05 y SPK07 la puntuación MOS ha sido infraestimada, mientras que para las voces SPK01 y SPK15 la puntuación MOS ha sido sobreestimada.

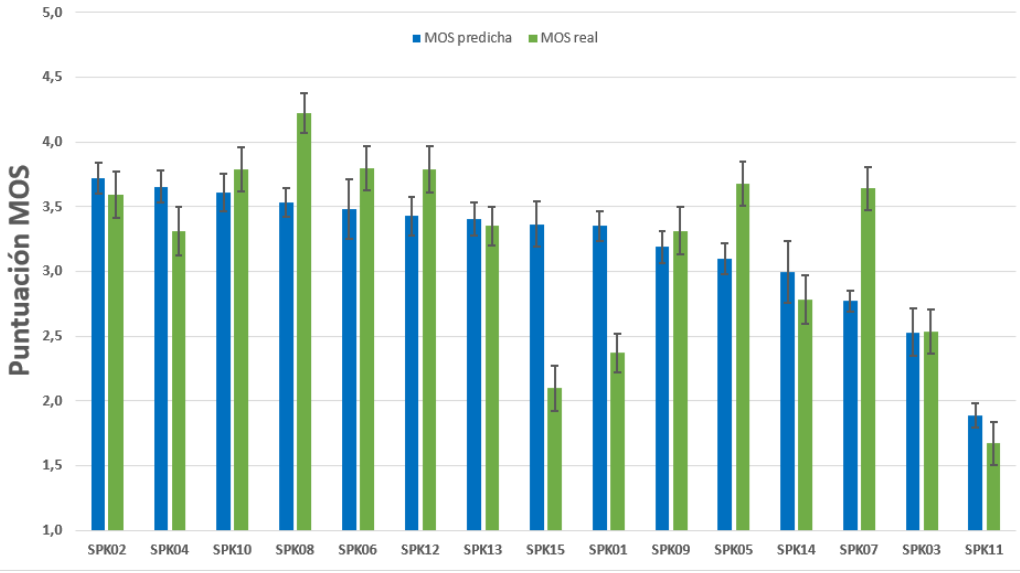


Figura 4.7: Puntuación media MOS real vs. MOS predicha e intervalo de confianza del 95 %.

Así, el estimador de puntuación MOS propuesto podría usarse para hacer una selección preliminar de las voces personalizadas con mayor calidad. Aunque aún es necesaria una revisión manual para eliminar posibles falsos positivos, usando este predictor se puede ahorrar tiempo a la hora de realizar el cribado de voces disponibles.

4.4 Aportaciones

En el presente capítulo se ha presentado la implementación y evolución del banco de voces ZureTTS - *ahoMyTTS* del laboratorio Aholab. Este sistema permite obtener una voz personalizada o donarla, tanto en castellano como en euskera. Se ha diseñado para que sea lo más cómodo de utilizar de cara al usuario o donante, siendo necesario grabar 100 frases en una o varias sesiones. Dada la popularidad y el éxito del banco de voces, ya se dispone de más de 1.000 voces de donantes, resultando complejo seleccionar aquéllas que presentan mejor calidad. Para poder seleccionar de manera automática un primer *subset* de las mejores voces, se ha estudiado el uso de las medidas objetivas STOI, ESTOI, SIIB y NISQA. Se ha comprobado que mediante una combinación de las puntuaciones objetivas de STOI, ESTOI y NISQA es posible predecir la puntuación MOS de una voz personalizada. La medida SIIB ofrece un desempeño inferior comparado con el de las otras medias estudiadas para el presente caso de uso.

Durante el transcurso de esta investigación se han publicado los siguientes trabajos científicos:

1. D. Erro, I. Hernaez, E. Navas, **A. Alonso**, H. Arzelus, I. Jauk, N.Q. Hy, C. Magariños, R. Perez-Ramon, M. Sulir, X. Tian, X. Wang, J. Ye, "ZureTTS: online platform for obtaining personalized synthetic voices", Proc. 10th International Summer Workshop on Multimodal Interfaces (eNTERFACE), pp. 17-25, Bilbao, June 2014.

2. D. Erro, I. Hernaez, **A. Alonso**, D. Garcia-Lorenzo, E. Navas, J. Ye, H. Arzelus, I. Jauk, N.Q. Hy, C. Magariños, R. Perez-Ramon, M. Sulir, X. Tian, X. Wang, "Personalized Synthetic Voices for Speaking Impaired: Website and App", Proc. Interspeech, pp. 1251-1254, Dresden, September 2015.

3. **A. Alonso**, V. García, I. Hernaez, E.Navas, J. Sanchez, "Automatic Speaker Adaptation Assessment Based on Objective Measures for Voice Banking Donors" In IberSPEECH. 2020.

4. **A. Alonso**, V. García, I. Hernaez, E. Navas, J. Sanchez. “Automatic Classification of Synthetic Voices for Voice Banking Using Objective Measures”. Applied Sciences, 12(5), 2473. 2022.

CAPÍTULO

5

Conclusiones y líneas futuras

En el presente capítulo se expone de forma global el trabajo investigador realizado a lo largo de la tesis, y se describen las posibles líneas futuras de investigación a seguir. Finalmente, se detalla la difusión de los resultados obtenidos, categorizándolos según se trata de artículos de revista, publicaciones en congreso o participación en campañas de evaluación y *workshops*.

5.1 Conclusiones

El objetivo principal de esta tesis ha consistido en el estudio de la generación de voces personalizadas en sistemas TTS, permitiendo que personas con alguna discapacidad oral los puedan utilizar contando con voces adaptadas que suenen lo más parecido posible a la suya.

De entre las tecnologías de síntesis existentes actualmente se ha escogido la síntesis estadístico paramétrica para el transcurso de toda la investigación. En este ámbito, se ha estudiado cómo obtener voces sintéticas usando dicha tecnología así como distintos algoritmos de adaptación del estado del arte que permiten obtener voces sintéticas nuevas a partir de una voz inicial ya entrenada y una pequeña cantidad de datos de adaptación. Además, se ha desarrollado un sistema TTS multilingüe para castellano, euskera, gallego, catalán e inglés empleando esta tecnología. Este sistema integra módulos lingüísticos de diferentes universidades y se han entrenado voces específicas para cada idioma usando bases de datos disponibles. El código se ha liberado bajo licencia de código abierto en un repositorio público.

Además, se ha diseñado un nuevo método de adaptación de locutor basado en FW+AS. Aunque este tipo de transformaciones suele emplearse en conversión de voz, se ha demostrado que también pueden emplearse en el ámbito de adaptación de voces estadístico paramétricas. La técnica presentada hace uso únicamente de fragmentos vocálicos de la voz objetivo para calcular las transformaciones que se aplican a todo el modelo. Se ha comparado con un método de adaptación del estado del arte bien establecido realizando diversas evaluaciones MOS. El material empleado para realizar las adaptaciones se ha obtenido usando grabaciones realizadas por locutores no profesionales, en entornos no controlados por el laboratorio y presenta una calidad media-baja. Los resultados demuestran que,

respecto a la calidad de la voz adaptada, el método propuesto es robusto cuando el material de adaptación no presenta una calidad óptima. También se ha demostrado que es robusto cuando se limita la cantidad de datos de adaptación. Sin embargo, respecto a la similitud de la voz adaptada con la voz objetivo, el método del estado del arte obtiene puntuaciones algo superiores con respecto al método propuesto.

Se ha probado este método de adaptación con grabaciones realizadas por una persona que presenta disartria. El objetivo era tratar de obtener una voz adaptada que sonara como lo haría su voz si no presentara dicha discapacidad oral. Lamentablemente, aunque la calidad de la voz resultante es bastante alta la similitud obtenida es muy baja. Dado que utilizar esta voz sintética frente a emplear una voz genérica o adaptada con material de otra persona no supone una ventaja, se ha descontinuado esta línea de investigación.

Finalmente se ha presentado el diseño, evolución y funcionamiento de un banco de voces para castellano y euskera desarrollado por el laboratorio Aholab. Este banco de voces puede emplearse en cualquiera de los dos idiomas tanto para obtener una voz personalizada para uso propio como para que donantes altruistas donen su voz. De esta forma, en el caso de que alguien desee una voz personalizada pero no pueda hacer las grabaciones necesarias, se le puede proveer de una voz sintética de un donante. El banco de voces ha sido un éxito, contando actualmente con varios miles voces disponibles para su uso. Sin embargo, las voces disponibles presentan una gran variabilidad en su calidad, dado que no se tiene control sobre el proceso de grabación y cómo se han realizado las grabaciones afecta en gran medida a la calidad final de la voz adaptada. Por ello, surge la necesidad de puntuar las voces adaptadas con el fin de seleccionar las mejores. La gran cantidad de voces donadas hace inviable plantear evaluaciones subjetivas, es por eso que se propone una estrategia de puntuación automática. Se ha estudiado el uso de cuatro medidas objetivas (STOI, ESTOI, SIIB y NISQA) para determinar su idoneidad a la hora de puntuar diferentes aspectos (inteligibilidad y naturalidad) de las voces sintéticas. Se han realizado diversas evaluaciones MOS para comparar las puntuaciones obtenidas por las medidas objetivas con la opinión real de los evaluadores. En base a los resultados obtenidos, se ha concluido que las puntuaciones correspondientes a las medidas STOI, ESTOI y NISQA correlan mejor con las puntuaciones MOS que la medida SIIB. Seguidamente, haciendo uso de las tres medidas objetivas con

mejor desempeño se ha diseñado un predictor lineal de puntuación MOS. Se ha observado que el predictor puede emplearse para estimar la puntuación MOS que obtendría una voz sintética en una evaluación subjetiva.

5.2 Líneas Futuras de Trabajo

Como posibles líneas de trabajo futuras, se proponen las siguientes opciones en base a los resultados obtenidos durante la investigación realizada.

Se ha demostrado que el método propuesto de adaptación haciendo uso de transformaciones basadas en FW+AS es robusto frente a escasez de datos. En el caso de la voz con disartria con la que se ha probado, aunque la calidad de la voz obtenida es alta, no ha sido capaz de captar correctamente la personalidad del usuario. No obstante, dado que solo se ha probado con un tipo concreto de patología, no se puede descartar que en otros casos sí sea capaz de adaptar la personalidad del usuario. Si se consiguen grabaciones de voz de personas que sufran diferentes tipos de impedimentos orales, se puede estudiar si para ciertos casos el método es capaz de obtener una voz sintética que suene como el usuario deseado.

Además, el método de selección de fragmentos vocálicos para ser usados durante la adaptación es bastante sencillo. Únicamente tiene en cuenta que la duración sea mayor que un umbral fijado de manera arbitraria y que la f_0 haya sido detectada correctamente. Una selección más restrictiva que fije condiciones más estrictas puede dar lugar a que los fragmentos seleccionados presentan mayor calidad. Dado que el método es robusto frente a escasez de datos, es mejor seleccionar mejores vocales que más vocales. Es necesario estudiar qué tipo de condiciones pueden dar lugar a una mejor selección.

En relación con las funcionalidades del banco de voces, actualmente el sistema permite obtener una voz sintética personalizada en dos idiomas: castellano y euskera. Sin embargo, en el caso de que un donante sea bilingüe, es necesario que grabe su voz para cada idioma de manera independiente. Además, un usuario con discapacidad oral que domina varios idiomas debe emplear la voz personalizada de diferentes donantes si no existe uno de su agrado que haya grabado en ambos idiomas. Una posible mejora consistiría en emplear técnicas de adaptación *cross-lingual* [63] para poder obtener voces personalizadas en diferentes idiomas empleando las grabaciones en un único idioma. De esta manera, los donantes solo

necesitarían grabar en un idioma, y los usuarios con discapacidad oral podrían utilizar la misma voz personalizada independientemente del idioma.

Respecto a la evaluación de voces adaptadas del banco de voces empleando medidas objetivas, se ha demostrado que su uso puede ser útil para realizar una predicción de la puntuación MOS que obtendrían en una evaluación subjetiva. En este aspecto, únicamente se han estudiado cuatro medidas objetivas (STOI, ESTOI, SIIB y NISQA), siendo posible que existan otras medidas que puedan aportar información complementaria sobre la calidad de las señales sintéticas. Se pueden estudiar otras medidas como por ejemplo DAU [16], *Glimpse Proportion* [19] o *Hearing-Aid Speech Perception Index* (HASPI) [53, 54]. El estudio de la predicción automática de la puntuación MOS es un tema que está siendo ampliamente estudiado en la actualidad, existiendo diversos *challenges* (como por ejemplo *The VoiceMOS Challenge*[44]) en los que se enfrentan diversos sistemas de predicción de MOS.

Además, el predictor desarrollado hace uso de un modelo muy sencillo basado en regresión lineal. Modelos más complejos pueden dar lugar a predicciones más precisas de la puntuación MOS. No obstante, para ser capaces de entrenar modelos más complejos se necesitan más datos, por lo que es necesario realizar evaluaciones subjetivas adicionales.

Por último, respecto a la tecnología de síntesis empleada durante el desarrollo de esta investigación, se decidió enfocarla a la síntesis estadístico paramétrica. Sin embargo, las redes neuronales presentan un gran futuro y una gran oportunidad para la síntesis de voz y la personalización de voces sintéticas. Aunque empleando esta tecnología se han conseguido voces sintéticas de gran calidad, aún presentan ciertos inconvenientes que deben ser estudiados de cara a su empleo en sistemas AAC. El mayor de ellos actualmente reside en su complejidad para poder integrarlas en sistemas embebidos o de gama baja. Las redes neuronales generadas tienen un tamaño de varios cientos de megabytes, lo cual es dos órdenes de magnitud más que los modelos estadísticos de las voces estadístico paramétricas. Además, la capacidad de computación necesaria para sintetizar audio en tiempo real con redes neuronales también es sensiblemente superior. Es por ello por lo que su

implementación en dispositivos con prestaciones bajas resulta tremendamente ardua y trabajosa. Una opción para aliviar este problema y reducir el tamaño de las redes consiste realizar un podado de la red resultante para reducir su tamaño. Sin embargo, ésta no es una tarea trivial dado que hay que conseguir un compromiso entre el tamaño de la poda y la degradación de la calidad de la voz resultante.

Respecto a la personalización de voces empleando redes neuronales, también existen alternativas a estudiar de cara a su implementación en el banco de voces de Aholab. Dado que puede obtener una similitud muy grande con poco material de adaptación, conviene estudiar su viabilidad de cara a su uso. Las redes de partida que se emplean en estos casos suele entrenarse a partir de varios locutores mediante varias horas de grabación. Se pueden emplear las 100 frases del corpus de adaptación del banco de voces para realizar un *fine tuning* de la red para que adquiera las características de la voz del usuario o donante. En el caso de las adaptaciones *zero-shot*, tienen a favor que requieren muy poco material de adaptación, siendo necesaria una única grabación de pocos segundos de duración. En contra, la red de partida necesita ser entrenada con miles de horas de grabación de una gran cantidad de locutores diferentes. Dado que es muy difícil obtener el material necesario para entrenar una red de estas características,

Con todos estos factores, se puede concluir que aún existen demasiadas incertidumbres respecto a la síntesis por redes neuronales para su aplicación en sistemas AAC. No obstante, que se están haciendo progresos y es posible que en un futuro cercano sea una opción viable a tener en cuenta.

5.3 Difusión de Resultados

A continuación, se listan todas las publicaciones científicas realizadas durante el transcurso de está investigación tanto en revistas como en congresos. Así mismo, se mencionan las participaciones en diferentes campañas de evaluación y *workshops*.

5.3.1 Artículos de Revista

2022 **Agustin Alonso**, Víctor García, Inma Hernaez, Eva Navas, Jon Sanchez, *"Automatic Classification of Synthetic Voices for Voice Banking Using Objective Measures"*, Applied Sciences, vol. 12, no 5, p. 2473, 2022

2016 **Agustin Alonso**, Daniel Erro, Eva Navas, Inma Hernaez, *"Study of the effect of reducing training data in speech synthesis adaptation based on Frequency Warping"*, Lecture Notes in Artificial Intelligence LNCS/LNAI (ISSN: 0302-9743), vol. 10077, pp. 3-13, 2016.

2015 Daniel Erro, **Agustin Alonso**, Luis Serrano, Eva Navas, Inma Hernáez, *"Interpretable parametric voice conversion functions based on Gaussian mixture models and constrained transformations"*, Computer Speech & Language (Q3 en 2015), vol. 30, no 1, pp. 3-15, 2015.

2014 **Agustin Alonso**, Daniel Erro, Eva Navas, Inma Hernaez, *"Fine Vocoder Tuning for HMM-Based Speech Synthesis: Effect of the Analysis Window Length"*, Lecture Notes in Artificial Intelligence LNCS/LNAI (ISSN: 0302-9743), vol. 8854, pp. 21-29, 2014.

2013 Daniel Erro, **Agustin Alonso**, Luis Serrano, Eva Navas, Inma Hernaez, *"Towards Physically Interpretable Parametric Voice Conversion Functions"*, Lecture Notes in Artificial Intelligence LNCS/LNAI (ISSN: 0302-9743), vol. 7911, pp. 75-82, 2013.

2013 **Agustin Alonso**, Iñaki Sainz, Daniel Erro, Eva Navas, Inma Hernaez, *"Sistema de conversión texto a voz de código abierto para lenguas ibéricas"* (in Spanish), Procesamiento del Lenguaje Natural (ISSN: 1135-5948), vol. 51, pp. 169-175, 2013.

5.3.2 Publicaciones en Congresos

2021 **Agustin Alonso**, Victor García, Inma Hernaez, Eva Navas, Jon Sanchez *"Automatic Speaker Adaptation Assessment Based on Objective Measures for Voice Banking Donors"*, In Proceedings of IberSPEECH 2021, Valladolid, Spain, pp. 210-214, 2021

2017 David Tavárez, Xabier Sarasola, **Agustin Alonso**, Jon Sánchez, Luis Serrano, Eva Navas, Inma Hernáez, *"Exploring Fusion Methods and Feature Space for the Classification of Paralinguistic Information"*, In Proceedings of INTERSPEECH 2017, Stockholm, Sweeden, pp. 3517-3521, 2017

2016 David Tavárez, Xabier Sarasola, Eva Navas, Luis Serrano, **Agustin Alonso**, Ibon Saratxaga, Inma Hernáez, *"Aholab Speaker Diarization System for Albayzin 2016 Evaluation Campaign"*, In Proceedings of IberSPEECH 2016, Lisboa, Portugal, pp. 9-18, 2016

2016 Daniel Erro, **Agustin Alonso**, Luis Serrano, David Tavárez, Igor Odriozola, Xabier Sarasola, Eder Del Blanco, Jon Sanchez, Ibón Saratxaga, Eva Navas, Inma Hernáez, *"ML Parameter Generation with a Reformulated MGE Training Criterion-Participation in the Voice Conversion Challenge 2016"*, In INTERSPEECH 2016, San Francisco, USA, pp. 1662-1666, 2016

2015 **Agustin Alonso**, Daniel Erro, Eva Navas, Inma Hernaez, *"Speaker Adaptation using only Vocalic Segments via Frequency Warping"*, In Proceedings of INTERSPEECH 2015, pp. 2764-2768, Dresden, Germany 2015.

2015 Daniel Erro, Inma Hernaez, Eva Navas, **Agustin Alonso**, Haritz Arzelus, Igor Jauk, Nguyen Quy Hy, Carmen Magariños, Rubén Pérez-Ramón, Martin Sulir, Xiaohai Tian, Xin Wang, Jianpei Ye, *"Personalized Synthetic Voices for Speaking Impaired: Website and App"*, In Proceedings of INTERSPEECH 2015, pp. 1251-1254, Dresden, Germany 2015.

2013 Daniel Erro, **Agustin Alonso**, Luis Serrano, Eva Navas, Inma Hernaez, *"New Method for Rapid Vocal Tract Length Adaptation in HMM-based Speech*

Synthesis", In Proceedings of 8th ISCA Workshop on Speech Synthesis (SSW8), pp. 125-128, Barcelona, 2013.

5.3.3 Campañas de Evaluación y Workshops

5.3.3.1 Albayzin

2016 *Best system in the Albayzin audio segmentation evaluation campaign*, In IberSPEECH 2016, Lisboa, Portugal.

2014 *Best system in the Albayzin audio segmentation evaluation campaign*, In IberSPEECH 2014, Las Palmas, Spain.

5.3.3.2 eNTERFACE

2014 *ZureTTS: online platform for obtaining personalized synthetic voices*, In eNTERFACE 2014, Bilbao, Spain.

5.3.3.3 RTTH

2013 RTTH Summer School on Speech Technology Evaluation, 2013, Vigo, Spain.

Bibliografía

[1] ITU-T Rec. P563, "Single-ended method for objective speech quality assessment in narrow-band telephony applications". 71

[2] Acapela Group. https://www.acapela-group.com/solutions/my-own-voice/, Accessed June 2023. 65

[3] ahoMyTTS (formerly ZureTTS). https://aholab.ehu.eus/ahomytts/, Accessed June 2023. 52, 77

[4] AhoTTS. https://sourceforge.net/projects/ahotts/, Accessed June 2023. 119

[5] Allen, Jonathan, Hunnicutt, Sharon, Carlson, Rolf, & Granstrom, Bjorn. 1979. MITalk-79: The 1979 MIT text-to-speech system. *The Journal of the Acoustical Society of America*, **65**(S1), S130–S130. 8

[6] Anastasakos, Tasos, McDonough, John, Schwartz, Richard, & Makhoul, John. 1996. A compact model for speaker-adaptive training. *Pages 1137–1140 of: Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP'96*, vol. 2. IEEE. 30

[7] Arik, Sercan, Chen, Jitong, Peng, Kainan, Ping, Wei, & Zhou, Yanqi. 2018. Neural voice cloning with a few samples. *Advances in neural information processing systems*, **31**. 11, 12

[8] Arık, Sercan Ö, Chrzanowski, Mike, Coates, Adam, Diamos, Gregory, Gibiansky, Andrew, Kang, Yongguo, Li, Xian, Miller, John, Ng, Andrew,

Raiman, Jonathan, *et al.* 2017. Deep voice: Real-time neural text-to-speech. *Pages 195–204 of: International Conference on Machine Learning*. PMLR. 10

[9] Azizah, Kurniawati, Adriani, Mirna, & Jatmiko, Wisnu. 2020. Hierarchical transfer learning for multilingual, multi-speaker, and style transfer dnn-based tts on low-resource languages. *IEEE Access*, **8**, 179798–179812. 11

[10] Cappé, Olivier, Laroche, J., & Moulines, Eric. 1995. Regularized Estimation of Cepstrum Envelope from Discrete Frequency Points. *IEEE ASSP Workshop on Applications of Signal Processing to Audio and Acoustics*, 213–216. 45

[11] Casanova, Edresson, Shulby, Christopher, Gölge, Eren, Müller, Nicolas Michael, de Oliveira, Frederico Santos, Junior, Arnaldo Candido, Soares, Anderson da Silva, Aluisio, Sandra Maria, & Ponti, Moacir Antonelli. 2021. Sc-glowtts: an efficient zero-shot multi-speaker text-to-speech model. *arXiv preprint arXiv:2104.05557*. 12

[12] CereVoice. `https://www.cereproc.com/en/products/cerevoiceme`, Accessed June 2023. 65

[13] Chen, Mingjian, Tan, Xu, Li, Bohan, Liu, Yanqing, Qin, Tao, Zhao, Sheng, & Liu, Tie-Yan. 2021. Adaspeech: Adaptive text to speech for custom voice. *International Conference on Learning Representations*. 11, 12

[14] Chen, Yutian, Assael, Yannis, Shillingford, Brendan, Budden, David, Reed, Scott, Zen, Heiga, Wang, Quan, Cobo, Luis C, Trask, Andrew, Laurie, Ben, *et al.* 2018. Sample efficient adaptive text-to-speech. *International Conference on Learning Representations*. 11, 12

[15] Chien, Jen-Tzung, Lee, Chin-Hui, & Wang, Hsiao-Chuan. 1997. Improved Bayesian learning of hidden Markov models for speaker adaptation. *Pages 1027–1030 of: 1997 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2. IEEE. 34

[16] Christiansen, Claus, Pedersen, Michael Syskind, & Dau, Torsten. 2010. Prediction of speech intelligibility based on an auditory preprocessing model. *Speech Communication*, **52**(7-8), 678–692. 94

[17] Coker, Cecil H. 1976. A model of articulatory dynamics and control. *Proceedings of the IEEE*, **64**(4), 452–460. 8

[18] Colomes, Catherine, Schmidmer, Christian, Thiede, Thilo, & Treurniet, William C. 1999. Perceptual quality assessment for digital audio: PEAQ-The new ITU standard for objective measurement of the perceived audio quality. *In: Audio Engineering Society Conference: 17th International Conference: High-Quality Audio Coding*. Audio Engineering Society. 71

[19] Cooke, Martin. 2006. A glimpsing model of speech perception in noise. *The Journal of the Acoustical Society of America*, **119**(3), 1562–1573. 94

[20] Cooper, Erica, Lai, Cheng-I, Yasuda, Yusuke, & Yamagishi, Junichi. 2020a. Can speaker augmentation improve multi-speaker end-to-end tts? *Pages 3979–3983 of: Interspeech 2020*. 11

[21] Cooper, Erica, Lai, Cheng-I, Yasuda, Yusuke, Fang, Fuming, Wang, Xin, Chen, Nanxin, & Yamagishi, Junichi. 2020b. Zero-shot multi-speaker text-to-speech with state-of-the-art neural speaker embeddings. *Pages 6184–6188 of: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 12

[22] Cotovia. `https://sourceforge.net/projects/cotovia/`, Accessed June 2023. 36, 119

[23] Creer, Sarah, Cunningham, Stuart, Green, Phil, & Yamagishi, Junichi. 2013. Building personalised synthetic voices for individuals with severe speech impairment. *Computer Speech & Language*, **27**(6), 1178–1193. 66

[24] ctbparser. `https://sourceforge.net/projects/ctbparser/`, Accessed June 2023. 120

[25] Dempster, Arthur P, Laird, Nan M, & Rubin, Donald B. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, **39**(1), 1–22. 24, 32

[26] Digalakis, Vassilions V, & Neumeyer, Leonardo G. 1996. Speaker adaptation using combined transformation and Bayesian methods. *IEEE transactions on speech and audio processing*, **4**(4), 294–300. 34

[27] Donahue, Jeff, Dieleman, Sander, Bińkowski, Mikołaj, Elsen, Erich, & Simonyan, Karen. 2020. End-to-end adversarial text-to-speech. *arXiv preprint arXiv:2006.03575*. 10

[28] eNTERFACE 14. `https://aholab.ehu.eus/eNTERFACE14/`, Accessed June 2023. 66

[29] Erro, Daniel, Alonso, Agustín, Serrano, Luis, Navas, Eva, & Hernáez, Inma. 2013. Towards physically interpretable parametric voice conversion functions. *Pages 75–82 of: Advances in Nonlinear Speech Processing: 6th International Conference, NOLISP 2013, Mons, Belgium, June 19-21, 2013. Proceedings 6*. Springer. 43

[30] Erro, Daniel, Zorilă, Tudor-Cătălin, & Stylianou, Yannis. 2014a. Enhancing the intelligibility of statistically generated synthetic speech by means of noise-independent modifications. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, **22**(12), 2101–2111. 11

[31] Erro, Daniel, Sainz, Iñaki, Navas, Eva, & Hernaez, Inma. 2014b. Harmonics plus Noise Model based Vocoder for Statistical Parametric Speech Synthesis. *IEEE Journal of Selected Topics in Signal Processing*, **8**(2), 184–194. 9, 14, 52, 76, 119

[32] Falk, Tiago H, Parsa, Vijay, Santos, Joao F, Arehart, Kathryn, Hazrati, Oldooz, Huber, Rainer, Kates, James M, & Scollie, Susan. 2015. Objective quality and intelligibility prediction for users of assistive listening devices: Advantages and limitations of existing tools. *IEEE signal processing magazine*, **32**(2), 114–124. 73

[33] Festcat. `http://festcat.talp.cat/`, Accessed June 2023. 36, 118, 119

[34] Festival. `https://www.cstr.ed.ac.uk/projects/festival/`, Accessed June 2023. 36, 119

[35] Freij, Ghassan J, & Fallside, Frank. 1988. Lexical stress recognition using hidden Markov models. *Pages 135–138 of: ICASSP-88., International Conference on Acoustics, Speech, and Signal Processing*. Los Alamitos, CA, USA: IEEE Computer Society. 23

[36] Fukada, Toshiaki, Tokuda, Keiichi, Kobayashi, Takao, & Imai, Satoshi. 1992 (March). An adaptive algorithm for mel-cepstral analysis of speech. *Pages 137–140 vol.1 of: [Proceedings] ICASSP-92: 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1. 15

[37] Gales, Mark JF, & Woodland, Philip C. 1996. Mean and variance adaptation within the MLLR framework. *Computer speech and language*, **10**(4), 249–264. 31

[38] Gibiansky, Andrew, Arik, Sercan, Diamos, Gregory, Miller, John, Peng, Kainan, Ping, Wei, Raiman, Jonathan, & Zhou, Yanqi. 2017. Deep voice 2: Multi-speaker neural text-to-speech. *Advances in neural information processing systems*, **30**. 10

[39] Godoy, Elizabeth, Rosec, Oliver, & Chonavel, Thierry. 2012. Voice Conversion Using Dynamic Frequency Warping With Amplitude Scaling, for Parallel or Nonparallel Corpora. *IEEE Transactions on Audio, Speech, and Language Processing*, **20**(4), 1313–1323. 29, 40, 43

[40] Gupta, Arjun K, & Varga, Tamas. 2012. *Elliptically contoured models in statistics*. Vol. 240. Springer Science & Business Media. 33

[41] Hall, Mark, Frank, Eibe, Holmes, Geoffrey, Pfahringer, Bernhard, Reutemann, Peter, & Witten, Ian H. 2009. The WEKA data mining software: an update. *ACM SIGKDD explorations newsletter*, **11**(1), 10–18. 120

[42] Hernaez, Inma, Navas, Eva, Murugarren, Juan Luis, & Etxebarria, Borja. 2001. Description of the AhoTTS system for the Basque language. *In: 4th ISCA Tutorial and Research Workshop (ITRW) on Speech Synthesis*. 36

[43] Hu, Qiong, Bleisch, Tobias, Petkov, Petko, Raitio, Tuomo, Marchi, Erik, & Lakshminarasimhan, Varun. 2021. Whispered and Lombard neural speech synthesis. *Pages 454–461 of: 2021 IEEE Spoken Language Technology Workshop (SLT)*. IEEE. 11

[44] Huang, Wen-Chin, Cooper, Erica, Tsao, Yu, Wang, Hsin-Min, Toda, Tomoki, & Yamagishi, Junichi. 2022. The voicemos challenge 2022. *arXiv preprint arXiv:2203.11389*. 94

[45] Hunt, Andrew J, & Black, Alan W. 1996. Unit selection in a concatenative speech synthesis system using a large speech database. *Pages 373–376 of: 1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*, vol. 1. IEEE. 9

[46] Itakura, Fumitada. 1975. Line spectrum representation of linear predictor coefficients of speech signals. *The Journal of the Acoustical Society of America*, **57**(S1), S35–S35. 15

[47] Janbakhshi, Parvaneh, Kodrasi, Ina, & Bourlard, Hervé. 2019. Pathological Speech Intelligibility Assessment Based on the Short-time Objective Intelligibility Measure. *Pages 6405–6409 of: ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 72, 73

[48] Jensen, Jesper, & Taal, Cees H. 2014. Speech intelligibility prediction based on mutual information. *IEEE/ACM transactions on audio, speech, and language processing*, **22**(2), 430–440. 73

[49] Jensen, Jesper, & Taal, Cees H. 2016. An algorithm for predicting the intelligibility of speech masked by modulated noise maskers. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, **24**(11), 2009–2022. 73

[50] Jensen, U, Moore, Roger K, Dalsgaard, Paul, & Lindberg, Børge. 1994. Modelling intonation contours at the phrase level using continuous density hidden Markov models. *Computer Speech & Language*, **8**(3), 247 – 260. 23

[51] Jia, Ye, Zhang, Yu, Weiss, Ron, Wang, Quan, Shen, Jonathan, Ren, Fei, Nguyen, Patrick, Pang, Ruoming, Lopez Moreno, Ignacio, Wu, Yonghui, *et al.* 2018. Transfer learning from speaker verification to multispeaker text-to-speech synthesis. *Advances in neural information processing systems*, **31**. 12

[52] Jørgensen, Søren, Cubick, Jens, & Dau, Torsten. 2015. Speech intelligibility evaluation for mobile phones. *Acta Acustica United with Acustica*, **101**(5), 1016–1025. 73

[53] Kates, James M, & Arehart, Kathryn H. 2014. The hearing-aid speech perception index (HASPI). *Speech Communication*, **65**, 75–93. 94

[54] Kates, James M, & Arehart, Kathryn H. 2021. The hearing-aid speech perception index (HASPI) version 2. *Speech Communication*, **131**, 35–46. 94

[55] Kawahara, Hideki, Masuda-Katsusue, Ikuyo, & de Cheveigne, Alain. 1999. Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous- frequency-based F0 extraction: possible role of a repetitive structure in sounds. *Speech Communication*, **27**, 187–207. 9, 14

[56] Kishimoto, Yuka, Zen, Heiga, Tokuda, Keiichi, Masuko, Takashi, Kobayashi, Takao, & Kitamura, Tadashi. 2003. Automatic estimation of postfilter coefficients for HMM-based speech synthesis. *Pages 243–244 of: Proc. Spring Meeting of ASJ*. 27

[57] Klatt, Dennis H. 1980. Software for a cascade/parallel formant synthesizer. *the Journal of the Acoustical Society of America*, **67**(3), 971–995. 8

[58] Kominek, John, & Black, Alan W. 2004. The CMU Arctic speech databases. *In: Fifth ISCA workshop on speech synthesis*. 118

[59] Kons, Zvi, Shechtman, Slava, Sorin, Alex, Rabinovitz, Carmel, & Hoory, Ron. 2019. High quality, lightweight and adaptable TTS using LPCNet. *Pages 176–180 of: Interspeech*. 12

[60] Lavan, Nadine, Mileva, Mila, & McGettigan, Carolyn. 2021. How does familiarity with a voice affect trait judgements? *British Journal of Psychology*, **112**(1), 282–300. 4

[61] Leggetter, Christopher J, & Woodland, Philip C. 1995. Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models. *Computer speech & language*, **9**(2), 171–185. 31

[62] Li, Aijun. 2002. Chinese prosody and prosodic labeling of spontaneous speech. *In: Speech Prosody 2002, International Conference*. 120

[63] Magariños Iglesias, Maria del Carmen. 2019. *Voice personalization and speaker de-identification in speech processing systems*. Ph.D. thesis, Universida de Vigo, Teoría do sinal e comunicacións. 93

[64] Masuko, Takashi, Tokuda, Keiichi, Kobayashi, Takao, & Imai, Satoshi. 1996. Speech synthesis using HMMs with dynamic features. *Pages 389–392 of: 1996 ieee international conference on acoustics, speech, and signal processing conference proceedings*, vol. 1. IEEE. 32

[65] Masuko, Takashi, Tokuda, Keiichi, Kobayashi, Takao, & Imai, Satoshi. 1997. Voice characteristics conversion for HMM-based speech synthesis system. *Pages 1611–1614 of: 1997 IEEE international conference on acoustics, speech, and signal processing*, vol. 3. IEEE. 31, 32

[66] Masuko, Takashi, Tokuda, Keiichi, & Kobayashi, Takao. 2003. A study on conditional parameter generation from HMM based on maximum likelihood criterion. *Pages 209–210 of: Autumn Meeting of ASJ*. 27

[67] Mittag, Gabriel, & Möller, Sebastian. 2019. Non-intrusive speech quality assessment for super-wideband speech communication networks. *Pages 7125–7129 of: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 74

[68] Mittag, Gabriel, & Möller, Sebastian. 2020. Deep Learning Based Assessment of Synthetic Speech Naturalness. *Proc. Interspeech 2020*, 1748–1752. 74

[69] Miyanaga, Keisuke, Masuko, Takashi, & Kobayashi, Takao. 2004. A style control technique for HMM-based speech synthesis. *In: Proc. ICSLP*, vol. 4. 21

[70] Model Talker. `https://www.modeltalker.org/`, Accessed June 2023. 65

[71] Moreno Bilbao, M Asunción, Poig, D, Bonafonte Cávez, Antonio, Lleida, Eduardo, Llisterri, Joaquim, Mariño Acebal, José Bernardo, & Nadeu Camprubí, Climent. 1993. Albayzin speech database: Design of the phonetic corpus. *Pages 175–178 of: EUROSPEECH 1993: 3rd European Conference on Speech Communication and Technology: Berlin, Germany: September 22-25, 1993*. . EUROSPEECH. 117

[72] Morise, Masanori, Yokomori, Fumiya, & Ozawa, Kenji. 2016. WORLD: a vocoder-based high-quality speech synthesis system for real-time applications. *IEICE TRANSACTIONS on Information and Systems*, **99**(7), 1877–1884. 9, 14

[73] Moss, Henry B, Aggarwal, Vatsal, Prateek, Nishant, González, Javier, & Barra-Chicote, Roberto. 2020. Boffin tts: Few-shot speaker adaptation by bayesian optimization. *Pages 7639–7643 of: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 12

[74] Moulines, Eric, & Charpentier, Francis. 1990. Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech communication*, **9**(5-6), 453–467. 9

[75] NISQA. `https://github.com/gabrielmittag/NISQA`, Accessed June 2023. 74

[76] Nose, Takashi, Yamagishi, Junichi, Masuko, Takashi, & Kobayashi, Takao. 2007a. A style control technique for HMM-based expressive speech synthesis. *IEICE TRANSACTIONS on Information and Systems*, **90**(9), 1406–1413. 21

[77] Nose, Takashi, Yamagishi, Junichi, Masuko, Takashi, & Kobayashi, Takao. 2007b. A style control technique for HMM-based expressive speech synthesis. *IEICE TRANSACTIONS on Information and Systems*, **90**(9), 1406–1413. 29

[78] Odell, Julian James. 1995. *The use of context in large vocabulary speech recognition*. Ph.D. thesis, University of Cambridge. 25

[79] Olive, Joseph. 1977. Rule synthesis of speech from dyadic units. *Pages 568–570 of: ICASSP'77. IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2. IEEE. 9

[80] Oord, Aaron van den, Dieleman, Sander, Zen, Heiga, Simonyan, Karen, Vinyals, Oriol, Graves, Alex, Kalchbrenner, Nal, Senior, Andrew, & Kavukcuoglu, Koray. 2016. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*. 10

[81] Paul, Dipjyoti, Shifas, Muhammed PV, Pantazis, Yannis, & Stylianou, Yannis. 2020. Enhancing speech intelligibility in text-to-speech synthesis using speaking style conversion. 1361–1365. 11

[82] Pierard, Arnaud, Erro, Daniel, Hernaez, Inma, Navas, Eva, & Dutoit, Thierry. 2016. Surgery of speech synthesis models to overcome the scarcity of training data. *Pages 73–83 of: International Conference on Advances in Speech and Language Technologies for Iberian Languages*. Springer. 66

[83] Ping, Wei, Peng, Kainan, Gibiansky, Andrew, Arik, Sercan Ömer, Kannan, Ajay, Narang, Sharan, Raiman, Jonathan, & Miller, John. 2017. Deep Voice 3: 2000-Speaker Neural Text-to-Speech. 10

[84] Ping, Wei, Peng, Kainan, & Chen, Jitong. 2018. Clarinet: Parallel wave generation in end-to-end text-to-speech. *arXiv preprint arXiv:1807.07281*. 10

[85] Pitz, Michael, & Ney, Hermann. 2005. Vocal Tract Normalization Equals Linear Transformation in Cepstral Space. *IEEE Trans. Speech & Audio Processing*, **13**, 930–944. 44

[86] Pucher, Michael, Zillinger, Bettina, Toman, Markus, Schabus, Dietmar, Valentini-Botinhao, Cassia, Yamagishi, Junichi, Schmid, Erich, & Woltron, Thomas. 2017. Influence of speaker familiarity on blind and visually impaired children's and young adults' perception of synthetic voices. *Computer Speech & Language*, **46**, 179–195. 4

[87] Rabiner, Lawrence, & Juang, Biinghwang. 1986. An introduction to hidden Markov models. *IEEE ASSP Magazine*, **3**(1), 4–16. 10, 23

[88] Recommendation, ITU-T. 2001. Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs. *Rec. ITU-T P. 862*. 71, 72

[89] Recommendation, ITU-T. 2018. Perceptual objective listening quality prediction. *Rec. ITU-T P. 863*. 71

[90] Ren, Yi, Ruan, Yangjun, Tan, Xu, Qin, Tao, Zhao, Sheng, Zhao, Zhou, & Liu, Tie-Yan. 2019. Fastspeech: Fast, robust and controllable text to speech. *Advances in Neural Information Processing Systems*, **32**. 10

[91] Ren, Yi, Hu, Chenxu, Tan, Xu, Qin, Tao, Zhao, Sheng, Zhao, Zhou, & Liu, Tie-Yan. 2020. Fastspeech 2: Fast and high-quality end-to-end text to speech. *arXiv preprint arXiv:2006.04558*. 10

[92] Ross, Kenneth N, & Ostendorf, Mari. 1999. A dynamical system model for generating fundamental frequency for speech synthesis. *IEEE Transactions on Speech and Audio Processing*, **7**(3), 295–309. 23

[93] Sainz, Iñaki, Erro, Daniel, Navas, Eva, Hernáez, Inma, Sanchez, Jon, Saratxaga, Ibon, & Odriozola, Igor. 2012. Versatile Speech Databases for High Quality Synthesis for Basque. *Pages 3308–3312 of: LREC*. Citeseer. 117

[94] Seeviour, P, Holmes, J, & Judd, M. 1976. Automatic generation of control signals for a parallel formant speech synthesizer. *Pages 690–693 of: ICASSP'76. IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1. IEEE. 8

[95] Shadle, Christine H, & Damper, Robert I. 2001. Prospects for articulatory synthesis: A position paper. *In: 4th ISCA Tutorial and Research Workshop (ITRW) on Speech Synthesis*. 8

[96] Shao, YQ, Sui, ZF, Han, JQ, & Wu, YF. 2008. A study on Chinese prosodic hierarchy prediction based on dependency grammar analysis. *Journal of Chinese Information Process*, **2**, 020. 120

[97] Shen, Jonathan, Pang, Ruoming, Weiss, Ron J, Schuster, Mike, Jaitly, Navdeep, Yang, Zongheng, Chen, Zhifeng, Zhang, Yu, Wang, Yuxuan, Skerrv-Ryan, Rj, *et al.* 2018. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. *Pages 4779–4783 of: 2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE. 10

[98] Shinoda, Koichi, & Lee, Chin-Hui. 2001. A structural Bayes approach to speaker adaptation. *IEEE Transactions on Speech and Audio Processing*, **9**(3), 276–287. 31, 34

[99] Siohan, Olivier, Myrvoll, Tor André, & Lee, Chin-Hui. 2002. Structural maximum a posteriori linear regression for fast HMM adaptation. *Computer Speech & Language*, **16**(1), 5–24. 34

[100] Speak Unique. `https://www.speakunique.co.uk/`, Accessed June 2023. 65, 66

[101] Stewart, John Q. 1922. An electrical analogue of the vocal organs. *Nature*, **110**(2757), 311–312. 8

[102] Sulír, M, & Juhár, J. 2013. Design of an optimal male and female slovak speech database for HMM-based speech synthesis. *Proceedings of the Redžúr*, 5–9. 118

[103] Sundermann, David, & Ney, Hermann. 2003. VTLN-based voice conversion. *Pages 556–559 of: Proceedings of the 3rd IEEE International Symposium on Signal Processing and Information Technology (IEEE Cat. No. 03EX795)*. IEEE. 29, 40, 43

[104] Taal, Cees H, Hendriks, Richard C, Heusdens, Richard, & Jensen, Jesper. 2010. A short-time objective intelligibility measure for time-frequency weighted noisy speech. *Pages 4214–4217 of: 2010 IEEE international conference on acoustics, speech and signal processing*. IEEE. 72

[105] Taal, Cees H, Hendriks, Richard C, Heusdens, Richard, & Jensen, Jesper. 2011. An algorithm for intelligibility prediction of time–frequency weighted

noisy speech. *IEEE Transactions on Audio, Speech, and Language Processing*, **19**(7), 2125–2136. 72, 73

[106] Tachibana, Makoto, Yamagishi, Junichi, Masuko, Takashi, & Kobayashi, Takao. 2005. Speech synthesis with various emotional expressions and speaking styles by style interpolation and morphing. *IEICE transactions on information and systems*, **88**(11), 2484–2491. 29

[107] Tachibana, Makoto, Yamagishi, Junichi, Masuko, Takashi, & Kobayashi, Takao. 2006. A style adaptation technique for speech synthesis using HSMM and suprasegmental features. *IEICE transactions on information and systems*, **89**(3), 1092–1099. 29

[108] Tamura, Masatsune, Masuko, Takashi, Tokuda, Keiichi, & Kobayashi, Takao. 1998. Speaker adaptation for HMM-based speech synthesis system using MLLR. *In: the third ESCA/COCOSDA Workshop (ETRW) on Speech Synthesis*. 31

[109] Tamura, Masatsune, Masuko, Takashi, Tokuda, Keiichi, & Kobayashi, Takao. 2001. Adaptation of pitch and spectrum for HMM-based speech synthesis using MLLR. *Pages 805–808 of: 2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 01CH37221)*, vol. 2. IEEE. 31

[110] The Voice Keeper. `https://thevoicekeeper.com/`, Accessed June 2023. 65

[111] Tian, Xiaohai, Wu, Zhizheng, Lee, Siu Wa, & Chng, Eng Siong. 2014. Correlation-based frequency warping for voice conversion. *Pages 211–215 of: The 9th International Symposium on Chinese Spoken Language Processing*. IEEE. 43

[112] Toda, Tomoki, & Tokuda, Keiichi. 2007. A speech parameter generation algorithm considering global variance for HMM-based speech synthesis. *IEICE TRANSACTIONS on Information and Systems*, **90**(5), 816–824. 27

[113] Tokuda, Keiichi, Zen, Heiga, Yamagishi, Junichi, Black, Alan W, Masuko, Takashi, Sako, Shinji., *et al. The HMM- based speech synthesis system (HTS)*. 21

[114] Tokuda, Keiichi, Kobayashi, Takao, Masuko, Takashi, & Imai, Satoshi. 1994. Mel-generalized cepstral analysis-a unified approach to speech spectral estimation. *Pages 18–22 of: ICSLP*, vol. 94. Citeseer. 15

[115] Tokuda, Keiichi, Yoshimura, Takayoshi, Masuko, Takashi, Kobayashi, Takao, & Kitamura, Tadashi. 2000. Speech parameter generation algorithms for HMM-based speech synthesis. **30**, 1315–1318. 9, 26

[116] Tokuda, Keiichi, Masuko, Takashi, Miyakazi, Noboru, & Kobayashi, Takao. 2002. Multi-Space Probability Distribution HMM. *IEICE Transactions on Information and Systems*, **E85-D**(03), 45–464. 23, 31

[117] Torcoli, Matteo, Kastner, Thorsten, & Herre, Jürgen. 2021. Objective measures of perceptual audio quality reviewed: An evaluation of their application domain dependence. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, **29**, 1530–1541. 72

[118] Tu, Tao, Chen, Yuan-Jui, Yeh, Cheng-chieh, & Lee, Hung-Yi. 2019. End-to-end text-to-speech for low-resource languages by cross-lingual transfer learning. *arXiv preprint arXiv:1904.06508*. 11

[119] Valbret, Hélene, Moulines, Eric, & Tubach, Jean-Pierre. 1992. Voice Transformation Using PSOLA Technique. *Speech Communication*, **11**(2-3), 175–187. 40

[120] Van Kuyk, Steven, Kleijn, W Bastiaan, & Hendriks, Richard C. 2017. An instrumental intelligibility metric based on information theory. *IEEE Signal Processing Letters*, **25**(1), 115–119. 73

[121] VocalID. `https://vocalid.ai/`, Accessed June 2023. 65, 66

[122] Wang, Cheng-Cheng, Ling, Zhen-Hua, Zhang, Bu-Fan, & Dai, Li-Rong. 2008. Multi-layer F0 modeling for HMM-based speech synthesis. *Pages 1–4 of: 2008 6th International Symposium on Chinese Spoken Language Processing*. IEEE. 27

[123] Wang, Yuxuan, Skerry-Ryan, RJ, Stanton, Daisy, Wu, Yonghui, Weiss, Ron J, Jaitly, Navdeep, Yang, Zongheng, Xiao, Ying, Chen, Zhifeng, Bengio, Samy, *et al.* 2017. Tacotron: Towards end-to-end speech synthesis. *arXiv preprint arXiv:1703.10135*. 10

[124] Xia, Risheng, Li, Junfeng, Akagi, Masato, & Yan, Yonghong. 2012. Evaluation of objective intelligibility prediction measures for noise-reduced signals in mandarin. *Pages 4465–4468 of: 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 73

[125] Xu, Jin, Tan, Xu, Ren, Yi, Qin, Tao, Li, Jian, Zhao, Sheng, & Liu, Tie-Yan. 2020. Lrspeech: Extremely low-resource speech synthesis and recognition. *Pages 2802–2812 of: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 11

[126] Yamagishi, Junichi. 2003. A training method of average voice model for HMM-based speech synthesis using MLLR. *IEICE Transactions on Information and Systems*. 30

[127] Yamagishi, Junichi, & Kobayashi, Takao. 2007. Average-voice-based speech synthesis using HSMM-based speaker adaptation and adaptive training. *IEICE TRANSACTIONS on Information and Systems*, **90**(2), 533–543. 30, 31

[128] Yamagishi, Junichi, Onishi, Koji, Masuko, Takashi, & Kobayashi, Takao. 2005. Acoustic modeling of speaking styles and emotional expressions in HMM-based speech synthesis. *IEICE TRANSACTIONS on Information and Systems*, **88**(3), 502–509. 29

[129] Yamagishi, Junichi, Kobayashi, Takao, Nakano, Yuji, Ogata, Katsumi, & Isogai, Juri. 2009. Analysis of speaker adaptation algorthims for Hmm-based speech synthesis and a constrained SMAPLR adaptation algorithm. *IEEE*

Transactions on Audio, Speech, and Language Processing, **19**(1), 66–83. 21, 31, 34

[130] Yamagishi, Junichi, Veaux, Christophe, King, Simon, & Renals, Steve. 2012. Speech synthesis technologies for individuals with vocal disabilities: Voice banking and reconstruction. *Acoustical Science and Technology*, **33**(1), 1–5. 66

[131] Yang, Jingzhou, & He, Lei. 2020. Towards Universal Text-to-Speech. *Pages 3171–3175 of: Interspeech*. 11

[132] Yoshimura, Takayoshi, Tokuda, Keiichi, Masuko, Takashi, Kobayashi, Takao, & Kitamura, Tadashi. 1999. Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis. *Pages 2347–2350 of: Sixth European Conference on Speech Communication and Technology, EUROSPEECH 1999, Budapest, Hungary, September 5-9, 1999*. 9, 25

[133] Yoshimura, Takayoshi, Tokuda, Keiichi, Masuko, Takashi, Kobayashi, Takao, & Kitamura, Tadashi. 2001. Mixed excitation for HMM-based speech synthesis. *Pages 2263–2266 of: Seventh European Conference on Speech Communication and Technology*. 27

[134] Young, S., Kershaw, D., Odell, J., Ollason, D., Valtchev, V., Woodland, P., *et al.* 2006. *The Hidden Markov Model Toolkit (HTK), version 3.4.* 21, 23, 119

[135] Zen, Heiga, Tokuda, Keiichi, Masuko, Takashi, Kobayashi, Takao, & Kitamura, Tadashi. 2007. A hidden semi-Markov model-based speech synthesis system. *IEICE Trans. Inf. Syst.*, **E90-D**(5), 825–834. 24

[136] Zen, Heiga, Tokuda, Keiichi, & Black, Alan W. 2009. Statistical parametric speech synthesis. *Speech Communication*, **51**(11), 1039–1064. 9, 20

[137] Zhang, Yu, Weiss, Ron J, Zen, Heiga, Wu, Yonghui, Chen, Zhifeng, Skerry-Ryan, RJ, Jia, Ye, Rosenberg, Andrew, & Ramabhadran, Bhuvana. 2019. Learning to speak fluently in a foreign language: Multilingual speech synthesis and cross-language voice cloning. *Pages 2080–2084 of: Interspeech*. 11

[138] Zhang, Zewang, Tian, Qiao, Lu, Heng, Chen, Ling-Hui, & Liu, Shan. 2020. Adadurian: Few-shot adaptation for neural text-to-speech with durian. *arXiv preprint arXiv:2005.05642*. 12

[139] Zorila, Tudor-Catalin, Erro, Daniel, & Hernaez, Inma. 2012. Improving the Quality of Standard GMM-Based Voice Conversion systems by Considering Physically Motivated Linear Transformations. *Communications in Computer and Information Science*, **328**, 30–39. 44, 45

ANEXO

A

Detalles de Implementación ZureTTS

En este anexo se detalla en qué consiste la implementación del banco de voces de Aholab ZureTTS. Para la primera versión se entrenaron voces promedio para los siete idiomas disponibles (castellano, euskera, gallego, catalán, inglés, eslovaco y chino mandarín) y se integraron módulos lingüísticos para cada uno de ellos.

A continuación se describen las bases de datos utilizadas para entrenar cada una de las voces promedio:

- Castellano: la voz promedio se ha entrenado usando el subset "phonetic" de la base de datos Albayzin [71]. Contiene 6800 frases y 204 locutores diferentes, cada uno con 160, 50 ó 25 frases grabadas.

- Euskera: La voz promedio se ha entrenado usando la base de datos descrita en [93]. Contiene 9 locutores (5 mujeres y 4 hombres) y consta de una hora de grabación de cada uno, excepto para una mujer y un hombre, de quienes se dispone de 6 horas.

- Inglés: la voz promedio se ha entrenado usando las voces disponibles en la base de datos CMU ARTIC [58]. Contiene 7 locutores (2 mujeres y 5 hombres) con 1132 frases cada uno

- Gallego: dada la dificultad de encontrar bases de datos con voces en gallego, se ha usado una única voz proporcionada por la universidad de Vigo, generada con un único locutor varón y 1.316 frases, aproximadamente 1h 15min de grabaciones.

- Catalán: la voz promedio se ha entrenado usando la base de datos liberada por el proyecto Festcat [33]. Contiene 10 locutores (5 mujeres y 5 hombres) con 10 horas de grabación para un locutor hombre y una mujer, y 1 hora de grabación para el resto de lcoutores.

- Eslovaco: la voz promedio se ha entrenado usando un total de 17.903 frases (más de 36h de grabación) de 18 locutores diferentes de tres bases de datos distintas.

    1. Una base de datos grande, compuesta por 4.526 frases fonéticamente balanceadas grabadas por dos locutores diferentes, 1 mujer y 1 hombre (unas 6h de grabación cada uno) [102].
    2. Una base de datos más pequeña compuesta por 330 frases fonéticamente balanceadas grabadas por una mujer y un hombre (40-50min de grabación cada uno).
    3. Una base de datos compuesta por 14 locutores (7 mujeres y 7 hombres) con un número variable de frases por locutor (entre 469 y 810).

- Chino: la voz promedio se ha entrenado usando dos bases de datos proporcionadas por iFLYTECK Co. Ltd. La primera, contiene 1.000 frases grabadas por un locutor varón, siendo la duración media de las grabaciones 6,6s siendo la duración total 110min. La segunda, consta de 1000 frases grabadas por una locutora mujer, con una duración media de 11,2s siendo la duración total 186min. Los textos de sendas bases de datos son diferentes.

En aquellas bases de datos en las que ya se dispone de segmentación fonética de las grabaciones, se ha utilizado. Cuando dicha segmentación no está disponible, se ha obtenido una mediante alineamiento forzado de HMM usando HTK [134].

Todas las grabaciones de todos los idiomas se han parametrizado usando Ahocoder [31]. Dado que Ahocoder trabaja a 16kHz, ha sido necesario re-muestrear aquellas bases de datos que no están grabadas a dicha frecuencia de muestreo.

La integración de lo módulos lingüísticos ha sido la siguiente:

- Castellano y euskera: se ha empleado el módulo lingüístico del TTS open source AhoTTS [4], desarrollado por Aholab.
- Gallego: se ha empleado el módulo lingüístico del TTS Cotovia, desarrollado por el grupo GTM de la Universidad de Vigo [22].
- Inglés: se ha empleado el sistema Festival [34] desarrollado por el grupo CSTR la Universidad de Edimburgo.
- Catalán: se ha usado el front-end compatible con Festival, desarrollado en el marco del proyecto Festcat [33]. Incluye un normalizador de palabras, lexicon y conversor *Letter-to-Sound* (L2S).
- Eslovaco: al igual que para catalán, el analizador está basado en Festival. Incluye un diccionario de más de 150k palabras problemáticas (cuya escritura y pronunciación canónica discrepan), un sistema de conversión L2S y un conjunto de reglas *token-to-word* para numerales (de cero hasta varios millones).
- Chino: para un correcto funcionamiento del analizador lingüístico en chino, éste debe parsear el texto de entrada no solo en una estructura fonema-tono por cada sílaba sino que también se debe especificar la estructura prosódica de toda la frase. Dado que no se ha encontrado disponible un analizador de texto con tales características, se ha construido un analizador específico. Dicho analizador se encarga de: realizar la segmentación de las palabras, el etiquetado POS, parseado gramatical, conversión L2S y predicción de la

prosodia. Para los tres primeros pasos, se emplea un parser de código abierto llamado ctbparser [24] para la conversión L2S, cada ideograma chino se puede convertir en una composición de fonema y tono, o *pinyin*. Sin embargo el chino es conocido por sus polífonos: la pronunciación de una sílaba puede variar dependiendo del contexto. Por ello se ha construido un diccionario jerárquico y se ha adoptado una estrategia de búsqueda simple: si una palabra o frase se puede encontrar en el diccionario, se utiliza la secuencia *pinyin* correspondiente; de lo contrario, el *pinyin* de todas las sílabas de la unidad gramatical se recupera y concatena en una única secuencia *pinyin*. Obtener la secuencia *pinyin* no es suficiente para la síntesis de voz de alta calidad en chino mandarín. Otro componente necesario es la jerarquía prosódica [62]: algunos caracteres adyacentes deben pronunciarse como una sola palabra prosódica y varias palabras prosódicas forman una sola frase prosódica. Tal jerarquía se parece a la estructura gramatical de una oración; por lo tanto, es posible derivar la estructura prosódica en función de los resultados del análisis gramatical mencionado anteriormente. En este caso, se ha adoptado características gramaticales similares a las mencionadas en [96] y se han empleado árboles de decisión [41] para construir el modelo de predicción prosódico. El modelo se ha entrenado a partir de 1900 oraciones del citado corpus. Las pruebas realizadas con las 100 oraciones restantes han mostrado que el rendimiento del modelo de predicción prosódica logra resultados similares a los informados en [96].

Declaration

I herewith declare that I have produced this work without the prohibited assistance of third parties and without making use of aids other than those specified; notions taken over directly or indirectly from other sources have been identified as such. This work has not previously been presented in identical or similar form to any examination board.

The dissertation work was conducted from 2014 to 2023 under the supervision of Inma Hernáez and Daniel Erro at the University of the Basque Country.

Bilbao, September 2023

This dissertation was finished writing in Zaragoza on Tuesday 12th September, 2023