UPV EHU

UNIVERSIDAD DEL PAÍS VASCO / EUSKAL HERRIKO UNIBERTSITATEA

TÉCNICAS PARA LA MEJORA DE LA INTELIGIBILIDAD EN VOCES PATOLÓGICAS

Tesis doctoral presentada por Luis Serrano García

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

Dirigida por la Dra. Inmaculada Hernáez Rioja

UPV EHU

UNIVERSIDAD DEL PAÍS VASCO / EUSKAL HERRIKO UNIBERTSITATEA

TÉCNICAS PARA LA MEJORA DE LA INTELIGIBILIDAD EN VOCES PATOLÓGICAS

Tesis doctoral presentada por Luis Serrano García

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

Dirigida por la Dra. Inmaculada Hernáez Rioja

El doctorando La directora

Bilbao, julio 2019

Técnicas para la mejora de la inteligibilidad en voces patológicas

Autor: Luis Serrano García

Directora: Dra Inmaculada Hernáez Rioja

Impreso en Bilbao

Primera edición, julio 2019

A todos los que hicieron esta tesis posible.

Resumen

Los laringectomizados son personas cuya laringe ha sido extirpada quirúrgicamente, normalmente como consecuencia de un tumor. Al tratarse éste de un órgano fundamental para la producción de la voz, pierden la capacidad de hablar. Sin embargo, muchas de ellas consiguen reaprender a hablar de una manera distinta. Este tipo de habla se conoce como voz esofágica y es bastante distinta de la voz sana. Su naturalidad e inteligibilidad es menor hasta el punto de que ciertos oyentes tienen que hacer un esfuerzo para comprender lo que se les está diciendo.

Esto supone un perjuicio en la vida de los laringectomizados ya que sus capacidades comunicativas se ven afectadas, no solo en las interacciones entre personas sino también en las interfaces hombre-máquina controladas por la voz. En esta tesis se abordan diferentes métodos para la mejora de la inteligibilidad de las voces alaríngeas de manera que palíen estos problemas.

Un aspecto importante ha sido por tanto analizar las características propias de la voz esofágica. No es fácil encontrar el material necesario para hacer este análisis. Los recursos disponibles son escasos y esta tesis ha querido llenar este vacío mediante la grabación de una base de datos paralela de locutores esofágicos.

Esta base de datos ha sido caracterizada acústicamente. Con este objetivo se ha comprobado los efectos que tiene el método de extracción de la frecuencia fundamental a la hora de analizar las características de las señales esofágicas. Se ha propuesto utilizar el análisis del residuo glotal ya que capta mejor las peculiaridades de este tipo de voces.

Es necesario también disponer de algún método para evaluar de manera objetiva el impacto que tienen los métodos propuestos para mejorar la

inteligibilidad. Con este propósito se ha implementado un reconocedor cuyas características y particularidades se recogen en este documento. Este ASR se validó participando en una evaluación de detección de términos hablados organizada por la Red Temática en Tecnologías del Habla.

Para la mejora de la inteligibilidad de las voces esofágicas primero se han analizado diferentes algoritmos basados en las técnicas de conversión de voz existentes aplicadas a voces sanas. Se ha evaluado tanto el comportamiento de técnicas clásicas basadas en mezclas de Gaussianas como el de técnicas de conversión basadas en aprendizaje profundo. Para ello se ha participado en un "*challenge*" internacional de conversión de voz.

Por último, se han adaptado con éxito estas técnicas de conversión a las voces esofágicas. Estas conversiones se han evaluado de manera objetiva mediante el ASR construido, y subjetivamente mediante tests de preferencia. Aunque los resultados de las pruebas subjetivas exponen que para los oyentes no hay diferencias significativas entre las voces convertidas y las esofágicas originales, los resultados del reconocimiento automático muestran que las técnicas de conversión aplicadas a este tipo de voces consiguen disminuir la tasa de error obtenida.

Abstract

The laryngectomees are people whose larynx has been surgically removed, usually as a result of a tumor. The larynx is a fundamental organ in the voice production process, so they lose the ability to speak. However, many of them re-learn to speak in a different way. This type of speech is known as esophageal speech and is quite different from healthy speech. Its naturalness and quality are affected in a way that some listeners have to make an effort to understand what they are being told.

This causes a detriment in the life of the laryngectomees because their communication skills are affected, not only in the interactions with other people but also when using the human-machine interfaces controlled by voice. In this thesis, different methods to improve the intelligibility of the esophageal speech are proposed in order to alleviate these problems.

An important aspect has therefore been to analyze the characteristics of the esophageal speech. It is not easy to find the necessary material to do this analysis. Speaking during a long time requires a lot of effort for the laryngectomees, so the available resources are scarce. This thesis wanted to fill this void by recording a parallel database composed by esophageal speakers voices.

This database has been characterized acoustically. The effects that the pitch extraction method has on the analysis of the characteristics of esophageal signals have been studied. The use of the glottal residue has been proposed to perform the analysis since it captures better the peculiarities of this voices.

It is also necessary to have some method to objectively evaluate the impact of the proposed intelligibility improving methods. For this purpose, a recognizer has been implemented. Its characteristics and peculiarities are detailed in this document. This ASR was validated participating in a spoken term detection evaluation campaign organized by the "'Red Temática en Tecnologías del Habla'.

To improve the intelligibility of the esophageal speech, first we have analyzed different existing voice conversion techniques applied to healthy speech. Both classical techniques based on Gaussian mixtures models and conversion techniques based on deep learning have been evaluated. In order to do this, we have participated in an international voice conversion challenge.

Finally, these conversion techniques have been successfully adapted to esophageal speech. These conversions have been evaluated objectively by the implemented ASR, and subjectively by preference tests. Although the results of the subjective tests show that for the listeners there are no significant differences between the converted voices and the original esophageal voices, the results of the automatic recognition show that the conversion techniques applied to this type of voices manage to reduce the word error rate.

Laburpena

Pertsona laringektomizatuak laringea operazio kirurgikoaren bidez kendu zaienak dira, normalean tumore baten ondorioz. Ahotsaren ekoizpenerako funtsezko organoa denez, hitz egiteko gaitasuna galtzen dute. Hala ere, horietako askok beste modu batean hitz egiten berrikastea lortzen dute. Hizketa mota horri ahots esofagiko esaten zaio eta ahots osasuntsu batekin konparatuta nahiko desberdina da. Naturaltasun eta ulergarritasunean galera handia gertatzen denez, entzule batzuek ahalegin berezia egin behar dute ulertu ahal izateko.

Horrek arazoak ekartzen dizkio laringektomizatuari bere komunikaziotrebetasunean eragina duelako; ez bakarrik pertsonen arteko elkarrekintzetan, baita ahots bidez kontrolatutako giza-makina interfazeetan ere. Tesi honetan laringerik gabeko ahotsen ulergarritasuna hobetzeko metodo desberdinak jorratzen dira arazo horiek arintzeko.

Horretarako ahots esofagikoen ezaugarriak aztertzea funtsezkoa izan da. Ez da erraza azterketa hori egiteko beharrezko materiala aurkitzea. Eskuragarri dauden baliabideak urriak dira. Tesi honek hutsune hau bete nahi izan du ahots esofagikoekin datu-base paralelo bat grabatuz.

Datu-base hori akustikoki karakterizatu da. Helburu horrekin aztertu da frekuentzia nagusiaren erauzketa-metodoak seinale esofagikoen analisian duen eragina. Horretan, hondar glotalaren analisia erabiltzea proposatu da ahots hauen berezitasunak hobeto jasotzen dituelako.

Era berean, irizpide bat eduki behar da ulergarritasuna hobetzeko proposatu diren metodoen eragina objektiboki ebaluatzeko. Helburu horrekin hizketa-ezagutzaile bat inplementatu da, zeinaren ezaugarri eta

berezitasunak dokumentu honetan definitu diren. ASR sistema hori balioztatu egin zen gero RTTH sareak (Hizketa Teknologien Sare Tematikoa) antolatutako hizketa-terminoen detekzioaren ebaluazio-lehiaketa batean parte hartuz.

Ahots esofagikoen ulergarritasuna hobetzeko, lehenik ahots osasuntsuetan erabiltzen diren hizketa-bihurketarako zenbait algoritmo aztertu dira. Gaussiarren nahasketetan oinarritutako teknika klasikoak ebaluatu dira baita ikasketa sakonean oinarritutako bihurketa-teknikak ere. Nazioarteko ahots-bihurketa "*challenge*" batean parte hartuz egin da hori.

Bukatzeko, bihurketa-teknika horiek ahots esofagikoetara arrakastaz egokitu dira. Bihurketa horiek objektiboki ebaluatu dira eraikitako ASR sistemaren bidez, eta subjektiboki ere hautatze-testak erabiliz. Nahiz eta proba subjektiboen emaitzek erakutsi entzuleentzat ez dagoela desberdintasun handirik bihurtutako ahotsak eta ahots esofagiko originalen artean, ezagutza automatikoaren emaitzek erakusten dute hizketa mota honetan aplikatuko bihurtze-teknikak errore-tasa murrizteko gai direla.

Agradecimientos

Parce que esto se termina. Han sido años de trabajo (algún maledicente diría que demasiados) para poder llegar hasta aquí. El camino recorrido ha requerido esfuerzo y no lo hubiera podido conseguir sin el apoyo de mucha gente a la que es de recibo agradecer.

En primer lugar, quisiera expresar todo mi agradecimiento a mi directora de tesis, Inma Hernáez. Muchas gracias por guiarme pacientemente y por poner orden y estructura dentro de mi caos. Ella fue la que me permitió formar parte de este magnífico grupo de investigación. Su energía y dedicación consigue que Aholab salga siempre adelante.

Aholab. ¿Qué decir de este maravilloso grupo de gente? ¿Cómo agradecer todo su apoyo? A Eva le ha tocado más de una vez revisar mis escritos e informes y darles algo de sentido. Mis presentaciones se han visto beneficiadas de su exquisito sentido estético. Cada vez que Ibon se pasa por el laboratorio le asaltamos con dudas y preguntas que nunca se niega a responder, por muy ocupado o falto de tiempo que esté. Además, gracias a sus habilidades de montaje hemos podido "jugar" con redes neuronales. Jon siempre está dispuesto a echar una mano con lo que sea. Sólo he visto flaquear su amabilidad al hablar de los árbitros. Igor tuvo la paciencia de iniciarme en el ingrato mundo del reconocimiento, donde hay que luchar por cada décima de mejora. Iñaki siempre será mi administrador de sistemas.

Dani me enseñó lo que implica ser investigador, además de compartir conmigo sus conocimientos sobre conversión. Siempre recordaré su capacidad de trabajo, sus lecciones y aquél golazo que marcó. Agustín entró conmigo al laboratorio y su compañía y consejos lo hicieron todo más fácil. Aunque haya aparcado su tesis espero que la retome

para darle el empujón final. Con David me encanta discutir. Sus ideas descabelladas me han empujado a afrontar retos y presentarme a "challenges" que se han demostrado indispensables para esta investigación. Espero que no se canse de "molestarme". Xabi siempre tiene alguna teoría interesante que merece la pena escuchar. Nuestras conversaciones sobre cualquier tema siempre me hacen replantearme las cosas.

Itxasne es una recién llegada, pero su encanto bermeano la ha hecho encajar rápidamente. Sneha es una joven brillante que puede contar muchas cosas interesantes, si es que no le toca estar viajando. Estoy convencido de que ambas van a conseguir su doctorado en un periquete.

A todos los componentes del grupo Aholab muchas gracias de corazón por todo lo que me habéis aportado.

También quisiera agradecer la ayuda prestada por la Asociación de Laringectomizados de Bizkaia y en especial a su presidente, Juan Toledo. Gracias por estar siempre dispuestos a colaborar con las grabaciones y hacer este trabajo posible.

Gracias también a mis amigos por animarme siempre. Su confianza en mi capacidad y sus palabras en momentos difíciles han sido esenciales.

Por último, y no por ello menos importante, quisiera darle las gracias a mi familia, especialmente a mi hermana y mis padres. Puede que no entiendan muy bien a que me dedico, pero eso nunca ha sido obstáculo para apoyarme. Sin sus ánimos en todas las etapas importantes de mi vida jamás podría haber llegado hasta aquí. Gracias por estar siempre ahí.

Eskerrik asko,

Luis Serrano

julio 2019

Índice general

Índice de figuras

Índice de tablas

Las palabras acercan. Los silencios destruyen.

André Maurois

CAPÍTULO

1

Introducción

La comunicación es algo inherente al ser humano. Las personas tienen necesidad de socializar y para este fin, la voz es una herramienta importantísima. Poder hablar y ser entendido es algo que damos por sentado, pero no es así para todo el mundo.

Los laringectomizados son personas que han sido sometidos a una intervención quirúrgica en la que se les ha extirpado la laringe (laringectomía total), normalmente como consecuencia de un tumor. Como resultado inmediato de la operación, estas personas se quedan sin capacidad de hablar. Sin embargo, con esfuerzo muchas consiguen aprender a hablar de una manera distinta. Esta nueva voz permite a los laringectomizados volver a comunicarse. Para comprender la importancia de este hecho no hay más que leer las siguientes palabras, escritas por un laringectomizado:

"Para un laringectomizado hablar es, sin lugar a dudas, lo más importante de su vida, una vez que esta la tiene "asegurada" superados los trámites postoperatorios." [1]

[1] Del blog personal de Juan Toledo, http://jtoledo.over-blog.es/2017/01/y-me-oyeron.html

1. INTRODUCCIÓN

La voz de los laringectomizados, denominada voz alaríngea, es bastante diferente a la voz sana. Su inteligibilidad y naturalidad sufren una disminución y, aunque es entendible, para ciertos oyentes comprender lo que les están diciendo requiere un esfuerzo. Por tanto, la comunicación de los laringectomizados se ve afectada.

Es por ello que esta tesis está dedicada a la mejora de las voces alaríngeas. En este capítulo primero se hará una breve descripción de este tipo de voces, así como de los problemas que esto supone para los laringectomizados en su día a día. A continuación se detallarán la motivación y objetivos que se persiguen con este trabajo. Para terminar con la introducción, se expondrá la estructura que sigue este documento.

1.1 Las voces alaríngeas

La laringe es un órgano fundamental en la producción del habla ya que contiene las cuerdas vocales. La vibración que se produce en ellas al pasar el aire es lo que produce la voz. La frecuencia de esta vibración es lo que determina el pitch del locutor y depende de la tensión de las cuerdas vocales, tensión controlada por los músculos de la laringe.

A pesar de no tener laringe, las personas sometidas a una laringectomía total son capaces de producir habla inteligible utilizando mecanismos de producción alternativos. Estos mecanismos incluyen la vibración del segmento faringo-esofágico como sustitución de las cuerdas vocales. Además, durante la operación el tracto vocal se separa del tracto respiratorio. Para que el paciente respire se practica un orificio llamado estoma que conecta la traquea con el exterior (figura 1.1). Por esta razón, el aire que hace vibrar el esfínter esofágico no puede provenir directamente de los pulmones y la tráquea como ocurre en la producción del habla normal. Por tanto, el aire necesario para general la vibración debe producirse de otros modos.

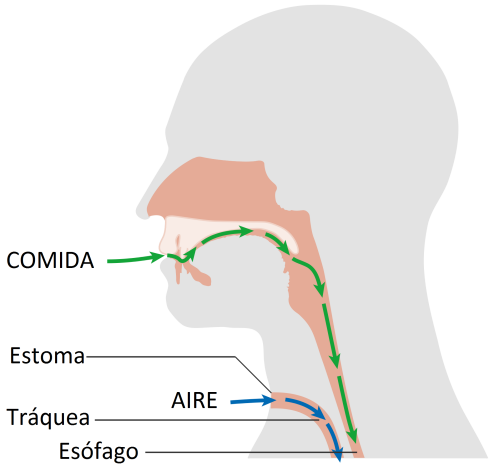

Figura 1.1: Cambios anatómicos debidos a una laringectomía. Cancer Research UK / Wikimedia Commons.

Debido a estas particularidades, después de la operación los laringectomizados tienen que volver a aprender a hablar utilizando un nuevo tipo de voz. Este nuevo tipo de voz se denomina voz alaríngea y, hoy en día, existen tres alternativas:

- Electrolaringe: La electrolaringe es un dispositivo que se coloca en el exterior de la garganta y produce vibraciones. Estas vibraciones se propagan por la cavidad bucal y la simple articulación de palabras produce la voz[9]. Este tipo de voz es fácil de usar, pero es muy monótona y está siempre acompañada del zumbido producido por el aparato. En palabras de los propios laringectomizados sobre la electrolaringe:

  "Es parecido a un pequeño micrófono que se coloca bajo la barbilla y al hacer el efecto del habla vibra produciendo un sonido metálico y monótono pero inteligible."[1]

- Voz esofágica: Este tipo de voz no requiere de ningún dispositivo. Para producir los sonidos, se engulle aire desde la cavidad bucal hacia el esófago y se va soltando de manera controlada. Este tipo de habla se aprende con la ayuda de un logopeda. Suele requerir entre 30 y 50 horas de intenso entrenamiento aprender a hablar de esta manera, aunque no todos los pacientes son capaces de conseguirlo [9].

- Voz traqueoesofágica (TE): En este tipo de habla, se practica una fístula quirúrgica denominada punción TE en la pared que separa la traquea y el esófago y que permite la colocación de una prótesis fonatoria. Esta prótesis TE actúa como una válvula unidireccional: el flujo de aire puede ir de la tráquea al esófago y llegar a las cavidades del tracto vocal permitiendo la producción de voz. Esta punción se hace de tal manera que el paso del esófago a la tráquea sea imposible para evitar que la comida o la bebida pueda entrar en la tráquea y vaya hacia los pulmones. A la hora de hablar, el paciente traqueoesofágico tapa la abertura de la válvula TE, lo que le permite tener un control del aire mucho mejor que el de los locutores de habla esofágica (figura 1.2), haciéndola más fácil de aprender.

El uso de una electrolaringe tiene la ventaja de que puede ser utilizada de forma inmediata tras la operación de laringectomía porque no requiere aprendizaje, pero produce un sonido metálico y monótono. En este sentido, el habla traqueoesfágica tiene una inteligibilidad superior al habla esofágica. Sin embargo, las posibles

[1] https://asbila.jimdo.com/preguntas-frecuentes/

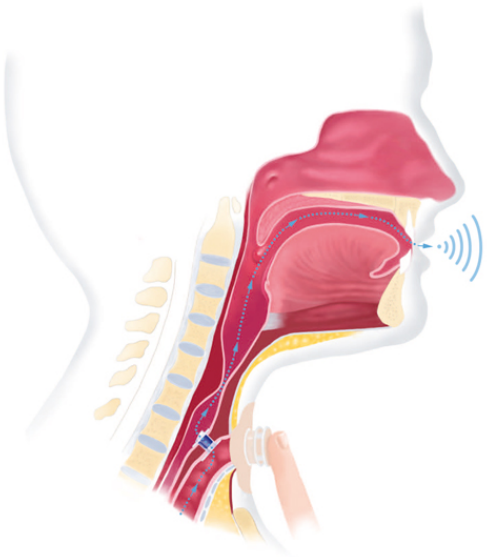

Figura 1.2: Mecanismo de habla con prótesis traqueoesofágica. Laryngectomy 2010 [CC BY-SA 3.0], de Wikimedia Commons.

complicaciones médicas asociadas al implante hacen que exista en la comunidad de médicos y logopedas una clara preferencia por impulsar el aprendizaje del habla esofágica [1]. Es por ello que este trabajo está centrado específicamente en el habla esofágica.

Las diferencias en el mecanismo de producción de la voz hace que las voces alaríngeas sean muy distintas de las voces sanas. La inteligibilidad de este tipo de voces es menor, los oyentes han de hacer un mayor esfuerzo para entender el mensaje. Además, el ritmo de habla suele ser menor, ya que la manera de controlar el aire de los locutores esofágicos hace que introduzcan mayor número de pausas y de silencios.

Otra diferencia importante está evidentemente en el valor de la frecuencia fundamental (f_0). En el habla sana, la curva de f_0 representa la entonación de la voz al pronunciar una frase. La frecuencia fundamental está presente cuando el sonido es *sonoro*, es decir, cuando hay vibración en las cuerdas vocales para producir el sonido. Por el contrario, no hay f_0 cuando el sonido es *sordo*, es decir, cuando no

[1] Al menos en el entorno hospitalario del País Vasco

se usan las cuerdas vocales. Al hablar de la frecuencia fundamental para los hablantes alaríngeos lo que se tiene en cuenta son las vibraciones producidas por el esófago. El valor de f_0 en este caso es en general muy irregular y su estabilidad y continuidad depende mucho del grado de dominio de la técnica de habla esofágica del locutor.

Estos factores llevan a una reducción de naturalidad y de inteligibilidad. Como consecuencia, la capacidad de comunicación de estas personas en su día a día se ve afectada. Además este tipo de voces no son tenidas en cuenta a la hora de diseñar los algoritmos de reconocimiento de voz utilizados en las tecnologías de interacción hombre/máquina que están cada vez más presentes en todas partes.

1.2 Motivación y objetivos

Aunque no hay registros unificados, se calcula que en España hay unos 10000 casos nuevos al año de cánceres de cabeza y cuello, concentrándose la mayor parte de ellos (38 %) en la laringe [93], [43]. Se valora que en 2018 se podrían haber practicado hasta 1200 laringectomías totales [18]. Esto supone que un gran número de personas han visto afectada su calidad de vida de manera importante. El perjuicio más evidente que sufren estas personas es el comunicativo. Incluso los que consiguen reaprender a hablar con voz alaríngea ven mermadas sus capacidades sociales.

Esta pérdida de habilidades comunicativas afecta a varios niveles:

- El primero y más claro es la comunicación cara a cara. La pérdida de inteligibilidad que experimentan las personas sometidas a este tipo de procedimiento quirúrgico es importante incluso cuando consiguen dominar las técnicas de voz sustituta. Esto hace que en sus conversaciones diarias la gente que les escucha tenga que hacer un esfuerzo importante para entenderlos. Aunque los familiares y personas cercanas a los laringectomizados están acostumbrados a este tipo de voz y son capaces de mantener conversaciones fluidas con ellos, la mayor parte de la sociedad no está habituada. Esto provoca que aparezca una barrera que dificulta la interacción normal de los laringectomizados en su vida diaria, creándose una comunicación poco ágil debido a la tensión que puede aparecer en el oyente al tener que hacer un esfuerzo que no está acostumbrado a hacer para entender lo que le dicen. En algunos casos, ciertas personas sienten rechazo frente este tipo de voces y evita mantener conversaciones con los laringectomizados por miedo a no entenderlos bien. Estos problemas afectan al laringectomizado en el plano social, pudiendo hacer que eviten el contacto con los demás y lleguen a aislarse.

- Un problema quizá menos obvio que afecta a los laringectomizados es la comunicación vía teléfonica. A los problemas ya descritos se añade la distorsión propia introducida por el canal telefónico y, sobre todo, la desaparición de toda comunicación no verbal. Los gestos, que en una conversación cara

a cara pueden ayudar a reducir ambigüedades y malentendidos, quedan fuera de escena, lo que hace el entendimiento más complicado. Esto que puede parecer un problema menor, afecta mucho a la calidad de vida del laringectomizado ya que hoy en día hay muchas gestiones que se deben hacer por teléfono. Quizá el ejemplo más crítico de todas sea llamar al servicio de emergencias y no ser capaz de hacerse entender.

- Además de la comunicación entre personas, cada día tiene mayor importancia el uso de la voz como interfaz hombre-máquina. El reconocimiento automático del habla es el primer paso de muchos nuevos sistemas (asistentes personales, sistemas de diálogo telefónico...). Sin embargo, estos sistemas no están diseñados para funcionar con voces patológicas. Los modelos acústicos de los reconocedores suelen estar entrenados sólo con voces sanas lo que hace que los hablantes esofágicos tengan problemas al hacer uso de este tipo de sistemas.

Todos estos problemas y limitaciones que se encuentran los laringectomizados son los que motivan el objetivo general de esta tesis, que es:

– Investigar y desarrollar técnicas que mejoren la inteligibilidad de las voces esofágicas y sean capaces de mejorar la calidad de vida de los laringectomizados.

Es cierto que se puede mejorar el comportamiento de un reconocedor automático de habla frente a los hablantes esofágicos si se adaptan los modelos acústicos a estas voces alaríngeas, pero esto supondría hacer modificaciones en sistemas externos. Por tanto, puede que estos cambios no sean viables. Además, esto no resolvería el problema que tienen estas personas en el resto de escenarios planteados y que implican comunicación con otras personas. Sin embargo, si las modificaciones se hacen sobre la voz esofágica la solución es aplicable a todos los problemas descritos.

Una de las campos que pueden hacer que la inteligibilidad de las voces esofágicas mejore es la conversión de voz. Estas técnicas consisten en aprender a convertir una voz en otra. Una vez se aprende la relación existente entre dos voces, basta con aplicar una función de conversión a la voz del locutor origen para convertirla en

la del locutor destino. Si se consigue desarrollar una función que convierta la voz de un locutor esofágico a una voz sana, todos los problemas vistos anteriormente se solucionarían. Además, las técnicas actuales de conversión de voz funcionan bien para las voces sanas, e incluso hay algunas experimentos diseñados para voces esofágicas.

Para poder evaluar si las técnicas aplicadas consiguen mejorar la inteligibilidad de la voces patológicas será necesario disponer de un reconocedor. Es importante destacar que no se pueden utilizar los reconocedores existentes como el de Google porque, además de no tener ningún control sobre ellos, suelen ser modificados con el tiempo. No es necesario que el reconocedor sea extremadamente bueno ya que se usará para comparar las distintas técnicas y métodos que se implementen, pero es importante que sea el mismo para todas las pruebas. Es por ello que para desarrollar esta tesis es necesario construir un reconocedor de habla continua para el castellano. Los modelos acústicos se entrenarán con audios de locutores de voz sana para que los resultados sean extrapolables.

Es importante también para poder desarrollar estas técnicas de mejora disponer de los datos necesarios. Es por tanto necesario investigar sobre los materiales utilizados en los distintos trabajos de investigación en voces alaríngeas y, si fuera preciso, realizar una base de datos propia adecuada para este trabajo en particular.

Teniendo en cuenta estos aspectos, el objetivo general de esta investigación puede dividirse en los siguientes:

- Analizar las principales diferencias existentes entre las voces sanas y las esofágicas.
  - Hacer un estudio de la literatura para ver qué parámetros y mediciones se utilizan para definir y caracterizar las voces patológicas.
  - Obtener un corpus adecuado que recoja las características más representativas de las voces alaríngeas.
- Investigar algoritmos basados en las técnicas de conversión de voz existente:
  - Evaluar el comportamiento de las técnicas clásicas basadas en mezclas de Gaussianas.

  - Investigar las nuevas técnicas de conversión basadas en aprendizaje profundo.

- Conseguir un sistema que permita evaluar las técnicas desarrolladas y los diferentes experimentos a realizar con las voces alaríngeas: desarrollo de un sistema de reconocimiento de voz de habla continua en castellano.

1.3 Estructura del documento

Tras esta breve introducción, el segundo capítulo está dedicado al estudio de los trabajos realizados con las voces alaríngeas. Este repaso a la literatura comienza revisando los distintos tipos de análisis que se han hecho para tratar de caracterizar y clasificar este tipo de voces. Después se profundiza sobre el comportamiento que tienen los reconocedores automáticos del habla al ser enfrentados a estas voces. Se exploran también los estudios dedicados a mejorar la inteligibilidad y naturalidad de las voces alaríngeas. El último punto de este capítulo se centra en repasar el material existente para poder realizar estos estudios y su disponibilidad.

El capítulo 3 describe con detalle la base de datos paralela de habla esofágica que se ha grabado para la realización de esta tesis. Esta base de datos es una de las aportaciones hechas por este trabajo, así que su obtención y análisis se expone con minuciosidad. En un primer momento se explica cuál es el diseño del corpus, así como las características del proceso de grabación llevado a cabo y de los locutores esofágicos que participaron en él. Posteriormente se detalla la composición del material grabado y el contenido de las transcripciones. Se aborda a continuación el proceso seguido para conseguir etiquetar las grabaciones realizadas. Para terminar este capítulo, se caracteriza acústicamente la base de datos en base a parámetros como la frecuencia fundamental, el jitter, el shimmer o la extracción de los formantes. También se hace un estudio de la duración de los diferentes sonidos grabados.

El capítulo 4 se centra en la descripción del sistema de reconocimiento de habla automático para el castellano desarrollado. En primer lugar, se explica cómo se han construido las diferentes partes que forman el reconocedor (modelos acústicos, diccionario, modelo de lenguaje). El reconocedor desarrollado fue evaluado en un *challenge*, y en este capítulo se presentan los resultados obtenidos por el reconocedor, que medirán su comportamiento con habla sana. Después se describe la adaptación del sistema para la evaluación de las grabaciones esofágicas. Al final del capítulo, se incluye también un estudio de la inteligibilidad de las voces esofágicas utilizando tanto reconocimiento automático como reconocimiento humano del habla.

En el capitulo 5 se abordan las técnicas de conversión estadística aplicadas a voces sanas. En primer lugar se explica en qué consiste la conversión utilizando

GMMs, explorándose cuatro técnicas diferentes. También se evalúa el funcionamiento de estas técnicas aplicadas a PMA, un método de habla silenciosa. Después se explora el uso de redes neuronales para llevar a cabo la conversión estadística, más concretamente el uso de LSTMs. Se describe el sistema construido basándose en esta técnica que se presentó al segundo Voice Conversion Challenge, analizándose los resultados obtenidos. Para finalizar el capítulo, se compara el sistema presentado con un sistema de conversión basado en GMMs.

El capítulo 6 trata sobre las técnicas de conversión para las voces esofágicas. En un primer lugar se abordan los problemas que presenta el alineamiento entre las voces esofágicas y las sanas. Después, se investigan distintas alternativas para la conversión de frecuencia fundamental de las voces alaríngeas. Seguidamente, se evalúan las técnicas de conversión basadas en GMMs aplicadas a este tipo de voces. Por último, se explora la utilización de redes neuronales basadas en LSTMs y PPGs para convertir de una voz origen esofágica a una voz destino sana.

Por último, en el capítulo 7 se exponen las conclusiones a las que se han llegado tras la realización de los trabajos realizados en esta tesis. Se describen las aportaciones hechas por este trabajo y los trabajos futuros a llevar a cabo. Para finalizar, se recoge la difusión de los resultados obtenidos en esta tesis.

Con el conocimiento se acrecientan
las dudas.

Goethe

CAPÍTULO

2

Estado del Arte

Las tecnologías del habla siempre suelen desarrollarse teniendo en cuenta sólo las voces sanas. Esto es lógico porque intentan dar cobertura al mayor número de hablantes posible. La cantidad de recursos disponibles para investigar la voz patológica es escaso y las grandes empresas suelen obviar la problemática de estas personas en el diseño de sus productos.

Sin embargo, hay investigadores que han dedicado sus esfuerzos para mejorar la calidad de vida de estas personas. En este capítulo presentamos un visión general de los trabajos realizados, clasificándolos en tres grandes grupos: por un lado, estudios que tratan de analizar y caracterizar las voces patológicas; por otro lado aquéllos trabajos orientados al reconocimiento del habla alaríngea; un tercer grupo lo forman las investigaciones que desarrollan técnicas para mejorar la naturalidad y la inteligibilidad de las voces alaríngeas; finalmente se dedica un apartado al análisis de las bases de datos existentes para este tipo de voces.

2.1 Análisis y caracterización de las voces alaríngeas

En la literatura se encuentran referencias a muchos trabajos previos relacionados con las voces alaríngeas. Hay muchos estudios que tratan el impacto en la calidad de vida que tiene la voz sustituta que utilizan las personas sometidas a una laringectomía. En los trabajos descritos en [84], [94], [115] y [98] se les pide a hablantes alaríngeos con voces esofágicas, traqueoesofágicas y de electrolaringe que autoevalúen la calidad de su voz respondiendo a las preguntas del índice de incapacidad vocal (voice handicap index - VHI) [57] o a las del instrumento de medida de calidad de vida relacionado con la voz (voice-related quality of life - V-RQOL) [53]. El VIH o el V-RQOL son herramientas diseñadas para evaluar la influencia que las patologías de la voz han tenido en el paciente en su vida cotidiana. En estos estudios se intenta comprobar las diferencias que los distintos tipos de habla implican, pero los resultados no son muy concluyentes.

Otros estudios se centran en caracterizar las voces alaríngeas. Es el caso de [12], donde se describe las características acústicas de las vocales castellanas producidas por personas laringectomizadas y se comparan con los resultados obtenidos por un grupo de control de hablantes con voces sanas. Los resultados muestran que en general los laringectomizados producen vocales con formantes en frecuencias más altas y duraciones mayores que el grupo de hablantes laríngeos. Otro estudio acústico de las voces alaríngeas en castellano se presenta en [16]. En este trabajo se analiza el Tiempo Máximo de Fonación (TMF), las sílabas por minuto (spm), la intensidad, la frecuencia fundamental (f_0) y los formantes para caracterizar este tipo de voces. Además se hace un análisis prosódico. La conclusión obtenida es que los datos calculados permiten relacionar aceptabilidad perceptiva con los datos objetivos de los parámetros acústicos. El mismo enfoque para el polaco se hace en [124]. Los parámetros que analizan para caracterizar las voces esofágicas son la desviación estándar y variación de la frecuencia fundamental, el jitter y el shimmer.

Ciertos trabajos intentan clasificar de alguna manera las voces alaríngeas. En el estudio desarrollado en [119] se expone un sistema de tipificación de las voces esofágicas basado en la inspección del espectrograma de banda estrecha de las señales de audio. También se analizan diversos parámetros para evaluar la calidad de la

voz traqueoesofágica: mediana y desviación estándar de la frecuencia fundamental, jitter, porcentaje de sonoridad (% voiced), relación entre armónicos y ruido (harmonics-to-noise ratio HNR), relación de excitación glotal a ruido (glottal-to-noise excitation GNE) y diferencia de energía de banda (band energy difference BED). Con estas medidas se propone clasificar la calidad de las voces en 4 tipos. Otro estudio que trata de clasificar las voces traqueoesofágicas es [30]. En este artículo se propone utilizar una escala dedicada para evaluar el habla alaríngea de manera automática llamada A4S. Esta escala se compone de cinco dimensiones normalizadas, relacionadas con la periodicidad, la regularidad, el ruido de alta frecuencia, el gorgoteo/ronquera de la voz, así como la frecuencia de habla.

También hay diversos estudios que tratan de predecir los resultados de una evaluación subjetiva de las voces esofágicas. En [54] se propone extraer una serie de parámetros de las voces TE para evaluar su calidad y relacionarlos de manera automática con un MOS (mean opinion score) subjetivo. Los parámetros propuestos son los provenientes de un análisis adaptativo temporal-frecuencial, los propuestos por los estándares de evaluación de calidad de voz ITU-T P.563 y ANIQUE+ y también se sugiere utilizar el reverberation-to-signal modulation energy ratio (RSMR). En [77] y [78] se utilizan dos métodos distintos para predecir la evaluación de la voz traqueoesofágica por parte de oyentes inexpertos: medidas espectrales y de predicción lineal y la creación de un modelo perceptual auditivo que intenta replicar la percepción de un oyente estándar. La conclusión alcanzada es que los modelos psicoacústicos funcionan mejor para predecir las evaluaciones de calidad.

2.2 Reconocimiento automático de voces alaríngeas

Hay distintos trabajos centrados en ver como afectan las voces alaríngeas a los reconocedores automáticos del habla. Las diferencias que existen entre estas voces y las voces sanas utilizadas para entrenar y probar los reconocedores hace que este tipo de voces sean un reto para las tecnologías de voz. Las voces de los laringectomizados no obtienen buenas tasas de reconocimiento al ser evaluadas por un ASR.

Estas voces son muy diferentes entre sí en términos de calidad y de inteligibilidad y su evaluación la hacen profesionales entrenados acostumbrados a los problemas que presentan este tipo de voces. Esto supone un problema porque se necesita personal cualificado para poder clasificar a los hablantes alaríngeos.

Surgen entonces estudios como los descritos en [101] y [102] en los que se usa un reconocedor entrenado con voces no patológicas para calcular la tasa de reconocimiento de palabras de locutores TE. Estos resultados intentan relacionarse con la puntuación de inteligibilidad otorgada por un grupo de expertos a estos mismos locutores. Los resultados obtenidos hacen pensar que la tasa de reconocimiento puede ser un buen indicador de la inteligibilidad de un locutor TE.

Otro sistema que propone evaluar automáticamente la voz, pero esta vez ampliada a diferentes desordenes, es el descrito en [72]. Además de utilizar un ASR para calcular la tasa de reconocimiento, se utiliza un módulo prosódico para extraer otros parámetros de las voces tanto a nivel de palabra (21 parámetros distintos) como de frase (16 parámetros). Con estos parámetros se entrena un SVR (support vector regression) para predecir las evaluaciones de inteligibilidad de un grupo de expertos sobre las frases de entrenamiento. Los resultados obtenidos en este trabajo muestran una correlación del 0.90 para la evaluación de las voces TE y del 0.87 para la de las voces de niños con labio y paladar hendido.

Hay diversos intentos de mejorar la tasa de reconocimiento para las voces esofágicas. En [47], el enfoque seguido es adaptar el reconocedor entrenado con voces no patológicas a los hablantes traqueoesofágicos. La adaptación se hace para cada locutor TE mediante interpolación de HMMs no supervisada. También se calcula la correlación entre la tasa de acierto a nivel de palabra de estos reconocedores adaptados y las puntuaciones de inteligibilidad otorgadas a cada locutor por un grupo de

expertos, llegándose a la conclusión de que existe una fuerte relación entre ambos resultados.

Algo parecido se hace en [42] pero para voz de electrolaringe. En este artículo se utiliza un ASR diseñado para voz sana para reconocer la voz sustituta y se evalúan los resultados aplicando diferentes tipos de adaptación: adaptación MLLR (Maximum Likelihood Linear Regression) dependiente del locutor y adaptación MLLR de dominio, basada en volver a entrenar el modelo utilizando nuevos datos del dominio de destino. También se aplican estrategias de reducción del ruido radiado por la electrolaringe (sustracción espectral y filtrado de modulación). Los resultados de esta investigación indican que la inclusión de audio de electrolaringe en el entrenamiento del reconocedor hace que los resultados del reconocimiento mejoren muchísimo.

El enfoque presentado en [65] es diferente. En este caso se propone aplicar un algoritmo de conversión de voz basado en GMMs para mejorar el reconocimiento de la voz esofágica. La voz destino a la que convertir se trata de una voz sana. Con este procedimiento se consigue una mejora del 3.40 % en la tasa de reconocimientos de fonemas respecto a la voz alaríngea sin convertir.

Otras aproximaciones son más sencillas. Por ejemplo, en [73] se construye un clasificador de vocales para la voz esofágica en castellano. Este clasificador está basado en HMMs y se entrena con MFCCs, coeficientes LPC y valores de formantes de los segmentos vocálicos de las voces esofágicas. Este es un primer paso para el trabajo desarrollado por los mismos autores en [74]. El sistema identifica los segmentos sonoros de voz alaríngea usando técnicas de reconocimiento de comandos aislados de voz y las reemplaza por los segmentos equivalentes de voz sana almacenados en un codebook, con lo que consiguen mejorar la inteligibilidad percibida.

Los estudios presentados en [87] y [88] también se centran en el reconocimiento de vocales de las voces esofágicas, pero además de la voz utilizan también imágenes de vídeo. Como parámetros acústicos se utilizan MFCCs y de la señal de vídeo se extrae el contorno labial. Para el reconocimiento de vocales se entrenan tres clasificadores: una red neuronal, Support Vector Machines y un clasificador Bayesiano ingenuo, aunque los mejores resultados se obtuvieron para la SVM: 75 % para el reconocimiento acústico y el 40 % para el visual.

Se ve que el problema del reconocimiento de las voces esofágicas es un tema complejo al que todavía no se ha conseguido dar respuesta ya que las soluciones más prometedoras consisten en modificar los modelos acústicos de los reconocedores ya entrenados, y esto no es siempre posible.

2.3 Mejora de las voces alaríngeas

Además del análisis y la evaluación, hay muchos trabajos cuyo objetivo es mejorar las voces alaríngeas, ya sea su naturalidad o su inteligibilidad. Muchos de estos trabajos se basan en técnicas de procesado de señal para conseguir estas mejoras.

Un ejemplo se presenta en [3] y [4], donde se trata de mejorar el habla de electrolaringe. Para ello, primero se usa un filtro adaptativo para limpiar el ruido de fondo intrínseco a la electrolaringe. Después, se identifican los segmentos sonoros y se sustituyen por otros equivalentes de voz sana almacenados en un codebook. La elección del segmento sustituto se hace utilizando un clasificador construido con una red neuronal. El trabajo se valida mediante una evaluación subjetiva que muestra una mejora respecto a las señales EL originales.

Otros trabajos que se basan en técnicas de procesado de señal para mejorar las voces alaríngeas son los descritos en [19]. En este artículo se propone remplazar el pulso glotal de los hablantes traqueoesofágicos por uno sintético para reducir el jitter y el shimmer del locutor. Además, se emplea un suavizado espectral y una corrección de inclinación para que la envolvente de las voces TE se parezca más a la de las voces sanas. Los mismos autores en [20] tratan de mejorar las voces traqueoesofágicas modificando las duraciones de la prosodia de estos hablantes. Algo similar se propone en [122], donde se manipulan tres aspectos problemáticos de la voz TE que afectan a la inteligibilidad de la voz percibida (estabilidad de amplitud, estabilidad de pitch y espectro de fuente mediante la forma del periodo de pitch).

El enfoque seguido en [55] opta por mejorar la voz esofágica utilizando un ecualizador de ganancia adaptativa (AGE - adaptive gain equalizer) para modificar la fuente sonora. Para estimar la fuente sonora se utiliza filtrado inverso adaptativo iterativo (IAIF - Iterative Adaptive Inverse Filtering), consiguiéndose una mejora en la relación armónico a ruido (HNR - Harmonic Noise Ratio).

Otro enfoque distinto para mejorar las voces esofágicas es utilizar técnicas de conversión de voz (VC - voice conversion). En la VC, durante la fase de entrenamiento se entrena una transformación para pasar de una voz origen a una voz destino. Para ello se parte de una base de datos paralela de la voz origen y la voz destino y se aprende de una manera estadística las relaciones entre ambas voces.

Con estos conocimientos se define una función de conversión que permite convertir de una voz a otra (figura 2.1).

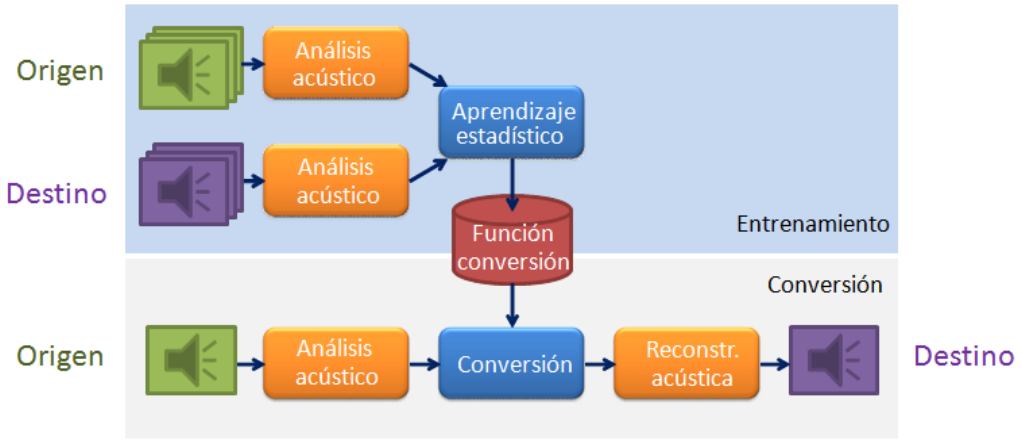

Figura 2.1: Esquema del proceso de conversión de voz.

Aunque la identidad del locutor también está contenida en las características supra-segmentales (prosodia) e incluso en las lingüísticas, la investigación en VC se ha centrado principalmente en el mapeo de las características espectrales [15, 31, 116].

El campo de la VC se investiga desde hace tiempo, una review extensa describiendo las distintas aproximaciones al problema puede leerse en [82]. A la hora de obtener la función de conversión se ha propuesto una gran variedad de enfoques diferentes: desde usar codebooks [2, 6] y modelos ocultos de Markov (HMM - Hidden Markov Models) [8, 66, 128] a mezclas de gaussianas (GMM - Gaussian Mixture Models) [7, 49, 58, 107, 116, 126] o procesos gaussianos [125]. En los últimos años, las soluciones basadas en redes neuronales han tomado gran relevancia [15, 22, 68, 81, 85, 108].

En el caso de utilizar la VC para mejorar el habla esofágica, se utiliza como origen al locutor esofágico y como destino una voz sana. Este es el caso del trabajo desarrollado en [24] y [27]. En él se prepara un sistema de conversión basado en modelos de mezcla de Gaussianas (GMM - Gaussian mixture models) con un algoritmo de máxima verosimilitud. Una evolución de este trabajo se presenta en

[26], [25], [28] y en [23]. Para poder dar cierta personalización de la voz convertida para los sujetos laringectomizados, el sistema de conversión estadístico basado en GMMs se cambia a un sistema de conversion denominado Eigenvoice (EVC - Eigenvoice conversion) y que permite una conversión uno-a-muchos sólo con el ajuste de unos parámetros. De todos moods, como la VC va a utilizarse en esta tesis, una descripción más profunda de estas técnicas se incluye en el capítulo 5.

Los métodos de conversión de voz aplicado a las voces alaríngeas han mostrado una mejora de la inteligibilidad y de la naturalidad. Sin embargo la literatura se centra en la conversión de la componente espectral, dejándose más desatendida la prosodia, lo que hace pensar que existe margen de mejora. Es por ello que esta tesis se centra en la conversión de voz como método de mejora de las voces alaríngeas, con la esperanza de conseguir una mejora del reconocimiento de este tipo de voces para los sistemas automáticos.

2.4 Bases de datos de voces patológicas

Para el desarrollo de las investigaciones descritas en los apartados anteriores se han grabado muy diversas bases de datos, en general adaptadas a las necesidades particulares de la investigación que se desea realizar. Muchas de estas bases de datos recogen diferentes tipos de patologías y se puede decir que pocas de ellas están enfocadas en las voces alaríngeas y aún menos en las voces esofágicas.

Aunque el trabajo de este tesis está enfocado en las voces alaríngeas, en este apartado presentamos una descripción de bases de datos de voces patológicas en general con objeto de mejorar el diseño y el desarrollo de nuestra propia base de datos.

Una base de datos muy citada al trabajar con voces patológicas ha sido la MEEI [38] (Massachusetts Eye and Ear Infirmary) que forma parte de la "Disordered Voice Database and Program" de la empresa KayPENTAX. Esta base de datos está en inglés y contiene dos muestras de habla de 53 locutores de habla sana y de 657 de habla con habla patológica de diferentes orígenes (orgánicos, neurálgicos, traumáticos y psicogénicos). Incluye los primeros 12 segundos de cada locutor enunciando un texto estándar, "The Rainbow Passage" [39] y también una fonación sostenida del fonema /a/ (como en la palabra inglesa "father"). Esta base de datos principalmente se comercializa como un recurso de entrenamiento para que los profesionales clínicos se familiaricen con la interpretación de los parámetros calculados por el software MDVP de KAYELEMETRICS. Sin embargo, ha sido ampliamente usada en el desarrollo y evaluación del funcionamiento de nuevos algoritmos, principalmente por su gran tamaño y su disponibilidad.

Otra base de datos es la SVD (Saarbruecken Voice Database) [90], realizada por el Instituto de Fonética de la Universidad de Saarland. Contiene grabaciones de las vocales /a/, /i/ y /u/ y de una frase en alemán hechas por más de 2000 personas. En la interfaz web en la que se presenta esta base de datos de uso libre se pueden elegir entre unas 70 patologías que van desde laringitis o disfonía a cordectomía o tumor de laringe.

También existen otras bases de datos más pequeñas grabadas por la misma época. Por ejemplo, está la base de datos Nemours [80], compuesta de 11 locutores

disártricos masculinos, o la base de datos Whitaker [21], compuesta de habla inglesa estadounidense de pacientes con parálisis cerebral.

Otro ejemplo muy usado es la base de datos conocida comúnmente como "Universal Access" [61], centrada en locutores de habla disártrica estadounidenses con parálisis cerebral. Esta base de datos se ha demostrado muy útil para muchos investigadores porque incluye grabaciones simultáneas de 8 micrófonos distintos y una videocámara, permitiendo el estudio de parámetros no acústicos y el efecto de algunas variaciones en el entorno acústico.

Otro par de bases de datos de voces patológicas son la TORGO [95] y la NKI CCRT Speech Corpus [121]. La primera contiene habla disártrica perteneciente a 8 pacientes (5 hombres y 3 mujeres) con parálisis cerebral o esclerosis lateral amiotrófica, y habla normal de 7 personas (4 hombres y 3 mujeres) como grupo de control. La segunda base de datos contiene audio a nivel de frase y su puntuación de inteligibilidad perceptual. Las grabaciones consisten en la lectura de 17 frases en holandés por parte de 55 pacientes de cáncer de cabeza y cuello. El audio está recogido en tres estados del tratamiento de quimio-radiación de los pacientes (CRT - Chemo-Radiation Treatment): antes del CRT, 10 semanas después y 12 meses después del CRT. La puntuación de inteligibilidad es la EWE (evaluator weighted estimator) de cada frase, que se calcula a partir de la evaluación de profesionales. EWE es la media ponderada de las puntuaciones de múltiples evaluadores donde el peso es el coeficiente de correlación entre la puntuación de un evaluador y la media no ponderada de todos los evaluadores. La evaluación la llevaron a cabo 13 patólogos del habla nativos holandeses.

Para el castellano existe la base de datos en Alborada-I3A [99], compuesta por más de 2 horas de voz de 14 locutores jóvenes (de 11 a 21 años) con distintos problemas (síndrome de Down, desorden cognitivo, desorden de crecimiento, hiperactividad, trastorno de privación, tetraplegia, ataxia motriz, polimarfolmaciones, parálisis cerebral, encefalopatía, discapacidad expresiva) y cerca de 9 horas de 232 locutores de la misma edad sanos.

Estas son las bases de datos de voces patológicas disponibles públicamente más extendidas, pero ninguna de ellas se centra específicamente en el tipo de habla alaríngea en el caso que llegue a contener material perteneciente a esta patología. Ante esta escasez de material con el que trabajar, en la mayoría de los casos los grupos

de investigación en habla alaríngea graban específicamente para cada estudio y las grabaciones obtenidas no son difundidas después.

Para el trabajo descrito en [65] se grabó una base de datos en francés denominada FPSD (French Pathological Speech Database) para mejorar el comportamiento de un reconocedor. Contiene 480 archivos de audio de voz esofágica acompañados por sus correspondientes transcripciones ortográficas. Las frases están pronunciadas por un único locutor laringectomizado y están compuestas por palabras de una sola sílaba, palabras de una y dos sílabas o palabras de tres sílabas, además de frases con entonaciones ascendente y descendente.

Otro ejemplo es la base de datos paralela ELHE grabada en alemán de Austria y que se utiliza en [42] para estudiar como se comporta un ASR diseñado para habla normal sana cuando se le aplica voz producida por una electro-laringe. Esta base de datos está compuesta por 500 frases diferentes. Cada frase se graba una vez por un locutor sano con habla normal y otra por el mismo locutor sano con la electro-laringe para poder ver las diferencias entre ambos tipos de habla. En total se compone de 7 locutores.

De todos modos, realizar grabaciones tan exhaustivas no es lo habitual. Lo normal es que varios locutores graben unas pocas frases que se ajusten a los experimentos a realizar. Hay multitud de ejemplos, como el descrito en [122], dónde se utilizan las grabaciones de 16 locutores TE masculinos de una lectura en alto de una historia corta en holandés. En total utilizaron 30 grabaciones de 16 pacientes para generar los estímulos de este experimento.

Para el estudio acústico de la voz traqueoesofágica llevado a cabo en [30], 36 pacientes sometidos a una laringectomía total y con una punción TE leyeron un texto fonéticamente equilibrado de 10 frases (cuya duración es de aproximadamente 40 segundos para un locutor sano).

De manera similar, el material utilizado en [47], [101], [102] y [72] consiste en las grabaciones hechas a 18 laringectomizados con voz traqueoesofágica. Las grabaciones son en alemán y cada locutor lee un texto fonéticamente equilibrado utilizado habitualmente en terapias del habla en los países de habla alemana y que consiste en 108 palabras (71 distintas). La duración de los 18 archivos de audio hace un total de 21 minutos. Estos trabajos están orientados al diagnóstico y evaluación de la voz traqueoesofágica.

También se utilizan 30 frases de voz de electro-laringe en español grabadas por pacientes laringectomizados mexicanos en los trabajos descritos en [3] y [4] con objeto de mejorar la calidad de la voz generada mediante electrolaringe.

El mismo tipo de material se utiliza en [54]. El audio recopilado consiste en grabaciones de 28 locutores traqueoesofágicos masculinos en las que cada locutor lee una frase en inglés. Este material, ampliado a 35 locutores se utiliza en trabajos de los mismos autores ([77] y [78]) para realizar estudios de diagnóstico y evaluación.

Otra práctica habitual es centrar el estudio y el análisis solamente a vocales. Es el caso de [119], donde las grabaciones consistieron en tres vocales /a/ sostenidas en un pitch cómodo y un tono de lectura en voz alta estándar de 40 pacientes laringectomizados holandeses con habla traqueoesofágica. El mismo caso para el castellano se explica en [55], trabajo para el cual se grabó a 6 pacientes con voz ES de un centro de rehabilitación 20 realizaciones de cada vocal de este idioma.

En otros enfoques, como el reconocimiento de vocales polacas aisladas mostrado en [87] o en [88], además de las grabaciones de audio se utilizaron también grabaciones de las caras de los locutores para extraer sus contornos labiales. 10 pacientes (4 esofágicos , 1 traqueoesofágico y 5 de habla pseudo-susurrada) pronunciaron 13 veces las 6 vocales polacas aisladas (/a/, /i/, /e/, /y/, /o/, /u/).

También se encuentran casos dónde se graban palabras aisladas. En [12], se hicieron grabaciones de 20 pacientes, de cada uno de los cuales se grabaron 24 palabras en castellano.

En muchas otras investigaciones, el tipo de audios grabados no se limita a frases o vocales, sino que se recopilan tanto frases como vocales y palabras aisladas. Un ejemplo de este tipo de material se recoge en [124]. Los autores grabaron a 33 pacientes polacos que sufrieron una laringectomía total. El grupo de estudio consistió en 31 hombres y 2 mujeres, con una edad media de 61 años y presentaban distintos grados de voz esofágica (muy buena, buena, suficiente). El material grabado consistió en vocales, palabras que contenían vocales, y una frase de test.

Otro ejemplo se puede encontrar en [16] o en [73]. En el primer artículo, el material utilizado consiste en en grabaciones sostenidas de la vocal /a/ y la lectura de un texto diseñado ad hoc con una oración declarativa y otra interrogativa hechas

por 10 hombres que habían sido sometidos a una laringectomía total pertenecientes a la Asociación Sevillana de Laringectomizados. En el segundo, utilizan 250 archivos de voz normal y 360 de voz esofágica en español que incluyen vocales aisladas, palabras aisladas y frases. La voz de los locutores esofágicos se recogió en el "Instituto de Rehabilitación Nacional" de México.

Por terminar, a veces además de vocales, palabras y frases se recogen otro tipo de materiales. Es el caso de [19] o [20]. Estos artículos utilizan como material las grabaciones de 13 pacientes traqueoesofágicos de habla inglesa. Las grabaciones consistieron en vocales sostenidas y frases a un nivel cómodo de pitch y de volumen. También recogieron las electroglotografías (EGG) al hacer las grabaciones.

Este apartado no pretende ser un análisis exhaustivo de los materiales que se usan en el campo de la investigación de las voces esofágicas, pero se puede deducir de su lectura que no existe una base de datos estándar a la que se suele recurrir. Lo más habitual a la hora de desarrollar algoritmos (o de realizar análisis) para la voz esofágica es que se entre en colaboración con terapeutas u hospitales que trabajen con pacientes laringectomizados para realizar grabaciones a medida.

Ante la inexistencia de una base de datos para el castellano que pudiese ser utilizada durante el desarrollo de esta tesis, se decidió realizar el esfuerzo de grabar una base de datos paralela de voces esofágicas. La intención final es liberar estas grabaciones con el propósito de que la comunidad investigadora tenga material suficiente para poder facilitar el desarrollo y evaluación de diversas técnicas y algoritmos que tengan como propósito mejorar la calidad de vida de los pacientes laringectomizados.

Tabla 2.1: Contenido de las distintas bases de datos.

Artículo	Área	Idioma	Contenidos			Nº locutores			
			vocales	palabras	frases	ES	TE	EL	HS
Blood (1984)	Análisis	Inglés	+ /a/ sostenida	+ 4	+ 2ª frase de "The Rainow passage"	10	10	-	10
Robbins (1984)	Análisis	Inglés	+ /a/ sostenida	-	+ Párrafo estándar	15	15	-	15
Kinishi & Amatsu (1986)	Análisis	ND	+ /a/ sostenida	-	-	5	20	-	10
Williams & Watson (1987)	Análisis	Inglés	-	+ Nombre + Contar hasta 25 + Palabras aisladas	+ Párrafo + Frases + Describir fotografía + Conversación	12	10	11	10
Debruyne et al (1994)	Análisis	-	+ /a/ sostenida	-	-	12	12	-	-
Miralles & Cervera (1995)	Análisis	Español	-	+ 24	-	10	20	-	10
Ng et al (1997)	Análisis	Chino	-	-	+ Historia (136 caracteres monosi-lábicos)	15	12	15	-
Merol et al (1999)	Análisis	Francés	/a/ sostenida	-	+ 1 frase de referencia	30	29	-	-
Finizia et al (1999)	Análisis	ND	+ /a/ sostenida	-	+ Historia (89 palabras)	-	12	-	10
Arias et al (2000)	Análisis	Español	+ /a/ sostenida	-	-	20	20	-	20
Cervera et al (2001)	Análisis	Español	-	+ 24 (2 sílabas)	-	10	10	-	10
Bellandese et al (2001)	Análisis	Inglés	+ /a/ sostenida	-	+ "The Rainbow passage"	9	7	-	10

ES: Hablante esofágico TE: Hablante traqueoesofágico EL: Habla con electrolaringe HS: Hablante sano ND: No definido

2. ESTADO DEL ARTE

Artículo	Área	Idioma	Contenidos			Nº locutores		
			vocales	palabras	frases	ES	TE	EL	HS
Haderlein et al (2004) Schuster et al (2005) Schuster et al (2006)	ASR	Alemán	-	-	+ Texto "North wind and sun" (108 palabras)	-	18	-	-
Aguilar et al (2004) Aguilar et al (2006)	Mejora	Español	-	-	+ 30 frases	-	-	-	ND
van As-Brooks et al (2006)	Análisis	Holandés	+ /a/ sostenida	-	-	-	40	-	-
Cuenca et al (2006)	Análisis	Español	+ /a/ sostenida	-	+ Frase enunciativa: "Mi calle es algo chica" + Frase interrogativa: "¿Quién puede seguir ese ritmo?"	10	-	-	-
Mantilla et al (2006)	ASR	Español	Sí (ND)	Sí (ND)	Sí (ND)	ND	-	-	ND
del Pozo & Young (2006)	Mejora	Inglés	+ Vocales sostenidas	-	+ Rainbow passage (28 frases) [a]	-	13	-	11
McDonald et al (2008)	Análisis	Inglés			+ 1 frase: "The rainbow is a division of white light into many beautiful colors"	-	35	-	-
Maier et al (2009)	ASR	Alemán	-	-	+ Texto "North wind and sun" (108 palabras)	-	41	-	-
Huang et al (2009)	Análisis	Inglés	-	-	1 frase: "The rainbow is a division of white light into many beautiful colors"	-	28	-	-
Carello & Magnano (2009)	Análisis	Italiano	+ /a/ sostenida	-	-	7	7	-	-

[a] más las señales EGG

Artículo	Área	Idioma	Contenidos			Nº locutores			
			vocales	palabras	frases	ES	TE	EL	HS
Law et al (2009)	Análisis	Chino cantonés	-	-	22 frases de 5 a 15 palabras	7	13	14	-
MacCallum el al (2009)	Análisis	Chino mandarín	+ /a/ sostenida	-	-	10	-	-	10
Van Son et al (2010)	Análisis	Holandés	-	-	+ Historia corta	-	16	-	-
Pietruch & Grzanka (2010)	ASR	Polaco	+ Vocales sostenidas [b]	-	-	4	1	-	5 [c]
Deore et al (2011)	Análisis	Inglés	+ /i/ normal + /i/ sostenida	-	-	-	30	-	30
Fuchs et al (2011)	ASR	Alemán de Austria	-	-	+ 500 frases [d]	-	-	7	7
Yamamoto et al (2012)	Análisis	Japonés	-	-	+ 50 frases	1	-	-	61
Ishaq & Zapirain (2013)	Mejora	Español	+ Vocales sostenidas	-	-	6	-	-	-
Širić et al (2013)	Análisis	Croatian	+ /a/ sostenida	-	+ Adapted paragraph	10	10	-	-
Doi PhD (2013)	Mejora	Japonés	-	-	+ 50 frases	1		2 [e]	41 [f]
Wszołek et al (2014)	Análisis	Polaco	+ Vocales sostenidas	+ 5	+ 1 frase: "/dzis'/ /jest/ /wadna/ /pogoda/"	33	-	-	35

[b] Más vídeo del contorno de los labios [c] Pseudo-susurro [d] Última versión: https://www.spsc.tugraz.at/tools/elhe-austrian-Alemán-parallel-electro-larynx-healthy-speech-corpus [e] 1 EL silencioso [f] 1 locutor sano con prosodia del locutor ES

Artículo	Área	Idioma	Contenidos			Nº locutores			
			vocales	palabras	frases	ES	TE	EL	HS
Drugman et al (2015)	Análisis	ND	-	-	+ 10 frases	-	36	-	-
Lachhab et al (2015)	ASR	Francés	-	-	+ 480 frases + Frases con palabras de 1, 2 y 3 sílabas + Frases con entonación ascendente y descendente	1	-	-	-
Shim et al (2015)	Análisis	Coreano	+ /a/ sostenida	-	-	20	-	-	20
Eadie et al (2016)	Análisis	Inglés	-	-	+ 6 frases de 5, 7, 9, 11, 13 y 15 palabras (60 palabras por locutor)	2	23	11	-
Membiela et al (2016)	Análisis	Español	+ /a/, /e/, /i/ sostenidas	+ 20	+ 9 frases	10	8	-	-
Croseti et al (2017)	Análisis	Italiano	-	+ 20 palabras (2/3 sílabas)	+ 5 frases (4 a 6 palabras)	-	12	-	12

2.5 Conclusiones

Como se ha visto en este capítulo, hay mucha investigación hecha en relación con las voces alaríngeas, pero aún queda mucho por hacer. Muchos de los estudios que tratan de analizar este tipo de voces se centran en vocales sostenidas y buscar cómo caracterizarlas. Se proponen varias medidas, pero las más habituales son la frecuencia fundamental y su desviación estándar así como el Jitter y el Shimmer.

Los trabajos relacionados con el reconocimiento automático del habla para las voces alaríngeas demuestran que es necesario modificar o adaptar los modelos acústicos de los reconocedores si se quiere mejorar la tasa de reconocimiento para estos locutores. Esta es una opción que no es posible siempre, así que no es la solución buscada en esta tesis. Otro de los campos del reconocimiento automático que se investiga relacionado con las voces esofágicas es la posibilidad de utilizar la tasa de reconocimiento como indicador de la inteligibilidad de este tipo de voces. Las conclusiones de estos estudios es que esta medida está correlada con la inteligibilidad y que puede utilizarse para evaluar a los locutores esofágicos.

Para la mejora de las voces alaríngeas se han encontrado distintos enfoques. Los que se basan en técnicas de procesado de señal utilizan filtros para limpiar la señal o cambiar la envolvente espectral o tratan de sustituir los pulsos glotales por pulsos sintéticos. Otros intentos tratan de modificar aspectos concretos de este tipo de voces: duraciones, estabilidad frecuencial y de amplitud... Otro enfoque es el de aplicar las técnicas de conversión de voz a las voces esofágicas. Hay varios trabajos ya realizados en este campo, pero en esta tesis se tratará de investigar estas técnicas en profundidad y conseguir mejorar las voces esofágicas para que sean más fácilmente reconocibles por los sistemas de reconocimiento automático. En los trabajos realizados se ha visto que los esfuerzos de conversión se aplican a las características espectrales, dejándose más de lado el problema de la frecuencia fundamental. Por ello, en esta tesis habrá que prestar un cuidado especial a la f_0.

En relación con las bases de datos de voces patológicas existentes se ha podido comprobar que hay poco material disponible para trabajar con voces esofágicas. La mayor parte de los trabajos analizados utilizan grabaciones propias hechas por los investigadores según las necesidades del estudio que llevan a cabo. Sin embargo, si se quiere desarrollar técnicas de conversión de voz es necesario disponer de una

base de datos de locutores esofágicos paralela. Es por ello que se decidió crear una base de datos de estas características. La descripción detallada de estas grabaciones se hará en el capítulo 3.

Antes que toda otra cosa la preparación es la clave para el éxito.

Alexander Graham Bell

CAPÍTULO

3

Obtención de los datos

Para poder llevar a cabo las diferentes técnicas que se explican en esta tesis, es fundamental disponer de una base de datos con grabaciones de habla esofágica debidamente etiquetada. Además, se necesita que las frases contenidas en esta base sean también enunciadas por un locutor de voz no patológica si se quiere convertir hacia voz sana. En este capítulo se describe el proceso de diseño y de obtención de la base de datos creada, y los trabajos realizados para su etiquetado y caracterización acústica.

3.1 Diseño de la base de datos

El principal objetivo a cumplir con la grabación de esta base de datos es obtener señales que permitan el entrenamiento y prueba de sistemas de conversión de voz. El fin último es el de obtener un sistema que mejore la inteligibilidad de la voz alaríngea. Como se expone en el capítulo 5, uno de los requisitos necesarios para entrenar ciertos sistemas de conversión de voz es disponer de una base de datos con grabaciones paralelas: La misma frase debe ser grabada por el locutor origen y el locutor destino para que durante el entrenamiento el sistema sea capaz de aprender cómo se relacionan las características de ambos locutores.

3.1.1 Corpus

El corpus de texto seleccionado para grabar es el corpus ZureTTS [35]. Este corpus ha sido utilizado en un proyecto para la obtención de un banco de voces, por lo que ya se dispone de grabaciones del mismo, pero de voces sanas. Al grabar las mismas frases que los locutores esofágicos, además de obtener las señales paralelas que permiten el entrenamiento de sistemas de conversión de voz, también se pueden extraer y comparar distintos parámetros que caractericen y distingan a las voces patológicas frente a las sanas.

El corpus ZureTTS utilizado se compone de 100 frases. Se trata de un corpus diseñado para ser fonéticamente equilibrado, es decir, contiene realizaciones de todos los fonemas que componen el idioma (el castellano en nuestro caso). El contenido fonético se muestra en la tabla 3.1.

La grabación de 100 frases con un contenido semántico relativamente complejo puede suponer entre 30 y 40 minutos a un hablante sano. Para un hablante esofágico, esta misma tarea puede resultar un gran esfuerzo. Por ello, las 100 frases a grabar se dividen en tres bloques de 33, 33 y 34 frases respectivamente. Cada bloque de frases está fonéticamente equilibrado, de tal manera que si un locutor se cansa tras grabar sólo el primer bloque y no puede grabar los siguientes o decide que no quiere seguir, lo grabado podrá seguir utilizándose.

Además de las 100 frases, cada hablante de voz esofágica grabó 4 realizaciones de vocales sostenidas (las 5 del castellano). Para tener además apariciones de vocales coarticuladas, se grabaron también cuatro palabras que contienen las 5 vocales

Tabla 3.1: Número de fonemas que contienen las 100 frases del corpus.

Fonema	Num apariciones	Fonema	Num apariciones
e	741	b	159
a	715	m	142
o	549	T	112
i	489	p	108
n	401	g	81
s	337	f	61
u	294	rr	58
d	280	x	46
l	279	L	39
r	267	tS	25
t	252	jj	23
k	195	J	17

del castellano. Por último, se incluyeron en la grabación una serie de 10 palabras aisladas que apareciesen en las frases del corpus ZureTTS para posibilitar la evaluación en tareas de STD (Spoken Term Detection).

El texto mostrado a los locutores durante la grabación se recoge en el anexo A.

3.1.2 Condiciones de la grabación

Las grabaciones se han llevado a cabo en la sala de grabaciones del grupo Aholab para que la calidad sea la mejor posible y el impacto del ruido ambiental mínimo. Se han utilizado 4 micrófonos distintos (estudio -Neumann TLM 103-, instrumentación -Behringer ECM8000-, de diadema -DPA 4066-F- y capacitivo -AKG C542BL- respectivamente) de forma simultánea, conectados a un PC (Dell Latitude E4200) utilizando una tarjeta de adquisición de audio (Fireface 400) mediante un cable Firewire, de tal manera que de cada frase grabada se obtienen 4 canales distintos (figura 3.1). Mediante un software específico [1] se va mostrando al locutor la frase o palabra a grabar.

[1] Nanny Record, cedido por la UPC

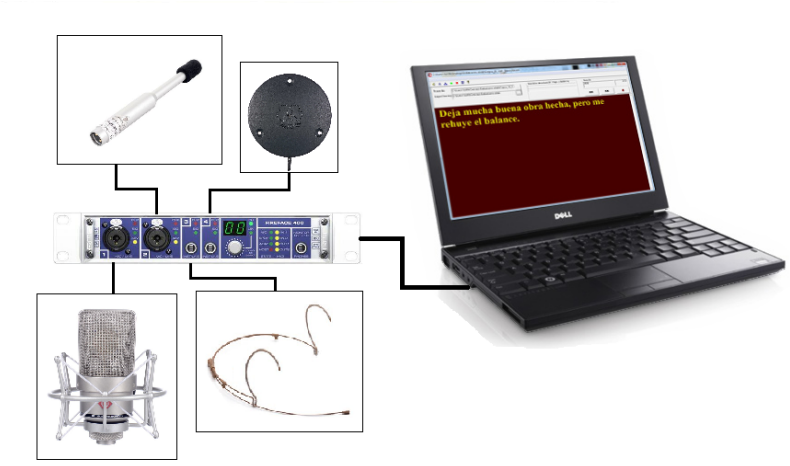

Figura 3.1: Montaje para la adquisición de las grabaciones

3.2 Captación de las personas

Para encontrar locutores dispuestos a hacer la grabación, se habló con la asociación de laringectomizados de Bizkaia (ASLABI). Se llevó a cabo una charla en la que se expuso la investigación en la que se está trabajando y se pidió su colaboración. La asociación animó a sus socios a colaborar y ayudó a coordinar los horarios con la distintas personas laringectomizadas. En la selección de las personas colaboradoras se tuvo en cuenta su estadio en el proceso de aprendizaje de habla esofágica, eligiendo fundamentalmente personas que han alcanzado ya el máximo en su proceso de aprendizaje. Así, la mayor parte de los participantes habla sin dificultad. Sin embargo, se seleccionaron también dos participantes que se encontraban en su fase inicial de aprendizaje, con objeto de obtener datos de las diferentes fases de este difícil proceso. Estas personas realizaron 3 sesiones de grabaciones en las fases intermedia y final de su proceso de aprendizaje.

Como la obtención y utilización de grabaciones es información sensible, se trabajó con la comisión de ética de la UPV/EHU para seguir los protocolos y permisos necesarios en términos de anonimización y protección de datos. Con este fin, se diseñó un formulario de consentimiento informado que se dio a leer a cada locutor antes de grabar su voz en el que de manera voluntaria cedieron sus datos para el desarrollo de la investigación. Los formularios a firmar por los locutores voluntarios se adjuntan en el anexo B.

3.3 Características de los participantes

En total, la base de datos está compuesta por 35 sesiones diferentes grabadas por 32 locutores distintos:

- 28 hombres y 4 mujeres.
- 30 locutores de habla esofágica en fase final, 29 de ellos en castellano y 1 en euskera.
- 4 locutores (castellano) que realizaron grabaciones en fase intermedia. 2 de ellos no realizaron la fase final.
- 1 locutor (masculino) de habla traqueoesfágica (incluido también en el grupo de habla esofágica) (en castellano).

En total, 30 de las sesiones pertenecen a locutores en la fase final del aprendizaje, una es de habla traqueoesofágica, y las cuatro restantes a locutores en la fase intermedia. Cabe decir que todas las sesiones están grabadas en castellano salvo una de las sesiones en fase final de aprendizaje, que contiene las 100 frases de ZureTTS en euskera.

La edad media de los locutores está en los 65 años y 4 meses, pero con una variabilidad muy alta. El más joven tenía 51 años y 4 meses a la hora de realizar la grabación, y el mayor, 82 años y 5 meses.

La variabilidad del tiempo pasado desde que se les realizó la laringectomía también es muy grande. La media es de 5 años y 8 meses, pero hubo varios locutores que se habían sometido a la operación sólo 10 u 11 meses antes de efectuar las grabaciones. Sin embargo, para un locutor concreto, este tiempo se extiende hasta los 32 años y 7 meses. En un caso, el locutor no dijo en qué fecha había sido operado.

Para identificar cada sesión se utiliza la siguiente nomenclatura:

- Dos números para identificar la identidad del locutor
- Una letra, para indicar si el locutor es hombre (M) o mujer (F).
- Un tercer carácter, que será “2” o “3” si el locutor está en la segunda o tercera fase del aprendizaje respectivamente, o una “T” si es un hablante traqueoesofágico.

La mayoría de las sesiones corresponden a hombres en la tercera fase del aprendizaje. Por tanto, y por razones de claridad, en la tabla 3.2 solo se indican las sesiones que difieren de estos parámetros.

Tabla 3.2: Sesiones que difieren de "locutor masculino en 3.ª fase de aprendizaje".

	Identificador de sesión
Voces femeninas	11F3, 15F2, 15F3, 25F3, 28F3
Hablante traqueoesofágico	09MT
Segunda fase de aprendizaje	13M2, 14M2, 15F2, 16M2

El resumen de las características de cada sesión se puede ver en el anexo C.

3.4 Grabación

Para facilitar la tarea posterior de procesado de los datos, es muy importante que las grabaciones obtenidas se correspondan fielmente con texto mostrado. Es muy habitual que en la lectura en voz alta de una frase el locutor omita alguna palabra, sustituya unas palabras por otras o inserte palabras nuevas, haga pausas innecesarias u omita las pausas que están explícitamente indicadas en el texto con signos de puntuación. Es por ello que para la obtención de esta base de datos, la persona responsable de la grabación ha estado en todo momento con el locutor, validando las grabaciones. Además, y con el fin de minimizar repeticiones innecesarias, errores, titubeos y las ambigüedades al pronunciar ciertas frases o palabras, el responsable ha leído primero la frase o palabra en voz alta e inmediatamente el locutor la ha repetido para ser grabada. No obstante, aún con todas estas precauciones se producen errores. Por tanto, para el etiquetado ortográfico final de la base de datos, las transcripciones originales del corpus se han modificado de acuerdo a las anotaciones que el responsable ha tomado durante la sesión de grabación. Los principales problemas que aparecen son la repetición de sílabas (el hablante necesita volver a empezar con una palabra), la desaparición de parte de algunas palabras (el hablante se queda sin aire para terminar la palabra) y la pronunciación de algunos fonemas (incapacidad para pronunciar algunos fonemas).

3.4.1 Material grabado y duraciones

La base de datos está compuesta por 35 sesiones diferentes. La descripción de las sesiones está explicada con detalle en el apartado 3.3. En cuanto al contenido de las sesiones, de las 35 sesiones hay 31 en las que se grabaron las 100 frases. De estas 31, 30 corresponden a locutores en la fase final de aprendizaje y la otra a un locutor en fase intermedia. Respecto a tres de las otras cuatro sesiones dónde no se consiguió grabar todas las frases, dos de ellas están compuestas del primer bloque de frases (33 frases), y la tercera tiene 91 frases. Estas tres sesiones fueron grabadas por locutores en la fase intermedia del aprendizaje. La sesión restante corresponde al locutor traqueoesofágico sin la válvula, y que está compuesta por las 33 frases del primer bloque.

En todas las sesiones se grabaron las vocales sostenidas. Las 14 palabras aisladas (4 palabras con vocales coarticuladas y 10 términos) se grabaron en todas las sesiones excepto en la repetida por el locutor traqueoesofágico sin usar la válvula.

En total se ha grabado unas 9 horas y 50 minutos de audio, aunque la duración de las frases es de 9 horas y 7 minutos. Los términos tienen una duración de 17 minutos y las vocales sostenidas unos 25 minutos. En la tabla 3.3 se puede apreciar un resumen por contenido y duración de cada sesión.

3.4.2 Errores encontrados

Como se ha explicado anteriormente, aunque se intentó controlar al máximo que los locutores leyesen exactamente la transcripción de cada frase, las grabaciones no están exentas de errores. Durante el proceso de grabación se anotaron todas las diferencias entre el texto que aparece en la pantalla y lo realmente dicho por el locutor. Las transcripciones de cada sesión fueron corregidas de acuerdo a estas anotaciones. Cada error cometido se clasificó como uno de estos tres tipos:

- Sustitución: En vez de la palabra transcrita se dice otra. Puede ser debido a una equivocación, la imposibilidad de pronunciar todos los fonemas, o a una pronunciación a medias debido a la falta de aire.

- Inserción: Se introduce una palabra que no está en la transcripción original. La mayor parte de las veces se produce por titubeos o repeticiones al intentar pronunciar bien una palabra, aunque a veces también aparecen por errores de lectura.

- Eliminaciones: Una palabra deja de pronunciarse. La causa más habitual es que el locutor se queda sin aire y la palabra es totalmente inaudible, pero también puede producirse por errores de lectura.

En total se grabaron 3290 frases, en las que en 426 se detectó al menos un error. Esto supone en torno al 13 % de las frases. El número de errores registrados es de 680, más concretamente, 362 sustituciones, 252 inserciones y 66 eliminaciones. En la tabla 3.4 se puede ver de manera más detallada los errores cometidos en cada sesión.

Tabla 3.3: Contenido y duración de cada sesión.

Sesión	Idioma	Fase aprendizaje	Duración total (s)	N.º frases	Duración (s)	Palabras	Duración	Rep. vocales sostenidas	Duración (s)
01M3	ESP	3	994,60	100	898,65	14	31,55	4	64,40
02M3	ESP	3	826,56	100	751,62	14	20,97	4	53,99
03M3	ESP	3	793,48	100	713,65	14	25,95	4	53,88
04M3	ESP	3	981,52	100	927,27	14	25,25	4	29,00
05M3	ESP	3	874,46	100	813,16	14	24,40	4	36,90
06M3	ESP	3	894,85	100	825,90	14	24,75	4	44,20
07M3	ESP	3	1.154,55	100	1.083,30	14	32	4	39,25
08M3	ESP	3	773,44	100	696,64	14	23,95	4	52,85
09M3	ESP	3	347,95	33	300,70	0	0	4	47,25
09MT	ESP	TE	797,10	100	695,35	14	39,15	4	62,60
10M3	ESP	3	1.237,18	100	1.151,23	14	39,95	4	46,00
11F3	ESP	3	1.337,50	100	1.277,55	14	27,05	4	32,90
12M3	ESP	3	830,15	100	762,75	14	24,10	4	43,30
13M2	ESP	2	3.139,75	91	2.990,95	14	55,95	4	92,85
14M2	ESP	2	1.808,10	100	1.692,40	14	52,00	4	63,70
15F2	ESP	2	488,29	33	432,94	14	29,45	4	25,90
15F3	ESP	3	1.241,62	100	1.185,42	14	22,70	5	33,50
16M2	ESP	2	434,60	33	354,75	14	34,85	4	45,00
16M3	ESP	3	1.337,00	100	1.252,45	14	26,95	5	57,60
17M3	ESP	3	775,12	100	712,55	14	31,07	4	31,50
18M3	ESP	3	825,33	100	759,43	14	33,50	4	32,40
19M3	ESP	3	979,71	100	895,81	14	32,25	4	51,65
20M3	ESP	3	1.023,68	100	937,53	14	33,75	4	52,40
21M3	ESP	3	1.320,30	100	1.249,85	14	30,80	4	39,65
22M3	ESP	3	761,47	100	714,42	14	22,00	4	25,05
23M3	ESP	3	956,45	100	889,35	14	21,50	4	45,60
24M3	ESP	3	784,20	100	716,10	14	29,70	4	38,40
25F3	ESP	3	832,85	100	785,10	14	22,05	4	25,70
26M3	ESP	3	599,69	100	552,84	14	22,50	4	24,35
27M3	EUS	3	1.170,96	100	1.087,61	14	36,85	4	46,50
28F3	ESP	3	1.180,75	100	1.094,70	14	33,35	4	52,70
29M3	ESP	3	825,20	100	743,85	14	23,20	4	58,15
30M3	ESP	3	765,22	100	712,62	14	22,30	4	30,30
31M3	ESP	3	1.224,40	100	1.137,15	14	40,50	4	46,75
32M3	ESP	3	1.101,40	100	1.036,55	14	33,20	4	31,65

Tabla 3.4: Desglose de los errores cometidos al grabar para cada sesión.

Sesión	Frases grabadas	Frases con errores	% Frases con errores	Errores totales	Sustituciones	Inserciones	Eliminaciones
01M3	100	1	1	1	1	0	0
02M3	100	0	0	0	0	0	0
03M3	100	5	5	6	4	2	0
04M3	100	4	4	5	3	2	0
05M3	100	1	1	1	0	1	0
06M3	100	4	4	5	1	4	0
07M3	100	10	10	11	6	3	2
08M3	100	15	15	18	7	6	5
09M3	33	3	9,1	4	0	3	1
09MT	100	6	6	7	5	1	1
10M3	100	15	15	22	12	10	0
11F3	100	27	27	43	26	13	4
12M3	100	7	7	9	4	4	1
13M2	91	10	11,0	15	7	1	7
14M2	100	8	8	14	4	9	1
15F2	33	13	39,4	22	5	17	0
15F3	100	18	18	29	14	14	1
16M2	33	3	9,1	3	1	1	1
16M3	100	13	13	13	9	2	2
17M3	100	4	4	7	1	6	0
18M3	100	20	20	32	18	12	2
19M3	100	3	3	9	3	6	0
20M3	100	44	44	81	43	31	7
21M3	100	7	7	8	2	2	4
22M3	100	10	10	13	6	6	1
23M3	100	23	23	33	23	8	2
24M3	100	17	17	34	21	13	0
25F3	100	6	6	8	6	2	0
26M3	100	41	41	72	36	27	9
27M3	100	28	28	44	22	22	0
28F3	100	3	3	3	3	0	0
29M3	100	3	3	3	2	1	0
30M3	100	10	10	13	8	5	0
31M3	100	16	16	17	15	1	1
32M3	100	28	28	75	44	17	14

3.5 Etiquetado

Hasta este momento, la base de datos se reduce a una serie de grabaciones de locutores esofágicos con sus correspondientes transcripciones (con las correcciones que se hayan podido anotar). Al ser el corpus grabado el mismo que ZureTTS, estas grabaciones tienen sus equivalentes sanas, permitiendo que sean usadas en técnicas de conversión de voz. Para esto, y como se verá en el capítulo 5, el primer paso es alinear los segmentos acústicos de las frases de los locutores origen y destino. En el proceso de conversión estándar entre voces sanas, este alineamiento no ofrece especiales dificultades, y se realiza mediante la conocida técnica de proyección temporal dinámica (Dynamic Time Warping - DTW). Sin embargo, a la hora de aplicarse a las frases esofágicas aparecen un conjunto importante de dificultades, debidas precisamente a las grandes diferencias entre las voces origen y destino (figura 3.2). Por tanto, para poder realizar un correcto alineamiento es necesario primeramente segmentar adecuadamente las grabaciones esofágicas. Estas etiquetas servirán posteriormente como punto de partida a la hora de alinear.

3.5.1 Etiquetado automático

Un etiquetado fonético manual de todos los locutores sería lo ideal, pero requeriría demasiado tiempo. La solución debe buscarse entonces utilizando métodos automáticos de segmentación y etiquetado. Con habla sana, esto no supone ningún cuidado especial. Sin embargo, con el habla patológica, aparecen multitud de problemas y desajustes debido a las diferencias entre la voz sana y la de los laringectomizados. A continuación se explican los dos métodos automáticos que se han utilizado para segmentar.

3.5.1.1 Kaldi: alineamiento forzado a partir de modelos de voz sana

La solución propuesta en un principio fue el empleo del sistema de reconocimiento disponible basado en Kaldi [89] entrenado con modelos acústicos de voz sana en castellano (ver el capítulo 4) para aplicar alineamiento forzado a las frases de voz patológica aprovechando que las transcripciones son conocidas. Para el modelo de lenguaje, dado que las frases son conocidas, se ha creado uno específico para

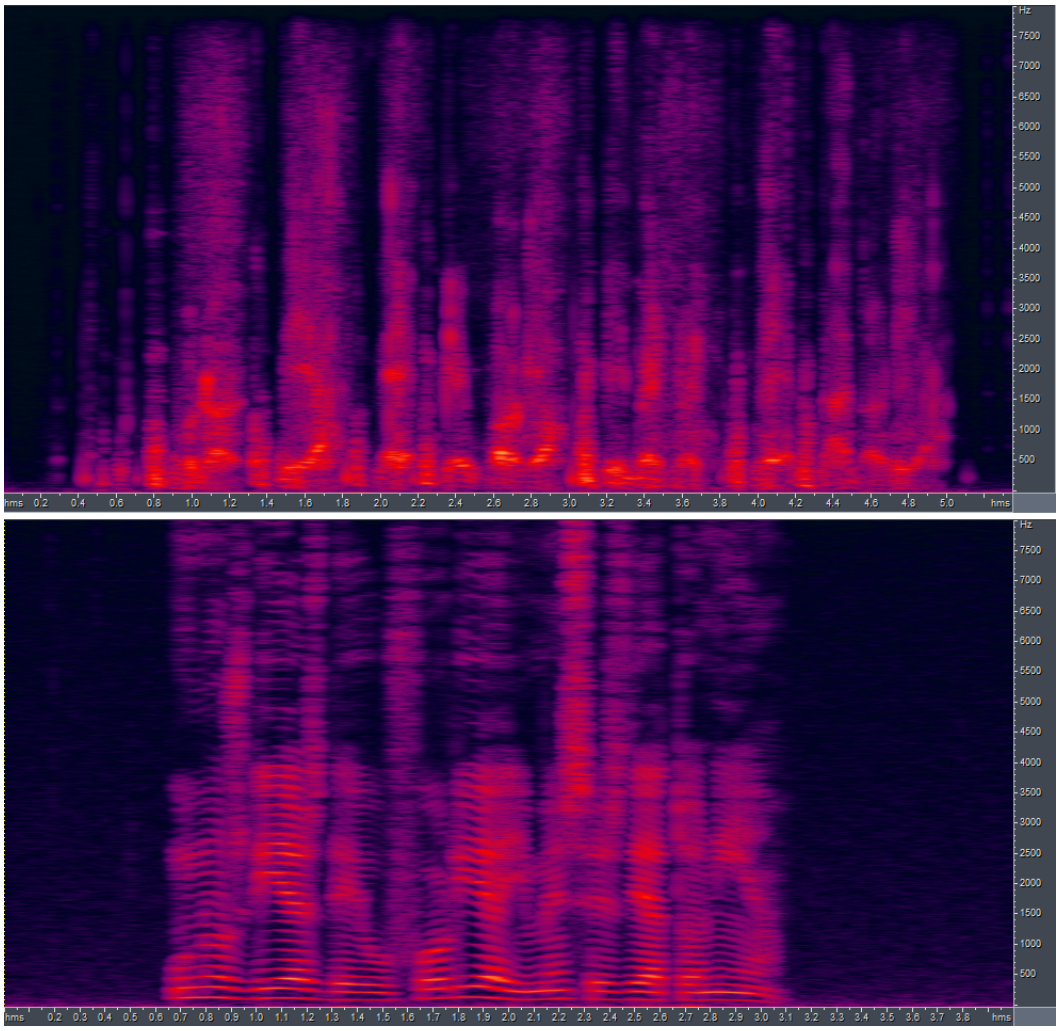

Figura 3.2: Diferencias entre un espectrograma de un locutor esofágico (arriba) y de un locutor de voz sana (abajo). En ambas imágenes la frase enunciada es la misma: *Unos días de euforia y meses de atonía*.

cada frase: se construye un único transductor de estados finito (FST - finite state transducer) a partir de la transcripción de cada frase.

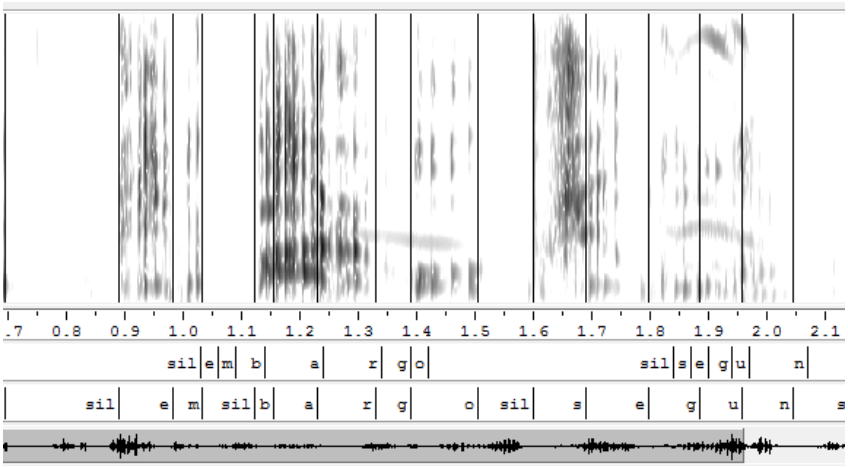

Figura 3.3: Resultado del alineamiento forzado utilizando Kaldi con FSTs.

La figura 3.3 presenta un segmento del resultado de alinear una frase del corpus. La parte superior de la figura muestra el espectrograma de la frase. Las etiquetas de la línea inferior corresponden al etiquetado manual y las de la línea superior al alineamiento automático. Se puede comprobar que los desajustes son grandes. Existen diversas razones que explican estos desajustes:

- Hay una diferencia muy importante entre las duraciones de los sonidos en uno y otro tipo de habla. De esta forma, el reconocedor entrenado con habla sana no es capaz de ajustar su comportamiento a las duraciones excesivas de los sonidos del habla esofágica.

- La velocidad de habla de los laringectomizados es menor que la de los hablantes sanos en parte debido a la constante introducción de pausas en el discurso que los hablantes esofágicos necesitan realizar para tomar aire. Estos silencios no se encuentran sólo entre palabras, sino también dentro de las propias palabras.

De esta forma, el uso del ASR con alineamiento forzado tal y como está no es una buena solución para la segmentación automática de las voces patológicas o para su alineamiento con las voces sanas paralelas.

Para mejorar el resultado, se realizaron modificaciones en los FST permitiendo que las pausas pudieran producirse en cualquier momento de la frase, incluso entre los fonemas de una misma palabra. Por lo tanto, la transcripción usada para el alineamiento debe ser una transcripción fonética y que permita la inserción del modelo acústico del silencio entre cualquier par de fonemas.

Un ejemplo de las mejoras obtenidas con esta estrategia simple puede observarse en la figura 3.4. Se siguen cometiendo fallos debido a las diferencias entre el habla esofágica y los modelos acústicos entrenados a partir de voces sanas, pero muchos errores se reducen.

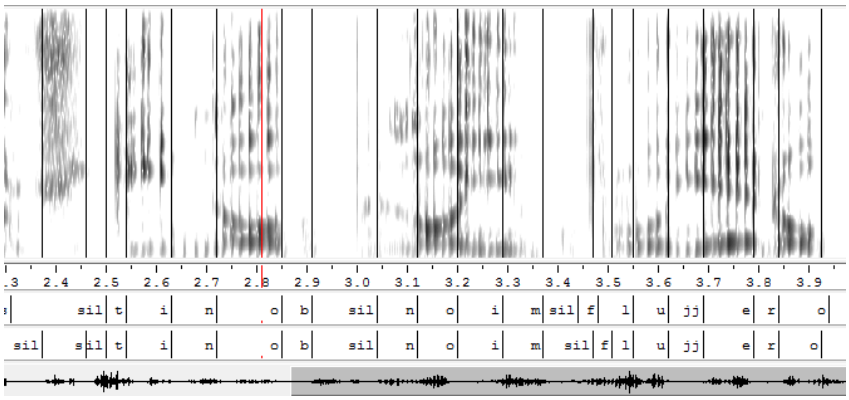

Figura 3.4: Resultado del alineamiento forzado utilizando Kaldi con FSTs permitiendo la inserción de silencios entre cualquier par de fonemas.

3.5.1.2 **Montreal: alineamiento con modelos patológicos**

Otra posible estrategia es entrenar nuevos modelos con el material disponible para después realizar el alineamiento forzado. Para ello se ha utilizado Montreal Forced Aligner [76]. Montreal es una herramienta basada en Kaldi. Para este proceso se utilizaron únicamente las sesiones de habla esofágica que contienen las 100 frases completas con sus correspondientes transcripciones fonéticas.

El resultado se ve en la figura 3.5. Se puede observar que los resultados son bastante buenos, aunque es lógico que se den algunos pequeños desajustes.

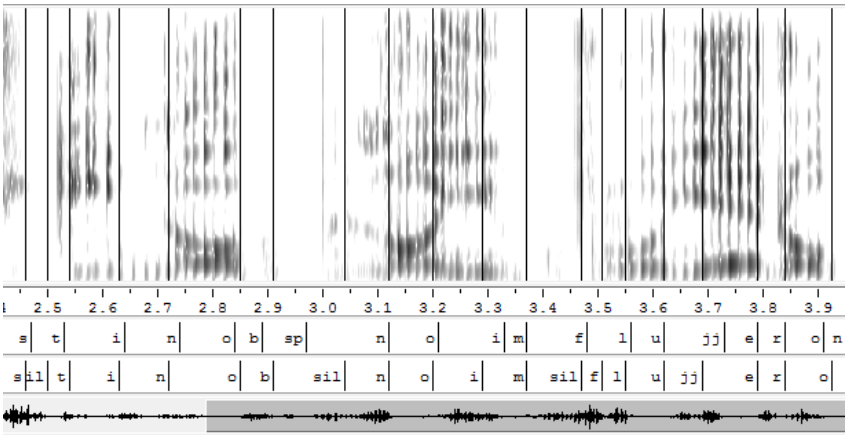

Figura 3.5: Resultado del alineamiento utilizando modelos de voz esofágica.

3.5.2 Evaluación del etiquetado

Para poder evaluar los resultados de estos métodos, es necesario partir de unos etiquetados de referencia. Con este objetivo, se eligieron dos sesiones de locutores esofágicos distintos y se etiquetaron manualmente a nivel de fonema. En una, la sesión 02M3, el locutor presentaba buenas características de locución (pocos errores en la pronunciación de las frases, habilidad para pronunciar todos los sonidos, velocidad por encima de la media, capacidad fonatoria correcta). En la otra, la sesión 05M3, el locutor presenta unas características de locución buenas, pero ligeramente peores que el otro locutor (ritmo de habla más lento, más problemas a la hora de pronunciar algunas palabras, peor control de la respiración).

Una vez se tiene la referencia con la cual comparar, se calculan las diferencias con los etiquetados automáticos. Para ello, las etiquetas correspondientes a los silencios se eliminan tanto del archivo segmentado manualmente como de los conseguidos con los métodos automáticos. De esta manera la correspondencia entre transcripciones será total. Una vez conseguido esto, lo que se hace es calcular el porcentaje de marcas que se alejan como máximo un cierto intervalo de tiempo de las marcas colocadas manualmente. Los resultados se muestran en la tabla 3.5. Cabe aclarar que el número de fonemas de cada sesión es distinto debido a que los fonemas que los locutores no han conseguido pronunciar son diferentes en cada sesión.

Para la sesión 02M3, parece que el alineamiento utilizando los modelos sanos funciona mejor que utilizando los modelos patológicos ya que en el primer caso el 46 % de los errores cometidos es menor de 10 ms, mientras que en el segundo esto sólo ocurre para el 25 % de los errores. Sin embargo, en ambos alineamientos el porcentaje de errores por debajo de los 50 ms es del 78

En la sesión 05M3, pasa lo contrario. Para el alineamiento hecho con modelos sanos el porcentaje de errores pequeños es mucho más pequeño que para el hecho con modelos patológicos: Para este locutor patológico el 83 % de las marcas realizadas con el alineamiento que usa los modelos patológicos presenta un error menor de 5 ms frente al 14 % que presenta la otra alternativa. Además, en el caso de usar los modelos patológicos casi la totalidad de errores cometidos son menores de 50 ms (el 97 %), mientras que usando los modelos sanos este número se queda en el 62 %.

Tales diferencias entre sesiones pueden parecer extrañas, pero si se escuchan las grabaciones de cada sesión puede comprobarse que mientras que la sesión 02M3 es claramente inteligible y con un ritmo de fonación rápido y constante, la sesión 05M3 es más titubeante y lenta, haciéndose más difícil de entender. Por ello, en la primera sesión lo enunciado se parece más a la voz sana, por lo que el alineamiento utilizando modelos sanos funciona muy bien (aunque, de todos modos, el porcentaje de errores que están por debajo de 50 ms es el mismo que al usar los modelos patológicos). En la otra sesión, por el contrario, el tipo de habla se aleja más de los modelos de voz sana. Al alinear con modelos sanos los resultados son claramente peores que al utilizar modelos patológicos puesto que en este último método el parecido con lo que se va a alinear es mayor.

Si se evalúan los errores cometidos en ambas sesiones de manera conjunta se puede comprobar que la utilización de modelos entrenados a partir de voces esofágicas consiguen los mejores resultados. El 50 % de los errores en las marcas son menores de 5 ms, y casi el 88 % es menor de 50ms. Esto significa que el sistema de alineamiento es capaz de detectar con una gran precisión más del 50 % de la posición de los fonemas y que la incidencia de errores grandes es muy pequeña.

Por último, se ha comparado el error medio cometido en cada alineamiento para cada fonema. En este caso, lo que se hace es calcular las diferencias absolutas entre

Tabla 3.5: Porcentaje de errores de etiquetado menores que varios valores de tolerancia (5, 10, 20 y 50 ms) para dos locutores esofágicos.

Sesión	Modelos sanos				Modelos patológicos			
	$<$ 5 ms	$<$ 10 ms	$<$ 20 ms	$<$ 50 ms	$<$ 5 ms	$<$ 10 ms	$<$ 20 ms	$<$ 50 ms
02M3 (5746 marcas)	41.23	45.98	56.89	78.21	17.37	25.70	52.54	78.40
05M3 (5740 marcas)	13.92	15.49	36.41	62.25	83.03	84.43	89.37	97.14
Total (11486 marcas)	27.58	30.74	46.66	70.23	50.18	55.05	70.95	87.77

cada fonema de referencia y el segmentado automáticamente, de tal forma que el error total Err será:

$$Err = \sum_{n=1}^{N} \left(\left| ph_{m_n}^{st} - ph_{a_n}^{st} \right| + \left| ph_{m_n}^{end} - ph_{a_n}^{end} \right| \right) \tag{3.1}$$

donde N es el número total de fonemas de la transcripción, $ph_{m_n}^{st}$ es la marca de tiempo inicial del fonema n en el etiquetado manual, $ph_{a_n}^{st}$ la marca de tiempo inicial del fonema n en el etiquetado automático (modelos sanos o patológicos), $ph_{m_n}^{end}$ es la marca de tiempo final del fonema n en el etiquetado manual y $ph_{a_n}^{end}$ la marca de tiempo final del fonema n en el etiquetado automático.

Para tener una idea del error más comparable, calculamos el error medio cometido por cada fonema, $meanErr$, de la siguiente manera:

$$meanErr = \frac{Err}{N} \tag{3.2}$$

En la figura 3.6 correspondiente a la sesión 02M3 se puede comprobar que el error es muy parecido para cada fonema, independientemente del método, excepto para los fonemas *jj* y *tS*. La figura 3.7 corresponde a la sesión 05M3. En este caso la diferencia del error cometido es mucho mayor utilizando el alineamiento forzado en Kaldi para un número mayor de fonemas. La manera de pronunciar estos fonemas por el locutor esofágico se alejan más del habla sana que el locutor de la primera sesión.

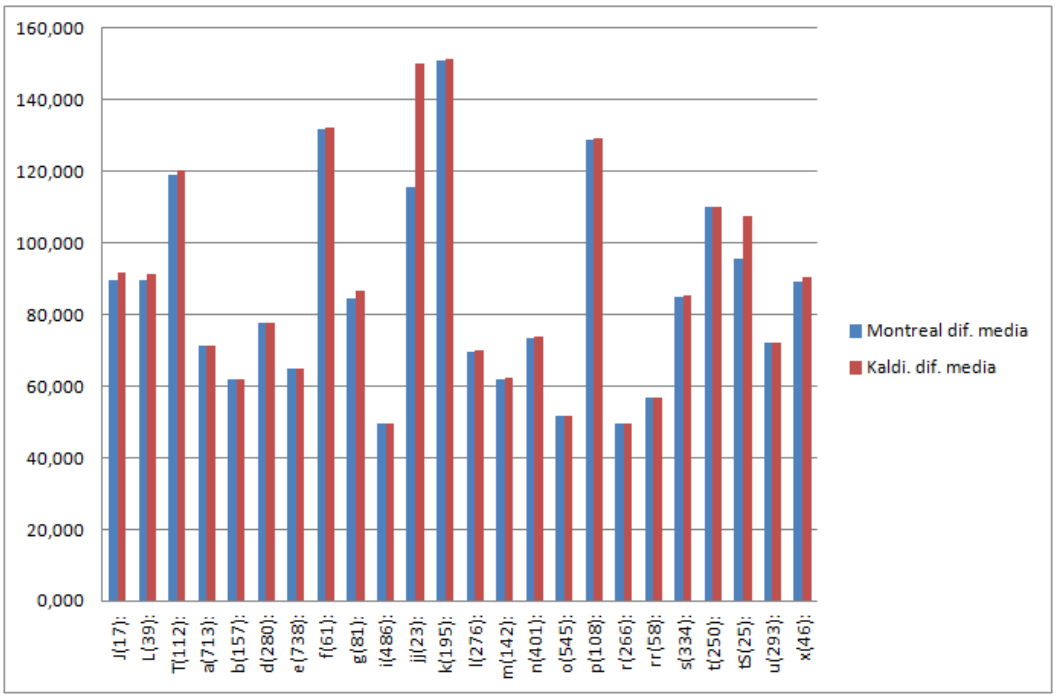

Figura 3.6: Error medio (en ms) cometido utilizando el alineamiento automático para cada fonema para la sesión 02M3.

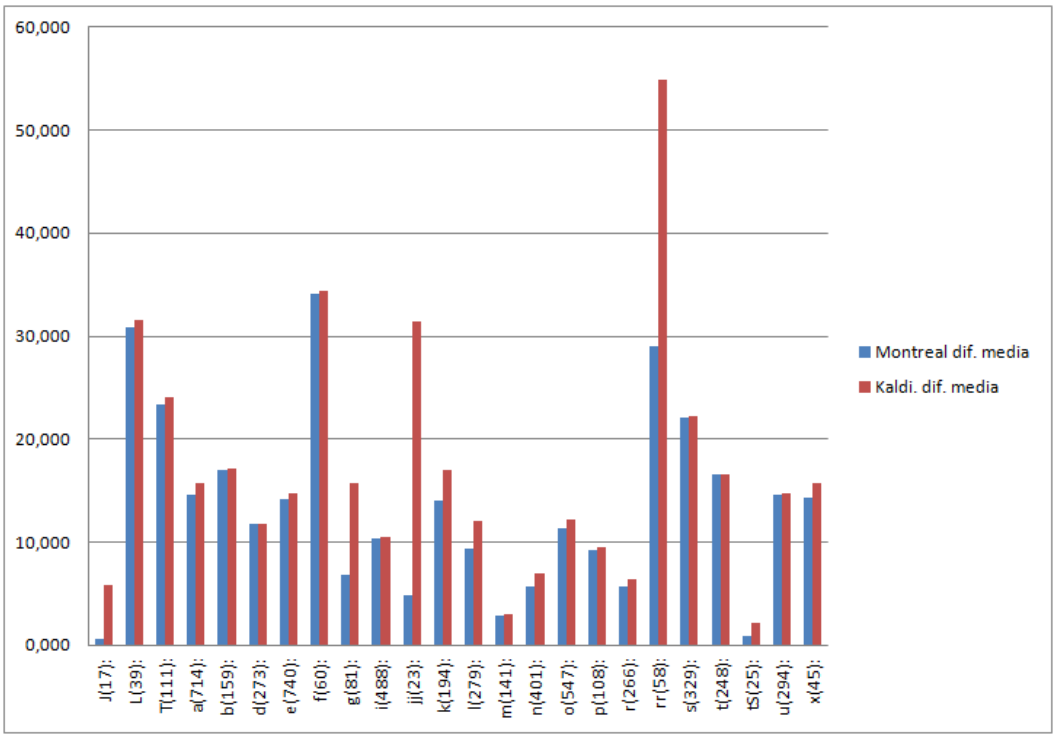

Figura 3.7: Error medio (en ms) cometido utilizando el alineamiento automático para cada fonema para la sesión 05M3.

3.6 Caracterización acústica de la base de datos

Para poder caracterizar las grabaciones recogidas en la base de datos, necesitamos extraer parámetros acústicos a partir de la señal, ya sean temporales, frecuenciales o cepstrales. Según lo propuesto en [48], algunos de los parámetros objetivos mas comúnmente usados para representar las voces patológicas basándose en el análisis de vocales sostenidas son los siguientes:

- Frecuencia fundamental media (f_0).
- Desviación estándar de la frecuencia fundamental (σf_0).
- Perturbación de pitch (jitter).
- Perturbación de amplitud (shimmer).

3.6.1 Frecuencia fundamental f_0

La frecuencia fundamental es la frecuencia más baja de una forma de onda periódica. Se trata de uno de los parámetros acústicos principales a la hora de caracterizar la señal de voz. Para calcularlo para voces sanas existen multitud de técnicas distintas. Sin embargo, a la hora de aplicarlo a las voces esofágicas los resultados obtenidos no son igual de buenos. Esto es debido a que la periodicidad de las señales alaríngeas no es tan consistente como en las voces sanas y los algoritmos utilizados normalmente no funcionan de la manera deseada. Por ello, a continuación se muestran tres métodos distintos a la hora de calcular la frecuencia fundamental, uno clásico y dos que intentan adaptarse a las características de estas señales. En la figura 3.8 se muestra la realización temporal de 100 ms de una /a/ sostenida para un hablante patológico (arriba) y otro sano (abajo). Se puede ver como la forma de onda es distinta en ambos casos. La señal patológica parece más simple que la otra, como si existiera una frecuencia predominante. Sin embargo, en la sana se puede apreciar que la señal es composición de varias frecuencias distintas.

Los métodos que se han comparado en este trabajo a la hora de extraer la f_0 son el método de la autocorrelación sobre la señal original, la autocorrelación sobre el residuo de la señal calculado mediante PSIAIF, y SRH sobre el residuo de la señal calculado mediante PSIAIF.

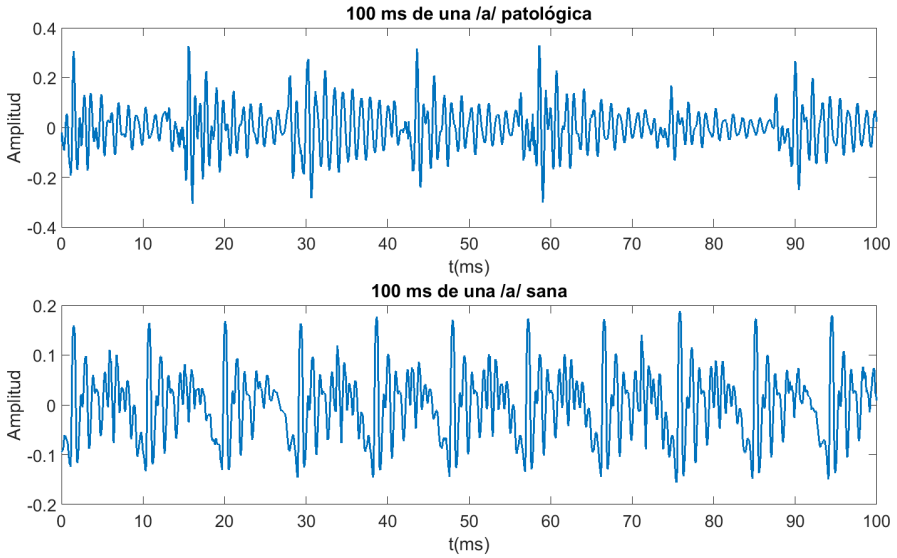

Figura 3.8: Aspecto en el tiempo de 100 ms de una /a/ para un hablante patológico y otro sano.

- Autocorrelación: La autocorrelación de una señal digital $y(n)$ se define de la siguiente manera:

$$R_{yy}(l) = \sum_{n \epsilon Z} y(n)y(n-l) \tag{3.3}$$

La correlación es una medida de similaridad entre dos señales. La correlación de una señal consigo misma permite por tanto encontrar las partes que se repiten. Para una señal periódica, la autocorrelación presentará un máximo cuando el desplazamiento corresponda con un múltiplo del periodo de esa señal. Esto implica que el periodo T_0 de una señal, y por tanto su frecuencia fundamental f_0 puede extraerse de su función de autocorrelación.

- Métodos basados en el análisis del residuo de la señal. Los métodos clásicos de extracción de pitch están diseñados para encontrar el pitch en los segmentos sonoros, es decir, cuando se producen vibraciones de las cuerdas vocales. Sin embargo esta definición no tiene sentido en el caso de los locutores laringectomizados. Para estas personas, la vibración se produce en el esófago, creándose unos sonidos de una frecuencia más baja. Por ello, cuando se uti-

lizan estos algoritmos con voces alaríngeas los resultados no son buenos, la curva de pitch es cero para la mayor parte de la señal. Por ejemplo, el método de la autocorrelación aplicado a una voz esofágica no devuelve ningún pico destacado dentro de las frecuencias razonables para la voz humana, por lo que el algoritmo no da buenos resultados la mayor parte del tiempo. Esto se puede comprobar observando la figura 3.9, en la que se muestran las autocorrelaciones de 100 ms de una /a/ para un hablante esofágico(arriba) y otro sano (abajo). Para el hablante sano aparecen picos en los múltiplos de la frecuencia fundamental, en torno a los 110 Hz. Sin embargo, para la señal patológica el máximo más importante aparece a los 2.125 ms, lo que corresponde a una frecuencia de 470 Hz, una frecuencia demasiado elevada para tratarse de la frecuencia fundamental. Por eso se adaptan los métodos de detección de pitch, aplicándolos a transformaciones de la señal esofágica en vez de a la señal en sí: sobre la señal de excitación o residuo calculada mediante filtrado inverso.

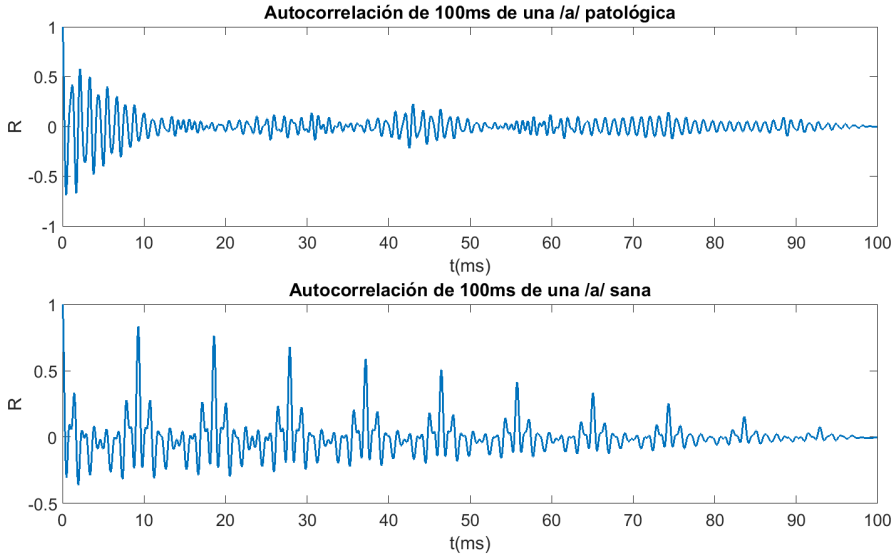

Figura 3.9: Autocorrelación de 100 ms de una /a/ para un hablante patológico y otro sano.

La señal de excitación es la señal fuente antes de ser filtrada por el tracto

vocal. Se modela idealmente como una serie de pulsos cuando el sonido es sonoro y ruido blanco cuando es sordo. Este tren de pulsos es el que contiene la periodicidad de la señal de voz, representada de una manera más simple.

La señal de excitación suele calcularse utilizando predicción lineal (LP), pero no funciona bien para señales alaríngeas. En su lugar, se ha utilizado el método Pitch Synchronous Iterative Adaptive Inverse Filtering (PSIAIF) propuesto en [5]. Este método aplica el análisis IAIF descrito en [56] dos veces a la misma señal. La primera iteración se aplica con una ventana de tamaño fijo para tener una primera estimación de los pulsos glotales y, por tanto, del periodo de la señal. La segunda iteración hace uso de este resultado para aplicar el análisis de manera síncrona entre cada par de pulsos glotales. A la señal de excitación calculada de esta manera se le aplican estos dos algoritmos:

- Autocorrelación: La periodicidad de la señal es más evidente en el residuo, así que es posible aplicarle el método de la autocorrelación para extraer la f_0.
- Summation of Residual Harmonics (SRH): En el SRH tradicional el residuo se calcula mediante análisis LPC. En este caso, el residuo utilizado es el calculado mediante PSIAIF. Para cada trama se calcula la amplitud del espectro $E(f)$. La envolvente de este espectro tiene una forma relativamente plana y presenta un máximo local en los múltiplos del pitch de la señal. Estos múltiplos se denominas armónicos. La suma de los armónicos residuales o SRH se define como

$$SRH(f) = E(f) + \sum_{k=2}^{N_{harm}} \left[E(k\Delta f) - E((k - \frac{1}{2})\Delta f) \right] \qquad (3.4)$$

  Esta ecuación se calcula para un intervalo [F_{0min}, F_{0max}] alrededor de la frecuencia fundamental buscada. El término $E(k\Delta f)$ del sumatorio se encarga de sumar todas las contribuciones de los armónicos de la frecuencia f. De esta manera, la función $SRH(f)$ tendrá un máximo en la frecuencia fundamental.

  La función $SRH(f)$ permite detectar el pitch eligiendo el máximo de la curva para cada trama. Habrá que decidir un umbral a partir del cual

se considera la señal sonora. Este valor se fijó en [29] en 0.07 porque daba el mejor compromiso entre falsos positivos y falsos negativos, y es el que se ha tomado en este apartado.

Para analizar las diferencias, se aplican los tres métodos distintos de extracción de pitch a las señales con vocales sostenidas para un locutor esofágico y uno sano. En vez de analizar todo el archivo, se seleccionan tramos estables de cada una de las cinco vocales grabadas. De este modo se podrán realizar medidas adicionales. La figura 3.10 muestra para un hablante esofágico (01M3) la f_0 (derecha) de cada tramo de vocal sostenida (forma de onda representada a la izquierda) para cada método. Se puede ver que para el método de la autocorrelación (linea negra continua) cuando da valor distinto de cero es un valor demasiado alto. Sin embargo, para los métodos de la autocorrelación del residuo PSIAIF (linea roja de puntos) y del SRH a partir del residuo PSIAIF (linea verde discontinua) los valores están en torno a los 50 Hz, valores que se ajustan más a lo percibido para la voz alaríngea.

En la figura 3.11 se muestra lo mismo pero para un locutor sano (S1). En este caso los tres métodos dan un resultado muy parecido, aunque es cierto que en algunos momentos el método basado en SRH del residuo PSIAIF produce algunas fluctuaciones. Por ello, se decide utilizar el método basado en la autocorrelación del residuo, ya que ofrece buenos resultados tanto para los voces esofágicas como para las sanas.

Por último, se utiliza los tres métodos para calcular la f_0 media para las vocales sostenidas y su desviación estándar. Los resultados se muestran el la tabla 3.6. Para la voz patológica (obviando los resultados del método tradicional puesto que da valores erróneos) se puede observar que el valor de f_0 es menor comparado con la voz sana. Pero la diferencia más significativa es la imposibilidad del hablante esofágico de mantener la frecuencia fundamental, como se desprende de los valores de desviación estándar.

Se ha calculado la frecuencia media y la desviación estándar sobre vocales sostenidas para todos los locutores de la base de datos, y también para un locutor sano (S1). Los resultados están recogidos en el anexo D.

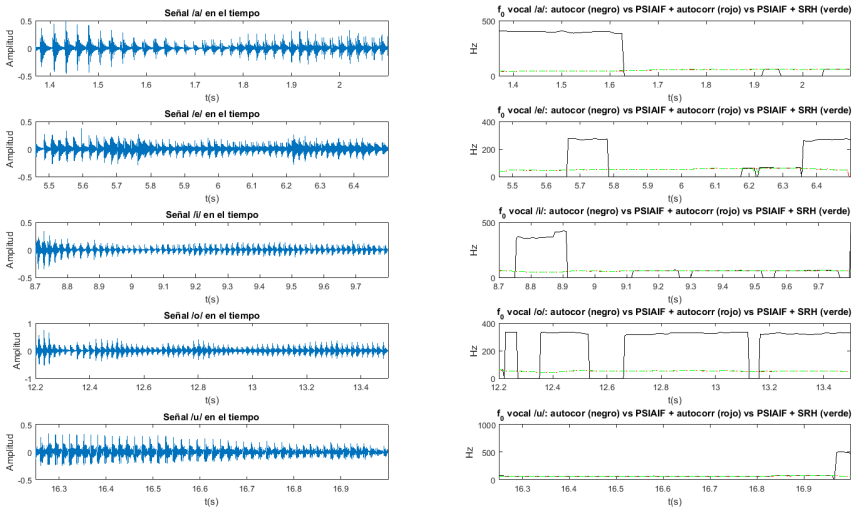

Figura 3.10: Cálculo de pitch para las 5 vocales sostenidas (/a/, /e/, /i/, /o/, /u/) de un locutor patológico.

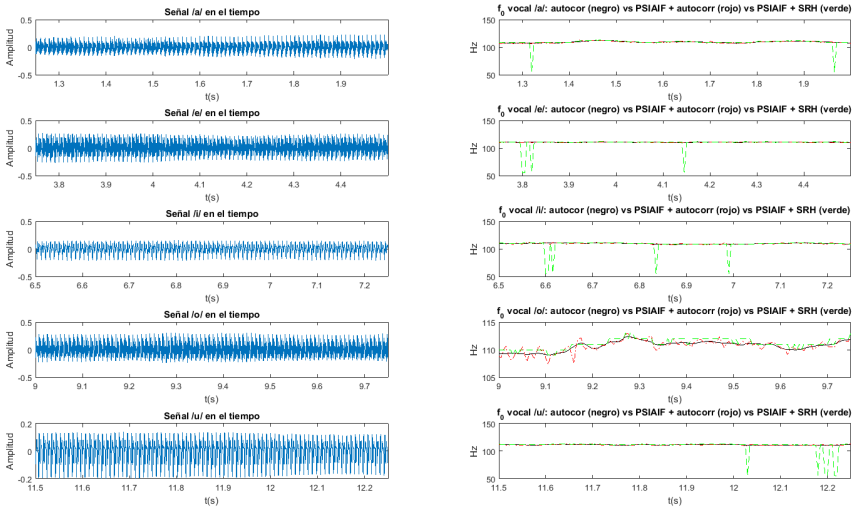

Figura 3.11: Cálculo de pitch para las 5 vocales sostenidas (/a/, /e/, /i/, /o/, /u/) de un locutor sano.

Tabla 3.6: f_0 media y desviación estándar para un locutor sano y otro esofágico calculada con tres métodos distintos sobre vocales sostenidas.

Método	Locutor patológico 01M3		Locutor sano S1	
	Media (Hz)	STD	Media (Hz)	STD
Autocorrelación	206.98	139.95	110.29	1.15
PSIAIF + Autocor	55.28	6.74	110.23	1.34
PSIAIF + SRH	55.54	6.92	109.51	8.28

3.6.2 Jitter

El jitter se define como la variación de la frecuencia fundamental de ciclo a ciclo. Se puede analizar sobre tramas de una vocal sostenida. Es uno de los parámetros que suelen utilizarse para caracterizar el habla patológica. La variación de este parámetro se ve afectada principalmente por la falta de control de las cuerdas vocales. Por ello, un valor alto de jitter es típico de hablantes con patologías.

Hay distintas maneras de calcular el jitter:

- El jitter absoluto representa la diferencia absoluta media entre dos periodos consecutivos. Se conoce como $jitta$ y se define como:

$$jitta = \frac{1}{N-1}\sum_{i=1}^{N-1}|T_i - T_{i-1}| \tag{3.5}$$

  donde N es el numero de ciclos a considerar y T_i es el periodo de cada ciclo i. Como T se ha cogido el inverso de la frecuencia fundamental calculada mediante los métodos explicados en 3.6.1. Según lo explicado en [46], el umbral para detectar patologías en adultos se establece en los 83.2 μs.

- El jitter local representa la diferencia absoluta entre dos periodos consecutivos dividida entre el periodo medio. Se conoce como $jitt$:

$$jitt = \frac{\frac{1}{N-1}\sum_{i=1}^{N-1}|T_i - T_{i-1}|}{\frac{1}{N}\sum_{i=1}^{N}T_i} \times 100 \tag{3.6}$$

  El límite del jitt para detectar patologías es del 1.04 %.

- Otra manera de calcular el jitter es el rap. Representa la media de la perturbación, es decir, la media de la diferencia absoluta entre un periodo y la media de ese periodo con sus dos adyacentes, dividido entre el periodo medio:

$$rap = \frac{\frac{1}{N-1}\sum_{i=1}^{N-1}\left|T_i - \left(\frac{1}{3}\sum_{n=i-1}^{i+1} T_n\right)\right|}{\frac{1}{N}\sum_{i=1}^{N} T_i} \times 100 \tag{3.7}$$

  El valor del umbral para detectar patologías es 0.68 %.

Con estas definiciones se calculan los diferentes valores para dos locutores patológicos (01M3 y 05M3) y uno sano (S1). El valor de N no está especificado a la hora de definir la medida, pero para que los resultados sean comparables se cogen las tramas de vocales sostenidas dónde la señal es estable. La tabla 3.7 muestra los resultados. Se puede comprobar que las técnicas basadas en análisis del residuo PSIAIF dan resultados erróneos para la voz sana. Para las voces esofágicas, los valores más estables se consiguen utilizando el análisis de pitch basado en la autocorrelación del residuo PSIAIF. Si se comparan los valores correspondientes a los locutores esofágicos obtenidos mediante este método, se puede comprobar que el locutor 01 presenta valores de jitter mejores que los del 05. Esto se corresponde con las apreciaciones subjetivas de que el locutor 01 es mejor locutor que el 05.

Tabla 3.7: Medidas de jitter para tres locutores, dos patológicos (01M3 y 05M3) y otro sano (S1).

Método	01M3			05M3			S1		
	jitta(μs)	jitt(%)	rap(%)	jitta(μs)	jitt(%)	rap(%)	jitta(μs)	jitt(%)	rap(%)
Autocorrelación	213.64	2.35	1.47	321.84	4.55	2.88	7.98	0.09	0.02
PSIAIF + Autocor	155.13	0.84	0.32	602.62	2.74	1.54	45.01	0.50	0.24
PSIAIF + SRH	261.31	1.43	0.87	1352.05	7.53	4.71	333.79	3.62	2.40

También se han calculado los valores de Jitter para todos los locutores esofágicos. Los resultados se recogen en el anexo D.

3.6.3 Shimmer

El shimmer se define como la variación de ciclo a ciclo de la amplitud de la señal. Para calcularlo se ha partido de los valores de f_0 extraídos mediante el método de la autocorrelación sobre el residuo PSIAIF. Con estos valores se han buscado los máximos (y los mínimos) de la amplitud de los tramos de vocales estables separados entre sí en torno a un período fundamental. En la figura 3.12 se puede ver el resultado de determinar de esta manera los máximos y los mínimos. Con estas amplitudes localizadas se puede pasar a calcular el Shimmer de una de las siguientes maneras:

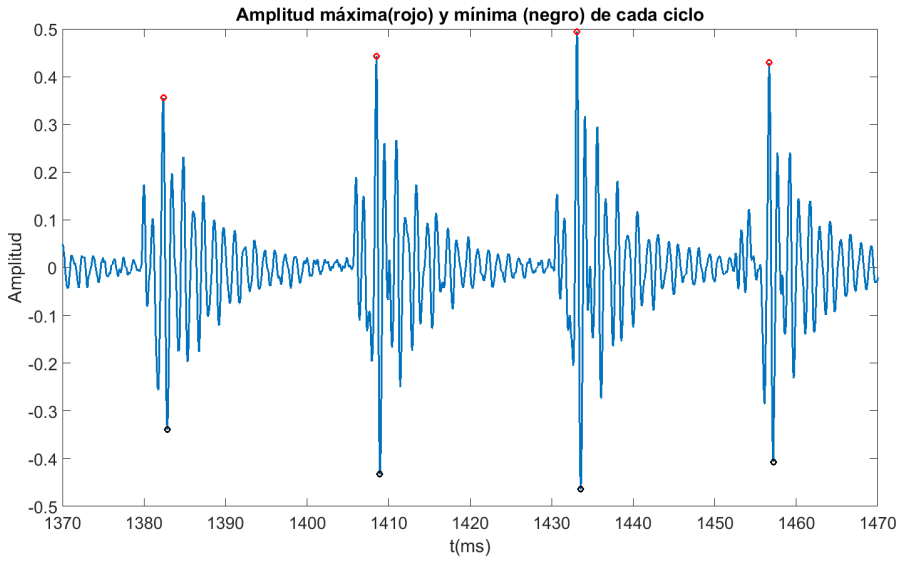

Figura 3.12: Detección de máximos y mínimos en 100ms de una /a/ patológica.

- Shimmer local: representa la media de la diferencia absoluta de amplitudes de dos periodos consecutivos dividida entre la amplitud media. Se conoce como $shim$ y se calcula:

$$Shim = \frac{\frac{1}{N-1}\sum_{i=1}^{N-1}|A_i - A_{i+1}|}{\frac{1}{N}\sum_{i=1}^{N}A_i} \times 100 \tag{3.8}$$

donde A_i representa la diferencia entre la amplitud máxima y la mínima del ciclo i de la señal y N el número de ciclos que se han tenido en cuenta a la hora de hacer el cálculo. El límite de este parámetro para detectar patologías está en el 3.81 %.

- Shimmer (local, dB): Representa la diferencia absoluta media del logaritmo decimal entre dos ciclos consecutivos. Se llama ShdB:

$$ShdB = \frac{1}{N-1}\sum_{i=1}^{N-1}\left|20log\left(\frac{A_{i+1}}{A_i}\right)\right| \tag{3.9}$$

  El límite para detectar patologías es de 0.350 dB.

Los valores de shimmer se han calculado sobre tramas estables de vocal para dos locutores esofágicos (01M3 y 05M3) y uno sano (S1). En la tabla 3.8 se muestran los resultados de $Shim$ y $ShdB$ tanto para todas las tramas como para cada trama de vocal por separado. Además de los valores, se muestra también el número de ciclos N tenidos en cuenta para su cálculo. Se puede ver que para el locutor sano los valores de $Shim$ se mantienen en todo momento por debajo del límite del 3.81 % y los del SHdB por debajo de los 0.85 dB. Sin embargo, cambian bastante de vocal a vocal. La /a/ es la que más cerca está del límite y hace que la media total valga un 2.22 %. Para los locutores esofágicos los valores de Shim y ShdB están siempre por encima de los umbrales. Las variaciones de valores entre vocales son también considerables, pero siempre están muy por encima. Aunque ambos locutores son esofágicos, los valores del locutor 05M3 son bastante peores que los del locutor 01M3. Si se escuchan frases de ambos locutores se puede comprobar que realmente el locutor 05M3 presenta una inteligibilidad peor que la del locutor 01M3.

En en anexo D se encuentra el shimmer se ha calculado para todos los locutores esofágicos.

3.6.4 Formantes

Para hacer el análisis de los formantes se toma como referencia el trabajo descrito en [12]. En este paper se hace un estudio acústico sobre las diferencias en las voca-

Tabla 3.8: Medidas de shimmer para dos locutores, dos patológicos (01M3 y 05M3) y otro sano (S1).

	01M3			05M3			S1		
	N	Shim(%)	ShdB(dB)	N	Shim(%)	ShdB(dB)	N	Shim(%)	ShdB(dB)
/a/	37	14.46	1.23	24	22.83	2.06	79	3.33	0.29
/e/	48	17.29	1.46	38	41.70	3.63	79	1.96	0.17
/i/	56	10.66	0.85	25	20.18	2.05	81	1.73	0.15
/o/	62	14.99	1.29	24	17.96	1.59	80	1.57	0.13
/u/	44	6.34	0.64	29	14.83	1.30	82	0.85	0.07
Total	247	14.78	1.22	140	24.59	2.32	401	2.22	0.19

les para grupos de hablantes sanos, esofágicos y traqueoesofágicos en castellano. Haciendo uso de un software comercial calculan el espectro del segmento estable central de las vocales y miden sus formantes, F1 y F2. La conclusión que se obtiene es que F1, F2, y la duración de las vocales en hablantes alaríngeos difieren de manera significativa de los valores normales. En general, los pacientes laringectomizados producen vocales con formantes mayores y duraciones más largas.

Con estas conclusiones en mente, en este apartado se muestra para un locutor patológico y uno sano las diferencias en los formantes de las vocales. Para ello hemos calculado los coeficientes cepstrales de tres maneras distintas y hemos representado la envolvente a partir de estos coeficientes para una trama central de la vocal. Las señales utilizadas han sido las vocales sostenidas de un locutor patológico (sesión 01M3) y un locutor sano.

Los métodos utilizados para obtener los coeficientes cepstrales se basan todos en Ahocoder [34], sólo difieren entre ellos en la manera de obtener la f_0. Todo el análisis cepstral se hace a partir de los valores de pitch para cada trama por lo que si los valores de la f_0 cambian, también habrá diferencias en los cepstrum obtenidos.

En la figura 3.13 se muestra para una señal sana la envolvente espectral de una trama correspondiente a una /a/. En azul, el análisis de f_0 se ha hecho con el método de la autocorrelación. La línea roja muestra los resultados cuando la f_0 se calcula haciendo la autocorrelación al residuo del algoritmo del método PSIAIF. En verde, el resultado si el análisis de f_0 se hace aplicando SRH al residuo del PSIAIF. Se puede ver que no hay grandes diferencias en las envolventes, todos los métodos

dan resultados muy parecidos. En este caso, los dos primeros formantes aparecen en 760 Hz y 1350 Hz.

Los resultados de hacer lo mismo para una /a/ emitida por un locutor esofágico se muestra en la figura 3.14. En este caso puede observarse que para la envolvente calculada a partir del f_0 obtenido con el método de la autocorrelación, para las frecuencias bajas la envolvente se diferencia mucho de una /a/ sana. Sin embargo, para los análisis con el pitch extraído a partir del residuo calculado con PSIAIF (líneas roja y verde) se puede ver que la envolvente se parece mucho más a la /a/ sana, sobre todo en la parte baja del espectro. En este caso se puede ver claramente que los dos primeros formantes están en 830 y 1240 Hz. Si se repite el mismo experimento, pero en otra trama donde aparezca la /a/ se obtiene la figura 3.15. En este caso, la diferencia es mucho menor entre métodos, pero al usar el análisis de pitch aplicado al residuo calculado con PSIAIF, los picos a bajas frecuencias aparecen más marcados. Cogiendo una tercera trama de /a/ distinta para la señal patológica se obtiene la figura 3.16. Para esta realización, las diferencias son mucho más visibles. Parece que para la señal patológica la diferencia entre codificaciones para extraer los coeficientes melcepstrales es muy dependiente de la realización de la trama.

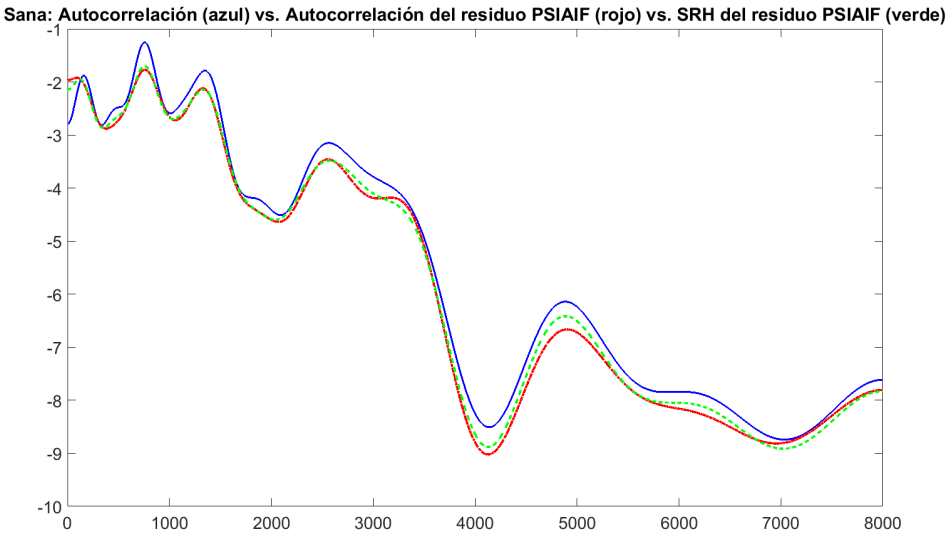

Figura 3.13: Envolvente para una trama de una /a/ para un hablante sano.

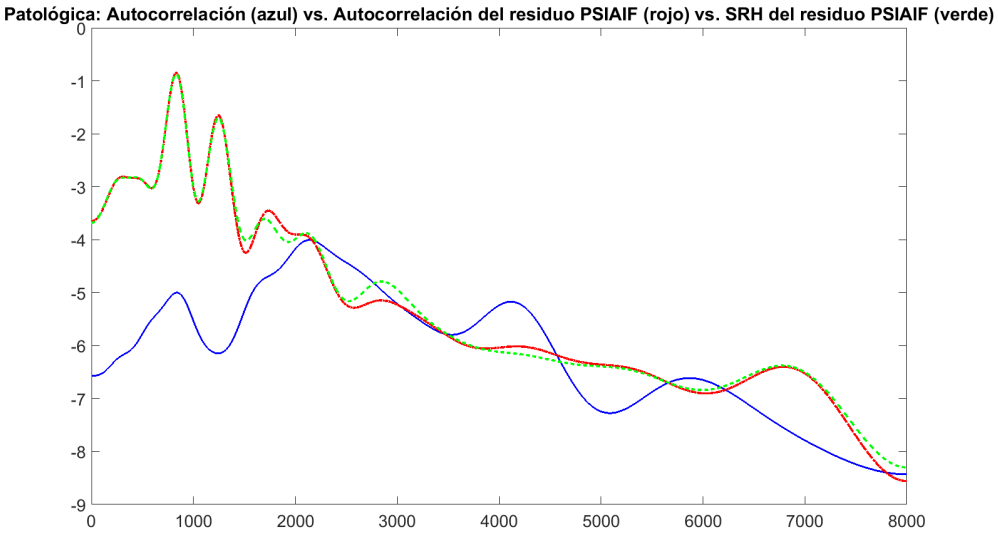

Figura 3.14: Envolvente para una trama de una /a/ para un hablante esofágico.

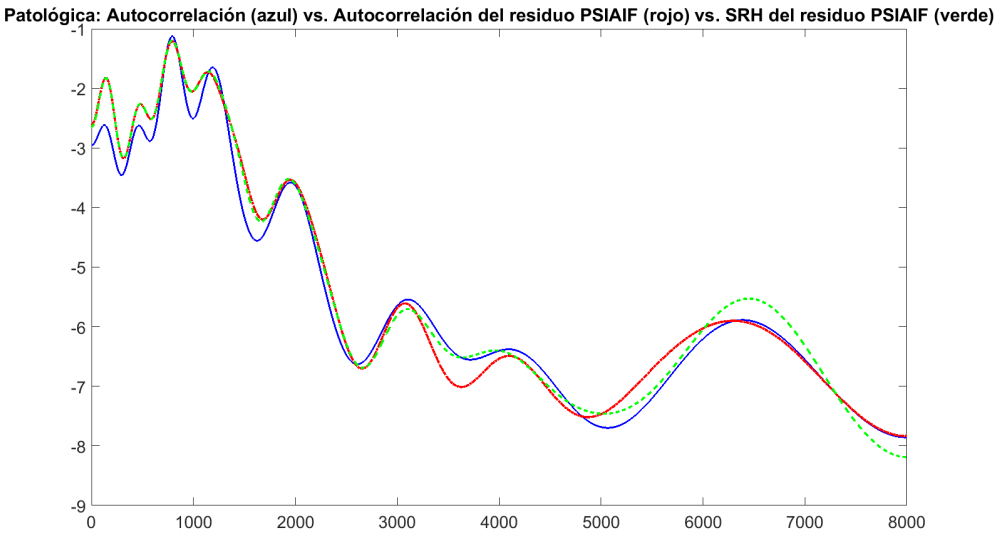

Figura 3.15: Envolvente para una trama distinta de una /a/ para un hablante esofágico.

En las figuras 3.17 y 3.18 se repite el mismo estudio, pero para una trama que contiene una /i/, para un hablante sano y uno esofágico respectivamente. En este

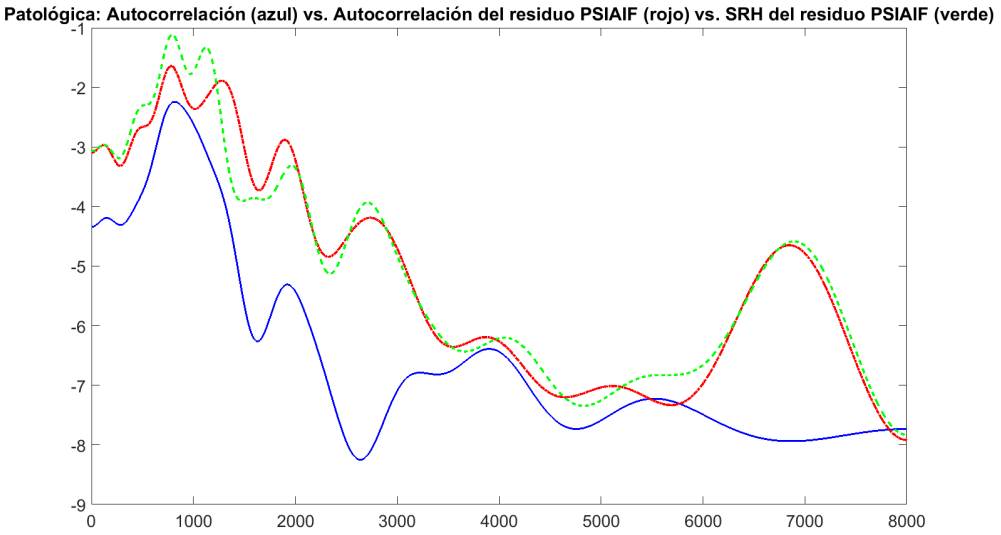

Figura 3.16: Envolvente para una tercera trama de una /a/ para un hablante esofágico.

caso los formantes aparecen en 260 y 2200 Hz para el hablante sano, y en 430 y 2460 Hz para el locutor esofágico.

El resultado de hacer lo mismo para una /u/ se muestra en las figuras 3.19 y 3.20. Los dos primeros formantes para la señal sana están en 320 y 690 Hz. Para el locutor esofágico, estos formantes aparecen en 475 y 805 Hz.

Para comprobar como de importante es el método de calcular la f_0, se han representado para todas las tramas estables de cada vocal sostenida su envolvente espectral en la misma gráfica. En las figuras 3.21 a 3.25 se muestran para cada fila la envolvente calculada extrayendo el pitch con un método distinto (de arriba a abajo, autocorrelación, autocorrelación sobre el residuo PSIAIF y SRH sobre el residuo PSIAIF), y para cada columna un locutor distinto(de izquierda a derecha, 01M3, S1 y 05M3).

Se puede comprobar que para el locutor sano todos los métodos muestran unas gráficas donde casi todas las tramas pasan por los mismos puntos, creándose una suma de líneas estrecha, con los formantes de cada vocal en las mismas frecuencias. Es cierto que para algunas vocales de este locutor el método del SRH sobre el residuo PSIAIF tiene algunas tramas bastante distintas al resto.

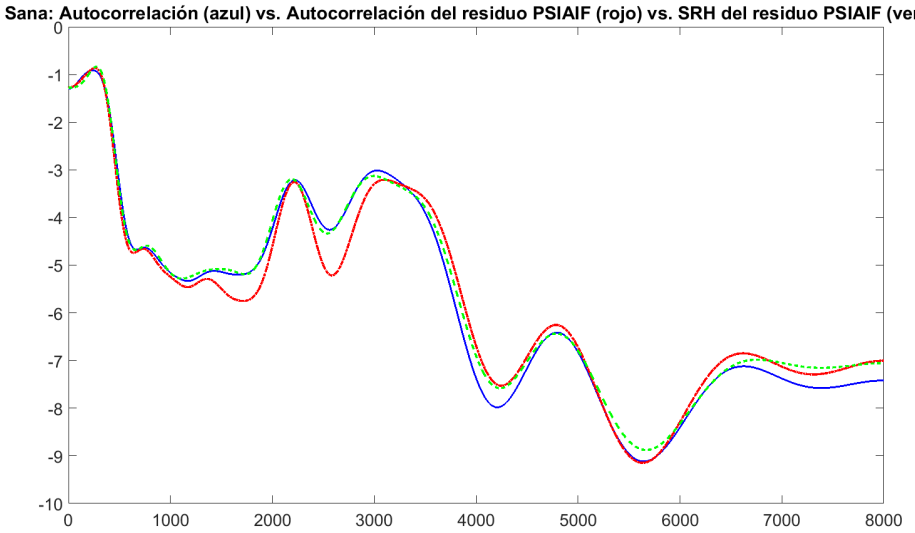

Figura 3.17: Envolvente para una trama de una /i/ para un hablante sano.

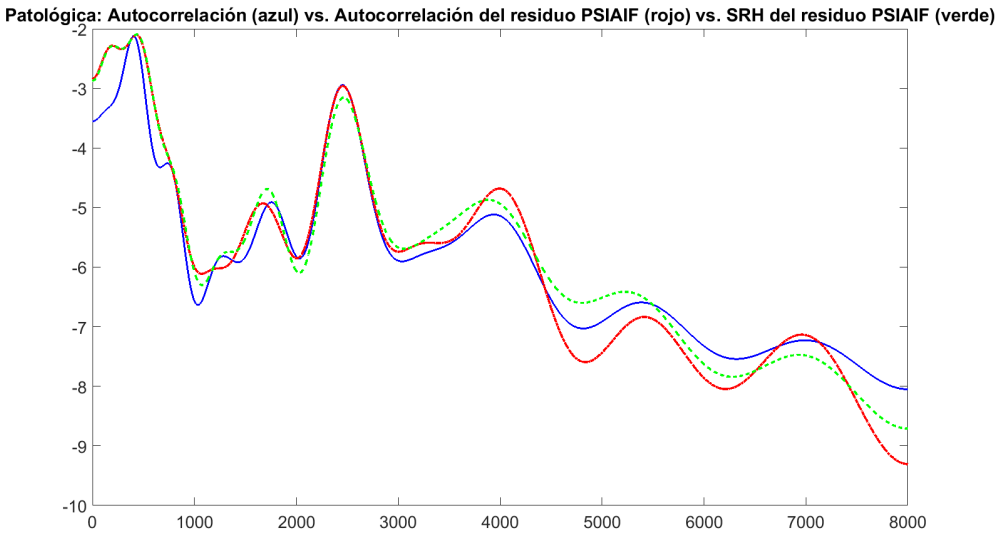

Figura 3.18: Envolvente para una trama de una /i/ para un hablante esofágico.

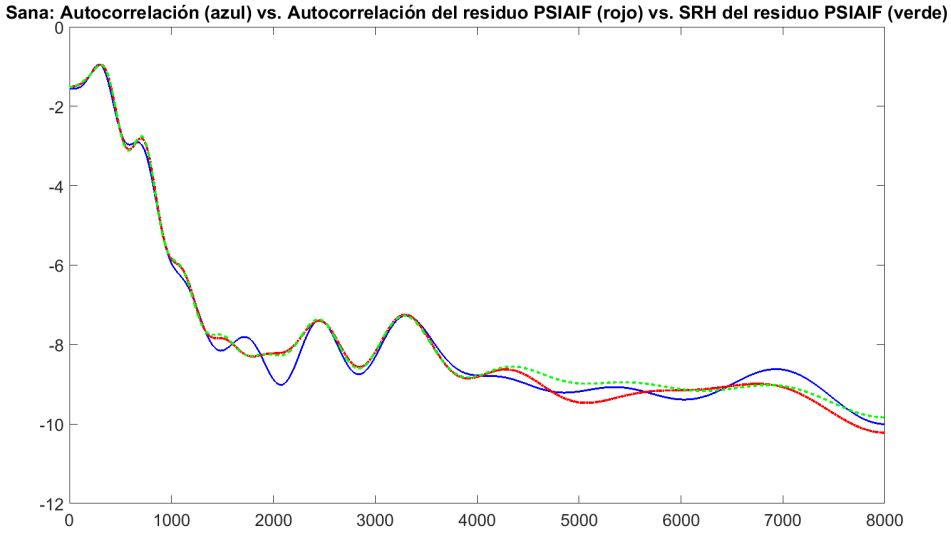

Figura 3.19: Envolvente para una trama de una /u/ para un hablante sano.

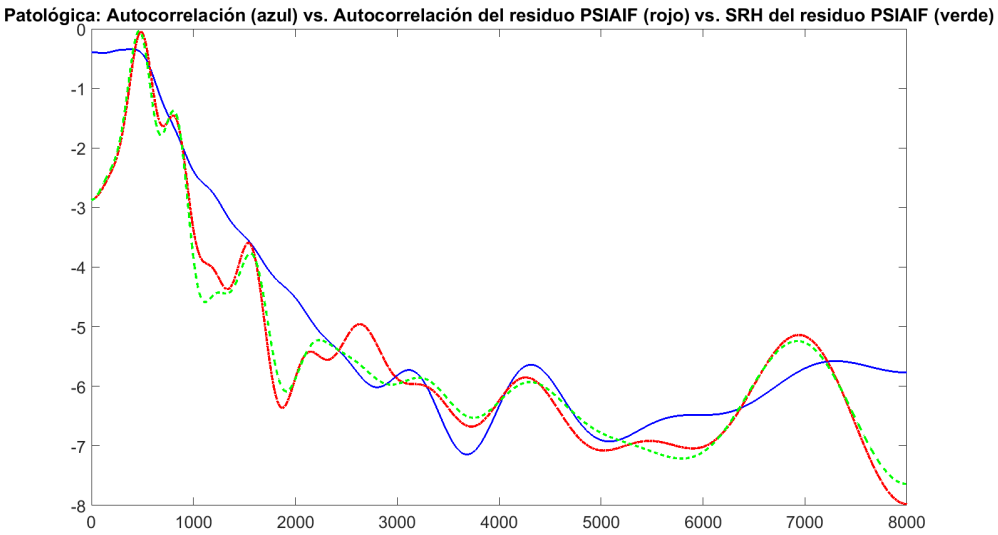

Figura 3.20: Envolvente para una trama de una /u/ para un hablante esofágico.

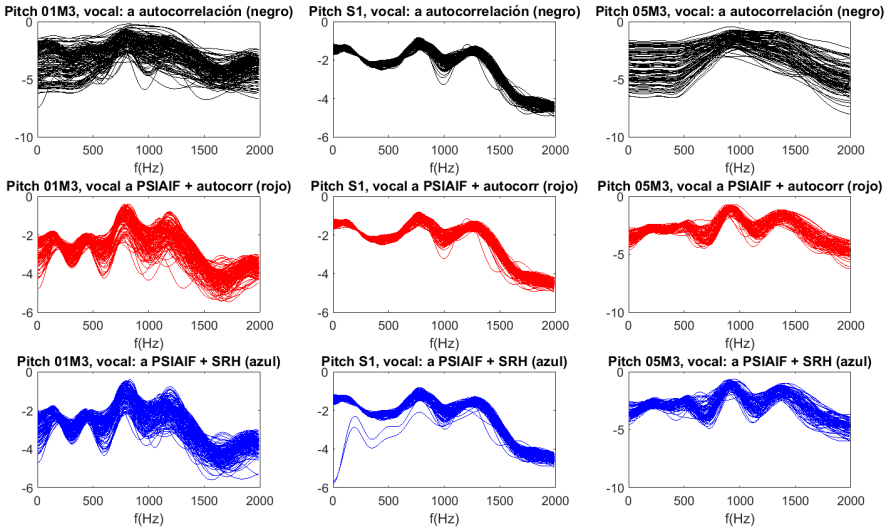

Figura 3.21: Envolventes para todas las trama de una /a/ para tres locutores distintos.

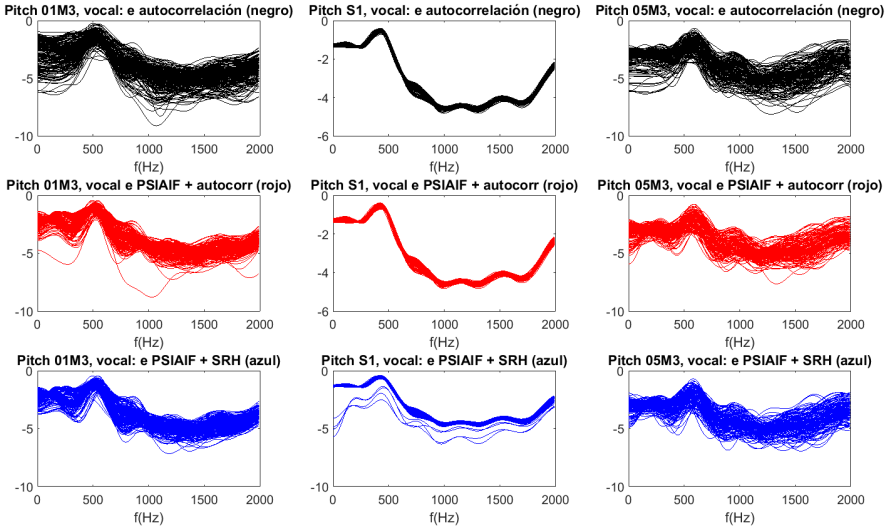

Figura 3.22: Envolventes para todas las trama de una /e/ para tres locutores distintos.

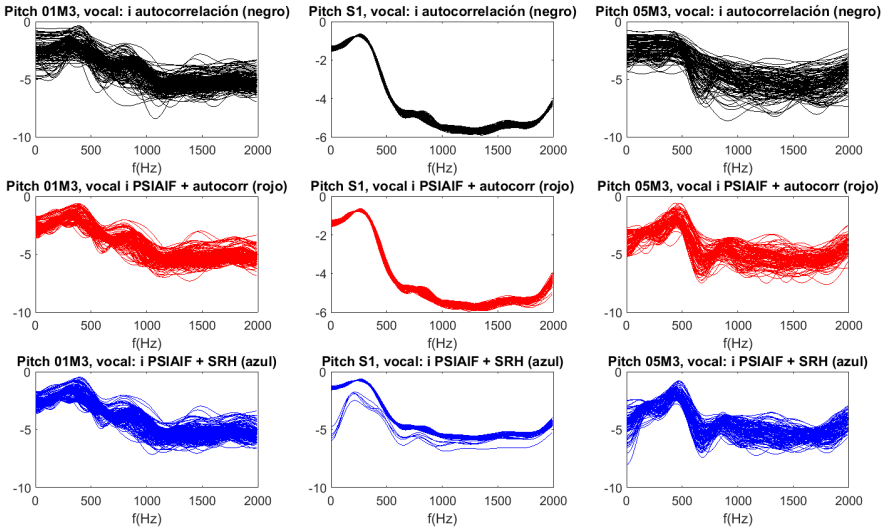

Figura 3.23: Envolventes para todas las trama de una /i/ para tres locutores distintos.

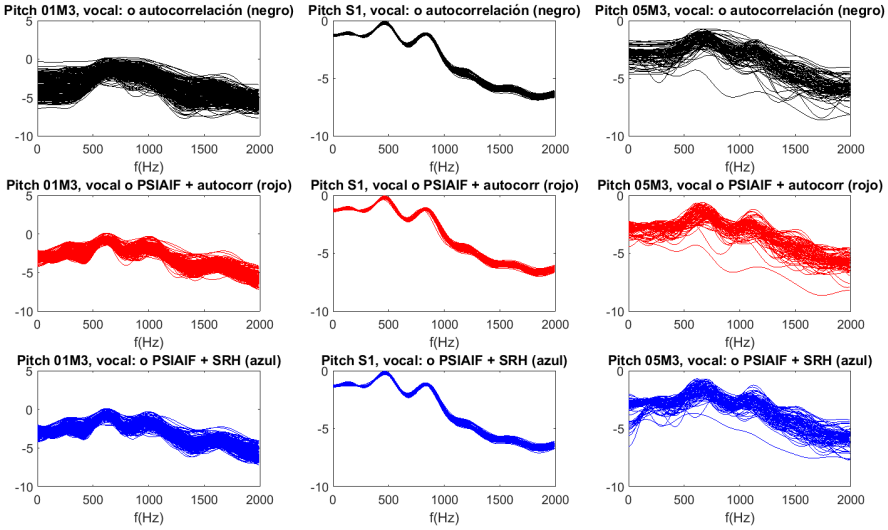

Figura 3.24: Envolventes para todas las trama de una /o/ para tres locutores distintos.

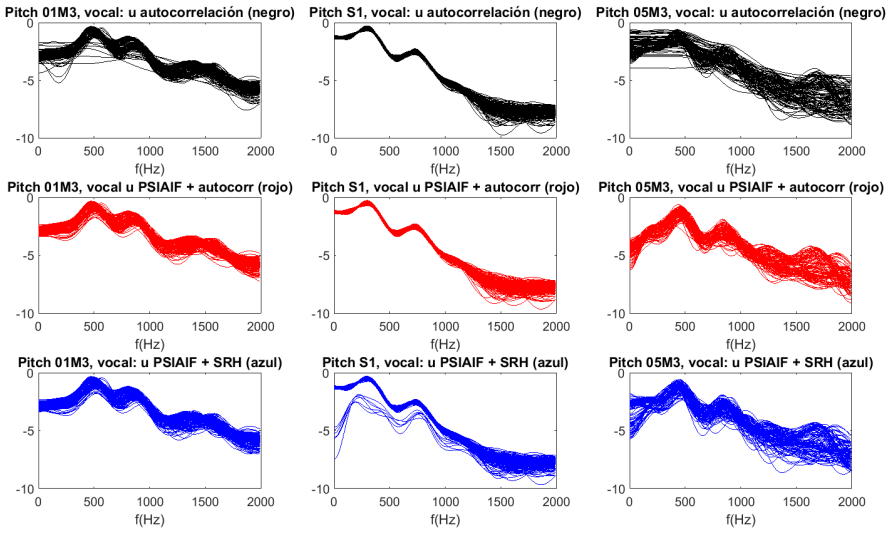

Figura 3.25: Envolventes para todas las trama de una /u/ para tres locutores distintos.

Para los locutores esofágicos, sin embargo, y aunque para todos los casos el conjunto de líneas obtenido es mucho más grueso y disperso, hay diferencias entre el método escogido para analizar la f_0. Por ejemplo, para la vocal /a/ u /o/ (figuras 3.21 y 3.24), el método tradicional (línea negra) hace que los formantes sean indistinguibles para cualquiera de los dos locutores esofágicos. Con los métodos aplicados al residuo PSIAIF de la señal, la posición de los formantes de una trama a otra sigue variando, pero parecen más fácilmente diferenciables.

Una vez vistas todas estas figuras, se llega a al conclusión de que para la señal de voz sana, no importa cuál de los métodos se utilice para calcular la f_0, pues no se aprecian diferencias significativas. Sin embargo, para la señal patológica, es mejor utilizar como punto de partida para la parametrización de estas señales el análisis de pitch basado en el cálculo del residuo mediante PSIAIF, sobre todo si nos fijamos en las frecuencias más bajas. En concreto, se decidió utilizar el método de la autocorrelación sobre el residuo PSIAIF puesto que parece que da resultados más estables que el método basado en SRH.

3.6.5 Duración de los sonidos

Una de las características de la voz esofágica es que la producción es en general más lenta. Los laringectomizados se ven obligados a hacer más pausas que los hablantes sanos para tragar saliva o tomar aire [30]. En este apartado se presenta un análisis de las duraciones de los sonidos de la base de datos grabada y los resultados se comparan con las duraciones obtenidas para hablantes sanos.

Para el cálculo de las duraciones de los sonidos ha sido necesario etiquetar la base de datos a nivel de fonemas, para lo que se ha utilizado el método explicado en 3.5.1.2 para los locutores esofágicos. Los silencios iniciales y finales de cada frase no se han tenido en cuenta.

3.6.5.1 Duración del habla

Utilizando las etiquetas, se ha calculado la duración de cada una de las sesiones de habla esofágica en las que se han grabado las 100 frases del corpus en castellano. Los resultados se pueden ver en la figura 3.26. Se puede ver que la duración de las grabaciones varía mucho de unos locutores a otros. El locutor al que menos tiempo le lleva grabar las 100 frases es el 09MT (traqueoesofágico) con 458 segundos. Entre las voces esofágicas, el locutor que habla más rápido es el 26M3 (468 segundos), y al que más, a 14M2 (segunda fase del aprendizaje), que emplea un tiempo de 1525 segundos, casi el triple.

Cuando se hace lo mismo con 9 locutores sanos que grabaron las mismas 100 frases se obtiene que la variabilidad es mucho menor, aunque el hecho realmente diferencial es que la velocidad de habla de los hablantes sanos es mayor que la de los esofágicos: todos los locutores tardan en enunciar las 100 frases en torno a los 400 segundos.

En la figura 3.27 se muestra una comparación entre las duraciones de las 30 sesiones esofágicas y las 9 hechas por locutores sanos. Las diferencias entre ambos grupos de hablantes son claras. La media de duración de un locutor esofágico es de 773.25 segundos (con una desviación estándar de 231.96 s), mientras que para un locutor sano este valor es de 405.11 segundos (y una σ de sólo 27.3 s).

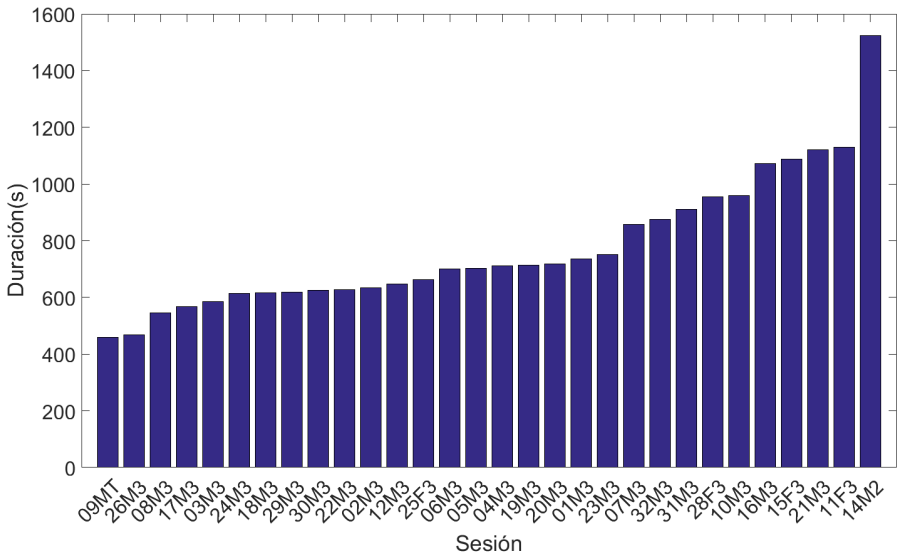

Figura 3.26: Tiempo empleado por cada locutor esofágico para grabar las 100 frases del corpus en castellano.

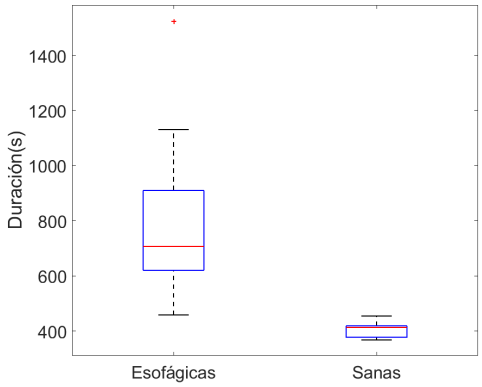

Figura 3.27: Comparación entre el tiempo tardado en emitir las 100 frases de 30 hablantes esofágicos y 9 locutores sanos. En cada caja,la línea central es la mediana, los bordes de la caja representan los percentiles 25 y 75, los bigotes se extienden a los valores más extremos no considerados outliers, y los outliers se muestran individualmente como una cruz roja.

3.6.5.2 Velocidad del habla

Otra manera de caracterizar la disfluencia de un hablante esofágico es calcular la *velocidad del habla*. Normalmente se calcula en número de fonaciones por unidad de tiempo (sílabas por segundo, palabras por minuto...). Para este trabajo se decidió calcularlo como:

$$SR = \frac{n^o\ sílabas}{duración} \tag{3.10}$$

Como se dispone de las transcripciones de las 100 frases grabadas por los distintos locutores, resulta sencillo calcular la velocidad. Mediante un transcriptor desarrollado por el grupo Aholab se extrae del los archivos de texto con las transcripciones el número de sílabas que compone cada frase grabada. Para la duración de cada frase se eliminan los silencios iniciales y finales de cada frase.

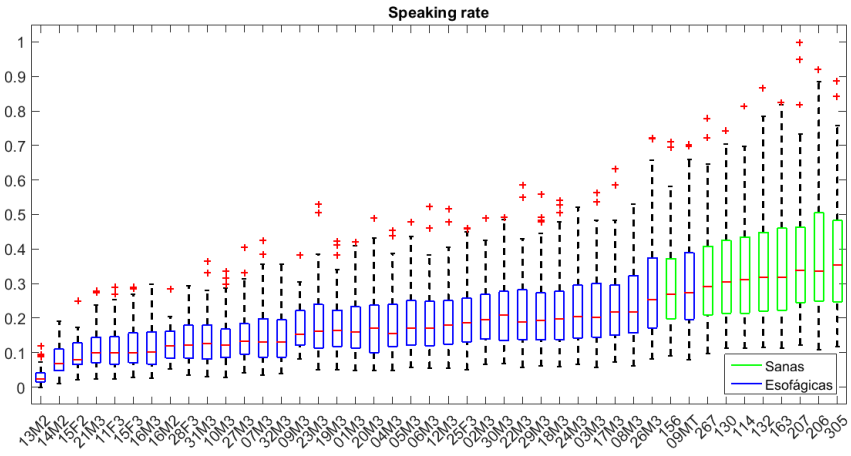

Figura 3.28: Velocidad del habla calculada para 35 sesiones de hablantes esofágicos (azul) y 9 de locutores sanos (verde). En cada caja,la línea central es la mediana, los bordes de la caja representan los percentiles 25 y 75, los bigotes se extienden a los valores más extremos no considerados outliers, y los outliers se muestran individualmente como una cruz roja.

En la figura 3.28 se muestran los resultados de calcular la velocidad del habla de las 35 sesiones de hablantes esofágicos (azul) y de 9 sesiones de locutores de habla sana (verde). Al analizar los resultados se puede ver que las voces sanas tienen

una velocidad de habla mayor que las esofágicas. Sin embargo hay una sesión, la 09MT, que tiene un valor mayor que un locutor sano. Esto se debe a que es la sesión grabada por el locutor 09 mediante habla traqueoesofágica, es decir, utilizando la válvula. Esto permite al locutor hablar de una manera más continua. Los resultados de la sesión del mismo locutor 09 hablando sin la válvula (09M3) se encuentran dentro del resto de valores de locutores esofágicos. Otra de las conclusiones que se pueden extraer al observar los resultados es que tres de los cuatro locutores esofágicos en segunda fase de aprendizaje tienen los valores de velocidad más bajos. El cuarto locutor, el 16, tiene curiosamente un speaking rate mayor en las 33 frases que grabó cuando se encontraba en la segunda frase de aprendizaje (16M2) que en las 100 que enunció cuando volvió al estudio a grabar unos meses más tarde (16M3). La explicación puede deberse a que al ser poco tiempo después, se produjo cierta mejoría en el locutor, pero no la suficiente para aguantar el cansancio al pasar a grabar de 33 a 100 frases.

3.6.5.3 **Duraciones de los sonidos**

La figura 3.29 muestra las duraciones de todos los sonidos del corpus para un hablante específico (se trata del locutor 02M3, que es un locutor experto). Se ha incluido el silencio entre palabras como un sonido más ('*sil*'). Para las 100 frases grabadas por este locutor aparecen 5670 fonemas y 656 silencios intermedios. La duración de los distintos fonemas es muy similar entre ellos, aunque la variabilidad es grande. Concretamente, la duración media de los fonemas es de 0.100 segundos con una varianza de 0.002; mientras que la duración media de los silencios es de 0.101 segundos con una varianza de 0.004.

En la figura 3.30 se comparan estas duraciones con las duraciones de los mismos sonidos para 9 locutores sanos que han grabado las mismas 100 frases. Para etiquetar estas grabaciones se ha utilizado el método descrito en 3.5.1.1. Estas 9 sesiones se componen de 900 frases con 50906 fonemas y 1342 silencios intermedios en total. La duración de los fonemas es de 0.068 segundos con una varianza de 0.001, mientras que la de los silencios es de 0.128 segundos con una varianza de 0.016.

La duración media de todos los fonemas del hablante esofágico es mayor que la de los hablantes sanos tomados como conjunto. En concreto, es una diferencia

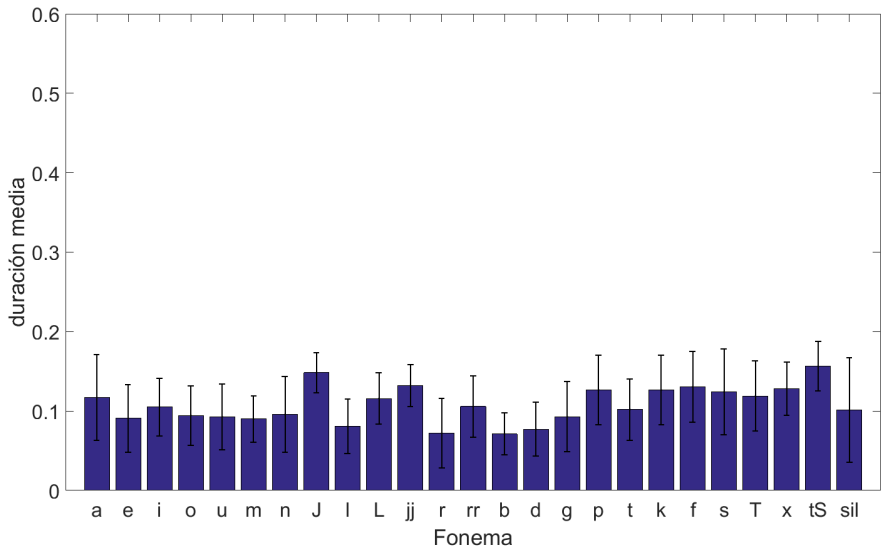

Figura 3.29: Duración de los sonidos del corpus para las 100 frases del hablante esofágico 02M3. Las lineas indican la desviación estándar para cada sonido.

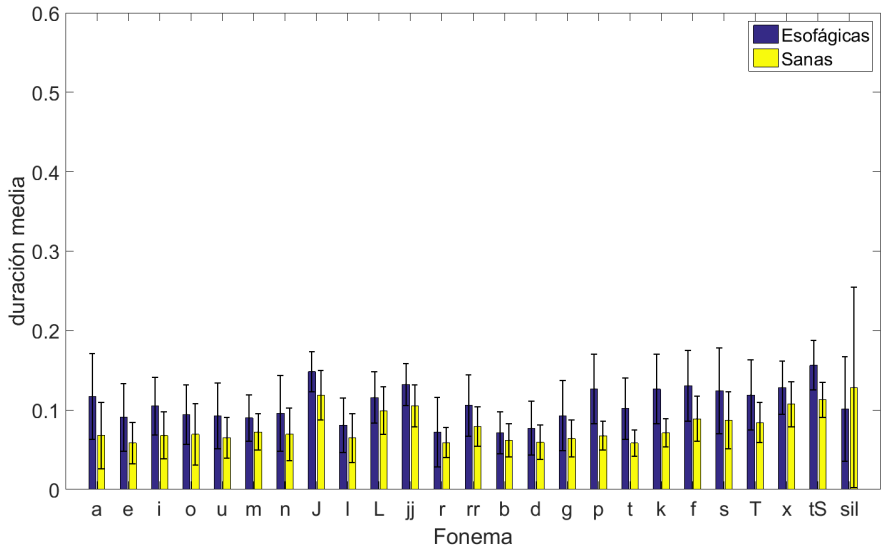

Figura 3.30: Duración de los sonidos del corpus para el hablante esofágico 02M3 y para 9 hablantes sanos. Las lineas indican la desviación estándar para cada sonido.

de unos 32 ms. La varianza en la duración de cada fonema también es mayor para el hablante esofágico que para el grupo de 9 hablantes sanos, casi el doble.

Sin embargo, a la hora de analizar el silencio estos resultados cambian. La duración media de las pausas intermedias de los hablantes sanas es 27 ms mayor que la del hablante esofágico. Se puede comprobar que la varianza de esta duración también es mucho mayor para los 9 hablantes sanos. La explicación a este hecho está en el tipo de frases grabadas. El corpus se compone de muchas frases largas y los locutores tienden a realizar pausas cuando leen. La duración de estas pausas varía mucho de un locutor a otro. El locutor 02M3 intenta leer las frases de una manera continua. Sin embargo, para las estadísticas de las voces sanas hay 9 locutores que intentar marcar bien las pausas cuando aparece una coma en la transcripción. Por ello tanto la duración como la variabilidad de los silencios es mayor.

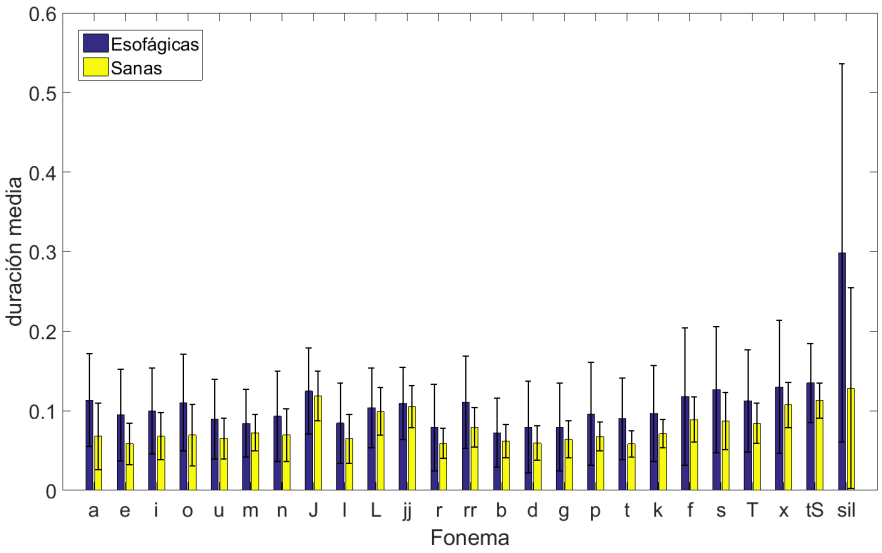

Figura 3.31: Duración de los sonidos del corpus para 30 locutores esofágicos y para 9 hablantes sanos. Las lineas indican la desviación estándar para cada sonido.

En la figura 3.31 se muestra la duración media de cada sonido tomando en conjunto 30 hablantes esofágicos (azul) y 9 hablantes sanos (amarillo), así como sus desviaciones estándar. En las frases esofágicas hay 170100 fonemas y 21423 silencios. La duración media de los fonemas es de 0.099 segundos con una varianza

de 0.004. La duración media de los silencios es de 0.299 segundos y una varianza de 0.057. Se puede ver que, al añadir más hablantes esofágicos, la duración media de los fonemas se mantiene en valores similares a los del locutor 02M3 (siguen durando ligeramente más que los fonemas de los locutores sanos) y una varianza un poco mayor. Sin embargo, la duración de los silencios se dispara y su variabilidad aumenta. Los nuevos hablantes esofágicos añadidos para hacer las estadísticas son muy distintos entre sí. Algunos tienen una calidad parecida al locutor experto 02M3, pero otros son menos inteligibles e introducen muchas más pausas y de una mayor duración a la hora de enunciar las frases.

3.7 Conclusiones

En este capítulo se ha descrito como se han obtenido los datos necesarios para desarrollar las técnicas de mejora de la inteligibilidad patológica que se describirán en los siguientes capítulos de esta tesis.

Como resultado, se ha obtenido una base de datos de voces esofágicas compuesta por 35 sesiones de 32 locutores diferentes. El grueso del contenido son 100 frases fonéticamente equilibradas, pero también se grabaron vocales sostenidas y palabras aisladas. En total se tiene unas 9 horas y 50 minutos de audio. Aunque el proceso de grabación ha sido supervisado, se han cometido errores. En total se han registrado 680 errores entre sustituciones, inserciones y eliminaciones, que han sido adecuadamente anotados.

También se ha descrito un método de etiquetado automático de las grabaciones esofágicas utilizando como herramienta básica el alineador Montreal. Con este alineador basado en Kaldi se crean modelos acústicos a partir de las propias grabaciones esofágicas a etiquetar. La evaluación de los resultados llevada a cabo ha evidenciado que con este método las etiquetas obtenidas se acercan más al etiquetado de referencia que otros métodos de alineamiento probados.

Por último, al analizar las grabaciones se han evaluado diferentes métodos de extracción de la frecuencia fundamental. Además del clásico basado en la autocorrelación de la señal, se estudia otro método basado en hacer la autocorrelación del residuo PSIAIF de la señal alaríngea. Al aplicar este algoritmo de extracción de pitch para analizar los formantes se ha comprobado que los formantes quedan mejor reflejados para los hablantes esofágicos que con los métodos tradicionales. Por tanto, este es el algoritmo de extracción de pitch que se usará para parametrizar las voces esofágicas durante el desarrollo de esta tesis.

Esta base de datos es una aportación importante ya que existe una escasez de señales alaríngeas con las que poder investigar. Su liberación llena un vacío existente. Parte del trabajo realizado en la grabación y análisis de esta base de datos se recoge en un artículo que ha sido enviado a la revista PLoS One y está a la espera de comunicación en el momento de depósito de esta tesis.

La información, si es bien transmitida y comprendida, conlleva inteligibilidad, primera condición necesaria para la comprensión, pero no suficiente.

Edgar Morin

CAPÍTULO

4

Preparación de un sistema ASR

Las características del habla esofágica la hacen fácilmente distinguible del habla sana. Sin embargo, en términos de inteligibilidad cualquier persona es capaz, con mayor o menor grado de esfuerzo, de entender el discurso de un locutor alaríngeo. En los sistemas de reconocimiento automático del habla (ASR), por otro lado, este tipo de habla no obtiene buenos resultados. Los modelos acústicos que utilizan estos sistemas están entrenados a partir de voces sanas lo que hace que estos locutores no puedan utilizarlo. Hoy en día los sistemas ASR están ganando más y más importancia en el día a día, así que se deben crear técnicas que permitan que los hablantes esofágicos no se vean excluidos.

El primer paso a tomar para poder conseguir que los laringectomizados puedan utilizar los sistemas ASR es construir un buen sistema de reconocimiento para hablantes de castellano sanos que pueda utilizarse como punto de partida. En este capítulo se explican los detalles que se han seguido para construir y evaluar ese sistema de reconocimiento automático de habla.

4.1 Preparación de un sistema de reconocimiento en castellano con Kaldi.

Para preparar el sistema de reconocimiento de habla se utilizaron las herramientas que proporciona Kaldi [89]. Kaldi es un toolkit para reconocimiento de voz escrito en C++ y que está distribuido bajo licencia Apache v2.0. Se eligió este conjunto de software y utilidades debido a que ofrece la posibilidad de utilizar de una manera relativamente sencilla redes neuronales a la hora de hacer el reconocimiento.

Se decidió construir un reconocedor de habla continua de gran vocabulario (LVCSR - Large Vocabulary Continuous Speech Recognizer). Para ello, Kaldi proporciona diversos ejemplos para desarrollar reconocedores que se denominan como "recetas". Para la construcción de este LVCSR se ha seguido la receta s5 diseñada para la base de datos Wall Street Journal, pero adecuándola a nuestras propias señales de audio en castellano.

Los parámetros utilizados para desarrollar el reconocedor son 13 coeficientes mel-cepstrales (MFCC - Mel-Frequency Cepstral Coefficients) extraídos a partir de los audios de entrenamiento a los que se les aplica una normalización en media y varianza (CMVN). Esta normalización sirve para mitigar los efectos que pudiera introducir el canal.

El entrenamiento de los modelos acústicos comienza con una inicialización desde cero de modelos ocultos de Markov (HMM - Hidden Markov Models) de fonemas sin contexto. Después se hace una serie de entrenamientos iterativos que van mejorando los modelos. Primero se añaden las diferencias de los parámetros, y luego se entrenan trifonemas con contexto izquierdo y derecho. En el último paso se entrena una red neuronal. Los parámetros de entrada a la red neuronal consisten en una serie de vectores de 40 dimensiones. La red ve una ventana de estos parámetros con 4 tramas de contexto a cada lado de la trama central. Estos parámetros son el resultado de procesar los 13 MFCCs iniciales. Los pasos necesarios están descritos en [92] y son los siguientes:

- Se aplica una sustracción de la media por locutor.
- Se concatena a los 13 parámetros resultantes los parámetros de las 4 tramas anteriores y las 4 posteriores hasta construir un vector de dimensión 117.

- Al resultado se le aplica análisis discriminativo lineal (LDA - linear discriminant analysis) para reducir la dimensionalidad a 40. Para la estimación LDA usa los estados HMM de fonemas con contexto.

- Al vector de dimensión 40 se le aplica una transformación lineal de máxima verosimilitud (MLLT - Maximum likelihood linear transform). Esta transformación es una transformación ortogonalizadora que hace que los parámetros sean modelados de manera más fiel por Gaussianas de covarianza diagonales.

- Por último, se aplica una regresión lineal de máxima verosimilitud al espacio paramétrico global (fMLLR - global feature-space maximum likelihood linear regression) que sirve para normalizar la variación entre locutores.

La salida final del LVCSR es una serie de "redes" de palabras o lattices que contienen la transcripción más probable de los audios introducidos. Sobre estas lattices se aplican las operaciones pertinentes para conseguir la transcripción final.

A la hora de describir un sistema de reconocimiento es importante explicar con qué material se han entrenado los modelos acústicos. También es necesario detallar cuál es la composición del diccionario de reconocimiento así como proporcionar información de si se hace uso de un modelo de lenguaje, y en caso afirmativo, cómo se ha construido.

4.1.1 Modelos acústicos

El corpus principal usado en el entrenamiento de los modelos acústicos es la parte en castellano de la base de datos del Parlamento Vasco. Este subset contiene las grabaciones de 47 sesiones parlamentarias que tuvieron lugar en el Parlamento Vasco con sus correspondientes transcripciones [1].

Se ha realizado un trabajo preliminar para separar las intervenciones de castellano de las de euskera. Como resultado se tienen 124 horas de voz pronunciadas por 84 locutores diferentes, 45 hombres y 39 mujeres. Sin embargo, el material perteneciente a los locutores masculinos es más del doble que el de las locutoras

[1] Esta base de datos se está desarrollando actualmente por el grupo de investigación GTTS de la UPV/EHU, contacto german.bordel@ehu.eus

femeninas. Además, hay que decir que la separación no es perfecta, algunas palabras en euskera aparecen en los audios de la parte castellana de la base de datos.

A la base de datos del Parlamento Vasco se añade como material de entrenamiento aproximadamente 4 horas de voz extraídas de 5 archivos de audio en castellano de los workshops MAVIR [83] que tuvieron lugar en 2007, 2008 y 2009.Este material se obtuvo al participar en la evaluación Spoken Term Detection (STD) Albayzin 2016 [111]. Junto con el audio los organizadores también proporcionaron las transcripciones correspondientes.

4.1.1.1 **Diccionario**

El diccionario usado por el LVCSR está compuesto únicamente por palabras. Estas palabras se extraen a partir de las transcripciones de los datos de entrenamiento 4.1.1. Para obtener la transcripción fonética de cada palabra se utilizó un transcriptor de castellano desarrollado en el laboratorio. Ciertas transcripciones se corrigieron manualmente (como palabras en otro idioma). Para evitar ambigüedades a la hora de evaluar los resultados del reconocimiento, todas las palabras se pasaron a mayúsculas. El resultado final consiste en un diccionario de 37723 entradas. El diccionario tiene el siguiente aspecto:

```
. . .
ABAJO a b a x o
ABANDERANDO a b a n d e r a n d o
ABANDERARSE a b a n d e r a r s e
ABANDONA a b a n d o n a
. . .
```

4.1.2 **Modelo de lenguaje**

Para entrenar el modelo de lenguaje que usa el reconocedor, se ha usado la parte en castellano del corpus paralelo de las Actas del Parlamento Europeo de los años 1996 a 2011 [62]. Este material consiste en 2123835 frases y 54806927. Se ha pasado todo el texto a mayúsculas (como en la sección del diccionario, 4.1.1.1), y se ha aplicado un proceso de normalización a los números. Después, el texto

se suministra a la herramienta SRILM [106] para crear un modelo de lenguaje de trigramas en formato arpa.

4.1.3 Evaluación del sistema ASR

Para evaluar el funcionamiento del sistema ASR implementado, se efectuó un reconocimiento sobre frases de test de la parte en castellano de la base de datos del Parlamento Vasco. En concreto, el material de test se compuso de 18478 frases enunciadas por 25 locutores distintos (13 hombres y 12 mujeres). Esto se traduce en algo más de 75 horas de audio de habla parlamentaria.

Con los modelos acústicos, el modelo de lenguaje y el lexicón descritos anteriormente, se consiguió un WER del 11.24 %. Este valor es relativamente pequeño. Es un número que permite afirmar que el reconocedor construido se comporta de una manera satisfactoria. Es cierto que el tipo de frases analizadas es el mismo que el utilizado para entrenar los modelos acústicos, pero el material de test es lo bastante variado para que esta prueba sea significativa. Por tanto, este resultado valida la decisión de utilizar este reconocedor para evaluar las técnicas de conversión que se desarrollan en esta tesis.

4.2 Evaluación Albayzin

Con el objetivo de validar el sistema de reconocimiento preparado, se decidió participar en la evaluación Spoken Term Detection (STD) Albayzin 2016 [111]. Esta evaluación consiste en crear un sistema para localizar una serie de términos dentro de archivos de audio. En concreto, la entrada al sistema es una lista de términos que son desconocidos a la hora de procesar el audio donde se tienen que buscar.

Un bloque fundamental de estos sistemas es la detección de términos hablados o STD, que es definido por el NIST como "la búsqueda de un término hablado en archivos sonoros grandes y heterogéneos" [40]. Es un área que ha tenido mucha atención últimamente, tal y como se muestra en[123], [1], [59], [86], [110], [112] o [114].

El sistema presentado a la evaluación se describe en [104] y se compone de dos módulos: Uno de ASR y otro de STD que se encarga de buscar el término sobre los audios reconocidos.

4.2.1 Módulo ASR

El reconocedor utilizado en el sistema desarrollado es el descrito anteriormente en 4.1, con unas pequeñas modificaciones. Los modelos acústicos son los mismos, sólo se cambian el diccionario y el modelo de lenguaje.

Del diccionario original se eliminan todas las palabras definidas como fuera de vocabulario (OOV - out of vocabulary) por los organizadores de la evaluación, según se exponía en las normas. Como resultado queda un lexicón con 37636 entradas. Por la misma razón, estas mismas palabras OOV se eliminan del modelo de lenguaje.

Otra de las diferencias es que se crea un segundo modelo de lenguaje con el objetivo de refinar la búsqueda, ya que como resultado del reconocimiento se crearán lattices distitas. Este modelo de lenguaje está basado en unigramas y se crea sólo con las palabras que aparecen en el lexicón. Además, utilizando SRLIM se hace que todas las palabras sean igual de probables.

4.2.2 Módulo STD

Al igual que el bloque de reconocimiento, el módulo STD también se construye utilizando las herramientas de Kaldi, aunque se han utilizado diferentes estrategias. Estas estrategias dependen de si el término a buscar aparece en el lexicón del sistema (INV) o no (OOV):

4.2.2.1 INV: Palabras en vocabulario

Para los términos recogidos en el vocabulario, se utiliza el módulo de búsqueda de palabras clave (KWS - Key Word Search) que incluye Kaldi. Este módulo procesa las lattices que genera el LVCSR aplicando las técnicas de indexación de lattices descritas en [11]: Las lattices resultantes del reconocimiento del audio que contiene los términos a buscar se convierten de transductores de estados finitos ponderados (WFST - weighted finite state transducers) en una única estructura de transductor de factor generalizado. Este transductor de factor es realmente un índice invertido de todas las secuencias de palabras que aparecen en las lattices. Está compuesta por el tiempo de inicio, el tiempo final y las probabilidades a posteriori de cada palabra. Para buscar un término en el índice se construye una máquina de estados finitos (FST - finite state transducer) con ese mismo término que se someterá a una operación de composición con el transductor de factor para obtener todas las apariciones del término en la lattice, así como la identidad de la frase, el tiempo de inicio y de final y las probabilidades a posteriori de cada aparición. Todas estas apariciones se ordenan en base a las probabilidades a posteriori, dándose una decisión SÍ/NO a cada instancia.

Se utiliza una estrategia de dos pasadas. En la primera, las lattices que se le pasan al módulo KWS de Kaldi se obtienen del reconocimiento usando el modelo de lenguaje de trigramas. De esta manera, se usa toda la información entre la relación que existe entre palabras, por si el término a encontrar está compuesto por varias palabras. Además, en esta primera pasada se aplica un postprocesado muy simple a los resultados de esta búsqueda: Si el número de ocurrencias de un término está por encima de un cierto umbral t, todas las apariciones que tengan una probabilidad mayor que un cierto valor s se etiquetan como aciertos aunque inicialmente estén

identificados por el módulo KWS como errores. Los valores de t y s se han elegido de manera empírica utilizando los datos de desarrollo: se intenta maximizar el número de términos detectados correctamente y minimizar las falsas aceptaciones introducidas.

Para encontrar los términos de los que no aparecen instancias en la primera búsqueda se hace una segunda pasada. En esta segunda repetición, las lattices que recibe el módulo KWS son las que se obtienen al reconocer los audios de test usando el modelo de lenguaje de unigramas construido con las entradas del diccionario. El objetivo es minimizar el efecto del modelo de lenguaje para favoreces la importancia de los modelos acústicos. Los términos detectados en la segunda pasada se unen a los resultados del primer pase.

4.2.2.2 **OOV: Palabras fuera de vocabulario**

La estrategia a la hora de encontrar los términos OOV se basa en el método de palabras cercanas descrito en [14]. Si una palabra no aparece en el diccionario del sistema, no aparecerá tampoco en las lattices resultantes del reconocimiento. Un enfoque razonable es buscar palabras que sean acústicamente parecidas a las OOVs, pero que aparezcan el el lexicón del LVCSR, es decir, utilizarlas como palabras clave cercanas en vez de las palabras clave OOV originales.

Como en el caso de los términos INV, se utiliza una estrategia de dos pasadas. Primero se sintetizan los términos OOV y se reconocen para crear las FSTs de palabras cercanas necesarias para usar el módulo KWS de Kaldi. Después, todos los términos de los que no se ha conseguido encontrar ninguna aparición después del primer pase se buscan mediante una descomposición silábica.

- Síntesis texto a voz.

  En esta estrategia, todos los términos OOV se sintetizan utilizando el sintetizador de texto a voz (TTS - Text-to-Speech) de Aholab [96]. La señales sintéticas generadas se pasan al LVCSR, que utiliza el modelo de lenguaje de unigramas para obtener las lattices correspondientes. De estas lattices se elige la mejor hipótesis para que sea el término clave a buscar. El objetivo es obtener el término INV más parecido acústicamente para cada término OOV. Del mismo modo que en el caso de los términos INV, las FSTs se construyen

a partir de los términos clave y se entregan al módulo KWS. Este módulo utiliza las lattices obtenidas de reconocer los audios de test con el modelo de lenguaje de unigramas como la otra entrada para realizar la detección de los términos claves.

La posibilidad de utilizar más de una hipótesis se descartó porque los resultados mostraban un número muy alto de falsas aceptaciones.

- Descomposición silábica.

  En este caso, el reconocimiento del audio de test se hace utilizando el modelo de lenguaje de unigramas, sin cambiar el diccionario de palabras. Sólo se utiliza la mejor hipótesis para calcular el alineamiento a nivel de palabras. De esta manera se tiene una transcripción con el instante de inicio y de final de cada palabra. El siguiente paso es descomponer las palabras de la transcripción es sílabas. Los términos OOV también se descomponen en sílabas y se calcula una medida de la diferencia entre las sílabas de cada término y las de toda la transcripción. Se desliza una ventana del tamaño de las sílabas del término a buscar sobre la transcripción y la diferencia se calcula en base a la transcripción fonética de las sílabas. Los lugares en los que esta diferencia es mínima se toman como una aparición del término buscado. Este proceso se muestra en la figura 4.1. Las sílabas del término OOV se comparan con las sílabas de la transcripción contenidas en la ventana. El tamaño de la ventana es el número de sílabas del término OOV buscado. Cuando el parecido es grande, la puntuación aumenta indicando una posible aparición del término OOV. La puntuación final obtenida para cada ventana de búsqueda se calcula como:

  $$1 - d_L \tag{4.1}$$

  donde d_L es la distancia de Levensthein.

  La distancia de Levensthein [67] es una métrica de cadenas de texto que mide la diferencia entre dos secuencias. En otras palabras, la distancia de Levensthein entre dos palabras es el mínimo número de ediciones de carácter (inserciones, eliminaciones y sustituciones). En nuestro caso, cada carácter representa un fonema. El coste o penalización asignado a cada inserción o

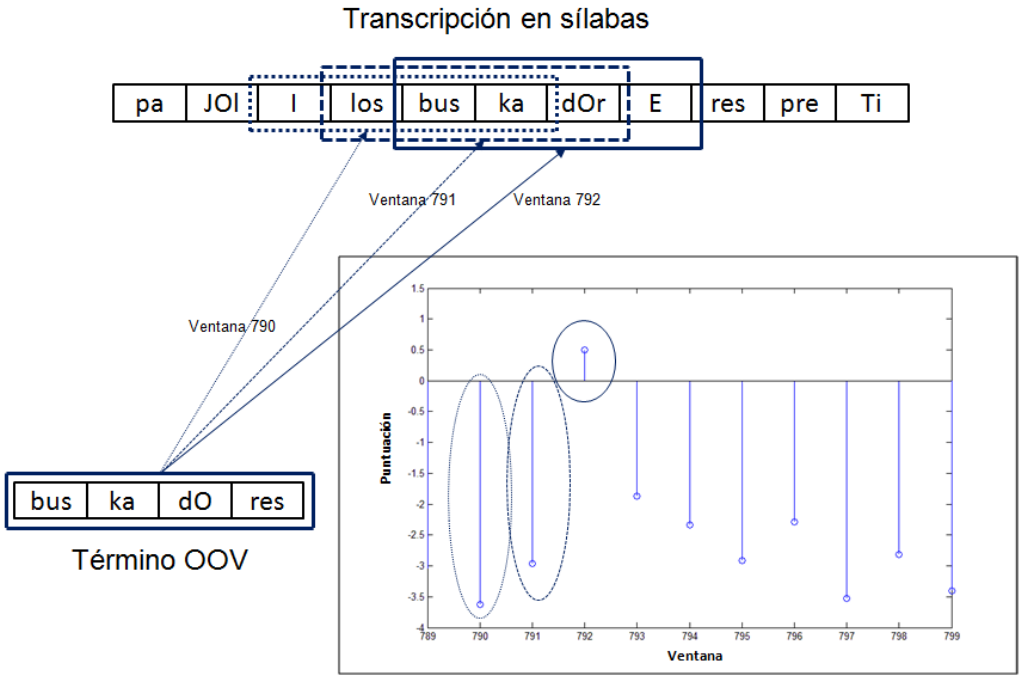

Figura 4.1: Búsqueda de términos OOV por descomposición silábica.

eliminación de un fonema de la sílaba se establece en $0{,}50$. La penalización por reemplazar un fonema por otro puede variar de 0 a 1 en base a la diferencia acústica estimada entre ambos fonemas calculada previamente. Este coeficiente se estima a partir de las bases de datos de entrenamiento con el siguiente procedimiento:

- Utilizando la herramienta llamada Hidden Markov Model Toolkit [127] se entrenan modelos HMM de monofonemas a partir de las bases de datos de entrenamiento. Los HMMs resultantes consisten en modelos de tres estados emisores y de dimensión 39 por cada fonema (24 en este caso).
- El vector de medias del estado central de cada modelo de fonema se elige para calcular la distancia euclídea entre fonemas. Este valor se toma como una estimación de la distancia acústica.
- Las distancias resultantes se normalizan en base a la máxima distancia. Obviamente, la diferencia entre dos fonemas iguales se establece como cero.

Una vez está calculada la distancia se debe establecer una distancia umbral para considerar un término como detectado. Este valor se eligió empíricamente basándose en la puntuación media y la desviación estándar de cada término evaluado sobre toda la transcripción. Este umbral se seleccionó utilizando los datos de desarrollo y se eligió un valor muy conservador para minimizar la inserción de falsas apariciones.

4.2.3 **Resultados**

La métrica utilizada para medir el funcionamiento del sistema es el llamado valor ponderado del término real (ATWV - Actual Term Weighted Value)[41]. Todas las decisiones tomadas durante la construcción del sistema han tenido como objetivo maximizar este valor.

4.2.3.1 Resultados sobre los datos de desarrollo

La evaluación final sobre los datos de desarrollo se muestran el las tablas 4.1, 4.2 y 4.3.

Tabla 4.1: Funcionamiento del sistema STD sobre los datos de desarrollo.

Ref	Corr	FA	Fallos	P(FA)	P(Fallo)	ATWV
1014	624	117	390	0.045	0.385	0.573

La tabla 4.1 muestra el funcionamiento global del sistema sobre los datos de desarrollo. Se encuentran correctamente (Corr) 624 ocurrencias de los términos a buscar y se introducen 117 falsas aceptaciones (FA). La probabilidad de detectar incorrectamente un término (P(FA)) es de 0.45, y la probabilidad de dejar de encontrar un término (P(Fallo))es de 0.358. El ATWV del sistema sobre los datos de desarrollo es de 0.573.

Tabla 4.2: Funcionamiento del sistema STD sobre los datos de desarrollo para los términos INV.

Términos INV	Ref	Corr	FA	Fallos	P(FA)	P(Fallo)	ATWV
Tras la 1ª pasada	668	545	68	123	0.023	0.184	0.721
Tras la 2ª pasada	668	554	85	114	0.029	0.171	0.756

La tabla 4.2 Muestra los resultados obtenidos sólo para los términos INV. Considerando sólo la primera pasada (i.e, el ASR utilizando el modelo de lenguaje de trigramas) el valor del ATWV es 0.721.Tras la segunda pasada(ASR con LM de unigramas) se consigue una mejora del $4,85\,\%$.

Tabla 4.3: Funcionamiento del sistema STD sobre los datos de desarrollo para los términos OOV.

Términos OOV	Ref	Corr	FA	Fallos	P(FA)	P(Fallo)	ATWV
Tras la 1ª pasada	346	62	27	284	0.008	0.821	0.213
Tras la 2ª pasada	346	70	32	276	0.009	0.798	0.272

En la tabla 4.3 se encuentran los resultados para la búsqueda de los términos OOV. No se encuentran un número importante de términos lo que afecta negativamente al resultado final del sistema. Con el segundo pase se detectan algunas nuevas ocurrencias y el ATWV mejora ligeramente.

Estos resultados reflejan el hecho de que el sistema se ha ajustado para minimizar las falsas aceptaciones. En consecuencia, el número de términos no encontrados es grande.

4.2.3.2 **Resultados sobre los datos de test final**

El resultado final de la evaluación se hizo utilizando dos bases de datos distintas llamadas *MAVIR* y *EPIC*. Los resultados calculados por los organizadores se muestran en la tabla 4.4. Son valores muy dispares entre sí. Aunque los resultados para EPIC conseguidos son muy buenos, para MAVIR son peores que los de desarrollo. Estos valores hicieron que el sistema presentado por Aholab quedase en segunda posición. Los resultados más detallados de esta evaluación se explican en el artículo [113].

Tabla 4.4: Funcionamiento del sistema STD sobre los datos de test final.

Base de datos	P(FA)	P(Fallo)	ATWV
MAVIR	0.00007	0.414	0.5090
EPIC	0.00004	0.156	0.8023

4.2.3.3 **Conclusiones**

El funcionamiento del sistema STD es altamente dependiente del módulo ASR. Si la transcripción del audio es fiel, la búsqueda de los términos será también más precisa, incluso para las palabras fuera de vocabulario: las palabras que aparecen en la transcripción sustituyendo las palabras que no aparecen en el diccionario serán más parecidas acústicamente a las palabras "cercanas" que vamos a buscar.

El uso de un modelo de lenguaje de unigramas para una segunda pasada en la búsqueda de los términos INV no encontrados en un primera pasada ha demostrado

ser una estrategia exitosa. Se encuentran nuevos términos con una subida en la tasa de falsas aceptaciones aceptable.

En la tarea de detección de términos OOV el uso del TTS da buenos resultados. Se detectan adecuadamente muchas palabras OOV manteniendo una baja tasa de FA. Parece que dejar al ASR seleccionar la palabra cercana más parecida acústicamente es una buena idea.

Los resultados del método de descomposición silábica no son tan buenos como lo esperado, principalmente porque la selección de umbral es un problema complicado. Al relajar su valor aumenta el número de términos detectados, pero al mismo tiempo sube la tasa de FA y empeora el funcionamiento del sistema. Por tanto, se decidió escoger un umbral muy estricto.

4.3 Reconocimiento de las voces esofágicas

El sistema de reconocimiento diseñado en la sección 4.1 se ha utilizado para evaluar el reconocimiento de voces esofágicas. Para ello ha sido necesario realizar ciertas adaptaciones que se describen en este apartado, junto con los resultados obtenidos para las voces esofágicas. Estos resultados servirán de base para evaluar el progreso realizado en los experimentos de mejora de la inteligibilidad, cuyo principal objetivo reside precisamente en mejorar el reconocimiento automático de estas voces.

4.3.1 Empleo de un ASR estándar

El primer experimento realizado consiste en utilizar el sistema previamente descrito con las señales de nuestra base de datos. Para tener una referencia, los resultados se comparan con los obtenidos para 9 locutores con voz sana y para el mismo conjunto de frases.

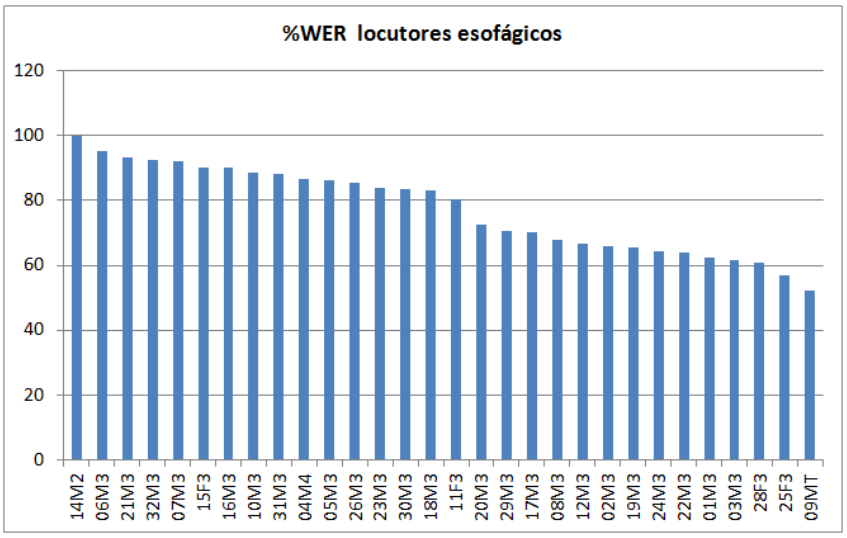

Figura 4.2: WER para las 29 sesiones de voz esofágica con las 100 frases en castellano.

Las figuras 4.2 y 4.3 muestran los resultados de reconocimiento para las 100 frases de ZureTTS de ambos grupos de hablantes. Los valores de WER son muy

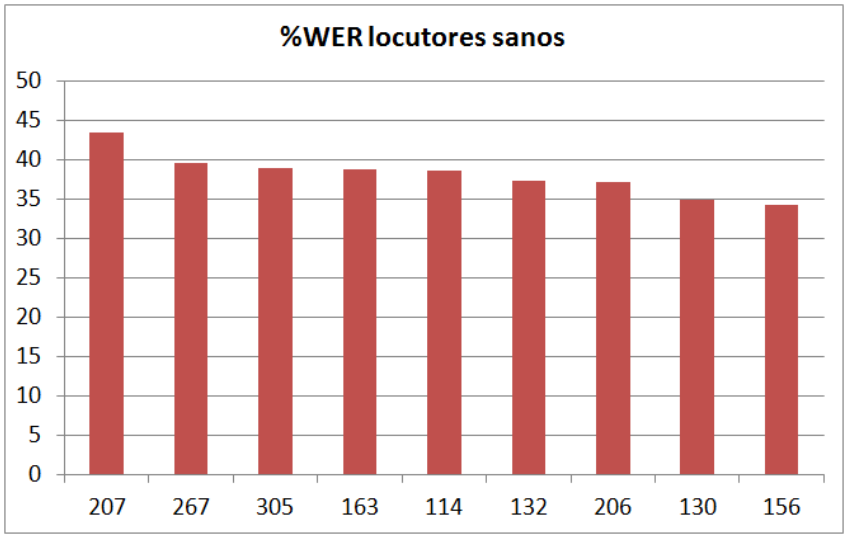

Figura 4.3: WER de 9 sesiones de voz sana con las mismas 100 frases en castellano utilizadas como referencia.

altos. Para las señales esofágicas la media del error es del 77.37 % con una desviación típica del 13.42. Esto era esperable, pero si se analiza también el error de las señales sanas se tiene que el error medio es del 38 % con una desviación del 2.7. Estos valores de error están muy por encima del obtenido al evaluar el ASR con las señales sanas pertenecientes a la base de datos del Parlamento Vasco. La explicación a este comportamiento extraño es que hay un gran número de palabras fuera de vocabulario, hasta un 23 %. Esto es debido a la naturaleza del corpus grabado. Se trata de un corpus fonéticamente equilibrado de sólo 100 frases, así que para conseguir este equilibrio se incluyen muchas palabras poco comunes y nombres propios de personas y lugares.

4.3.2 Empleo de un ASR con diccionario reducido

Con el objetivo de reducir el impacto de las OOV, se decidió realizar un segundo experimento. Se cambió el lexicón por otro que contenía únicamente las palabras que aparecen en las frases a reconocer. En total son 701 palabras. Este cambio en el diccionario obliga a cambiar también el modelo de lenguaje usado. Se decidió

crear un modelo de lenguaje basado en unigramas equiprobables que contenía sólo las palabras del lexicón.

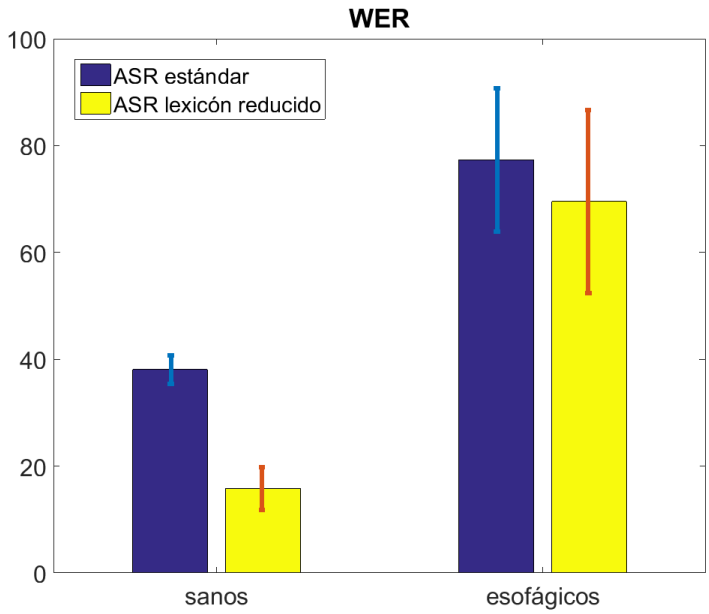

Figura 4.4: WER medio para las 9 sesiones de voz sana y las 30 esofágicas obtenido con el ASR estándar y el que hace uso del lexicón reducido. Las líneas sobre las barras muestran la desviación estándar.

El WER resultante de volver a reconocer las mismas 100 frases para los locutores sanos y esofágicos se puede ver el la figura 4.4. Como era de esperar, el WER medio de los locutores sanos se ha reducido del 38 % al 15.80 %, una reducción cercana al 23 % de palabras OOV del reconocedor estándar. Este valor está más cerca del obtenido para las frases de datos del Parlamento con estos mismos modelos acústicos, pero diferente lexicón y modelo de lenguaje. Para los locutores esofágicos también se da una reducción del error (del 77.37 % al 69.53 %), pero la variabilidad en los resultados de los diferentes locutores sigue siendo muy alta.

La figura 4.5 muestra la diferencia de WER para cada experimento (la diferencia positiva indica que hay una mejora). Como se puede comprobar, para algunos locutores no ha habido ninguna mejora pese a reducir el diccionario, lo que implica

que para el ASR la señal que entra no se parece en nada a los modelos acústicos que usa.

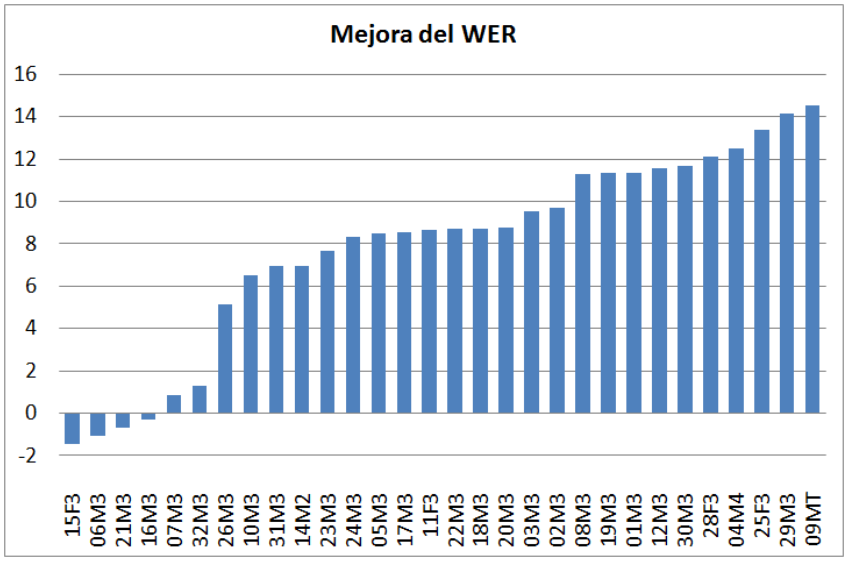

Figura 4.5: Diferencia entre el WER obtenido con el ASR estándar y el ASR con el lexicón reducido para cada sesión de habla alaríngea. Valores positivos implican una mejora del WER.

4.3.2.1 **Análisis del error**

Con el fin de realizar un mejor análisis del WER de las voces esofágicas se decidió separar a los hablantes patológicos en dos grupos diferentes. Para ello , a partir del WER de cada uno de los locutores se utilizó el algoritmo k-means [71] para hacer dos grupos. El resultado son 2 clústeres, uno con los mejores 14 locutores y el otro con los 16 restantes (los que presentan peor WER).

En la figura 4.6 se puede ver que utilizando esta separación entre los mejores y los peores hablantes esofágicos la variabilidad del WER de cada grupo está más acotada. Parece que la distinción hecha entre ambos grupos de “buenos” y “malos” hablantes esofágicos tiene sentido, aun en el caso de que todos estén clasificados en la misma fase del aprendizaje.

En el grupo de “malos” locutores el WER (con el diccionario pequeño) se mueve de un máximo del 96.21 % a un mínimo del 71.61 %. El WER medio tiene un

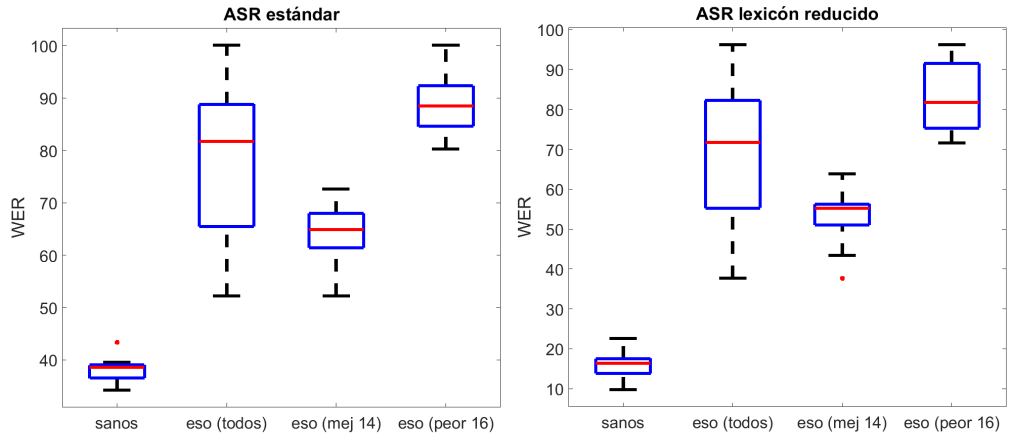

Figura 4.6: WER para los grupos de hablantes sanos y esofágicos obtenidos para el ASR estándar (izquierda) y con lexicón reducido (derecha). Para el caso de los locutores esofágicos se muestran también los resultados al ser separados en 2 grupos distintos. En cada caja,la línea central es la mediana, los bordes de la caja representan los percentiles 25º y 75º, los bigotes se extienden a los valores más extremos no considerados outliers, y los outliers se muestran individualmente como una cruz roja.

valor de 83.63 %. La inteligibilidad de estos locutores es muy mala, es muy difícil entender lo que están diciendo. Hacen un esfuerzo grande para decir las frases y, en muchos casos, se quedan sin aire, Algunas palabras se articulan, pero mediante voz susurrada y otras se dejan incompletas, perdiéndose su final.

El grupo con los 14 mejores locutores tiene un WER máximo del 63.89 % y uno mínimo del 37.7 % (también para el lexicón reducido). El locutor que obtiene el error de reconocimiento más bajo es el traqueoesofágico, siendo el siguiente mejor error de 43.38 %, un 5.7 % peor. Dejando de lado el locutor de este grupo con mayor WER (que cometió muchos errores de lectura), el resto de locutores presentaron un grado razonable de inteligibilidad. Estos locutores controlan mejor la respiración, y tanto su articulación como su velocidad de habla son mejores. El WER medio para este grupo es 53.42 %.

Los resultados de reconocimiento de todas las sesiones se incluyen en el anexo E.

4.3.3 Empleo de un ASR con modelos acústicos esofágicos

Otro de los experimentos que se llevaron a cabo fue entrenar nuevos modelos acústicos con las grabaciones esofágicas realizadas. El objetivo de este experimento es ver cómo funciona el reconocedor cuando se entrena con este tipo de voces. Esta solución consiste en modificar el reconocedor por lo que no es la buscada en esta tesis.

Para construir este nuevo reconocedor se vuelve a usar Kaldi. El diccionario vuelve a ser el reducido con las 701 palabras del vocabulario de las 100 frases y el modelo de lenguaje el de unigramas equiprobables. El cambio se produce en los modelos acústicos. Para entrenarlos se utilizan las 30 sesiones en las que se han grabado las 100 frases en castellano. Como el material es escaso y se quiere tener 100 frases sobre las que evaluar el reconocedor se ha recurrido a una validación cruzada. En este caso, se crean tres bloques de 10 locutores y se seleccionan 2 de estos bloques para entrenar y 10 para evaluar. Este proceso se repite tres veces para poder evaluar todas las sesiones.

La división en bloques se hace observando los resultados de WER obtenidos para estas mismas 30 sesiones con el reconocedor con el diccionario reducido (apartado 4.3.2) tratando de que haya locutores con resultados de WER similares en todos los bloques. La composición de los bloques se puede ver en la tabla 4.5.

Con esta división, los resultados de WER pasan de una media del 69.53 % al 47.47 % (figura 4.7). Si se toma otra vez la separación en locutores "buenos" y "malos" hecha en el Apartado 4.3.2.1, se tiene que para el primer grupo la media del WER es del 33.70 %, mientras que para el segundo es del 59.52 %. Como era de esperar, la mejora es palpable. Para el grupo de locutores "buenos" los valores de WER no están demasiado alejados de los resultados del grupo de voces sanas (que era del 16 %). Si se dispusiera de más material de habla esofágica con la que entrenar los modelos acústicos no es de extrañar que estos valores de WER se redujesen aún más. Excepto para un locutor considerado outlier, todos los resultados de WER mejoran, pero no ocurre que mejoren más para uno de los dos grupos de locutores, son bastante heterogéneas (figura 4.8). Estas mejoras varían desde un 8.93 % (del 37.70 al 28.77 %) a un 36.71 % (96.21 % al 59.50 %).

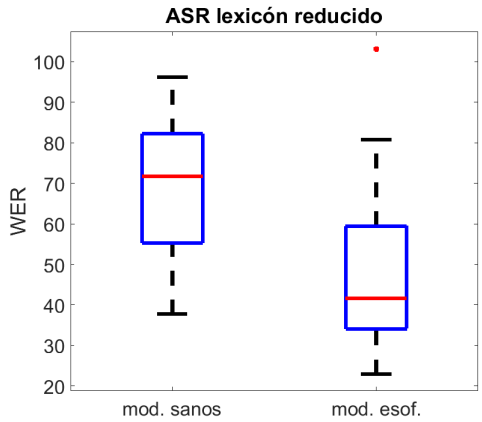

Figura 4.7: WER obtenido para las 30 sesiones utilizando modelos acústicos entrenados con voces sanas (izquierda) y con voces esofágicas (derecha). En cada caja,la línea central es la mediana, los bordes de la caja representan los percentiles 25º y 75º, los bigotes se extienden a los valores más extremos no considerados outliers, y los outliers se muestran individualmente como una cruz roja.

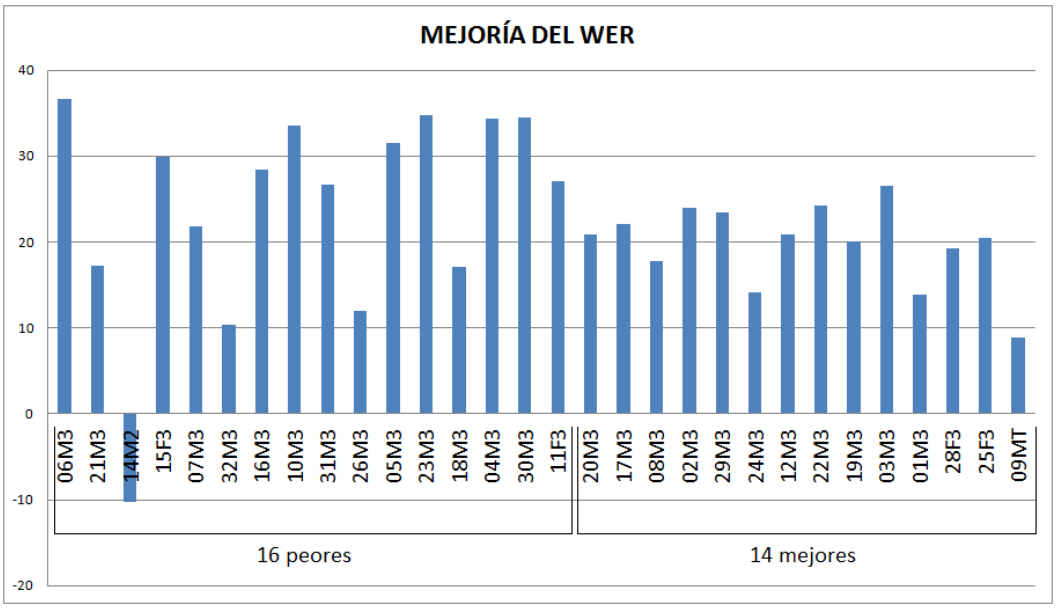

Figura 4.8: Mejora en el WER al cambiar los modelos acústicos del ASR por uno entrenado con voces esofágicas.

Tabla 4.5: Composición de los bloques utilizados para hacer validación cruzada al utilizar un ASR con modelos acústicos de voces esofágicas.

Bloque 1	Bloque 2	Bloque 3
04M3	02M3	01M3
05M3	03M3	06M3
07M3	11F3	08M3
15F3	12M3	09MT
16M3	17M3	10M3
18M3	24M3	14M2
26M3	22M3	19M3
30M3	25F3	20M3
31M3	28F3	21M3
32M3	29M3	23M3

Si se aplica el algoritmo k-means para separar otra vez los locutores en 2 grupos a partir de estos resultados de reconocimiento con los modelos esofágicos se obtienen un grupo de 20 “buenos” locutores y otro de 10 “malos” que son distintos a la separación anterior. Para estos grupos se muestran los resultados en la figura 4.9: el grupo de locutores con mejores resultados tiene un WER medio del 36.51 % y el de peores 10 del 69.38 %.

Este experimento demuestra que cambiar los componentes de los ASR haría que los locutores alaríngeos fuesen mejor entendidos por los reconocedores automáticos. Lamentablemente, hacer dichos cambios no es una opción válida en la mayoría de los casos, así que hay que buscar otras soluciones.

Los resultados de este experimento se encuentran recogidos en el anexo E

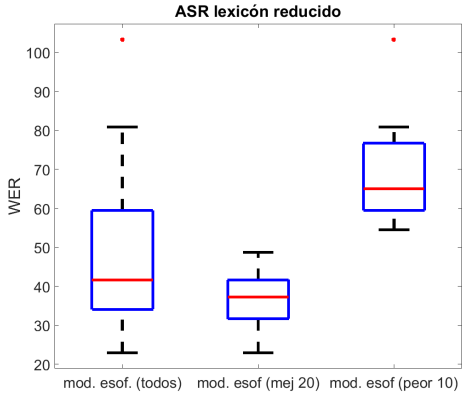

Figura 4.9: WER usando el ASR con modelos acústicos entrenados con voces esofágicas (izquierda). Se muestra también los resultados al separar los 30 sesiones en 2 grupos con los 20 mejores resultados (centro) y los 10 peores (derecha). En cada caja,la línea central es la mediana, los bordes de la caja representan los percentiles 25º y 75º, los bigotes se extienden a los valores más extremos no considerados outliers, y los outliers se muestran individualmente como una cruz roja.

4.4 Evaluación de la inteligibilidad de los locutores esofágicos

Las voces esofágicas son más difíciles de entender que las sanas. En el artículo presentado en [91] hemos intentado cuantificar esta dificultad a la hora de entender el habla esofágica midiendo la inteligibilidad y el esfuerzo de escucha[79]. La inteligibilidad se puede medir en el contexto del reconocimiento automático del habla (ASR) o del reconocimiento humano del habla (HSR - human speech recognition).

La principal ventaja del ASR es que la medida es objetiva y fácil de implementar. Sin embargo, sólo evalúa la inteligibilidad que percibiría la máquina, no como de entendible es para una persona. Además, este tipo de medida no tiene en cuenta otros factores importantes como la agradabilidad, la aceptabilidad o el esfuerzo de escucha. Sin embargo, no hay trabajo hecho con el HSR a nivel de frase para voces esofágicas en castellano.

La desventaja de las medidas de inteligibilidad es que sólo indican cuantas palabras han sido identificadas correctamente, pero no cómo de difícil ha sido identificarlas. Por eso se ha intentado medir la inteligibilidad de las voces esofágicas no sólo en términos de ASR y HSR, sino también de esfuerzo de escucha. Además, podría ser que la inteligibilidad fuese evaluada de manera diferente por oyentes expertos e inexpertos. Esta posibilidad se sugiere en [17], aunque se concluye que los resultados de ambos grupos son similares. Por ello, también se ha querido investigar si para el conjunto de datos recogido en esta tesis los resultados son similares para oyentes familiarizados y para no familiarizados con el habla esofágica. Se consideran oyentes familiarizados a amigos, familia y parientes cercanos de los locutores esofágicos.

4.4.1 Metodología y experimentos realizados

Para hacer los experimentos se utilizó la base de datos descrita con detalle en el capítulo 3. Más concretamente, se escogieron 4 locutores esofágicos de dicha base de datos, tres masculinos (01M3, 02M3, 03M3) y uno femenino (25F3). También se escogió dos locutores de voz sana, uno masculino (114) y otro femenino (207).

4.4.1.1 Reconocimiento automático del habla

Este experimento consiste en utilizar el reconocimiento automático de las 100 frases de los locutores seleccionados. Para ello se utilizó el reconocedor desarrollado en la sección 4.3.2. Este reconocedor hace uso de un diccionario reducido para evitar problemas de palabras fuera de vocabulario. El modelo de lenguaje está construido con unigramas equiprobables para que la importancia del reconocimiento recaiga sobre los modelos acústicos.

4.4.1.2 Reconocimiento humano del habla

La principal tarea de este experimento consistió en lo que se denomina recordar y transcribir: Los participantes escuchan una frase y escriben lo que han entendido. Con estas transcripciones se calcula el WER y, en consecuencia, la inteligibilidad.

Para calcular el esfuerzo de escucha, los propios participantes evalúan el esfuerzo que han hecho para entender cada frase en una escala de 5 puntos. Las opciones son “muy poco”, “un poco”, “algo”, “bastante” y “mucho”. Para evitar el sesgo producido por escuchar más de una vez la frase o por el orden en que se presenta, las oraciones se reproducen sólo una vez y en orden aleatorio.

Para elegir las frases a escuchar se buscaron dentro del corpus aquellas cuya longitud las hiciese transcribibles. También se buscó que estuvieran fonéticamente equilibradas. Como resultado se obtuvo un conjunto de 30 frases, cada una de las cuales tiene un máximo de 10 palabras. Todas las frases se normalizaron a un valor pico común (0.8) para lograr un nivel de sonoridad homogéneo y cómodo.

Para el test se crearon seis conjuntos de frases mutuamente excluyentes. Cada conjunto contiene 30 frases diferentes, 5 frases de cada locutor. Como resultado, con 6 participantes se escucharían las 180 frases (30 frases de 6 locutores). Esto asegura igual cobertura para todas las frases y locutores. Cada participante en el test escucha uno de estos bloques en orden aleatorio, pero primeramente se le reproducen al participante dos frases (una de voz esofágica y otra de voz sana) a modo de entrenamiento para que se familiarice con la tarea. Estas frases son también de la base de datos, pero no forman parte de las 30 a evaluar.

4.4.2 Resultados y análisis

En el test participaron 57 evaluadores, hablantes nativos de castellano. 15 de ellos eran gente cercana a larigectomizados, de tal manera que estaban familiarizados con las voces esofágicas.

Antes de calcular el WER se hace una limpieza inicial de las transcripciones recogidas. Esto incluye eliminar signos de puntuación, caracteres especiales y errores tipográficos (vocales con tildes, uso de mayúsculas-minúsculas, ortografía de nombres propios o nombres extranjeros...).

4.4.2.1 Resultados del WER para el HSR

En la tabla 4.6 se muestran los resultados de las transcripciones del test para los oyentes familiarizados y no familiarizados con las voces esofágicas. El WER se ha calculado utilizando la distancia de Levenshtein, cuantificando las inserciones, sustituciones y eliminaciones que aparecen al comparar la transcripción de la frase hecha por los participantes en el test con la referencia original.

Tabla 4.6: WER del experimento HSR.

WER(%)	Esofágicas	Sanas
Familiarizados	17.39	7.42
No familiarizados	18.35	4.85
Media total	**17.87**	**6.16**

Como era de esperar, el WER es mucho más alto para el habla esofágica. No hay gran diferencia entre el WER calculado para el grupo de familiarizados y el de no familiarizados, lo que corrobora el estudio hecho en [17]. Para el habla sana hay una diferencia de cerca de 3 puntos entre ambos grupos, pero esta diferencia no es significativa. Esto puede comprobarse en la figura 4.10. En ella se muestra el WER medio para cada locutor con los intervalos de confianza al 95 %.

4.4.2.2 Esfuerzo de escucha

Como se ha explicado anteriormente, a la vez que el participante en el test transcribía la frase a escuchar autoevaluaba el esfuerzo hecho para entender la frase. Los

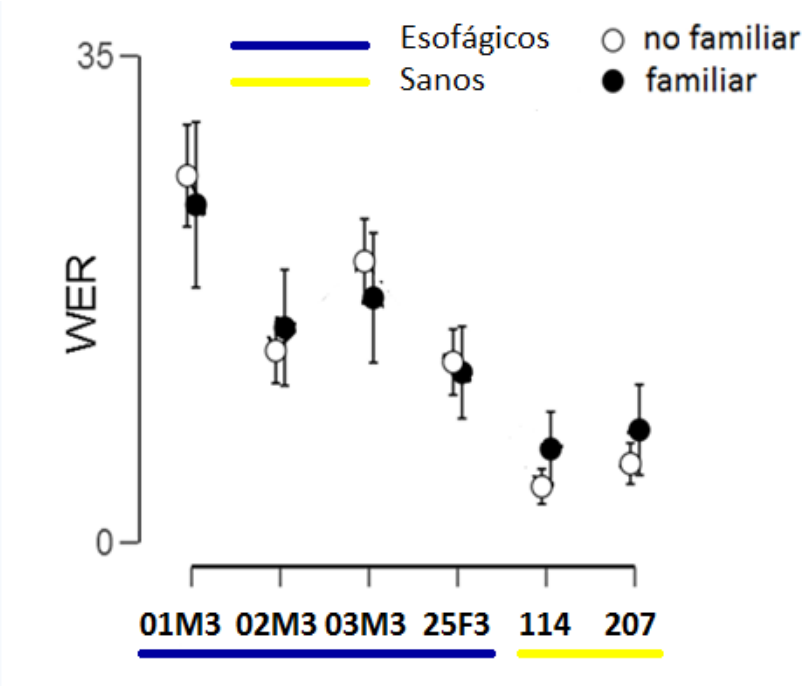

Figura 4.10: WER promedio por locutor para locutores esofágicos (01M3, 02M3, 03M3, 25F3) y saludables (114, 207). Las barras muestran los intervalos de confianza al 95 %.

resultados de este experimento se muestran el la tabla 4.7. La figura 4.11 muestra los mismos resultados pero por locutor. Como era de esperar, el esfuerzo percibido por los oyentes familiarizados es significativamente menor que para los que este tipo de habla no les es familiar.

Tabla 4.7: Esfuerzo de escucha del experimento HSR.

Esfuerzo	Esofágicas	Sanas
Familiarizados	2.61	1.25
No familiarizados	3.54	1.26
Media total	**3.07**	**1.255**

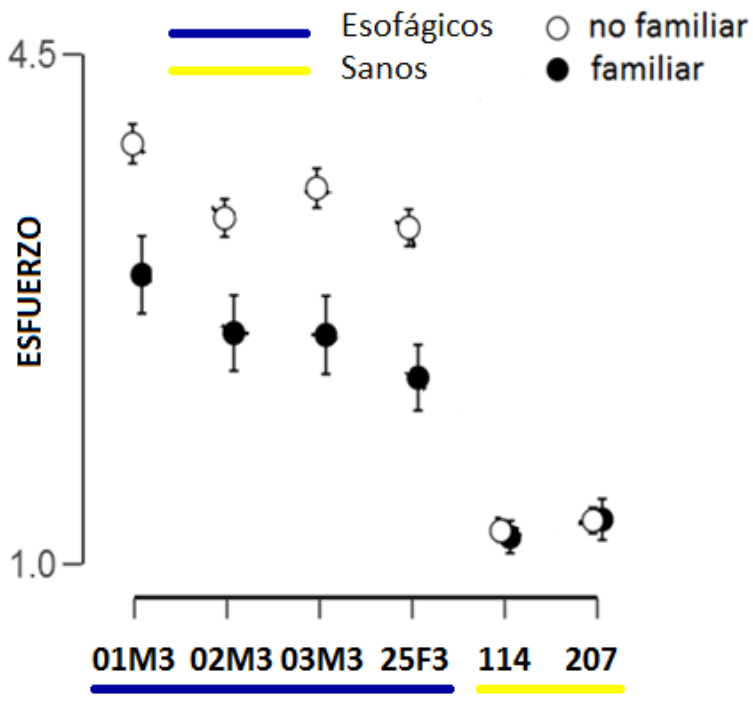

Figura 4.11: Esfuerzo de escucha promedio por locutor para locutores esofágicos (01M3, 02M3, 03M3, 25F3) y saludables (114, 207). Las barras muestran los intervalos de confianza al 95 %.

4.4.2.3 Correlación entre la inteligibilidad y el esfuerzo de escucha

La correlación entre la inteligibilidad (WER) y el esfuerzo de escucha autoevaluado es de $0{,}479$ (coeficiente de Pearson, $p < 0{,}001$). Es una correlación débil pero significativa que indica que las frases con mayor número de errores de transcripción se perciben como más difíciles de escuchar. Esta relación se muestra en la figura 4.12.

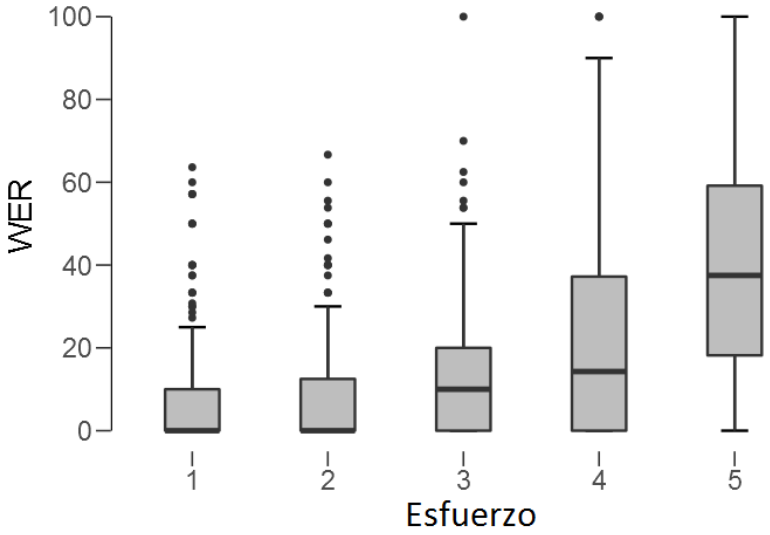

Figura 4.12: Correlación entre el WER y el esfuerzo de escucha.

4.4.2.4 Resultados del WER para el ASR

El experimento ASR se hace con las 100 frases disponibles de cada locutor. Esto hace que la medida de WER sea más fiable. En la figura 4.13 se puede ver que las tasas de error calculadas con el ASR son peores que las obtenidas en el experimento HSR tanto para las voces esofágicas como para las sanas. El hecho de que el reconocedor use un modelo de lenguaje de unigramas es uno de los motivos para este funcionamiento. Como era de esperar, el WER para los locutores esofágicos es significativamente mayor que el de los locutores de voz sana. Se puede ver que que el HSR y el ASR funcionan de manera distinta para cada locutor. Sin embargo el

número de locutores es demasiado pequeño como para llegar a ninguna conclusión extrapolable.

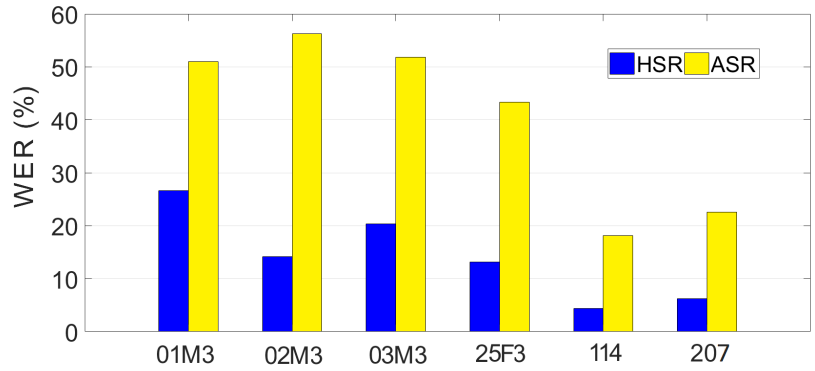

Figura 4.13: WER para HSR y ASR.

4.4.2.5 **Análisis de los resultados**

Según los resultados del HSR, las voces sanas son de media tres veces más inteligibles que las esofágicas. la esfuerza de escucha auto-evaluada también es tres veces mayor para los hablantes esofágicos que para la voz sana. Hay una correlación significativa entre inteligibilidad y esfuerzo. Aunque para ambos grupos de oyentes se tiene la misma inteligibilidad, el grupo de oyentes familiarizados con el habla alaríngea reportaron que hicieron menos esfuerzo al escuchar las frases que el grupo de no familiarizados.

4.5 Conclusiones

En este capítulo se ha explicado cómo se ha construido un reconocedor para el castellano basado en redes neuronales utilizando Kaldi. Este reconocedor ha demostrado tener una buena tasa de reconocimiento (WER del 11.24 %) para el reconocimiento de habla continua. También se ha demostrado su validez presentándose a la evaluación Albayzin STD 2016. Utilizando el reconocedor aquí diseñado como bloque fundamental del sistema presentado se consiguió quedar en segunda posición en dicha evaluación. Este trabajo ha dado origen a dos publicaciones, recogidas en [104] y [113].

Sin embargo, a la hora de reconocer las frases de la base de datos grabada, los resultados del reconocimiento no son buenos, incluso para hablantes de voz sana. Esto es debido a las características del corpus grabado: las frases contienen mucho vocabulario poco usual y con multitud de nombres propios que no están recogidos ni en el diccionario ni en el modelo de lenguaje. Por ello se ha modificado el diccionario del reconocedor limitándolo a las palabras que aparecen en las frases grabadas. También se ha cambiado el modelo de lenguaje a uno de unigramas equiprobables para centrar la importancia en los modelos acústicos.

Una vez hechos estos cambios, se ha calculado el WER de los hablantes esofágicos y se ha comparado los resultados con los obtenidos para un grupo de 9 locutores de voz sana. El WER de los hablantes laríngeos, además de ser evidentemente mucho más pequeño que el de los laríngeos (15.80 % frente a 69.53 %), está mucho menos disperso. Esta gran variabilidad ha llevado a separar a los hablantes esofágicos en dos grupos de "buenos" y "malos" locutores. Con esta separación se han conseguido valores de WER para cada uno de los grupos bien diferenciados, siendo el de los "buenos" locutores del 53.42 % y el de los "malos" del 83.63 %. Mientras que el primer grupo produce un habla con un mejor ritmo y una buena inteligibilidad, el segundo enuncia las frases de un modo más atropellado y controla peor la respiración. El WER ha servido por tanto para caracterizar a los locutores esofágicos.

Por último, se ha llevado a cabo un experimento para evaluar la inteligibilidad de 4 locutores esofágicos y dos sanos por parte de oyentes familiarizados y no familiarizados con las voces alaríngeas, así como el esfuerzo necesario para

ser entendidos. También se han evaluado a dichos locutores mediante ASR. Como era de esperar, los resultados del HSR muestran que la inteligibilidad de las voces laríngeas es mejor que la de las voces patológicas. Además, estos valores son iguales para el grupo de oyentes familiarizados con las voces esofágicas y para los no familiarizados. Sin embargo la autoevaluación del esfuerzo hecho por los familiarizados con el habla alaríngea por entender este tipo de voces ha sido significativamente menor que el esfuerzo reportado por los no familiarizados. También se ha encontrado una débil correlación ente el esfuerzo de escucha y los resultados de WER obtenidos en el test HSR. Para finalizar, decir que los resultados del ASR han sido peores que los del HSR aunque el experimento no ha sido lo suficientemente extensivo como para llegar a conclusiones relevantes al comparar ambas tasas de WER. Este experimento ha dado lugar a la publicación descrita en [91].

Comprender las cosas que nos rodean es la mejor preparación para comprender las cosas que hay mas allá.

Hipatia de Alejandría

CAPÍTULO

5

Preparación del sistema de conversión

En este capítulo se explican los esfuerzos realizados para implementar diferentes técnicas de conversión de voz para voces sanas ya que es el primer paso necesario para poder utilizar la conversión de voz para mejorar las voces esofágicas. A continuación se exponen los métodos desarrollados utilizando dos técnicas diferentes para afrontar la conversión: técnicas basadas en el entrenamiento estadístico de modelos de mezclas de gausianas (GMMs) y técnicas basadas en redes neuronales profundas. Para las técnicas estadísticas se han implementado y evaluado diferentes alternativas. Con estas técnicas se ha participado en competiciones internacionales, para evaluar su comportamiento. Además, también se han evaluado en el contexto de conversión aplicada a las interfaces silenciosas. Con respecto al empleo de redes neuronales profundas para conversión de voz, se describe un sistema que utiliza una red LSTM, que se ha evaluado y comparado con el sistema basado en GMM utilizando voces sanas. Como se verá en el capítulo 6, para la conversión de voces esofágicas se ha implementado otro sistema basado en el empleo de posteriorgramas fonéticos (PPGs). Al no haber sido evaluado formalmente con voces sanas, este sistema se describe en dicho capítulo.

5.1 Conversión con GMMs

Una de las técnicas más exploradas a la hora de realizar conversión de voz basada en métodos estadísticos es la utilización de GMMs: Se utiliza una mezcla de gaussianas que mapean las características del locutor origen con las del locutor destino. Las características a mapear en este caso han sido los coeficientes mel-cepstrales (MCEP) obtenidos mediante el vocoder Ahocoder. Para este trabajo se han implementado cuatro métodos diferentes para poder comparar los resultados:

- Joint-density modeling (JDM)
- GMM-weighted linear regression (WLR)
- Maximum-likelihood parameter generation (MLPG)
- MLPG with minimum generation error training (MGE)

Todos estos métodos tienen en común que utilizan como punto de partida una GMM conjunta de G componentes gaussianas entrenada a partir de los vectores de características del locutor origen, denominados $\{\mathbf{x}_t\}$, y los del locutor destino, denominados $\{\mathbf{y}_t\}$. Se construye una serie de vectores concatenados:

$$\mathbf{z}_t = [\mathbf{x}_t^\top \ \ \mathbf{y}_t^\top]^\top \tag{5.1}$$

Con ellos, se utiliza el algoritmo de esperanza-maximización (EM - expectation-maximization) para entrenar la GMM definida por los pesos $\{\alpha_g\}$, los vectores de medias $\{\boldsymbol{\mu}_g^{(z)}\}$y las matrices de covarianzas completas $\{\boldsymbol{\Sigma}_g^{(zz)}\}$, para $1 \leq g \leq G$. Como se explica en [58], los elementos de la GMM se pueden descomponer en sus partes origen y destino de la siguiente manera:

$$\boldsymbol{\mu}_g^{(z)} = \begin{bmatrix} \boldsymbol{\mu}_g^{(x)} \\ \boldsymbol{\mu}_g^{(y)} \end{bmatrix} \quad , \quad \boldsymbol{\Sigma}_g^{(zz)} = \begin{bmatrix} \boldsymbol{\Sigma}_g^{(xx)} & \boldsymbol{\Sigma}_g^{(xy)} \\ \boldsymbol{\Sigma}_g^{(yx)} & \boldsymbol{\Sigma}_g^{(yy)} \end{bmatrix} \tag{5.2}$$

De este modo, la probabilidad de que un vector origen $\mathbf{x}_t$ pertenezca a la clase g, se calcula a partir de $\{\alpha_g\}$, $\{\boldsymbol{\mu}_g^{(x)}\}$ y $\{\boldsymbol{\Sigma}_g^{(xx)}\}$, y se denomina como $\gamma_g(\mathbf{x}_t)$.

Una vez explicada la parte común a todos los métodos se describen cada uno de forma individual.

5.1.1 Joint-density modeling (JDM)

Este método transforma el vector de características fuente trama a trama. Fue propuesto originalmente en [58] y sigue la siguiente ecuación:

$$F(\mathbf{x}_t) = \sum_{g=1}^{G} \gamma_g(\mathbf{x}_t) \left[\boldsymbol{\mu}_g^{(y)} + \boldsymbol{\Sigma}_g^{(yx)} {\boldsymbol{\Sigma}_g^{(xx)}}^{-1} (\mathbf{x}_t - \boldsymbol{\mu}_g^{(x)}) \right] \tag{5.3}$$

en la que los vectores y matrices son los definidos en la ecuación 5.2. Como se puede deducir, una vez se tiene la GMM conjunta entrenada este método no necesita de entrenamiento extra.

5.1.2 GMM-weighted linear regression (WLR)

Esté método fue originalmente propuesto en [107]. Sin embargo, la formulación usada para explicar el método es la utilizada en [126], dónde la función de mapeo se define como:

$$F(\mathbf{x}_t) = \sum_{g=1}^{G} \gamma_g(\mathbf{x}_t) \left[\mathbf{A}_g \mathbf{x}_t + \mathbf{b}_g \right] \tag{5.4}$$

Las matrices $\{\mathbf{A}_g\}$ y vectores $\{\mathbf{b}_g\}$ desconocidos de $F(\cdot)$ se obtienen minimizando el error, calculado como

$$\sum_t \| F(\mathbf{x}_t) - \mathbf{y}_t \|^2 \tag{5.5}$$

sobre todo el material de entrenamiento.

Si se agrupan todas las incógnitas en una única matriz de la siguiente manera,

$$\boldsymbol{\Omega} = [\mathbf{A}_1 \ \mathbf{b}_1 \ \ldots \ \mathbf{A}_G \ \mathbf{b}_G]^\top \tag{5.6}$$

el problema puede formularse como un sistema de ecuaciones lineales:

$$\mathbf{U} \cdot \boldsymbol{\Omega} = [\mathbf{y}_1 \ \ldots \ \mathbf{y}_T]^\top \tag{5.7}$$

en el que

$$\mathbf{U} = \begin{bmatrix} \gamma_1(\mathbf{x}_1)\hat{\mathbf{x}}_1^\top & \cdots & \gamma_G(\mathbf{x}_1)\hat{\mathbf{x}}_1^\top \\ \vdots & \ddots & \vdots \\ \gamma_1(\mathbf{x}_T)\hat{\mathbf{x}}_T^\top & \cdots & \gamma_G(\mathbf{x}_T)\hat{\mathbf{x}}_T^\top \end{bmatrix} \quad , \quad \hat{\mathbf{x}}_t^\top = [\mathbf{x}_t^T \ \ 1] \tag{5.8}$$

La solución que minimiza el error de conversión se puede obtener mediante mínimos cuadrados:

$$\mathbf{\Omega} = (\mathbf{U}^\top \mathbf{U})^{-1} \mathbf{U}^\top [\mathbf{y}_1 \ \dots \ \mathbf{y}_T]^\top \tag{5.9}$$

5.1.3 Maximum-likelihood parameter generation (MLPG)

El algritmo MLPG que se describe en [116] se planteo inicialmente en el contexto de la síntesis de voz paramétrica [118]. Este método consiste en lo siguiente: Se tiene una secuencia de vectores de medias de dimensión $2p$ $\{\boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_T\}$ y $2p \times 2p$ matrices de covarianzas $\{\boldsymbol{\Sigma}_1, \dots, \boldsymbol{\Sigma}_T\}$ que modelan no sólo las características acústicas de una frase sino también sus derivadas de primer orden. Lo que se busca es calcular la secuencia de vectores acústicos p dimensional $\{\mathbf{y}_1, \dots, \mathbf{y}_T\}$ más probable. Normalmente se asume que la relación entre los vectores acústicos y sus derivadas se basa en la siguiente expresión:

$$\Delta \mathbf{y}_t = (\mathbf{y}_{t+1} - \mathbf{y}_{t-1})/2 \tag{5.10}$$

El problema se formula en términos de supervectores. Se construye el supervector que contiene las incógnitas:

$$\bar{\mathbf{y}} = [\mathbf{y}_1^\top \ \cdots \ \mathbf{y}_T^\top]^\top \tag{5.11}$$

Este supervector tiene dimensión $Tp \times 1$. Del mismo modo se construye un supervector de medias $\bar{\mathbf{u}}$ de dimensión $2Tp \times 1$ y una supermatriz de covarianzas diagonal por bloques $\bar{\bar{\mathbf{D}}}$ de dimensión $2Tp \times 2Tp$:

$$\bar{\mathbf{u}} = \begin{bmatrix} \boldsymbol{\mu}_1 \\ \vdots \\ \boldsymbol{\mu}_T \end{bmatrix} \quad , \quad \bar{\bar{\mathbf{D}}} = \begin{bmatrix} \boldsymbol{\Sigma}_1^{-1} & & 0 \\ & \ddots & \\ 0 & & \boldsymbol{\Sigma}_T^{-1} \end{bmatrix} \tag{5.12}$$

Sin tener en cuenta un término aditivo que no depende de $\bar{\mathbf{y}}$, la log-verosimilitud de un candidato $\bar{\mathbf{y}}$ dados $\bar{\mathbf{u}}$ y $\bar{\bar{\mathbf{D}}}$ se puede expresar como:

$$L = -\tfrac{1}{2}(\mathbf{W}\bar{\mathbf{y}} - \bar{\mathbf{u}})^\top \bar{\bar{\mathbf{D}}} (\mathbf{W}\bar{\mathbf{y}} - \bar{\mathbf{u}}) \tag{5.13}$$

donde $\mathbf{W}$ es una matriz que añade las derivadas a los vectores individuales contenidos en $\bar{\mathbf{y}}$. De acuerdo a la ecuación 5.10, $\mathbf{W}$ se puede describir matemáticamente como

$$\mathbf{W} = \mathbf{V} \otimes \mathbf{I} \ , \quad \mathbf{V} = \begin{bmatrix} 1 & 0 & \cdots & & \\ 0 & 1/2 & \cdots & & \\ 0 & 1 & 0 & \cdots & \\ -1/2 & 0 & 1/2 & \cdots & \\ \cdots & 0 & 1 & 0 & \cdots \\ \cdots & -1/2 & 0 & 1/2 & \cdots \\ & \vdots & \vdots & \vdots & \end{bmatrix} \tag{5.14}$$

donde $\otimes$ es el producto de Kronecker y $\mathbf{I}$ es la matriz identidad de orden p. Se puede demostrar que la solución que maximiza la ecuación 5.13 es

$$\bar{\mathbf{y}} = (\mathbf{W}^\top \bar{\bar{\mathbf{D}}} \mathbf{W})^{-1} \mathbf{W}^\top \bar{\bar{\mathbf{D}}} \bar{\mathbf{u}} \tag{5.15}$$

Por razones de eficiencia, las matrices de covarianzas (y la supermatriz resultante) se suelen forzar a ser diagonales. Con estas condiciones, los componentes de los vectores acústicos son mutuamente independientes y el problema se puede resolver de manera separada para cada componente:

$$\bar{\mathbf{y}}^{(i)} = (\mathbf{V}^\top \bar{\bar{\mathbf{D}}}^{(i)} \mathbf{V})^{-1} \mathbf{V}^\top \bar{\bar{\mathbf{D}}}^{(i)} \bar{\mathbf{u}}^{(i)} \ , \quad 1 \leq i \leq p \tag{5.16}$$

donde $\bar{\mathbf{u}}^{(i)}$ y $\bar{\bar{\mathbf{D}}}^{(i)}$ contienen las estadísticas del componente i-ésimo y sus derivadas.

En el problema específico de conversión que se afronta en este trabajo, los vectores de medias y las matrices de covarianzas usadas para generación se obtienen así:

$$\begin{aligned} \boldsymbol{\mu}_t &= \boldsymbol{\mu}_{\hat{g}}^{(y)} + \boldsymbol{\Sigma}_{\hat{g}}^{(yx)} {\boldsymbol{\Sigma}_{\hat{g}}^{(xx)}}^{-1} (\mathbf{x}_t - \boldsymbol{\mu}_{\hat{g}}^{(x)}) \\ \boldsymbol{\Sigma}_t &= \text{diag} \left(\boldsymbol{\Sigma}_{\hat{g}}^{(yy)} - \boldsymbol{\Sigma}_{\hat{g}}^{(yx)} {\boldsymbol{\Sigma}_{\hat{g}}^{(xx)}}^{-1} \boldsymbol{\Sigma}_{\hat{g}}^{(xy)} \right) \end{aligned} \ , \quad \hat{g} = \arg\max \gamma_g(\mathbf{x}_t) \tag{5.17}$$

En este caso, los vectores y matrices necesarias para calcular estos vectores y matrices de 5.17 no se pueden coger directamente de la descomposición explicada en la ecuación 5.2 ya que la GMM aquí descrita se ha entrenado a partir de las características acústicas sin sus correspondientes derivadas. Por tanto, utilizando $\{\gamma_g(\mathbf{x}_t)\}$ como inicialización del algoritmo EM se reestiman los parámetros de la GMM tras añadir las derivadas a los vectores acústicos origen $\{\mathbf{x}_t\}$ y destino $\{\mathbf{y}_t\}$.

Es importante reseñar que el algoritmo MLPG propuesto en [116] considera no solo la verosimilitud de las características acústicas sino también su varinza global (GV - global variance), es decir, su varianza a nivel de frase. Esta adición ha demostrado mediante test subjetivos internos al laboratorio tener mucha importancia en la calidad percibida de la conversión. La mejora es tal que todas las pruebas con MLPG llevan además la utilización de la GV. En este trabajo siempre que se haga referencia a las pruebas de conversión con este método se dará por implícito el uso de GV a no ser que se afirme lo contrario.

5.1.4 MLPG with minimum generation error training (MGE)

Este método se presenta en [36]. Se basa en que los vectores de medias utilizados para el MLPG se estiman minimizando el error entre los vectores generados y los datos del locutor destino. Básicamente se mantienen las matrices de covarianza obtenidas por el método anterior mientras que los vectores de media se estiman de la siguiente manera:

$$\boldsymbol{\mu}_t = \sum_{g=1}^{G} \gamma_g(\mathbf{x}_t) \left(\begin{bmatrix} \mathbf{A}_g \\ \mathbf{A}'_g \end{bmatrix} \mathbf{x}_t + \begin{bmatrix} \mathbf{b}_g \\ \mathbf{b}'_g \end{bmatrix} \right) \tag{5.18}$$

De manera similar a lo descrito en 5.1.2, las incógnitas pueden agruparse en una única matriz $\breve{\boldsymbol{\Omega}}$ definida como

$$\breve{\boldsymbol{\Omega}} = [\mathbf{A}_1 \ \mathbf{b}_1 \ \mathbf{A}'_1 \ \mathbf{b}'_1 \ \ldots \ \mathbf{A}_G \ \mathbf{b}_G \ \mathbf{A}'_G \ \mathbf{b}'_G]^\top \tag{5.19}$$

La i-ésima columna de $\breve{\boldsymbol{\Omega}}$, denominada $\breve{\boldsymbol{\omega}}_i$, está relacionada con el i-ésimo componente del vector acústico. Teniendo en cuenta las restricciones presentadas en la ecuación 5.18, la ecuación del MLPG (5.16) se convierte en

$$\bar{\mathbf{y}}^{(i)} = \mathbf{Q}^{(i)} \breve{\boldsymbol{\omega}}_i \ , \quad \mathbf{Q}^{(i)} = (\mathbf{V}^\top \bar{\bar{\mathbf{D}}}^{(i)} \mathbf{V})^{-1} \mathbf{V}^\top \bar{\bar{\mathbf{D}}}^{(i)} \breve{\mathbf{U}} \ , \quad 1 \leq i \leq p \tag{5.20}$$

donde

$$\breve{\mathbf{U}} = \begin{bmatrix} \gamma_1(\mathbf{x}_1)\breve{\mathbf{X}}_1^\top & \cdots & \gamma_G(\mathbf{x}_1)\breve{\mathbf{X}}_1^\top \\ \vdots & \ddots & \vdots \\ \gamma_1(\mathbf{x}_T)\breve{\mathbf{X}}_T^\top & \cdots & \gamma_G(\mathbf{x}_T)\breve{\mathbf{X}}_T^\top \end{bmatrix} \ , \quad \breve{\mathbf{X}}_t^\top = \begin{bmatrix} \mathbf{x}_t^\top & 1 & \mathbf{0} & 0 \\ \mathbf{0} & 0 & \mathbf{x}_t^\top & 1 \end{bmatrix} \tag{5.21}$$

En un escenario genérico con $N \geq 1$ frases para adaptar (ahora se include el índice de frase n en la notación), la solución que minimiza el error de generación respecto a los datos de entrenamiento es

$$\breve{\boldsymbol{\omega}}_i = \left(\sum_{n=1}^{N} \mathbf{Q}_n^{(i)\top} \mathbf{Q}_n^{(i)} \right)^{-1} \left(\sum_{n=1}^{N} \mathbf{Q}_n^{(i)\top} \bar{\mathbf{y}}_n^{(i)} \right) \quad , \quad 1 \leq i \leq p \tag{5.22}$$

5.1.5 Alineamiento

Como punto de partida para poder aplicar cualquiera de estas técnicas es necesario contar con una base de datos paralela con frases del locutor origen y del locutor destino. Las frases deben ser las mismas porque con la GMM se están mapeando las relaciones que existen entre las características acústicas origen y destino para los mismos sonidos. Es evidente por tanto que se necesitará realizar un alineamiento entre las frases origen y destino para poder realizar el entrenamiento.

Las características acústicas elegidas son los vectores de coeficientes MCEP obtenidos al pasar las señales de audio por el vocoder Ahocoder [34]. Antes de realizar cualquier entrenamiento, el primer paso es alinear cada vector de MCEP origen con su correspondiente vector MCEP destino. La estrategia seguida es utilizar un algoritmo de deformación dinámica del tiempo (DTW - dynamic time warping) de una manera iterativa:

1. Se realiza DTW clásica en cada par de frases después de añadir las derivadas de primer orden a los vectores de entrada.

2. Para el alineamiento actual, se aplica el método explicado en [33] para obtener el factor de normalización del tracto vocal (VTLN - vocal tract length normalization) α que consigue que los vectores origen estén más cerca de los vectores del locutor destino.

3. Si $|\alpha|$ es lo suficientemente pequeña, el alineamiento actual se toma como definitivo. Si no lo es, se reemplazan los vectores originales por sus equivalentes normalizados en función del tracto vocal y se vuelve al paso 1.

A la hora de hacer el alineamiento las tramas se repiten tantas veces como sea necesario tanto en el vector MCEP origen como en el destino. Con los resultados obtenidos de las características espectrales se alinea también los valores de frecuencia fundamental.

5.1.6 Conversión de f_0

Una de las características de las voces esofágicas es el valor bajo e irregular de la frecuencia fundamental. Si se quiere mejorar la agradabilidad de las voces, una de las estrategias posibles consiste en restaurar la señal glotal, y en particular la evolución continua de la frecuencia fundamental. Para ello, se han utilizado modelos basados en GMMs para estimar curvas de frecuencia fundamental ($\log f_0$) partiendo de los coeficientes cepstrales.

Los experimentos iniciales se han realizado con voz sana, con objeto de validar la estrategia. Para realizar las pruebas se utilizó de nuevo el corpus de 100 frases obtenidas para un locutor con voz sana. Las características del experimento son las siguientes:

- El corpus fue dividido en entrenamiento (66 %) y test (34 %).
- Parametrización hecha con Ahocoder, extrayéndose 25 coeficientes mel-cepstrales (MCEP), la $\log f_0$ y la MVF por trama cada 5 ms.
- Entrenamiento de una Joint GMM con 16, 32 y 64 Gaussianas.

Se han evaluado las mismas cuatro técnicas de transformación que se han explicado en el apartado 5.1: Joint-Density modeling, JDM, GMM-weighted linear regression,WLR, máximum-likelihood parameter generation, MLPG, y MLPG with minimum generation error training, MGE.

Para el entrenamiento, el vector origen $\{\mathbf{x}_t\}$ se compone de los 25 coeficientes MCEP del locutor origen sano (incluyendo en este caso el coeficiente 0), mientras que el vector destino $\{\mathbf{y}_t\}$ tiene dimensión 1 al estar compuesto de los valores de $\log f_0$ del mismo locutor sano. Obsérvese que en este caso no es necesario realizar ningún tipo de alineamiento puesto que tanto los coeficientes cepstrales como los valores de pitch provienen de la misma frase.

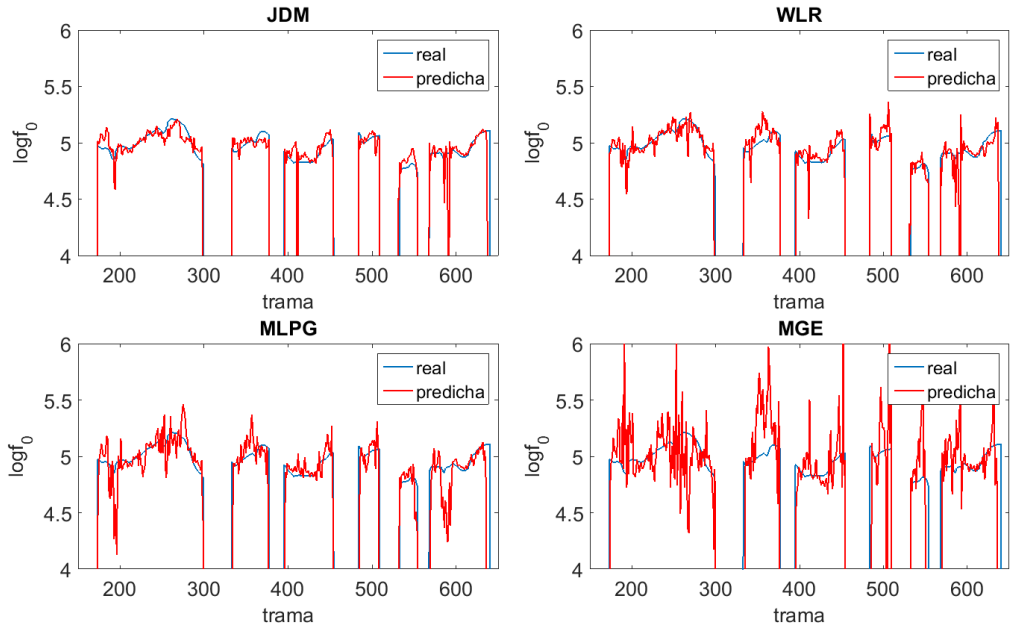

Figura 5.1: Comparación de la $\log f_0$ extraída por Ahocoder para una señal de voz sana de test (azul) y la $\log f_0$ predicha por los métodos de conversión basados en GMMs

La figura 5.1 muestra los resultados obtenidos para una de las frases de test del corpus (JGMM de 32 gaussianas). En ella, la línea azul corresponde con la curva de entonación real mientras que la roja muestra los valores estimados.

Se puede ver que en el cepstrum hay información suficiente para predecir con cierta fidelidad el $\log f_0$ correspondiente. Puede observarse también que los tres primeros métodos de conversión realizan un mejor ajuste que el último método, en el que la variabilidad final de la curva obtenida es muy elevada.

Para evaluar empíricamente los resultados, se han calculado los siguientes valores:

- El error cuadrático medio (MSE) cometido en la estimación de la curva, teniendo en cuenta únicamente los segmentos sonoros (tanto en la curva de origen como en la curva estimada):

$$MSE_{logf_0} = \frac{1}{N}\sum_{n=1}^{N}\left(logf_{0_n} - \widehat{logf_{0_n}}\right)^2 \tag{5.23}$$

  siendo $\widehat{logf_0}$ el valor de pitch predicho y n el número de trama sonora.

- El grado de acierto que se tiene a la hora de clasificar las tramas como sordas o sonoras. La estimación sordo/sonoro se hace a partir de la f_0 predicha por la red: si el valor de pitch para una trama está por encima de un cierto umbral (en el caso del experimento este valor fue de 60 Hz), la trama se considera sonora. Si está por debajo, la trama se considera como sorda.

  En un problema de estas características se definen una serie de conceptos para poder calcular métricas que evalúen la clasificación: se define el número de positivos verdaderos (TP - true positives) como el número de casos clasificados correctamente pertenecientes a la clase positiva (en este caso tramas sonoras clasificadas como sonoras), el número de falsos positivos (FP - false positives) como el número de casos clasificados erróneamente como pertenecientes a la clase positiva (tramas sordas clasificadas como sonoras), el número de negativos verdaderos (TN - true negatives) como el número de casos correctamente descartados como pertenecientes a la clase positiva (tramas sordas clasificadas como sordas), y el número de falsos negativos (FN

- false negatives) como el número de casos descartados erróneamente como pertenecientes a la clase positiva cuando realmente corresponden a dicha clase (tramas sonoras clasificadas como sordas). Usando estos conceptos, se ha decidido utilizar tres medidas para evaluar el funcionamiento de cada uno de los métodos:

  – Precisión: se define como la relación entre el número de casos identificados correctamente y las hipótesis realizadas por el sistema:

$$P = \frac{TP}{TP + FP} \tag{5.24}$$

  – Exhaustividad (Recall): se define como la relación entre el número de casos identificados correctamente y el número total de muestras de la clase positiva en las marcas de referencia. Se calcula de la siguiente manera:

$$R = \frac{TP}{TP + FN} \tag{5.25}$$

  – Fscore: para llevar a cabo una mejor comparación entre precisión y exhaustividad, se utiliza generalmente el valor de Fscore, definido como la media armónica de dichos términos:

$$Fscore = \frac{2RP}{R + P} \tag{5.26}$$

La figura 5.2 muestra el valor del error cuadrático medio obtenido en la predicción de las frases de test de un locutor sano. Se hicieron pruebas con los 4 métodos de conversión, entrenando mezclas de 16, 32 y 64 gaussianas. El mejor método parece ser el más simple, el JDM, y alcanza un error mínimo para 32 gaussianas.

En la tabla 5.1, se muestran los valores obtenidos para la precisión, exhaustividad y Fscore para la clasificación sorda/sonora utilizando los distintos métodos de predicción de f_0 basados en GMMs. En la figura 5.3 se representan estos mismos resultados.

El número total de tramas sonoras (V) fue 21030 y las tramas sordas totales (UV) 21170. Todos los sistemas presentan resultados muy parecidos. La precisión de las tramas sonoras está en torno al 0.99 para todos los métodos, mientras que

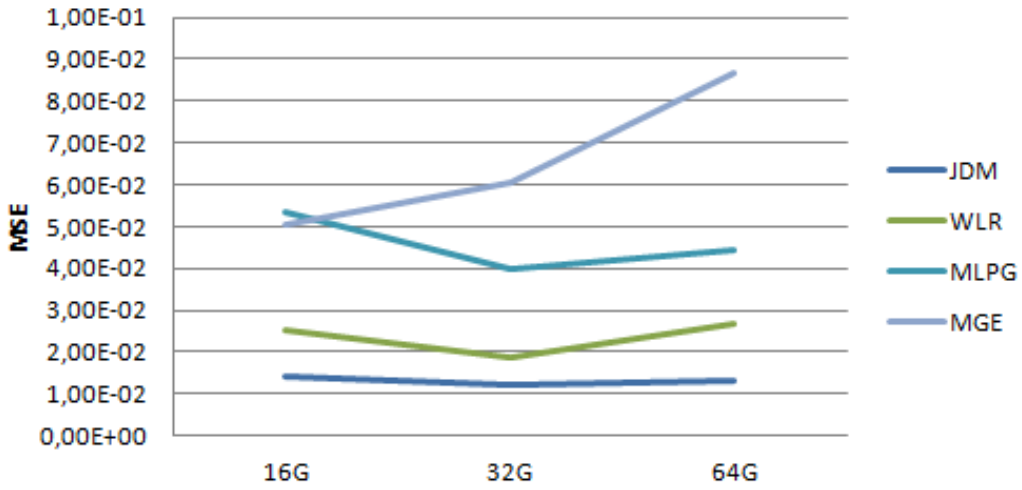

Figura 5.2: Error en la predicción de $\log f_0$ para los distintos métodos de conversión basados en GMMs para un locutor sano.

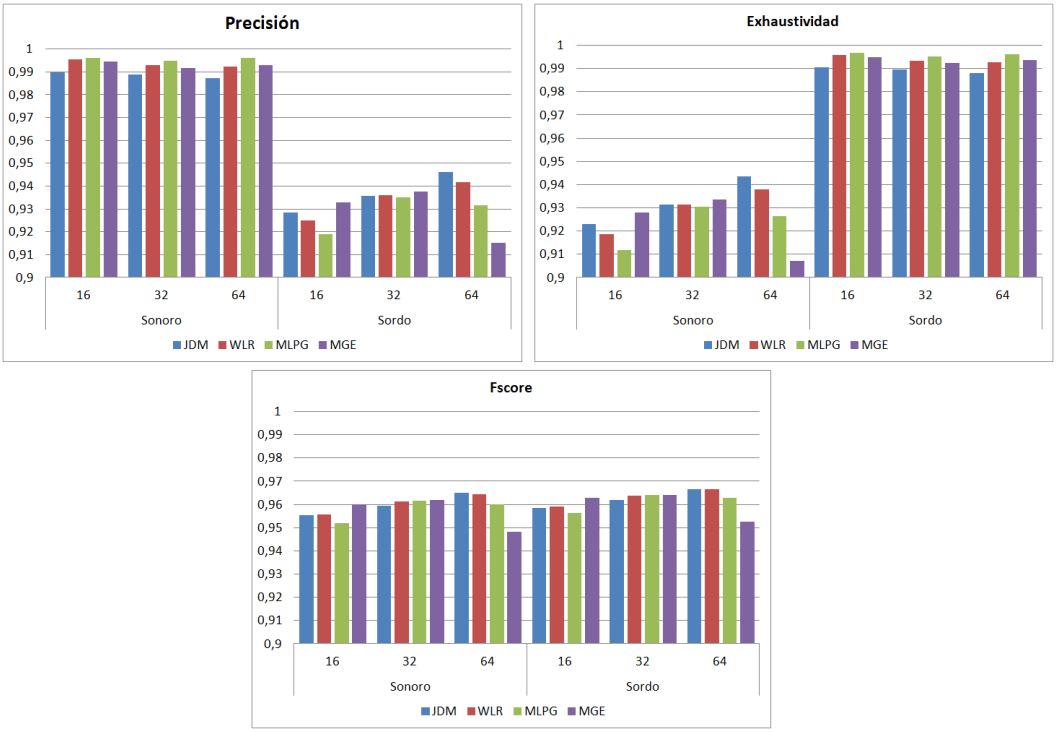

Figura 5.3: Precisión, exhaustividad y Fscore al clasificar las tramas como sordas o sonoras para los cuatro métodos de estimación basados en GMMs para un locutor sano.

Tabla 5.1: Precisión, exhaustividad y Fscore al clasificar las tramas como sordas o sonoras para los cuatro métodos de estimación basados en GMMs.

		JDM			WLR			MLPG			MGE		
	Num Gauss	Precisión	Exhaus.	Fscore	Precisión	Exhaus.	Fscore	Precisión	Exhaus.	Fscore	Precisión	Exhaus.	Fscore
Sonoro	16	0,990	0,923	0,955	0,996	0,919	0,956	0,996	0,912	0,952	0,994	0,928	0,960
	32	0,989	0,932	0,959	0,993	0,931	0,961	0,995	0,930	0,962	0,992	0,934	0,962
	64	0,987	0,943	0,965	0,992	0,938	0,964	0,996	0,926	0,960	0,993	0,907	0,948
Sordo	16	0,928	0,991	0,958	0,925	0,996	0,959	0,919	0,997	0,956	0,933	0,995	0,963
	32	0,936	0,990	0,962	0,936	0,993	0,964	0,935	0,995	0,964	0,938	0,992	0,964
	64	0,946	0,988	0,967	0,942	0,993	0,967	0,932	0,996	0,963	0,915	0,994	0,953

para las tramas sordas se consiguen valores más bajos (en torno al 0.93). Sin embargo, con la exhaustividad pasa justo lo contrario: a la hora de clasificar tramas sonoras el valor está en torno al 0.93, pero para las tramas sordas el valor sube hasta al 0.99. En todo caso son valores muy buenos que demuestran que cualquiera de los métodos sirve para predecir la $\log f_0$ de un locutor de voz sana a partir de sus valores cepstrale.

Si se utiliza el Fscore (que combina los valores de precisión y exhaustividad en una misma métrica) para comparar los distintos métodos, se puede observar que los resultados son prácticamente idénticos, incluso cuando varía el número de gaussianas. Por tanto, visto que todos los sistemas clasifican de manera muy parecida las tramas sordas y sonoras, como el menor error cuadrático medio a la hora de estimar el valor de f_0 se consigue con la transformación JDM con 32 gaussianas, ésta es la que se escogerá en los experimentos descritos en el siguiente capítulo para hacer las pruebas de conversión de $\log f_0$ basada en GMMs.

5.1.7 Conversión con GMMs aplicada a PMA

Otra de las aproximaciones a la hora de mejorar la inteligibilidad de las voces son los interfaces de habla silenciosa. En el habla silenciosa lo que se hace es establecer un mapeo entre bioseñales captadas por los sensores que forman la interfaz y las características acústicas de la voz. Las universidades de Hull y Sheffield proponen en [13, 44, 52] utilizar como estas bioseñales las variaciones del campo magnético producido por varios imanes adecuadamente colocados en articulatorios importantes. Esta técnica se llama PMA (permanent magnet-articulography). Sin embargo, debido a la disposición de los sensores en los articuladores, las señales PMA no continen información sobre la frecuencia fundamental. Las voces esofágicas, aun teniendo pitch, está muy deteriorado. Por ello, disponer de esta base de datos supuso una oportunidad para validar el funcionamiento de los métodos de conversión basados en GMMs descritos previamente.

En este apartado describimos el sistema de conversión entre señales PMA obtenidos al pronunciar una frase y sus componentes espectrales para un hablante sano utilizando grabaciones paralelas de PMA + voz recogidas por las Universidades de Hull y Sheffield obtenidas con hablantes sanos, articulando normalmente.

5.1.7.1 Descripción de los experimentos

En este caso, el conjunto de datos usado contiene 420 grabaciones paralelas de PMA y voz de un locutor británico. Cada grabación corresponde a una frase corta de la base de datos CMU Arctic [63]. Todas fueron grabadas en una única sesión. Se utilizaron 9 sensores para capturar las señales PMA. La frecuencia de muestreo para las grabaciones de voz es de 16 kHz y de 100 Hz para cada uno de los canales PMA. Este conjunto de datos se dividió en dos partes: 212 frases para entrenamiento y 208 frases para test.

Primero se hizo un análisis acústico de la voz cada 5 ms utilizando Ahocoder [34]. Como resultado, se obtienen tres datos acústicos diferentes: los coeficientes Mel-cepstrales (MCEP), el logaritmo de la frecuencia fundamental ($\log f_0$) y la máxima frecuencia sonora (MVF - maximum voiced frequency). Para esta primera evaluación de la conversión PMA sólo se tiene en cuenta los coeficientes MCEP ya que la predicción de f_0 a partir de las señales PMA no es algo trivial, y las características de excitación como la MVF son relativamente menos importantes. Por lo tanto, el vector de los MCEPs del locutor destino será el que se denomine como $\{\mathbf{y}_t\}$.

Las señales PMA en bruto capturadas por los sensores se someten al siguiente procesado:

- Primero, se normaliza cada uno de los canales PMA en términos de media y varianza.
- Como el periodo de muestreo de las señales PMA es el doble de el del vocoder, se interpolar por un factor de 2.
- Para evitar las fluctuaciones espúreas de las señales PMA, cada vector de PMA se combina con los dos adyacentes a la izquierda y también a la derecha, lo que incrementa la dimensión de 9 a 45.
- Esta dimensión se reduce mediante análisis de componente principal (PCA - principal component analysi). La dimensión final es 15 y captura el $99{,}93\,\%$ de la variabilidad.

Los vectores PMA resultantes son considerados los vectores origen y se denominan como$\{\mathbf{x}_t\}$. Hay que tener en cuenta que estos vectores contienen implícitamente información sobre la variación de las señales PMA en el tiempo. Por ello sería redundante considerar sus derivadas en el tiempo durante el procesado.

Se construye entonces una serie de vectores concatenados $\mathbf{z}_t = [\mathbf{x}_t^\top \quad \mathbf{y}_t^\top]^\top$ y con el conjunto de datos de entrenamiento se entrena una GMM conjunta tal y como se ha expuesto en el apartado 5.1.

5.1.7.2 **Evaluación y resultados**

Para poder evaluar el funcionamiento de los cuatro diferentes métodos se entrenar las respectivas funciones de mapeo para una GMM con un número variable de componentes gaussianas: $G = \{16, 32, 64, 128\}$. Una vez hecho esto, se utiliza la parte de test del conjunto de datos y se calcula la distorsión Mel-cepstral (MCD - Mel-cepstral distortion) entre los vectores MCEP destino y los obtenidos convirtiendo los vectores PMA. Esta medida es fidedigna sólo cuando se cumplen una serie de condiciones. Por ejemplo, se ha comprobado que aplicar la GV a la conversión es perjudicial en términos de MCD aunque perceptualmente produce una mejora. Sin embargo está comunmente aceptado que el MCD puede usarse para comparar diferentes variantes de un mismo método. El MCD calculado para esta comparación se formula de la siguiente manera:

$$\text{MCD}\left(\{\mathbf{c}_t\}, \{\hat{\mathbf{c}}_t\}\right) = \frac{10}{\log 10} \cdot \frac{1}{T} \sqrt{2 \sum_{t=1}^{T} \|\mathbf{c}_t - \hat{\mathbf{c}}_t\|^2} \tag{5.27}$$

El elemento 0 de los vectores de MCEP se elimina antes de calcular el MCD. Los resultados se muestran en la figura 5.4. Se puede ver que el número de componentes G óptimo depende del método en particular. En su punto óptimo, los dos algoritmos basados en MLPG (los marcados como MLPG y MGE) presentan un mejor funcionamiento que los algoritmos que se aplican tramo a tramo. Esto guarda consistencia con lo expuesto en trabajos previos [116, 117]. También, y aunque son más propensos al sobreajuste (*overfitting*) los métodos basados en la minimización del error, WLR y MGE, mejoran los resultados de sus equivalentes de máxima ve-

rosimilitud. Esto tiene sentido porque están entrenados optimizando casi la misma métrica que se está usando para la evaluación.

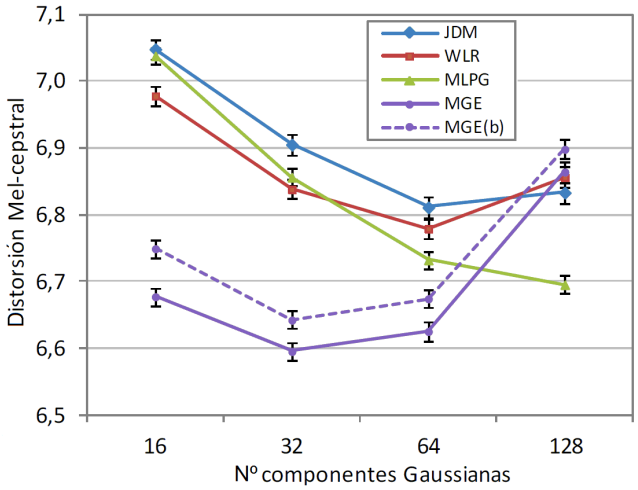

Figura 5.4: Puntuaciones de la MCD media [dB] e intervalos de confianza al 95 % para GMMs de diferente número de componentes y distintos métodos de mapeo.

En general, el método MGE en propuesto en [36] consigue los mejores resultados de MCD. En comparación con el método que consigue la segunda posición, MLPG, requiere un menor número de componentes gaussianas, lo que se traduce en una reducción del coste computacional de la clasificación al hacer uso de la GMM. Sin embargo puede argumentarse que el cálculo de los vectores de medias es mucho más ligero en la ecuación del MLPG (5.17) que en la ecuación del MGE (5.18). Es un matiz importante ya que un sistema de conversión PMA a voz comercial debería operar en tiempo real. Por ello, se volvió a entrenar el método MGE teniendo en cuenta sólo la transformación de la gaussiana más probable, como se hace en el método MLPG. La curva resultante, etiquetada como MGE(b) en la Figura 5.4, muestra que aunque esta simplificación implica una ligera penalización en términos de MCD, sigue siendo el método que mejor resultados tiene. Por tanto, se puede afirmar que el MGE proporciona un intercambio interesante entre precisión y eficiencia.

Aunque no se ha realizado un test perceptual para validar estas técnicas si que se realizaron test informales dentro del laboratorio en los que se conservó la f_0 y las

características de la excitación originales y se reemplazó el espectro por la salida de la conversión PMA a MCEP.

Se observó que se conserva la identidad del locutor a pesar de que se ha descartado totalmente la información espectral original, al menos para los métodos basados en MLPG con GV. La inteligibilidad de las señales puede definirse como moderada. Sin embargo, se debe decir que las grabaciones del conjunto de datos usados se han hecho con una articulación media/baja. En otras palabras, es posible que grabaciones hechas con un mayor esfuerzo de articulación consiguiesen señales sintéticas más inteligibles. En principio esto no debería ser una limitación importante de esta técnica ya que es aceptable asumir que los usuarios finales del sistema (los laringectomizados) serán cuidadosos con este aspecto. En todo caso, los resultados parecen prometedores dada la dificultad de la tarea.

5.2 Conversión con LSTMs

Además de la conversión estadística basada en GMMs, se ha desarrollado un sistema de conversión basado en redes neuronales profundas (DNN - deep neural network). Hay muchas arquitecturas distintas a la hora de implementar una DNN, así que se pensó en utilizar una solución que se beneficiara de la fuerte dependencia temporal existente entre las tramas consecutivas de la señal de voz, afirmación especialmente cierta para la curva de f_0. Por esto se optó por utilizar una arquitectura de larga memoria a corto plazo (LSTM - Long Short-Term Memory). Este tipo de NN permite a la red mantener (y olvidar de manera gradual) información de instantes de tiempo previos. La LSTM es un tipo de red neuronal recurrente (RNN - recurrent neural network). En una RNN las conexiones recurrentes permiten aprender de las dependencias temporales y de la información contextual de los datos de entrada [10]. Sin embargo, el reto está en cómo entrenar una RNN de una manera efectiva para evitar el problema de la desaparición del gradiente [50]. Las LSTMs solucionan este problema por su diseño: Una capa LSTM está compuesta por un conjunto de "bloques de memoriarecurrentemente conectados. Cada bloque está compuesto de una o más celdas de memoria recurrentemente conectadas y tres puertas, entrada, salida y olvido. La red interactúa con las celdas sólo a través de las puertas. Gracias a este diseño interno de la arquitectura, el error se mantiene bajo control dentro de cada celda, superándose así el problema del entrenamiento de las RNNs [51].

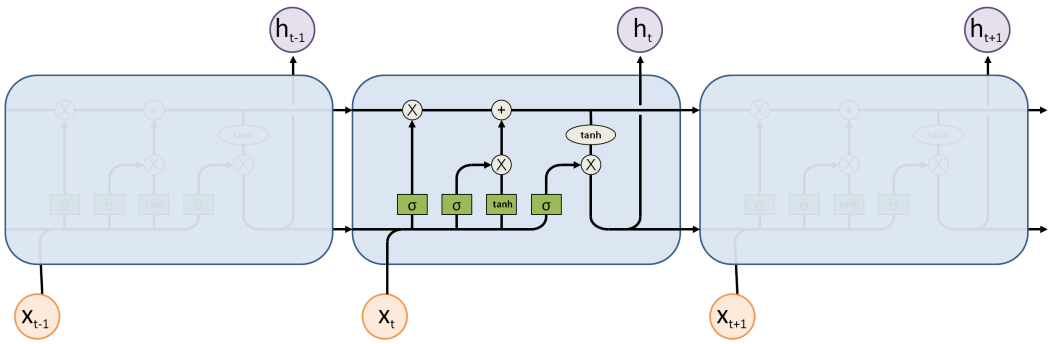

Figura 5.5: Arquitectura interna de una red LSTM. Imagen adaptada del blog de Colah

En la figura 5.5 se puede ver la arquitectura interna de una red LSTM. x_t es el vector de entrada y h_t la salida producida por el bloque actual. Como entrada al bloque se tiene la memoria del bloque anterior representada por la flecha superior izquierda, y la salida del bloque anterior h_{t-1} en la flecha inferior izquierda. La flecha superior derecha que sale del bloque actual es la memoria del bloque.

5.2.1 Arquitectura del sistema

El sistema propuesto sigue el diagrama de la figura 5.6. El primer paso utiliza un vocoder para extraer los parámetros acústicos de la señal. Para obtener la función de conversión se decidió utilizar dos DNNs: una para convertir los valores de pitch y otra para convertir las características espectrales. Ambas redes tienen la misma arquitectura global basada en LSTMs. También se aplica un postprocesado al igual que en [36].

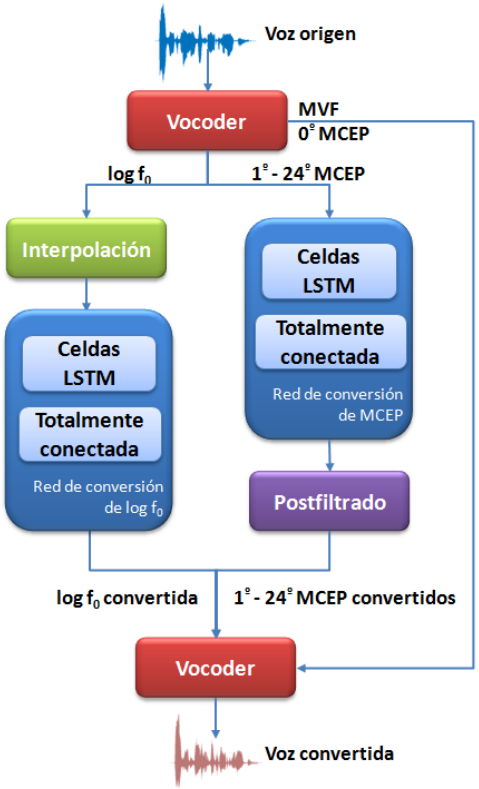

Figura 5.6: Arquitectura del sistema de conversión basado en LSTMs.

El análisis acústico se hace utilizando Ahocoder [34] extrayendo la representación Mel-cepstral (MCEP) de la envolvente espectral, la frecuencia fundamental en forma de $\log f_0$ y la máxima frecuencia sonora (MVF - maximum voiced frequency). La MVF está relacionada con la harmonicidad de la señal. Se extraen 25 coeficientes MCEP. El coeficiente 0 (cuyo valor está relacionado con la energía de la señal) no se convierte. En vez de eso se copia directamente del locutor origen al de destino. Del mismo modo, la MVF del origen no se modifica y se copia directamente para el locutor destino.

Este sistema es dependiente de la pareja de locutores origen-destino que vaya a utilizarse, es decir, debe ser entrenado de forma separada para cada pareja. Además, es necesario alinear las frases paralelas de ambos locutores.

5.2.1.1 Entrenamiento de los coeficientes MCEP

Para el entrenamiento del sistema, los vectores MCEP de dimensión 24 de todas las frases se concatenan uno a continuación de otro. Además, se añade un valor correspondiente a la decisión sordo/sonoro. En cada dimensión del vector cepstral se aplica normalización de media y varianza.

La red dará a la salida la predicción de los 24 coeficientes MCEP del locutor destino y el vector de decisiones sordo/sonoro. La métrica a minimizar es el MSE:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\sum_{d=0}^{24}(\widehat{y_{T_{d_i}}} - y_{T_{d_i}})^2 \tag{5.28}$$

donde $\widehat{y_{T_{d_i}}}$ es el parámetro destino y_{T_d} predicho por la red para cada trama i. n es el número de tramas de la matriz de características y $\widehat{y_{T_d}}$ es el d-ésimo coeficiente MCEP predicho cuando d va desde 1 a 24, y la decisión sordo/sonoro cuando d es 0.

En lo que respecta a la arquitectura de la red, se utilizan 100 celdas LSTM y un tamaño de secuencia de 50 tramas.Para implementar la red neuronal se utiliza la librería de python Keras. El optimizador utilizado en el entrenamiento es el RMS-Prop (dividir el gradiente entre una media actualizada de su valor reciente) con una tasa de aprendizaje de 0.0001. Para evitar el overfitting se utilizó un dropout con tasa de 0.5. Se entrenó la red durante 50 epochs.

5.2.1.2 **Entrenamiento de la logf_0**

Para la predicción de f_0 se entrenó una red LSTM con una arquitectura similar a la descrita para la conversión espectral. Con el fin de evitar cambios abruptos en los valores de la curva de logf_0 de las frases en las transiciones sordo-sonoro que pudieran crear dificultades en el proceso de aprendizaje de la red, se aplica una interpolación lineal (en la escala lineal) entre el último valor del segmento sonoro previo y el primer valor del próximo segmento sonoro para cada segmento sordo (figura 5.7). Sobre los valores de logf_0 se aplica normalización de media y varianza.

Para ayudar al proceso de aprendizaje, la decisión sordo/sonoro se introduce también a la red como un vector de 0s y 1s. La salida es el vector de logf_0 del locutor destino. Aunque se hicieron pruebas incluyendo también la decisión sordo/sonoro en la salida, no se incluyó este valor en la configuración final, tras evaluaciones informales subjetivas.

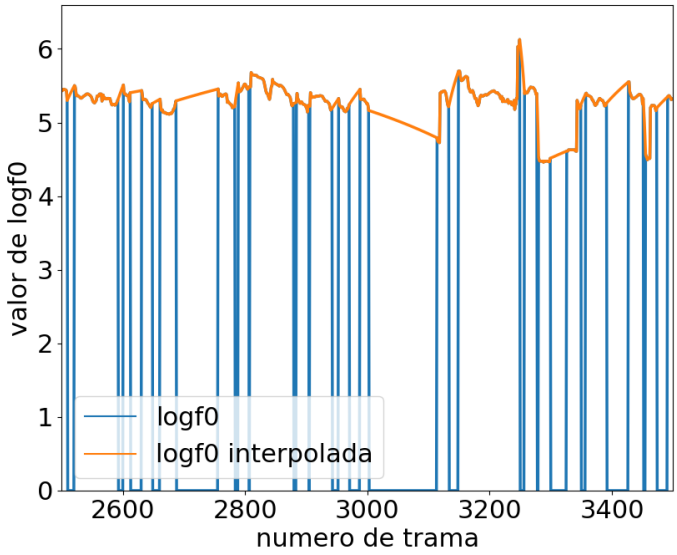

Figura 5.7: Ejemplo de la interpolación de los segmentos sordos de la logf_0.

La matriz resultante se divide en secuencias de 50 tramas antes de entrar a la red, que tiene, al igual que la red utilizada para conversión espectral, 100 celdas LSTM.

La función de pérdidas utilizada para entrenar la red es el error cuadrático medio (MSE - mean squared error):

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(\widehat{logf_{0_{T_i}}} - logf_{0_{T_i}})^2 \tag{5.29}$$

donde $\widehat{logf_{0_{T_i}}}$ es la $\log f_0$ del locutor destino predicha por la red para cada trama i, y n es el número de tramas del vector $\log f_0$.

El número de celdas de la capa LSTM se fija en 100. Como se ha dicho anteriormente, el número de tramas que compone cada lote (o batch sequence) es 50 y la dimensión de la entrada es igual a 2. La salida se obtiene de una capa totalmente conectada y consiste en una secuencia de 50 tramas y una única dimensión por cada secuencia que se introduce a la red.

El optimizador es RMSProp con una tasa de aprendizaje de 0.0001 y una tasa de dropout de 0.5. Esta vez se entrena la red durante 50 epochs.

En la figura 5.8 se muestra la conversión de una de las frases del set de validación desde un locutor origen masculino (linea verde) hacia un locutor destino femenino (línea azul). Se puede apreciar que la $\log f_0$ predicha (línea naranja) trata de seguir las variaciones de pitch de la locutora destino. Como resultado se tiene un pitch predicho en el rango de valores de la locutora destino, pero suavizado. Nótese que las $\log f_0$ origen y destino mostradas están alineadas, pero los segmentos sordos se copian directamente en el pitch predicho desde el pitch origen.

5.2.1.3 **Conversión**

Una vez que están entrenadas las redes LSTM de $\log f_0$ y de conversión espectral, se utiliza Ahocoder para extraer los parámetros correspondientes de las frases de evaluación, y se introducen como entrada a ambos sistemas.

Durante las pruebas de validación se observó que según aumenta el orden de los coeficientes, el parecido entre los coeficientes MCEP predichos y los del locutor destino desciende. Esto se puede apreciar calculando el MSE entre cada coeficiente cepstral normalizado predicho y el coeficiente MCEP destino normalizado (figura 5.9). Para los coeficientes de mayor orden la salida de la red presenta cierto suavizado de los valores destino. Los valores de estos coeficientes de alto orden son más

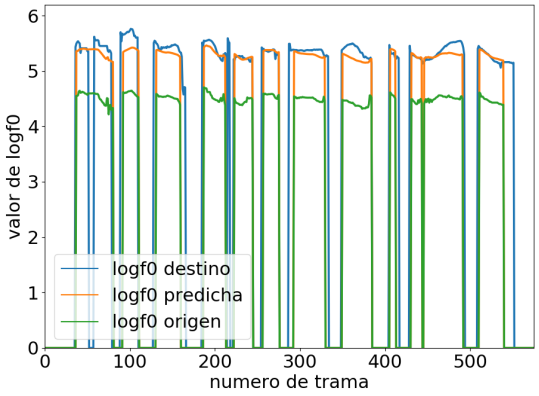

Figura 5.8: Ejemplo de la red de conversión de la $\log f_0$ de un locutor masculino origen a un locutor destino femenino para una frase del set de validación. La línea azul es la $\log f_0$ destino, la línea verde es la $\log f_0$ origen y la línea naranja es la $\log f_0$ predicha por la red.

pequeños y menos importantes que los de bajo orden, pero este suavizado produce un sonido "amortiguado" o *muffled*. El módulo de postprocesado [36] aplicado (ver figura 5.6) trata de minimizar este efecto.

El postfiltrado aplicado se puede formular como una transformación lineal en el dominio cepstral:

$$\mathbf{y}' = \mathbf{P}\mathbf{y} + \mathbf{e} \tag{5.30}$$

La matriz **P** implementa un postfilrado radial bibanda de factores $\{1.03, 1.05\}$ y una frecuencia de corte de 1 kHz (los detalles de este filtro están explicados en [32]).

El efecto que se busca usando **P** es hacer más prominentes los picos espectrales y el enfoque bibanda permite una acentuación más intensa en las frecuencias medias/altas, donde el efecto del suavizado es más apreciable. **e** es un término aditivo cepstral que implementa un filtro que eleva las frecuencias medias/altas 10 dB. Como se describe en [64], mejora la inteligibilidad de la voz sintética sin producir una degradación significativa en la calidad.

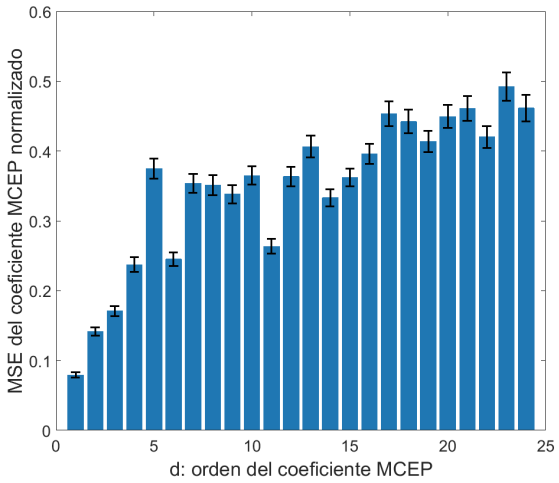

Figura 5.9: MSE de cada coeficiente MCEP normalizado para la conversión de un locutor origen masculino a un locutor destino femenino para los datos de validación. Las líneas indican los intervalos de confianza al 95 %

Finalmente, la síntesis se hace utilizando Ahocoder. Junto con la $\log f_0$ y los coeficientes MCEP convertidos se usa la MVF proveniente del locutor origen.

5.3 Evaluación del sistema

Para poder poner el sistema basado en LSTMs a punto y compararlo con la conversión basada en GMMs, se decidió tomar parte en la segunda edición del Voice Conversion Challenge [69] organizado en 2017. Los esfuerzos se centraron en la tarea principal de este challenge, denominada HUB. Esta tarea consistió en utilizar datos paralelos de locutores origen y destino: 81 frases en inglés de 4 locutores origen y 4 locutores destino (con 2 hombres y 2 mujeres en cada grupo). Como datos de evaluación, se proveyó de un set de 25 frases adicionales de los 4 mismos locutores origen que los surtidos para el entrenamiento.

Se prepararon sistemas VC como el de la figura 5.6 para cada combinación de locutores origen-destino (16 pares de locutores en total).

En la reconstrucción de la señal convertida, se aplica un modificación de la duración de la misma forma que en [36], reajustando el frame rate del vocoder utilizando el factor D (que debe ser calculado durante el entrenamiento).

5.3.1 Resultados generales

Los sistemas presentados al VC Challenge 2018 fueron evaluados mediante una campaña de evaluación subjetiva. En esta campaña se le presentaron las frases convertidas a 106 hablantes nativos de inglés (49 mujeres, 57 hombres) mediante una interfaz web. El test consistió en dos secciones, la primera preguntaba sobre la calidad de la voz sintética y la segunda sobre la similitud con el locutor destino.

Para evaluar la calidad, se les presentó a los oyentes el estímulo y se les pidió que puntuaran su naturalidad sin tener en cuenta la gramática o el contenido de la frase. La respuesta podía estar entre estas cinco opciones diferentes: “Completamente antinatural”, “Mayormente antinatural”, “Tan natural como antinatural”, “Mayormente natural” y “Completamente natural”.

En lo que respecta a la similitud, a los oyentes se les dio un par de estímulos y debían puntuar su parecido entre estas cuatro categorías: “El mismo, absolutamente seguro”, “El mismo, no seguro”, “Diferentes, no seguro”, “Diferentes, absolutamente seguro”. La puntuación de similitud se definió como %(El mismo: absolutamente seguro) + %(El mismo: no seguro), es decir, el porcentaje correcto comparado con el destino.

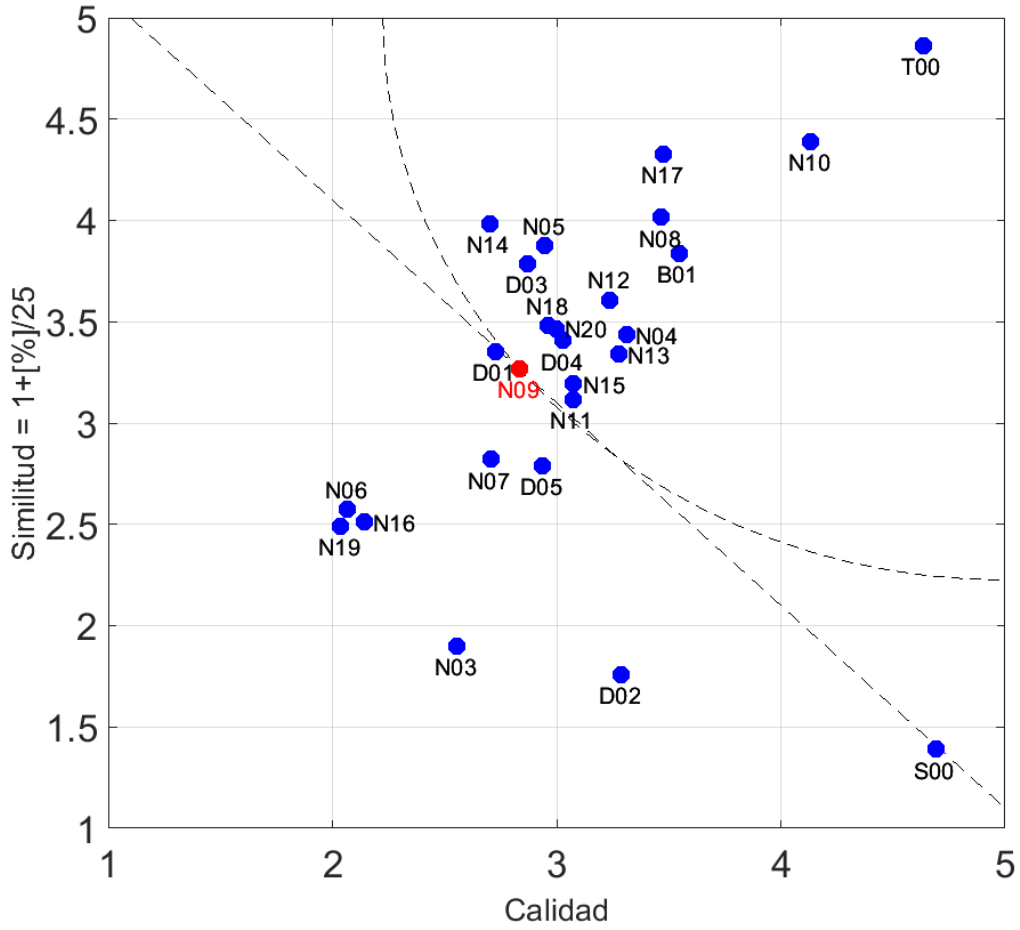

Figura 5.10: Resultados del VC Challenge en un plano calidad vs. similitud. El sistema presentado, denominado N09, aparece en rojo. La linea recta representa los puntos con la misma puntuación media que N09. La curva representa los puntos con la misma distancia a (5,5) que N09.

En la figura 5.10 se presenta una perspectiva general de los resultados del challenge. Los resultados de todos los participantes se muestran en un plano calidad vs. similitud. El sistema presentado, N09, está marcado en rojo. Los resultados de similitud han sido reescalados a una puntuación estilo MOS por motivos de comparación. La linea de puntos representa los puntos con la misma puntuación media que el sistema presentado (la suma de las puntuaciones de calidad y similitud). El arco de puntos representa los puntos con la misma distancia a (5,5), la puntuación perfecta. Las frases de evaluación destino esstán etiquetados como T00, y las frases origen como S00.

Tabla 5.2: Resultados obtenidos para similitud. F indica locutor femenino y M masculino.

	Diferente seguro	Diferente no seguro	Mismo no seguro	Mismo seguro	Puntuación
Todos	13.7 %	29.6 %	41.5 %	15.2 %	56.7 %
F - F	20.6 %	33.3 %	33.4 %	12.7 %	46.1 %
F - M	6.9 %	20.7 %	54.0 %	18.4 %	72.4 %
M - F	11.6 %	39.6 %	37.2 %	11.6 %	48.8 %
M - M	13.2 %	28.9 %	39.5 %	18.4 %	57.9 %

La tabla 5.2 muestra los resultados detallados para la similitud obtenido por el sistema presentado. También se dan las puntuaciones inter-género e intra-género. La puntuación general de similitud es casi del 57 %. No es una puntuación muy alta, pero sitúa al sistema una posición media con respecto al resto de sistemas presentados. La mejor puntuación en lo referente a la similitud se consigue en la conversión de hombre a mujer, con un 72 %. Es un buen resultado, especialmente si se compara con el 46 % obtenido para la conversión mujer a mujer. Parece que este sistema funciona peor al cuando el locutor destino se trata de una mujer. Sin embargo, un análisis detallado de los resultados del challenge muestra que en 21 de los 24 sistemas presentados, los peores resultados de similitud se obtienen cuando el destino de la conversión es una voz femenina (tanto para el caso F2F como el M2F). Parece por tanto que existe un claro desequilibrio en los resultados en relación con el género del destino de la conversión.

Tabla 5.3: Puntuaciones MOS sobre la calidad del sistema. F indica locutor femenino y M masculino.

	Media	Desv. St.	Frases evaluadas	Mediana
Todos	2,83	1,14	1344	3
F - F	2,84	1,12	336	3
F - M	2,66	1,14	329	3
M - F	2,91	1,13	340	3
M - M	2,91	1,17	339	3

La puntuación relativa a la calidad del sistema se muestra en la tabla 5.3. La puntuación media obtenida es de 2.83. El valor de la mediana es 3, compartida por 15 de los sistemas presentados. Como con los resultados de la similitud, se da también la puntuación para la conversión inter e intra género. En este caso no hay muchas diferencias entre las cuatro direcciones de conversión, aunque la conversión de mujer a hombre, que en el caso de la similitud tiene las mejores puntuaciones, es la que menor calidad ofrece a los evaluadores, con una puntuación de 2.66. Se puede ver un intercambio entre la similitud de un sistema VC y su naturalidad.

Además de la evaluación perceptual, también se proporcionó una medida objetiva en forma de la tasa de error de palabra (WER - word error rate) obtenida utilizando el sistema de reconocimiento de voz automático (ASR) de iFlytek [70]. La figura 5.11 presenta una descripción general de los resultados para todos los sistemas. El sistema aquí descrito obtuvo un valor de WER del 12.45 %, lo que significa que el proceso de conversión mantiene la inteligibilidad de las oraciones.

5.3.2 Evaluación Interna

Además de los resultados proporcionados por los evaluadores,se realizó un test de múltiples estímulos con referencia y ancla ocultas (MUSHRA - MUltiple Stimuli with Hidden Reference and Anchor) con el objetivo de comparar el sistema presentado al challenge de 2018 basado en LSTMs con el presentado en 2016 basado en GMMs [36]. Este test subjetivo está definido por la recomendación BS.1534-

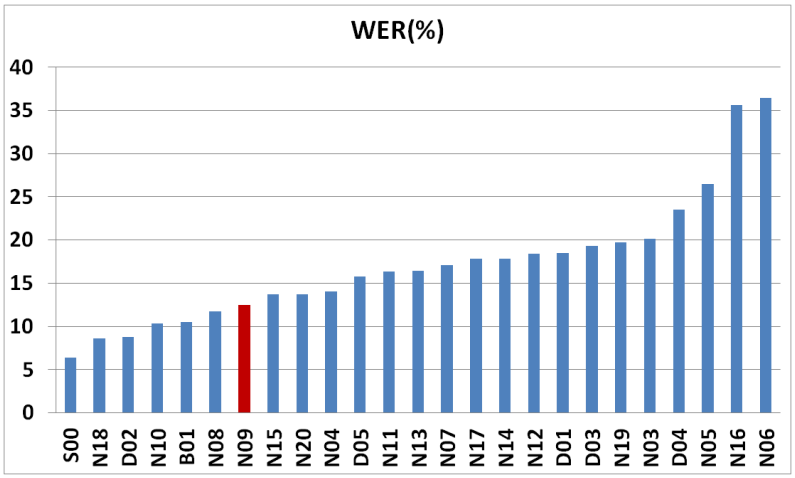

Figura 5.11: Resultados generales de WER del VC Challenge 2018. El sistema presentado aparece en rojo. S00 hace referencia a las frases origen

3 de la ITU-R. Para el experimento realizado se uso el software [100] disponible en <https://www.audiolabs-erlangen.de/resources/webMUSHRA>. Se personalizó para poder evaluar la similitud además de la calidad como un factor adicional.

Además, con este test también se quiso evaluar como la conversión de los coeficientes MCEP y de la $\log f_0$ afectan de manera separada a los resultados, por lo que se incluyó en la evaluación un tercer sistema con los MCEPs obtenidos a partir de una red LSTM y la $\log f_0$ convertida únicamente con transformación lineal de media y varianza, como en el sistema presentado al primer VC Challenge.

En el MUSHRA, se le presenta al evaluador la referencia (marcada como tal), un cierto número de muestras de test, una versión oculta de la referencia y una o más anclas. El propósito del ancla es hacer la escala más próxima a una "escala absoluta". La principal ventaja que tiene sobre la metodología de puntuación de opinión media (MOS - mean opinion score) es que requiere de un menor número de participantes para obtener resultados estadísticamente significativos. Los métodos que se evaluaron fueron:

- Método 1: El sistema presentado al VC Challenge 2018, con predicción de MCEP y $\log f_0$ que hace uso de RNNs basadas en celdas LSTM.

- Método 2: Las frases convertidas se sintetizan con los coeficientes MCEP provenientes de la red LSTM, pero la conversión de la prosodia se hace rescalando media y varianza tal y como se hacía en el sistema presentado al challenge anterior.

- Método 3: El sistema presentado al primer VC Challenge, descrito en [36] y basado en GMMs y MLPG.

- Método 4: Una referencia para delimitar por arriba el test MUSHRA. Esta referencia fue la señal original.El evaluador debía adivinar que era la referencia y puntuarla con un 100. De otra manera, sus puntuaciones eran descartadas.

- Método 5: Un ancla para delimitar por abajo el test MUSHRA. El ancla seleccionada fue una frase sintetizada a partir de los MCEP estimados por el sistema descrito en [36] y $\log f_0$ constante con el valor del pitch medio del locutor destino.

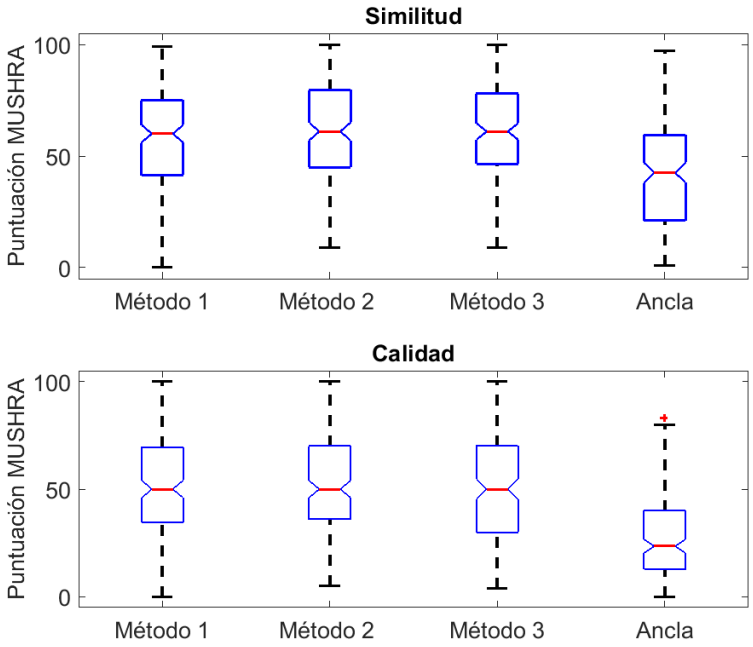

Figura 5.12: Resultados de similitud y calidad para el test MUSHRA.

11 personas participaron en el test, 9 de ellas con experiencia en el campo de la síntesis de voz. La figura 5.12 muestras los resultados para la similitud y la calidad. Como puede verse no hay diferencias significativas entre los tres sistemas evaluados. Los resultados detallados para todas las direcciones de conversión se muestran el las tablas 5.4 y 5.5.

Tabla 5.4: Resultados de similitud del test MUSHRA para la comparación de los sistemas de conversión. F indica locutor femenino y M masculino.

SIMILITUD	Método 1	Método 2	Método 3	Ancla
Todos	58.43	59.73	60.57	42.44
F-F	54.52	60.84	60.14	38.50
F-M	60.55	62.77	65.45	45.64
M-F	52.20	48.39	49.41	36.66
M-M	66.43	66.93	67.27	48.98

Tabla 5.5: Resultados de calidad del test MUSHRA para la comparación de los sistemas de conversión. F indica locutor femenino y M masculino.

CALIDAD	Método 1	Método 2	Método 3	Ancla
Todos	49.93	50.65	50.23	26.81
F-F	47.48	47.66	48.93	23.77
F-M	49.48	54.07	50.02	28.11
M-F	49.59	43.25	44.95	24.45
M-M	53.16	57.64	57.00	30.91

Por un lado, no se produce mejora a la hora de introducir la red para predecir el pitch. No hay contribución a la identidad del locutor destino. Aunque no es peor en términos de calidad que el pitch convertido linealmente, suena más monótono lo que lleva a una puntuación menor a la hora de evaluar la calidad. Por otro lado, no hay mejora tampoco en el uso de la red LSTM en vez del enfoque clásico basado en GMMs. Esto probablemente sea debido a la poca cantidad de datos utilizada para entrenar.

5.3.3 Análisis de los resultados

En este apartado se ha descrito el sistema de VC basado en LSTMs presentado al segundo VC Challenge. Convierte los coeficientes MCEP y la $\log f_0$. Para compensar el suavizado de la componente espectral se aplica un postfiltrado, obteniéndose una conversión con calidad intermedia y similitud aceptable, como muestran los resultados obtenidos en el challenge.

Con el test MUSHRA llevado a cabo para comparar este sistema con un sistema clásico basado en GMMs indica que no existen diferencias significativas entre ellos. Con la escasez de datos disponibles para entrenar las redes (sólo se han usado 81 frases para entrenar cada par de locutores origen-destino) el funcionamiento del sistema LSTM es el mismo que el del sistema con un enfoque más clásico basado en GMMs.

Los resultados muestran que el sistema que usa la red LSTM para convertir la $\log f_0$ está ligeramente peor considerado que el que utiliza conversión lineal. Esto lleva a pensar que quizás la arquitectura LSTM propuesta no es la mejor y que necesita algunos ajustes, como por ejemplo cambiar el tamaño de secuencia para captar la prosodia del locutor de una mejor manera.

5.4 Conclusiones

En este capítulo se han explorado distintas técnicas de conversión de voz que se utilizan para las voces sanas. En un primer lugar se han explicado las bases de los sistemas de conversión estadísticos basados en GMMs. Se han implementado cuatro métodos diferentes. Estos métodos se han adaptado para convertir de PMA (interfaz de habla silenciosa) a MCEP y se han comparado los resultados. Para esta comparación se ha utilizado como medida la dispersión Mel-cepstral. Se ha llegado a la conclusión de que los métodos basados en MLPG en combinación con un criterio de entrenamiento basado en MGE dan los mejores resultados.

También se ha explorado la utilización de estos métodos de conversión basados en GMMs para predecir la $\log f_0$, comprobándose que en las características espectrales hay suficiente información para obtener el pitch de una manera aceptable. El error cuadrático medio más pequeño a la hora de predecir la $\log f_0$ original se consigue utilizando la técnica JDM, usando una GMM conjunta de 32 gaussianas. Además, el valor del Fscore tanto para tramas sordas como sonora es siempre cercano a 0.96 para todos los métodos implementados, valor que evidencia el buen funcionamiento de la clasificación.

Después se ha hecho uso de DNNs para hacer la conversión de voz. Más concretamente, se ha construido un sistema basado en redes LSTM. Consta de dos bloques compuestos por una red neuronal cada uno, uno para convertir los coefcientes MCEP y otro para la conversión de f_0. Este sistema se presentó a la segunda edición del Voice Conversión Challenge, obteniéndose unos resultados de similitud y calidad que lo sitúan en una posición media entre los sistemas presentados. Además, se ha organizado una evaluación interna mediante un test MUSHRA para comparar el sistema basado en GMMs y el basado en DNNs. Los resultados de ambos sistemas no presentan diferencias significativas. Con la escasa cantidad de datos utilizada para entrenar ambos sistemas (81 frases) el funcionamiento del sistema basado en LSTMs es el mismo que el del sistema de GMMs. Otra de las conclusiones es que la conversión de f_0 utilizando LSTMs gusta ligeramente menos que el simple reescalado. Esto podría deberse a que a la salida de la red neuronal la curva de f_0 ha sufrido un suavizado que hace que se escuche con menos energía.

Los trabajos relatados en este capítulo relacionados con la conversión de GMMs se publicaron en [36] y en [37].

La imitación es la forma más sincera de adulación.

Charles Caleb Colton

CAPÍTULO

6

Técnicas de conversión para voces esofágicas

Las técnicas de conversión descritas en el capítulo 5 han demostrado funcionar razonablemente bien cuando tanto el locutor origen como el destino son voces sanas. Aplicar estas mismas técnicas para mejorar la calidad (naturalidad y/o inteligibilidad) de una voz esofágica, no es en absoluto inmediato. Debido a las características específicas de las voces esofágicas, aparecen un conjunto de dificultades que es necesario superar. Especialmente difícil es el tratamiento de la frecuencia fundamental (f_0). El pitch es una de las diferencias de las voces esofágicas, así que se le debe presentar especial atención. También cambia la forma de realizar el alineamiento entre las señales sanas (señales destino) y las patológicas (señales origen) principalmente debido a las irregularidades de las esofágicas.

En este capítulo se describen tres métodos de conversión investigados: utilizando GMMs (6.2), utilizando redes LSTM (6.3) y utilizando Phonetic Posteriorgrams (6.4). Esta última técnica se ha evaluado únicamente para la conversión de señales esofágicas, por lo que no ha sido descrita en el capítulo anterior.

Un esquema general del proceso de conversión puede verse en la figura 6.1. Este esquema es el que se seguirá tanto para la conversión basada en GMMs como para la conversión que usa redes neuronales (NN - neural networks).

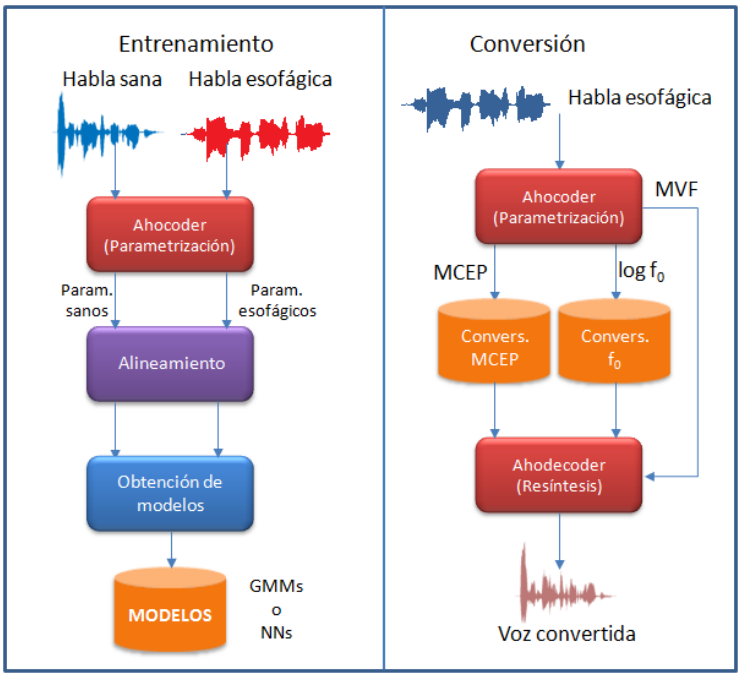

Figura 6.1: Esquema general del proceso de conversión.

En un primer paso es necesario hacer un análisis de las señales patológicas origen y de las sanas destino para obtener los parámetros de ambas voces. En el trabajo presentado en esta tesis, esta parametrización se ha hecho utilizando Ahocoder [34],que se encarga de extraer la representación Mel-cepstral (MCEP) de la envolvente espectral, la frecuencia fundamental en forma de $\log f_0$ y la máxima frecuencia sonora (MVF - maximum voiced frequency), tal y como se ha visto en el apartado 5.3.

Después, se debe realizar un entrenamiento para aprender a convertir de los parámetros esofágicos a los equivalentes de la voz sana. Para poder entrenar estos modelos de conversión es necesario alinear las frases sanas y las esofágicas (aunque en el caso de la conversión mediante PPGs se verá que este paso no es necesario).

Una vez aprendidas las relaciones existentes entre los parámetros esofágicos y los de la voz sana, se aplican los modelos para hacer la conversión. Estos parámetros convertidos serán los que se utilicen para hacer la resíntesis y conseguir la señal de voz con las características del locutor sano destino.

En este capítulo se explica cómo se han abordado estas tareas en esta tesis.

6.1 Estrategias de alineamiento

Uno de los requisitos de los métodos de conversión de voz investigados es que se necesitan bases de datos paralelas. El primer paso es alinear frase a frase de los locutores origen a los locutores destino. Para el proceso de conversión entre voces sanas, esto se hace mediante Dynamic Time Warping (DTW), tal y como se ha visto en el apartado 5.1.5. Sin embargo, a la hora de aplicarse a las frases esofágicas se ha constatado que los resultados no son buenos, principalmente debido al fuerte desalineamiento de las señales, las grandes diferencias de duración de los sonidos, etc, tal y como se ha visto en los capítulos anteriores. La solución propuesta es etiquetar adecuadamente la voz esofágica para poder conseguir un alineamiento entre las frases sanas y sus equivalentes patológicas a partir de estas marcas de tiempos.

Este proceso de etiquetado de las voces esofágicas es el descrito con detalle en el capítulo 3, dentro del apartado 3.5. La técnica de etiquetado que se utilizó finalmente consistió en entrenar un modelo con voces patológicas para realizar el alineamiento forzado (3.5.1.2).

Las voces sanas destino también se etiquetan. Para ello, se utiliza el reconocedor de voz entrenado con voces sanas que se ha descrito con detalle en el capítulo 4. Al igual que con las voces esofágicas, a partir de las transcripciones se lleva a cabo un proceso de alineamiento forzado con el que se tienen las marcas de tiempo.

Una vez se tienen todas las frases correctamente etiquetadas a nivel fonético, se puede proceder con el alineamiento. El enfoque tomado ha sido utilizar las marcas de inicio y final de cada par de fonemas esofágico/sano y utilizarlos como puntos de anclaje. Entre estos puntos se realiza una DTW del modo que aparece descrito en el apartado 5.1.5. De esta manera se pretende conseguir que la transición entre fonemas sea captada de una manera más gradual que si se eliminasen directamente las tramas sobrantes de cada fonema según los tiempos de las etiquetas.

6.2 Técnicas de conversión basadas en GMMs

Para hacer las pruebas de evaluación preliminares, se tomó como origen una de las sesiones alaríngeas (02M3) y como destino un locutor sano (130). Las 100 frases de las que está compuesta cada sesión se dividieron en 66 frases para el entrenamiento de los métodos y las 34 frases restantes para la evaluación. La evaluación se llevó a cabo utilizando reconocimiento automático del habla, utilizando el sistema de reconocimiento basado en Kaldi con diccionario de reconocimiento que sólo contiene las palabras de las transcripciones de las grabaciones y sin modelo de lenguaje y que se describe en el apartado 4.

En el siguiente apartado se explica qué aspectos se han evaluado.

6.2.1 Condiciones de los experimentos

A la hora de diseñar los experimentos a realizar, hay que tener en cuenta que hay muchas opciones posibles. Existen muchos factores a tener en cuenta que afectan al resultado final de la conversión.

- Alineamiento: Para poder entrenar la conversión ya se ha explicado que se necesitan alinear las frases del locutor origen con las del locutor destino. Se hicieron pruebas con dos tipos de alineamiento:
  - DTW: En un primer momento el alineamiento entre los cepstrum de las frases patológicas y de las sanas se hizo únicamente mediante DTW (Dynamic Time Warping). Tras estudiar los resultados, se vio que el alineamiento introducía errores a la hora de entrenar la conversión: la DTW funciona muy bien entre señales de dos locutores sanos, pero para los patológicos presenta grandes desajustes.
  - Manual: Para comprobar la mejora que se puede obtener en la conversión, se hizo un alineamiento manual entre la sesión alaríngea y la sana, y se volvieron a entrenar los métodos de transformación del cepstrum. Cabe destacar que el alineamiento manual de toda el habla esofágica es una tarea inviable, así que este experimento sólo marca el límite superior que puede conseguirse con esta técnica. Para automatizar la labor

del alineamiento para que se pueda aplicar a todos los locutores, se utilizará la técnica descrita en 6.1.

- Conversión cepestral: Con los alineamientos ya conseguidos, se pueden evaluar las distintas técnicas de conversión estadística de voz basadas en GMMs descritas en 5.1.

  Estas técnicas mapean el espectro de la señal alaríngea hacia el espectro de la voz sana. Los cuatro métodos de mapeo estadístico implementados son:

  - GMM Joint density modeling (JDM)
  - GMM-weighted linear regression (WLR)
  - Maximum-likelihood parameter generation (MLPG)
  - MLPG with minimum error generation (MGE)

  Como se puede comprobar, hay cuatro métodos posibles para la conversión espectral, seis al utilizar la varianza global (GV - global variance) con la conversión MLPG y MGE. En un primer test informal realizado en el laboratorio, se comprobó que las conversiones que se percibían como mejores eran las realizadas con el método MLPG añadiendo Global Variance (GV). Teniendo en cuenta que hay muchas opciones posibles para las pruebas (método de extracción de los coeficientes cepstrales, alineamientos a utilizar, distintas procedencias de los valores de pitch...), para evitar que el número de tests a hacer crezca de una manera inmanejable se decidió evaluar todas las variaciones de la conversión basada en GMMs sobre este método concreto.

- Conversión de f_0: Para poder resintetizar la señal con Ahocoder a partir de los parámetros espectrales convertidos, es necesario utilizar unos valores de $\log f_0$.

  - Un método sencillo que suele utilizarse tradicionalmente en la conversión entre voces sanas, es aplicar una transformación lineal para acercar la $\log f_0$ de la señal origen a la del locutor destino de la siguiente manera:

$$log f_0^{(conv)} = \frac{\sigma^{(tgt)}}{\sigma^{(src)}} \left(log f_0^{(src)} - \mu^{(src)} \right) + \mu^{(tgt)} \tag{6.1}$$

Sin embargo, cuando el locutor origen se trata de un hablante esofágico, este método pierde el sentido ya que ambas voces están muy alejadas entre sí.

- Otra posibilidad que se exploró fue dar el mismo valor de $\log f_0$ a todas las tramas de la señal convertida. Este valor será la $\log f_0$ media de las frases de test sanas de referencia. El problema es que la frase convertida suena muy robótica, haciendo difícil al oyente abstraerse de este hecho a la hora de evaluar la calidad de la conversión.
- Predicción de $log f_0$: Se predice la $\log f_0$ para los cepstrum convertidos a partir de cepstrums, tal y como se explica en el apartado 5.1.6. La predicción de $\log f_0$ se realizará a partir de los valores de cepstrum obtenidos de la conversión MLPG más GV. El entrenamiento de las funciones de conversión se habrá realizado en cambio con valores de cepstrum de la voz sana (la voz original del locutor objetivo). En este caso se han evaluado todos los métodos de conversión implementados: JDM, WLR, MLPG y MGE. Además de la $\log f_0$ que sale directamente de cada método de conversión, se evaluó también la resíntesis con la $\log f_0$ predicha y suavizada mediante un filtro de mediana.

- Cálculo de f_0: Como ya se ha visto en el apartado 3.6.1, en este trabajo se han investigado diferentes métodos de extracción de pitch para los locutores esofágicos. Por ello, para los experimentos se han usado dos métodos distintos de cálculo de la $log f_0$:

  - Método de la autocorrelación: Los parámetros se extrajeron utilizando Ahocoder, que extrae el pitch con un algoritmo basado en el método de la autocorrelación.
  - PSIAIF: Viendo que la manera de extraer la $\log f_0$ de Ahocoder podría no ser el método más adecuado para las voces alaríngeas, se probó a extraer de nuevo los parámetros utilizando el algoritmo basado en PSIAIF (Pitch Synchronous Iterative Adaptive Inverse Filtering) descrito en 3.6.1 en vez del basado en la autocorrelación. Con las nuevas

componentes espectrales extraídas de esta manera, se volvieron a entrenar los distintos métodos de conversión.

- MVF: Para poder resintetizar la señal, Ahocoder necesita para cada trama además de valores de MCEP y de$logf_0$ valores de MVF. Se hicieron distintas pruebas de conversión de este parámetro (usando los mismos métodos de predicción basados en GMMs como con la $logf_0$), pero los test informales realizados llevaron a la conclusión de que su impacto era mínimo en la calidad de la conversión. Por ello, para todos los experimentos realizados se trasplantó la MVF del locutor original.

6.2.2 Resultados de los experimentos

Los resultados de los experimentos se muestran en la tabla 6.1. Se puede ver claramente como la combinación de alineamiento manual y extracción de parámetros mediante PSIAIF consiguen los mejores resultados. El suavizado de la predicción de f_0 también tiene un efecto positivo para todas las conversiones.

Tabla 6.1: Resultados de los experimentos de conversión en base a GMMs. La conversión de MCEP utiliza el método MLPG + GV. Se muestran los valores de WER para diferentes alineamientos y valores de f_0 al utilizarse distintos métodos de extracción de parámetros.

	WER (%)			
Referencia sano				11.34
Referencia esofágico				63.03
Extracción	Autocorr.		PSIAIF	PSIAIF + Suavizado
Alineamiento	DTW	Manual	Manual	Manual
Conv f_0: JDM	52.31	45.59	45.38	44.12
Conv f_0: WLR	52.31	46.43	45.17	43.49
Conv f_0: MLPG	53.57	45.59	46.64	43.49
Conv f_0: MGE	52.10	46.43	44.54	43.28

Los valores de WER obtenidos para las conversiones mejoran el original del locutor esofágico, pero todavía están lejos del valor del locutor sano que se utiliza como destino. Como era de esperar, los mejores resultados se obtienen cuando el alineamiento es manual, el análisis se hace extrayendo el pitch mediante el método basado en PSIAIF y se aplica un suavizado a la f_0 predicha. Con estas condiciones se consigue una mejora absoluta en el WER del 19.75 %.

6.3 Técnicas de conversión basadas en LSTMs

Al igual que para el caso de las voces sanas explicado en 5.2, se propuso utilizar para hacer la conversión una arquitectura con dos sistemas paralelos basados en DNNs para convertir la envolvente espectral y la curva de la frecuencia fundamental f_0 [105].

El análisis acústico con el que obtener las características para entrenar el sistema vuelve a ser Ahocoder [34]. Los parámetros (coeficientes Mel-cepstrales, $\log f_0$ y MVF) se obtienen cada 5 ms. La MVF no se utiliza en la conversión, se trasplanta la de la frase del locutor origen. Para las señales alaríngeas, el método de extracción de f_0 es el explicado en 3.6.1, más adecuado para las señales patológicas. El número de coeficientes MCEP utilizados en la conversión espectral es 25, a los que además se añaden sus derivadas de primer orden. El coeficiente 0 (relacionado con la energía) no se convierte directamente, sino que al igual que con la MVF, se copia directamente de la fuente al destino.

Para la predicción de pitch se utilizan 40 coeficientes MCEP. Este número se eligió empíricamente en base a criterios de calidad observados durante el desarrollo del entrenamiento.

6.3.1 Datos de entrenamiento y de test

Debido a las características intrínsecas de los locutores esofágicos, la cantidad de datos disponible para realizar el entrenamiento y el test del sistema de conversión es escasa. Se utilizan por tanto 100 frases paralelas de un locutor esofágico del corpus descrito en el capítulo 3 y de un locutor sano. De las 100 frases, 90 se utilizan para entrenamiento y las 10 restantes para evaluar el sistema, utilizándose validación cruzada de 10 iteraciones (o “folds”).

6.3.2 Conversión espectral

Antes de entrenar la red LSTM es necesario alinear las frases paralelas de los locutores origen y destino. Debido a las características del habla esofágica, existe un desajuste importante con las frases sanas, así que se utiliza la técnica de alineamiento descrita en el apartado 6.1.

Con el resultado del alineamiento se construyen las entradas y las salidas de la red. A los 24 coeficientes cepstrales alineados del locutor origen se les añaden sus correspondientes derivadas de primer orden (se excluyen c_0 y c'_0).

Se concatenan todos los vectores de todas las frases uno detrás del otro, obteniéndose como resultado dos matrices para cada par de locutores origen-destino. Después se aplica normalización en media y varianza a cada dimensión para facilitar el entrenamiento de la red. Además, se añade un vector de decisiones sonoro/sordo en forma de 0 o 1 obtenido directamente de la $\log f_0$. Esto hace que la dimensión final de la entrada sea 49. La matriz resultante se divide en secuencias de 50 tramas antes de entrar a la red.

La red predecirá los 48 coeficientes MCEP del locutor destino. Para ello, la métrica a minimizar durante el entrenamiento es el MSE. El número de celdas de la capa LSTM es 100. El número de tramas que compone cada lote de secuencias es 50. La salida se obtiene de una capa totalmente conectada (*"fully connected"*). Esta capa da una secuencia de 50 tramas de dimensión 48 por cada secuencia que se introduce a la entrada. Para entrenar se utilizó un optimizador RMSProp (dividir el gradiente entre una media actualizada de su valor reciente) con una tasa de aprendizaje de 0.0001. Así mismo, se utilizó una tasa de dropout de 0.5 para evitar el sobreentrenamiento. Se entrenó la red durante 60 epochs.

Para analizar el comportamiento de la red se compara el resultado predicho con los datos del locutor destino. En la figura 6.2 se muestra para todas las tramas de 10 de las frases de test el MSE entre los coeficientes MCEP normalizados predichos por la red y los originales del locutor destino. Como puede verse, según aumenta el orden de los coeficientes la similitud entre el cepstrum predicho y el cepstrum destino disminuye. Hay que tener en cuenta que para poder hacer este cálculo es necesario alinear las frases test origen con las de destino, ya que su longitud será diferente.

Uno de los problemas observados en la conversión espectral realizada con la LSTM es que los coeficientes obtenidos presentan un suavizado debido al efecto de promediado que aplica la red neuronal. Para paliar este efecto se decidió aplicar el método MLPG [116], ya explicado en detalle en el apartado 5.1.3, ya que el uso del postfiltrado descrito en 5.2.1.3 no conseguía buenos resultados para esta conversión. El modo de aplicar esta técnica es el siguiente:

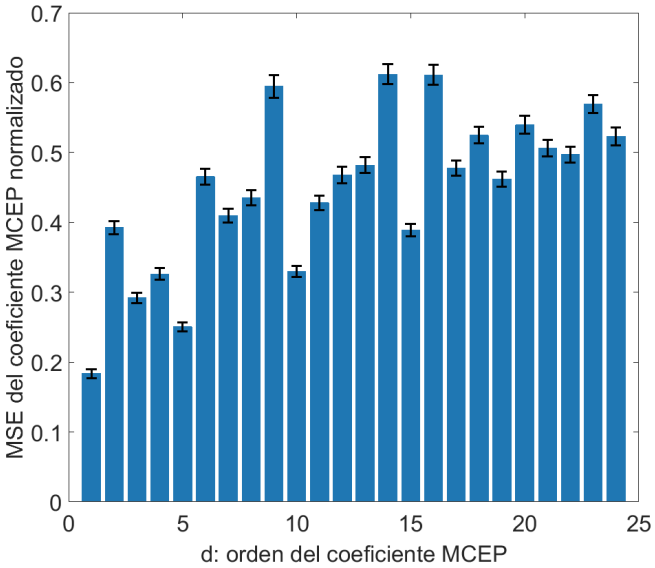

Figura 6.2: MSE con los intervalos de confianza del 95 % de cada coeficiente MCEP normalizado para un fold (10 frases de test).

- Se consideran los MCEP obtenidos de la conversión espectral llevada a cabo por la red LSTM como los vectores de medias de una distribución Gaussiana.

- Al igual que en [45], la matriz de covarianzas definida en la ecuación 5.12 se obtiene a partir del error cuadrático medio existente entre los vectores de características originales de los datos de entrenamiento y los predichos por la red.

- La varianza global (GV global variance) se calcula a partir de los datos de entrenamiento del locutor destino y se utiliza en el método MLPG para hacer la conversión espectral junto a las medias y covarianzas para cada frase de test.

En test internos realizados en el laboratorio se pudo apreciar que la calidad de la voz obtenida mediante esta técnica era superior que la de la voz resintetizada con la conversión salida directamente de la red neuronal.

6.3.3 Estimación de la frecuencia fundamental

Se han utilizado dos redes diferentes para llevar a cabo la estimación de la frecuencia fundamental, una para la predicción de la f_0 y otra para la estimación de la decisión sordo/sonoro. En la figura 6.3 se puede ver la arquitectura básica del sistema. En ambos casos, el modelo LSTM se entrena para mapear la relación entre los coeficientes MCEP sanos y su secuencia de f_0 correspondiente.

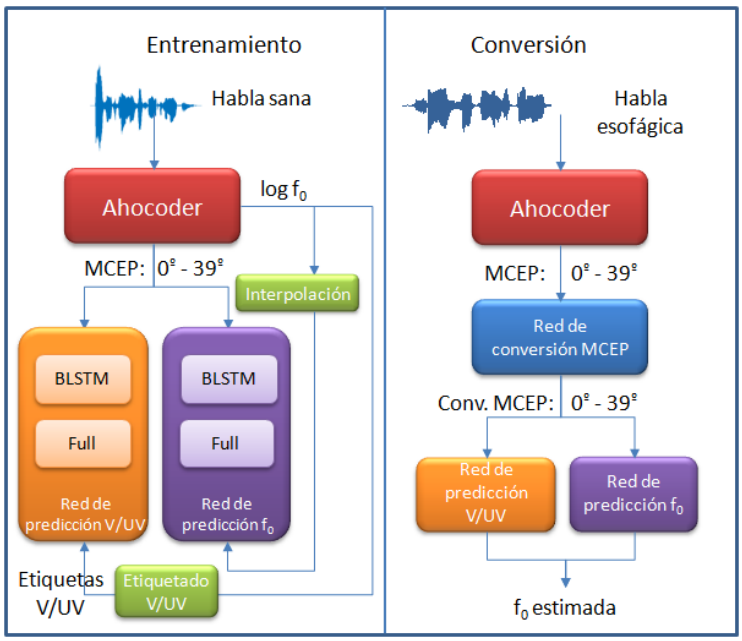

Figura 6.3: Arquitectura del sistema de estimación de f_0 para las etapas de entrenamiento y conversión.).

Se aplica normalización de media y varianza a cada dimensión del vector de cepstrums. La $\log f_0$ se interpola linealmente en las tramas sordas y también se le aplica normalización de media y varianza. Experimentos de desarrollo internos mostraron que la utilización de 40 coeficientes MCEP daban una estimación mejor de la decisión sordo/sonoro que al usar sólo 25 coeficientes. Por tanto, para el modelado de la entonación se utilizaron 40 coeficientes cepstrales (sin derivadas en este caso).

Para la decisión sordo/sonoro se utiliza una capa LSTM Bidireccional (BLSTM) con 64 celdas. La salida proviene de una capa totalmente conectada con una fun-

ción de activación sigmoide. La red se optimiza utilizando el algoritmo Adam con minilotes de tamaño 10 en un entrenamiento de 60 epochs. Como regularización se aplica una tasa de dropout de 0.2. La función de pérdidas utilizada es la entropía cruzada.

La misma configuración se utiliza para la predición de f_0: una capa BLSTM de 64 celdas seguida por na capa totalmente conectada, en este caso con activación lineal. Ahora la métrica a minimizar es el MSE, como ocurría en el caso de la conversión espectral.

En la etapa de conversión se necesitan 40 coeficientes MCEP. Para obtenerlos se ha entrenado otra red, similar a la utilizada para la conversión espectral. Por último, se utilizan estas dos redes BLSTM para estimar la $\log f_0$a partir de los MCEPs convertidos.

6.3.4 Evaluación

Para poder evaluar los resultados de la conversión, se hizo una evaluación objetiva y otra subjetiva.

6.3.4.1 Evaluación objetiva

Para tener una medida objetiva con la que poder evaluar el funcionamiento del sistema se decidió medir la tasa de palabras errónea (WER - word error rate) haciendo uso de un reconocedor automático de habla. El reconocedor usado ha sido el reconocedor explicado en el apartado 4.3.2: Los modelos acústicos están entrenados con voces sanas, pero el diccionario de reconocimiento y el modelo de lenguaje se ajustan al experimento llevado a cabo.

Se han llevado a cabo tres experimentos de reconocimiento diferentes:

- El primero reconoce las 100 frases originales grabadas por el locutor laringectomizado, la sesión denominada 02M3.
- El segundo evalúa las señales resintetizadas mediante Ahocoder utilizando los cepstrum esofágicos originales y la $\log f_0$ estimada por la red.
- El último utiliza como entrada al reconocedor las 100 señales resintetizadas con Ahocoder a partir del cepstrum convertido y la $\log f_0$ estimada.

Tabla 6.2: Valores de WER para los diferentes experimentos.

Caso	WER (%)
Original sano (destino)	10.15
Original esofágico (origen)	56.93
Cepstrum original esofágico + f_0 estimada	57.91
Cepstrum convertido + f_0 estimada	41.48

Los resultados pueden observarse en la tabla 6.2. También se muestra el WER de las 100 frases sanas del lcoutor destino utilizado en la conversión. Este valor, 10.15 %, se toma como una cota "superior" del funcionamiento del sistema, es decir, el menor valor que tomaría el WER si la conversión fuese perfecta. Se puede comprobar la dificultad de la tarea viendo que el valor del WER de las 100 frases del locutor esofágico origen es 56.93 %, más de un 46 % peor.

Como era de esperar, si sólo se cambia la f_0 el sistema se comporta de manera muy similar. La pequeña variación de menos del 1 % en el WER probablemente se deba al proceso de resíntesis y al cálculo de parámetros desde la forma de onda que realiza Kaldi. Sin embargo, cuando las señales reconocidas son las resintetizadas con el cepstrum convertido y la f_0 estimada, el WER cae en un 15 % en términos absolutos. Este hecho muestra que el sistema de conversión es capaz de corregir algunos de los problemas presentes en el espectro esofágico. De todos modos, el WER de las señales convertidas sigue estando muy lejos del obtenido para las señales sanas.

6.3.4.2 **Evaluación subjetiva**

Se llevó a cabo un test perceptual para evaluar el grado de preferencia que tiene cada método de conversión, incluyendo las señales originales sin ningún tipo de procesado, entre los evaluadores. En este test para cada evaluador se eligieron 18 frases de las 100 de manera aleatoria, y de cada frase se le presentaron dos de los tres métodos (esofágica original (ORIGINAL), espectro original + f_0 estimado (F0) y espectro convertido + f_0 estimado (F0 & SPEC)). Se pidió a las personas que escucharon las frases que evaluaran su preferencia entre los dos casos dando una puntuación en una escala de 5 valores.

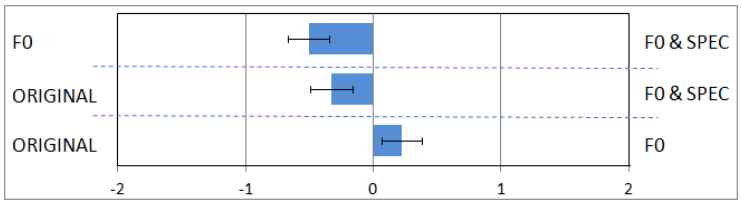

Figura 6.4: Resultado del test de preferencia con intervalos de confianza. -2:Prefiero claramente la frase 1, -1:Prefiero la frase 1, 0:No tengo preferencia, 1:Prefiero la frase 2, 2: Prefiero claramente la frase 2

Un total de 31 evaluadores hablantes nativos de castellano participaron en el test. En la figura 6.4 se puede ver el promediado de los resultados obtenidos al comparar los tres sistemas en pares. Como puede apreciarse, el sistema preferido es el que conserva el cepstrum esofágico original y tiene la entonación modificada. El sistema que convierte sólo la f_0 es preferido claramente sobre el sistema que convierte tanto el espectro como la f_0. Además, las señales esofágicas originales son preferidas sobre las señales con el espectro convertido. Por tanto, las señales que menos convencen a los evaluadores son las que presentan conversión de cepstrum. Esto puede ser debido a la pérdida de calidad introducida en la conversión de los coeficientes MCEP. La figura 6.5 muestra el grado de preferencia para cada par de sistemas comparados. Los resultados dicen que la opción de "preferencia clara" es la opción menos escogida en los tres pares. Además, se puede ver que el método preferido es el que sólo modifica la f_0 seguido por las señales originales (sin modificaciones).

6.3.5 Resultados

Gracias al experimento propuesto se han podido inspeccionar distintas facetas del sistema de conversión. Se han evaluado los efectos que tiene la conversión en términos de inteligibilidad y agradabilidad en comparación con las señales esofágicas originales. Se ha podido comprobar el efecto de únicamente convertir la prosodia separadamente de la conversión completa (MCEPs y f_0 conjuntamente).

Los resultados de comprobar la inteligibilidad por medios de un ASR muestran que la utilización de un locutor sano como destino de la conversión hace que los

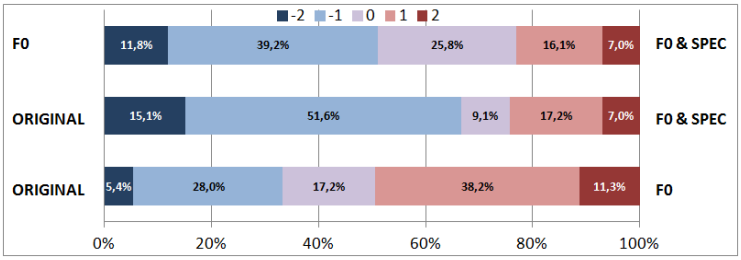

Figura 6.5: Resultados detallados del test de preferencia. -2:Prefiero claramente la frase 1, -1:Prefiero la frase 1, 0:No tengo preferencia, 1:Prefiero la frase 2, 2: Prefiero claramente la frase 2

resultados mejoren notablemente: los valores de WER experimentan una mejora absoluta del 15 % en el experimento realizado. Este valor sigue estando alejado del conseguido para las señales sanas, pero demuestra que la conversión de los parámetros espectrales está ayudando a las señales esofágicas a parecerse más a las señales sanas con las que se ha entrenado el ASR.

Si se analiza el efecto de la estimación de f_0 se puede comprobar que no tiene un impacto significativo en la tasa de reconocimiento. Sin embargo, es curioso ver que esta modificación fue la preferida por los evaluadores participantes en el test perceptual, incluso por encima de las señales esofágicas originales sin ninguna modificación. Este es un resultado importante porque corrobora la importancia de tener una señal fuente restaurada sin modificar las características del locutor. Por último, destacar que aunque la conversión cepstral conseguida por este sistema arregla problemas relacionados con la inteligibilidad de la señal desde el punto de vista de un sistema de reconocimiento automático, modifica aspectos de la señal que hace que sea percibida por la mayoría de evaluadores como poco natural o extraña.

6.4 Técnicas de conversión basadas en PPGs

Las técnicas de conversión descritas en apartados anteriores han necesitado una base de datos paralela para poder ser entrenados. En este apartado se presenta un sistema que no requiere de esta condición, y que está basado en el trabajo descrito en [109]. En la figura 6.6 se ha representado de forma esquemática el sistema descrito en dicho trabajo. Como puede verse, en la fase de conversión se propone utilizar una red neuronal que obtiene los coeficientes espectrales a partir de los posteriorgramas fonéticos (PPG - phonetic posteriorgrams). Estos PPGs representan las probabilidades a posteriori de cada clase fonética, y se extraen utilizando un ASR independiente de locutor (SI-ASR - speaker independent ASR). La red neuronal se ha entrenado utilizando los coeficientes espectrales del locutor destino. Así, este sistema es teóricamente independiente del locutor origen, ya que los PPGs están obtenidos de un sistema independiente del locutor (sistema de tipo *many-to-one*). En el trabajo descrito, la red neuronal está construida con 4 capas BLSTM, y el vocoder utilizado es STRAIGHT [60].

Como ya se ha visto en el apartado 6.1, el alineamiento supone un problema a la hora de afrontar la conversión de voces esofágicas. Por lo tanto, este sistema de conversión se presenta como una buena solución. A continuación se explica cómo se ha implementado dicho sistema para las voces esofágicas.

6.4.1 Conversión espectral

Como se ha explicado, para hacer la conversión espectral descrita en [109], se necesita disponer de dos bloques principales: un ASR capaz de calcular los PPGs, y una red neuronal para relacionar estos PPGs con sus correspondientes características acústicas.

Al utilizar un ASR independiente de locutor se conseguiría obtener PPGs también independientes de locutor siempre que el funcionamiento del reconocedor sea lo suficientemente bueno. En el caso de un reconocedor entrenado para voz sana, esto se cumple cuando la entrada al sistema se trata de un locutor sano cualquiera. Sin embargo, como se ha evidenciado en esta tesis, cuando la voz a reconocer se trata de una voz esofágica, aparecen problemas. Al implementar el sistema se comprobó que al utilizar el SI-ASR estándar, los PPGs obtenidos eran muy diferentes

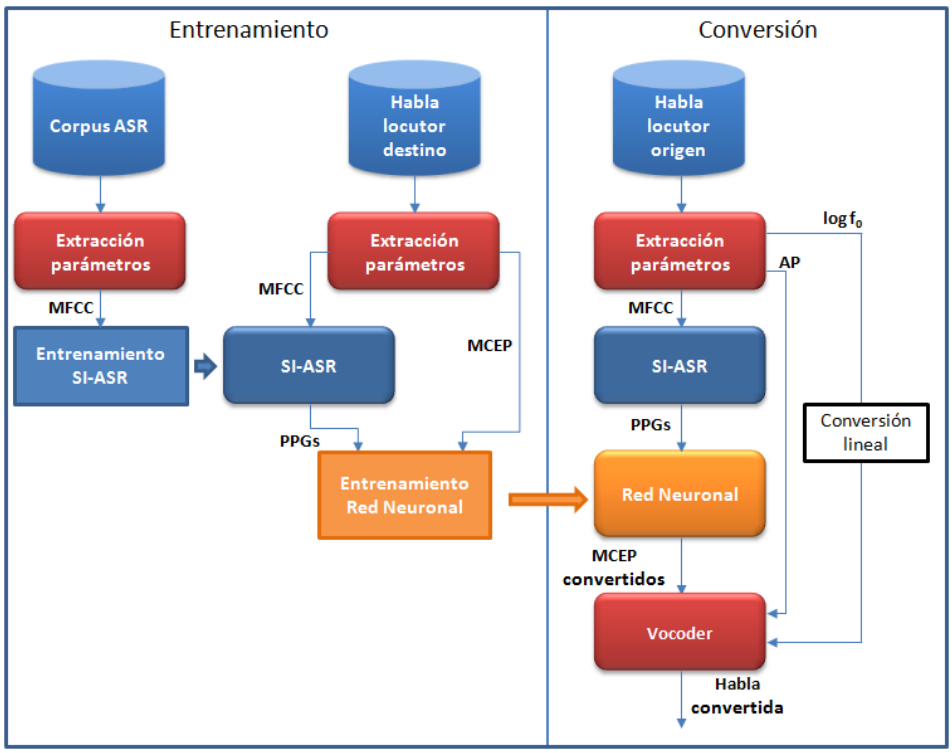

Figura 6.6: Esquema del sistema de conversión de voz con PPGs para voces sanas.

a los obtenidos de la voz sana usados para entrenar la red de conversión. Esto es debido a que los parámetros acústicos de las señales alaríngeas difieren mucho de los de las voces laríngeas. Por tanto, el enfoque seguido ha sido diseñar un sistema ASR específico para esta conversión.

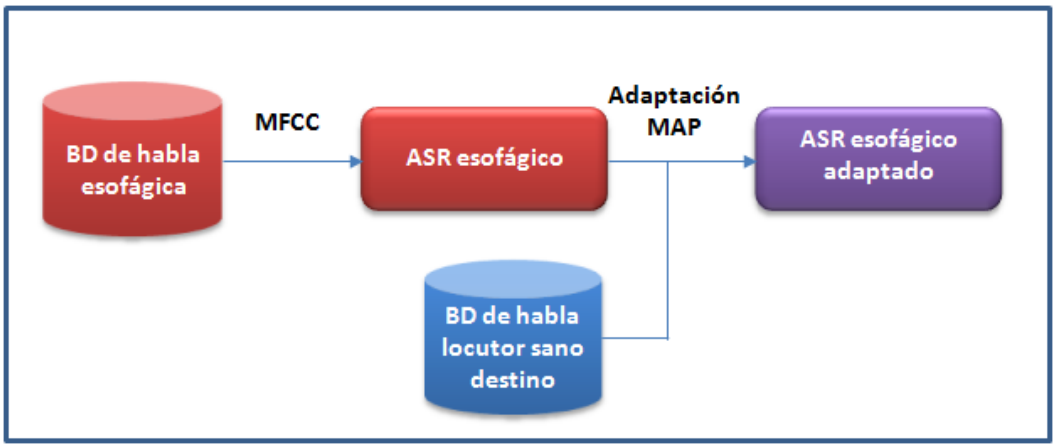

Figura 6.7: Entrenamiento del reconocedor para extraer los PPGs.

En un primer paso, se ha desarrollado un reconocedor esofágico, es decir, se ha entrenado utilizando sólo señales esofágicas. El esquema del proceso puede verse en la figura 6.7. Durante el entrenamiento, se han adaptado los modelos GMM al locutor destino de voz sana específico utilizado adaptación maximum a posteriori (MAP). El motivo de aplicar esta adaptación MAP fue que los PPGs obtenidos al utilizar el reconocedor esofágico directamente con la voz sana llevaban a un entrenamiento infructuoso de la red que predice los coeficientes MCEP. Estos modelos adaptados son los que se usan para obtener los PPGs del locutor sano, PPGs que se utilizan para entrenar la red de conversión, tal y como se muestra en la figura 6.8.

La red que se encarga de convertir los PPGs en coeficientes MCEP del locutor destino contiene 4 capas LSTM bidireccionales (BLSTM - bidirectional LSTM) seguidas de una capa totalmente conectada.

En la fase de conversión, las frases de test del locutor esofágico origen pasan por el ASR esofágico para extraer sus PPGs. Estos PPGs entrarán en la red de conversión y se tendrán los MCEPs convertidos (ver la figura 6.9). En el vocoder entrarán estos parámetros junto con la $\log f0$ estimada, que se explica a continuación.

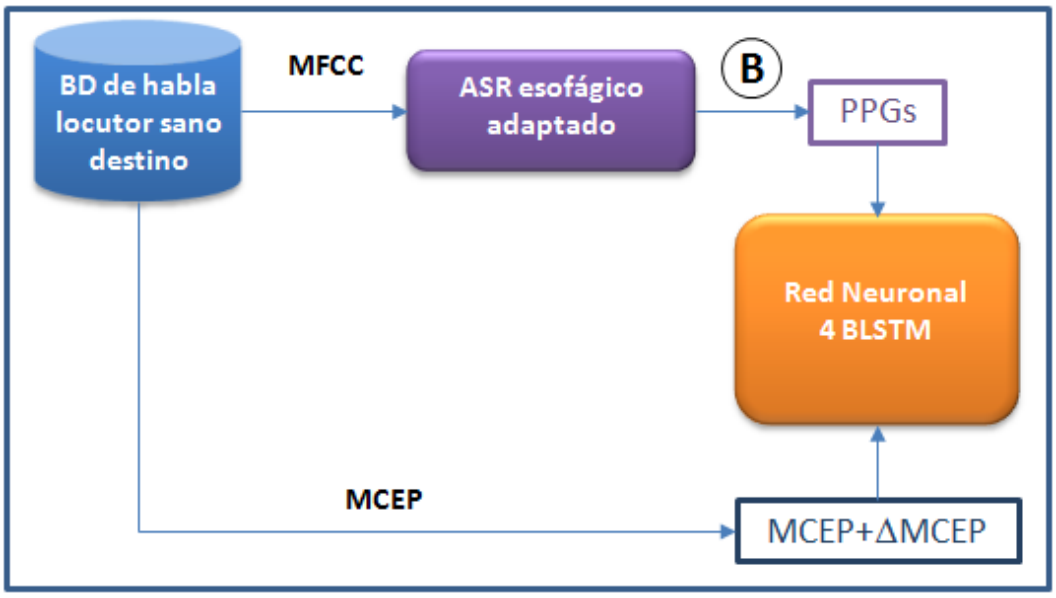

Figura 6.8: Entrenamiento de la red de predicción de los coeficientes MCEP a partir de los PPGs.

6.4.2 Estimación de la frecuencia fundamental

Se han utilizado dos redes para estimar la curva de f_0, una para obtener el valor de $\log f_0$ y otra para estimar la decisión sonoro/sordo. Ambos parámetros se predicen a partir de los coeficientes MCEP obtenidos del locutor destino. La DNN utilizada para estimar la $\log f_0$ es una BLSTM de 4 capas seguida por una capa totalmente conectada con activación lineal, mientras que la arquitectura de la red para la decisión sonoro/sordo se basa en 1 capa BLSTM seguida por una capa totalmente conectada con activación sigmoide.

6.4.3 Configuración de los experimentos

6.4.3.1 Datos de entrenamiento y de test

En este caso, se han utilizado dos conjuntos de datos distintos:

- La base de datos paralela de voz esofágica grabada que se ha descrito en el capítulo 3. Esta base de datos es la que se utilizará para entrenar el ASR

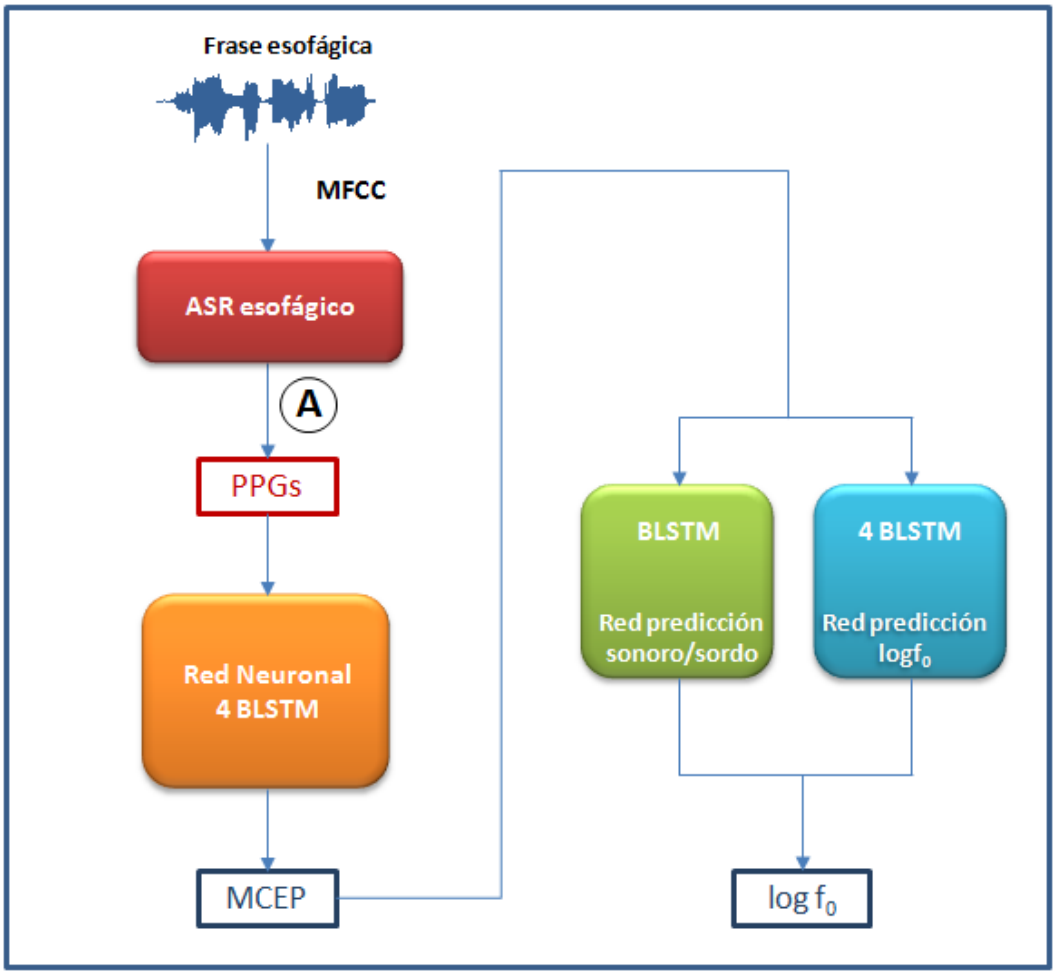

Figura 6.9: Esquema de la conversión utilizando PPGs.

necesario para la extracción de los PPGs. La sesión 02M3 se deja fuera del entrenamiento para ser utilizada como el locutor origen.

- 3995 frases grabadas por un locutor de voz sana, denominado KJC, y que provienen del corpus descrito en [97]. De estas 3995 frases, hay 100 frases que son las que se eligieron para ser grabadas en la base de datos esofágica, así que se descartan. Las 3895 frases restantes (que no han sido utilizadas en el entrenamiento del ASR) son las que se utilizarán para entrenar la red que convierte de PPGs a MCEPs.

6.4.3.2 **Entenamiento del sistema ASR**

Para obtener los PPGs se ha utilizado Kaldi, más concretamente la arquitectura basada en redes neuronales descrita en [129].

Este reconocedor se entrena utilizando las grabaciones de 29 locutores esofágicos. Para evitar que las frases del locutor origen (02) aparezcan en el entrenamiento del reconocedor, de las 100 frases disponibles de cada locutor, se utilizan sólo 90 para entrenar el ASR, lo que resulta en 2610 frases.[1]

El proceso de entrenamiento es el explicado en el capítulo 4, pero con las siguientes particularidades:

1. En la última iteración de la fase de GMMs, se limita a 150 el número de clases fonéticas (senones en este caso).

2. El modelo resultante es el punto de partida para entrenar la fase de la red neuronal del reconocedor que extraerá los PPGs del hablante esofágico (punto A de la figura 6.9).

3. Se aplica adaptación MAP a los modelos GMM del punto 1 con todo el material de entrenamiento que se dispone del locutor sano (3895 frases). Este modelo adaptado es el que se usa para entrenar la fase de red neuronal de otro ASR que dará los PPGs de las frases del locutor de voz sana destino (punto

[1]Para poder tener las 100 frases del locutor de test convertidas y analizar los resultados, se aplica un proceso de validación cruzada 90/10: se entrenan 10 sistemas, cada uno con un conjunto de 90 frases distintas.

B de la figura 6.8). Estos parámetros son los que se usarán para entrenar la red de conversión de PPG a MCEP.

6.4.3.3 Entrenamiento de la red de conversión espectral

Para entrenar el sistema que convierte de PPGs a coeficientes cepstrales, se construye una red neuronal con 4 capas BLSTM, conteniendo cada capa 64 celdas. Como se ha indicado anteriormente, la entrada a esta red son los PPGs que provienen del ASR adaptado al locutor sano destino, más concretamente, los PPGs extraídos las 3895 frases de este locutor (Figura 6.8). La salida la da una capa totalmente conectada. Consiste en un vector de dimensión 48 que contiene los coeficientes MCEP del 1º al 24º (al igual que en la conversión paralela descrita en 6.3, el c_0 se copia del locutor origen en la etapa de conversión) y sus respectivas derivadas de primer orden. Los coeficientes MCEP se extraen utilizando Ahocoder sobre las mismas 3895 frases del hablante sano de las que se han calculado los PPGs.

Para el entrenamiento de la red, se eligió un tamaño de lote de 50 tramas. La función de pérdidas a minimizar es el MSE. El sistema se entrena durante 25 epochs, con una tasa de dropout de 0.2 y el optimizador $Adam$.

6.4.3.4 Estimación de la frecuencia fundamental

Se han entrenado dos redes para obtener la curva de entonación: una para la decisión sonoro/sordo, y otra para los valores de la $\log f_0$. Ambas redes se entrenan con 25 coeficientes MCEP (incluido el c_0) normalizados en media y varianza. Estos coeficientes cesptrales provienen del conjunto de frases del locutor sano destino. En total se utilizan 3895 frases para entrenar y 100 para validar.

Para la red de decisión sonoro/sordo, la capa BLSTM está compuesta por 64 celdas. El vector sonoro/sordo se obtiene directamente de la $\log f_0$ de la voz sana destino. La red se optimiza utilizando el algoritmo Adam con un tamaño de lote de 50, entrenando durante 50 epochs, siendo esta vez la función de pérdidas la entropía cruzada binaria. Se aplica una tasa de dropout de 0.2.

La red de predicción de $\log f_0$ está compuesta por 4 capas BLSTM de 64 celdas cada una. Para el entrenamiento, la curva de $\log f_0$ del locutor destino se interpola linealmente para las tramas sordas. Después se calcula su derivada y se añade a la

$\log f_0$, aplicándose normalización de media y varianza. La configuración del entrenamiento es la misma que para la red de decisión sonora/sordo, pero en este caso la métrica que se busca minimizar es el MSE.

En la etapa de conversión, los 25 coeficientes MCEP obtenidos de la red de conversión espectral se usan para predecir los valores de la $\log f_0$ y la decisión sordo/sonoro, tal y como se muestra en la figura 6.9.

6.4.4 Evaluación

Para evaluar este sistema, se decidió compararlo con el sistema de conversión LSTM descrito en la sección anterior (6.3) y que, a diferencia de este método, necesitaba de una base de datos paralela para su entrenamiento. Ambos sistemas utiliza la red de predicción de $\log f_0$ descrita en el apartado anterior (6.4.3.4).

Para hacer la evaluación se han hecho medidas objetivas como el WER y la distorsión Mel-cepstral (MCD - Mel cepstral distortion), y otra subjetiva mediante un test de preferencia.

6.4.4.1 Evaluación objetiva

Para tener un valor objetivo de la inteligibilidad de las frases convertidas se calcula su WER con el ASR para castellano con el que se ha evaluado los distintos sistemas de conversión y que se describe en el capítulo 4.

La tabla 6.3 muestra los resultados para 4 conjuntos de 100 frases: el locutor esofágico fuente original, el locutor sano destino y las señales convertidas, tanto el sistema LSTM como el basado en PPGs. El WER del locutor esofágico origen es del 56.93 %, muy por encima del 11.88 % de este nuevo locutor sano destino.

Tabla 6.3: Valores de WER para los diferentes experimentos.

Caso	WER (%)
Original sano (destino)	11.88
Original esofágico (origen)	56.93
Cepstrum convertido con LSTM + f_0 estimada	40.58
Cepstrum convertido con PPGs + f_0 estimada	57.91

Al igual que en el experimento anterior (sección 6.3), vemos que el sistema de conversión paralelo LSTM consigue solucionar ciertos problemas del cepstrum esofágico y los acerca a los de la voz sana, mejorándose en este caso los resultados del reconocimiento en un 16 %, resultados que son iguales a los obtenidos en el experimento descrito en el apartado 6.3.4.1 para otra voz sana destino diferente.

Sin embargo, el sistema de conversión que utiliza los PPGs para hacer la conversión y que no necesita de datos paralelos, obtiene resultados de WER ligeramente peores que los del locutor esofágico fuente original. Esto se puede explicar porque aunque el reconocedor usado para extraer los PPGs está entrenado con habla esofágica, las puntuaciones acústicas en el reconocedor del locutor esofágico están muy lejos de las del locutor sano. Esto hace que aparezca un error acústico al obtener los PPGs que se propaga hasta la fase de conversión y se traduce en una mala pronunciación o incluso desaparición de ciertos fonemas en la frase convertida. Este error no tiene la misma relevancia en la conversión descrita en [109], dónde la conversión es entre locutores de voz sana ya que los espacios acústicos están más próximos que la voz sana y la voz esofágica.

Además del WER, se ha calculado también la MCD [75] entre la voz destino y las dos versiones de la conversión de la voz esofágica para tener una idea de la distancia espectral existente entre las voces convertidas y la destino. Los resultados están recogidos en la tabla 6.4. Como se puede comprobar, ambas estrategias de conversión acercan la señal esofágica original a la destino, aunque no hay diferencia significativa entre ambos métodos.

Tabla 6.4: Valores de MCD.

Caso	MCD
Destino y origen	7.840 ± 0.376
Destino y convertidas LSTM	5.025 ± 0.314
Destino y convertidas PPG	5.021 ± 0.313

6.4.4.2 **Evaluación subjetiva**

Para determinar qué versión es la preferida por los oyentes, si la voz esofágica origen o la convertida producida por el método de conversión paralelo o no paralelo,

se realizó un test perceptual similar al descrito en 6.3.4.2.

De las 100 frases disponibles, se escogió para el test de preferencia un set con las 30 frases más inteligibles. Las frases escogidas se calificaron como las más inteligibles utilizando el ASR utilizado previamente para calcular el WER sobre los tres tipos de habla (original, convertida LSTM y convertida PPG).

De estas 30 frases, cada evaluador tuvo que calificar 24 pares de frases elegidas de manera aleatoria, 8 comparando la voz original con la voz convertida mediante el sistema LSTM, 8 comparando la voz original con la voz convertida con el sistema de PPGs y 8 más comparando la voz generada por ambos sistemas de conversión. Los evaluadores escucharon cada par de frases y expresaron su preferencia en una escala de 5 puntos: Prefiero claramente la frase 1 (-2), Prefiero la frase 1 (-1), No puedo decidirme por ninguna de las 2 (0), Prefiero la frase 2 (1), Prefiero claramente la frase 2 (2).

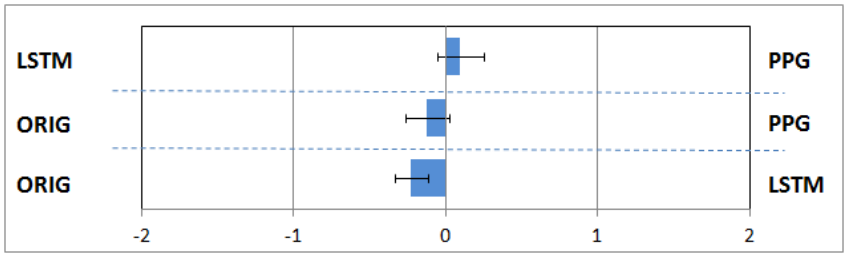

Figura 6.10: Resultado del test de preferencia con intervalos de confianza.

En el test participaron 35 evaluadores hablantes nativos de castellano. Los resultados se pueden ver en la figura 6.10, donde se muestra el promedio de las puntuaciones con los intervalos de confianza al 95 %. Se prefieren las frases originales sobre las convertidas por el sistema LSTM por un margen pequeño pero estadísticamente significativo. Estos resultados corroboran los resultados obtenidos en el apartado anterior (6.3). Cuando se comparan las frases originales con las que provienen del sistema de conversión con PPGs, los oyentes no muestran una preferencia significativa. Los mismos resultados se observan al comparar las dos versiones de la conversión. En este último caso, los evaluadores muestran una ligera preferencia por el sistema de PPGs, pero sin significación estadística.

La figura 6.11 muestra con más detalle el grado de preferencia para cada par de sistemas. Se puede observar que al comparar el sistema PPG frente al LSTM, la mayor parte de los evaluadores no pueden decidirse por ninguno de los dos (39.6 %). Por el contrario, en el caso del sistema PPG frente al habla esofágica original, sólo el 13.6 % los considera equivalentes, el 48.9 % prefiere la versión original y el 37.5 % restante prefiere las frases convertidas con el método de los PPGs. En el caso del sistema LSTM frente a las frases originales, un 15 % se muestran indecisos, el 50.7 % prefieren la versión original y el 34.4 % prefieren el habla proveniente del sistema LSTM.

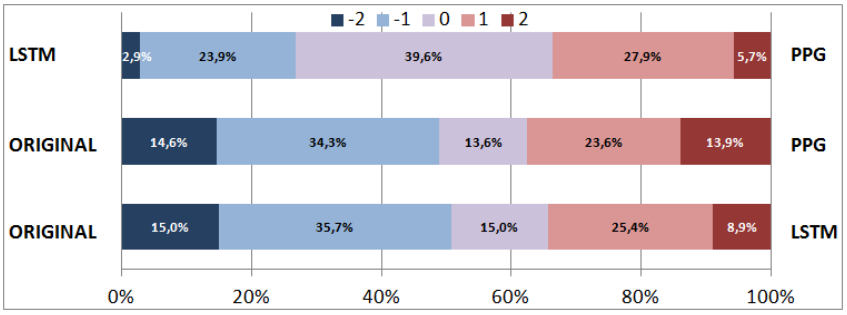

Figura 6.11: Resultados detallados del test de preferencia.

Los resultados del test subjetivo contrastan con los obtenidos en la evaluación objetiva. El sistema con mejor WER (LSTM) no es el preferido por los evaluadores, probablemente porque, aunque es más inteligible, sus señales suenan menos naturales que las otras dos versiones.

6.4.5 Resultados

Con este experimento se ha podido evaluar el funcionamiento de una arquitectura de conversión de voz que no necesita de una base de datos paralela al hacer uso de PPGs adaptada al problema de mejorar la calidad e inteligibilidad de la voz esofágica. Este sistema se ha comparado con la aproximación más clásica consistente en una red LSTM que necesita de una base de datos paralela para su entrenamiento y que se ha descrito en la sección 6.3.

Se ha medido la inteligibilidad mediante el WER utilizando un ASR. Los resultados muestran que mientras la red LSTM consigue mejorar la tasa de reconocimiento, el sistema basado en PPGs no lo hace, consiguiendo unos resultados ligeramente peores que los obtenidos por la voz original. En términos de distorsión MCD, ambos métodos reducen la distancia cepstral al locutor destino en igual medida.

Los resultados del test subjetivo indican que no hay una preferencia clara a favor de las frases convertidas. La mejora de la calidad esperada al utilizar la estrategia no paralela se ve emborronada por la pérdida de inteligibilidad de las frases convertidas.

6.5 Conclusiones

En este capítulo se han evaluado las distintas técnicas de conversión basadas en GMMs utilizando el ASR desarrollado en el capítulo 4. Utilizando como origen los coeficientes cepstrales de un locutor esofágico se ha utilizado el algoritmo MLPG para convertirlos en los cepstrum de un locutor de voz sana. Para resintetizar la señal se necesita además del cepstrum convertido valores de f_0. Estos valores se han obtenido de distintas maneras. Se han hecho diversos experimentos y se ha podido comprobar que el alineado de los locutores afecta mucho al resultado. El método de extracción de pitch a la hora de hacer el análisis de la voz esofágica también produce variaciones en la tasa de error. Los mejores resultados se obtienen cuando el alineamiento de las tramas para entrenar los GMMs se hace de manera manual y se utiliza la autocorrelación sobre el residuo del análisis PSIAIF. Para este experimento, si la f_0 es predicha mediante el algoritmo MGE y se le aplica un suavizado, las señales de test consiguen un WER del 43.28 %. Esto supone una mejora absoluta del 19.75 % si se compara con el WER obtenido para las mismas frases de test del locutor esofágico original.

También se construyó un sistema basado en redes neuronales para hacer la conversión espectral y para estimar la f_0 a partir del cepstrum convertido. Como en el caso de las GMMs, el locutor origen es alaríngeo, mientras que la voz destino es sana. En este caso, para que haya material suficiente para entrenar las redes neuronales se ha hizo una validación cruzada usando 10 iteraciones (90 frases para entrenar y 10 para evaluar). Para la conversión espectral, además de una red BLSTM se ha aplicado MLPG a los MCEP convertidos. La conversión obtenida se ha evaluado tanto objetivamente mediante ASR como subjetivamente mediante tests de preferencia. El análisis de los resultados muestra que para esta conversión se consigue una tasa de error del 41.48 % (mejora absoluta del 15.45 % frente al locutor esofágico origen de la conversión), demostrando que la conversión está arreglando algunos problemas presentes en el espectro de la voz esofágica. El efecto de la estimación de la f_0 no impacta de manera significativa en el WER. Sin embargo, esta modificación es la que resultó más agradable para los oyentes participantes en el test. Cabe señalar que aunque la conversión espectral hace que mejore la inteligibilidad para el sistema ASR es percibida por los oyentes como poco natural.

Por último, se adaptó un sistema de conversión basado en PPGs para convertir de un hablante esofágico origen a una voz sana destino. La característica de este sistema es que no requiere utilizar una base de datos paralela para su entrenamiento. Para implementarlo, primero se entrenó un ASR específico para el habla esofágica. Este ASR se utiliza para extraer los PPGs del hablante destino. Con estos PPGS, se entrena una red neuronal con 4 capas BLSTM para convertirlos en los coeficientes MCEP correspondientes al locutor sano. Después basta con reconocer las frases del locutor origen y entregarle los PPGs extraídos a la red neuronal para convertirlos en coeficientes MCEP con las carácterísticas acústicas del locutor destino. La evaluación objetiva de este sistema mediante un ASR muestra que no se consigue una mejora en el WER respecto al habla esofágica original. También se hizo una evaluación subjetiva en la que se compararon dos a dos este sistema, la voz esofágica original y el sistema LSTM descrito anteriormente. Los resultados de este test indican que no hay una preferencia clara entre ninguno de los sistemas. La conclusión principal extraída es que la mejora de la calidad esperada al utilizar la estrategia no paralela se ve lastrada por la pérdida de inteligibilidad de las frases convertidas.

Las investigaciones explicadas en este capítulo se han publicado en [105] y [103].

Volver donde empezaste no es lo mismo que nunca haber marchado.

Terry Pratchett

CAPÍTULO

7

Conclusiones

En esta tesis se han abordado diversos aspectos con el objetivo de mejorar la inteligibilidad de las voces alaríngeas. Los esfuerzos se han centrado mayoritariamente en las técnicas de conversión de voz, pero también se ha profundizado en otras facetas relacionadas con esta investigación. Con los trabajos realizados se ha conseguido mejorar la tasa de reconocimiento de los locutores esofágicos. Se ha participado en distintas evaluaciones para poder evaluar el alcance de este trabajo de una manera contrastable.

En este capítulo se recogen las conclusiones derivadas de la realización de esta tesis. En un primer apartado se listarán las aportaciones hechas en las distintas áreas exploradas durante este trabajo, así como las líneas futuras que pudieran derivarse del mismo. Finalmente, se presentan las publicaciones y conferencias hechas para difundir los resultados obtenidos.

7.1 Aportaciones de la tesis y trabajos futuros

Bases de datos

Al iniciar este trabajo se investigó sobre las diferentes grabaciones existentes relacionadas con el habla esofágica. El material encontrado no fue muy abundante. En

la mayoría de la literatura revisada, cada trabajo de investigación utiliza sus propias grabaciones hechas a medida para el aspecto en concreto a investigar. Estas grabaciones suelen ser escasas, muchas de ellas se reducen a vocales sostenidas o unas cuantas palabras o frases.

Para poder llevar a cabo las técnicas de conversión de voz propuestas en esta tesis, es necesario de disponer de una base de datos paralela con material suficiente. Ante la falta de una base de datos con estas características que incluyese voz alaríngea, se decidió grabar una propia. **Con la base de datos creada se cubre un vacío existente**.

Además de la grabación de la base de datos, se ha etiquetado a nivel de fonema, de forma manual (parcialmente) y automática. Se probaron distintas técnicas, pero se llegó a la conclusión que la que daba mejores resultados es utilizar alineamiento forzado entrenando los modelos acústicos con las propias frases esofágicas de la base de datos. El etiquetado resultante está evaluado de forma objetiva y se proporciona junto con las señales.

También se ha investigado el impacto que tiene el método de extracción de la f_0 en los resultados del análisis acústico. El método tradicional de la autocorrelación funciona muy bien con las voces sanas, pero con las voces alaríngeas presenta más problemas. El método finalmente elegido está basado en la autocorrelación del residuo del análisis PSIAIF y es el que se ha usado para parametrizar posteriormente la voz alaríngea. Se ha calculado la frecuencia fundamental y su desviación, así como el jitter, el shimmer y la velocidad de habla de distintas maneras para todos los locutores, quedando así caracterizados. Otro punto importante realizado es el análisis de los formantes, viéndose que la manera propuesta de calcular la frecuencia fundamental funciona mejor para las voces alaríngeas.

En el momento de finalizar este documento, **se están realizando las gestiones para distribuir la base de datos a través de un repositorio público (ELRA)**. Además, parte del trabajo relacionado con esta base de datos **se ha enviado para su publicación y está en espera de recibir respuesta**.

Alineamiento entre voces sanas y voces esofágicas

A la hora de hacer conversión estadística de voz es fundamental contar con una base de datos paralela. Además, es necesario realizar un proceso de alineamiento para cada par de frases locutor origen - locutor destino con el objetivo de aprender las relaciones entre las realizaciones acústicas de ambos locutores. Para el caso en que tanto la voz origen como la destino son sanas, el alineamiento puede hacerse de una manera sencilla utilizando DTW.

Sin embargo, cuando se necesita alinear las frases de un locutor de voz patológica con las de un locutor de voz sana, la manera de conseguirlo no es tan directa. Hay muchas diferencias entre las frases: la velocidad de habla del locutor esofágico es más lenta, se introducen muchas más pausas y silencios, la pronunciación de los distintos fonemas varía, etc. Es por ello que la estrategia elegida ha sido partir del etiquetado de las grabaciones hecho con las transcripciones y con las marcas de tiempo alinear cada par de fonema esofágico y fonema sano mediante DTW. Las pruebas realizadas en este sentido muestran que **el método de alineamiento diseñado mejora los resultados** de la conversión con respecto al alineamiento estándar.

ASR

Para poder evaluar los resultados de los algoritmos investigados, uno de los criterios utilizados ha sido utilizar la tasa de error de un sistema de reconocimiento como valor de comparación. Por ello, **se ha desarrollado un reconocedor automático del habla continua para el castellano basado en redes neuronales** con un WER de un 11 %, que además ha sido modificado para permitir la evaluación específica de los algoritmos propuestos. El sistema desarrollado ha sido validado con la **participación en una competición de detección de términos hablados** obteniendo el segundo puesto entre los 5 sistemas presentados. **Este trabajo de evaluación se ha publicado en [113]**.

Conversión de voz

El principal esfuerzo de esta tesis se ha centrado en conseguir técnicas que permitan mejorar la inteligibilidad de las voces alaríngeas. Para ello se han investigado principalmente técnicas estadísticas de conversión de voz, tanto utilizando GMMs como DNNs.

En primer lugar se han trabajado estas técnicas con voces sanas, con el objetivo principal de desarrollar los métodos y realizar una evaluación inicial, que sirviera como objetivo de referencia. En la conversión estadística con mezclas de gaussianas se han revisado cuatro métodos distintos para convertir desde parámetros de habla silenciosa (PMA) a parámetros espectrales (MCEP) del mismo locutor. Los resultados obtenidos (distrosión mel-cepstral) muestran que los algoritmos basados en MLPG en combinación con un criterio de entrenamiento MGE obtiene las distorsiones más bajas. **Los resultados de este trabajo se publicaron en [37]**.

Estas técnicas de conversión PMA-voz han sido la base para la conversión posterior de señales esofágicas. Por otro lado, no es descabellado pensar que el uso de las señales PMA obtenidas de una persona laringectomizada pueda mejorar la inteligibilidad de sus voces. Este aspecto no se ha abordado en la tesis, pero presenta mucho interés, y ha generado una posible línea de trabajo futuro en el grupo.

Además de los enfoques tradicionales, se han investigado sistemas basado en redes neuronales. En general, para entrenar redes neuronales se requieren una gran cantidad de datos, por lo que las expectativas en este sentido no eran muy altas. Sin embargo, era necesario experimentar con ellas. Para aprovechar la fuerte dependencia temporal existente entre las tramas consecutivas de la señal de voz, se optó por utilizar arquitecturas basadas en LSTM. El sistema finalmente diseñado utiliza dos redes, una para convertir los valores de pitch y otra para convertir las características espectrales. Con este sistema **se participó en el segundo Voice Conversion Challenge y los resultados obtenidos sitúan al sistema construido en una posición intermedia entre todos los sistemas presentados**, tanto en valores de similitud como de naturalidad. Se han comparado los resultados de este sistema con el presentado al primer Voice Conversion Challenge, basado en GMMs. El test MUSHRA llevado a cabo indica que no existen diferencias significativas entre

ellos. Se puede concluir que cuando los datos son escasos para el entrenamiento, el sistema basado en LSTMs y el basado en GMMs producen resultados similares.

Tanto para el caso de señales PMA como para las señales esofágicas es necesario estimar el valor de la f_0 -algo que no es necesario en la conversión con voces sanas. Las mismas técnicas utilizadas para la conversión espectral se han evaluado en la estimación de la frecuencia fundamental, dando muy buenos resultados (**con un Fscore de 0.96 tanto para la clasificación de las tramas sordas como las sonoras**).

Con las técnicas de conversión implementadas se ha conseguido mejorar el valor del WER del 63 % de la señal esofágica inicial a valores cercanos al 43.3 % para la señal resintetizada a partir del espectro convertido y de la f_0 obtenida con distintos métodos, lo que supone una mejora relativa de un 31 %). Para llegar a este valor, es necesario utilizar los datos del alineamiento manual. Con alineamiento automático se ha llegado a obtener un WER de 45.6 % (**mejora relativa de un 27,6 %**). Los diferentes experimentos han demostrado la importancia de un buen alineamiento y una buena estimación del valor de f_0 para obtener una buena parametrización.

En los experimentos realizados **con redes LSTM la mejora conseguida ha sido de un 27.1 %** con alineamiento manual, es decir, algo inferior al sistema GMM. Sin embargo, en el grupo el uso de redes neuronales comenzó con el desarrollo descrito en esta tesis, y se ha realizado un esfuerzo muy importante en la obtención de un sistema funcional. El ajuste de los hiperparámetros creemos que puede ser mejorado. La mejora del WER obtenida es importante, pero siguen siendo valores muy alejados de los obtenidos para las voces sanas. Se necesita por tanto probar nuevas arquitecturas de redes, como usar autoencoders para poder utilizar más cantidad de datos con los que se consigan entrenamientos mejores.

Con la idea de utilizar arquitecturas que puedan ser entrenadas con datos de voces sanas (más abundantes) y posteriormente adaptadas a las voces esofágicas, se implementó el sistema basado en PPGs, que tiene la gran ventaja de no requerir datos paralelos para su entrenamiento. Además, la calidad final que se obtuvo en pruebas informales para voces sanas nos animaron a explorar esta línea. Sin embargo, los resultados obtenidos no fueron satisfactorios, no mejorando el WER del locutor esofágico original. La solución pasaría por conseguir un ASR que funcionase mejor con hablantes esofágicos (utilizándose mayor cantidad de datos, por

ejemplo), pero dado que conseguir un reconocimiento mejor de la voz esofágica sin modificar un reconocedor estándar es lo que se está buscando al utilizar diferentes técnicas de conversión, usar esta técnica es incompatible con los objetivos planteados en esta tesis.

La evaluación de todas estas técnicas por métodos perceptuales revela que aunque las distintas técnicas de conversión consiguen mejores tasas de reconocimiento, los evaluadores no tienen una preferencia clara cuando se comparan con la voz esofágica original. La conclusión más sorprendente ha sido que de **entre todas las conversiones, la preferida por algunos oyentes utiliza únicamente conversión de la frecuencia fundamental**, sin modificar los coeficientes espectrales. Esto **demuestra la importancia de la restauración de la prosodia**. En esta tesis no se han investigado técnicas que modifiquen la duración y el ritmo de las señales, y este puede ser un camino futuro interesante. En este sentido, la aparición reciente de otro tipo de vocoders como Wavenet [120] que no hacen uso de la f_0 para resintetizar las señales abre nuevas posibilidades.

7.2 Difusión de resultados

Se enumeran a continuación las contribuciones a revistas y conferencias internacionales realizadas durante la elaboración de esta tesis.

ARTÍCULOS DE REVISTA

2019 Luis Serrano, Inma Hernaez, Eva Navas, Sneha Raman Jon Sanchez, *"A new multispeaker database of esophageal speech"*, **Enviado** a PLoS One (Q2 en 2018).

2017 Javier Tejedor, Doroteo Toledano, Paula Lopez-Otero, Laura Docio-Fernandez, Luis Serrano, Inma Hernaez, Alejandro Coucheiro-Limeres, Javier Ferreiros, Julia Olcoz, Jorge Llombart, *"ALBAYZIN 2016 spoken term detection evaluation: an international open competitive evaluation in Spanish"*, EURASIP Journal on Audio, Speech, and Music Proc. (Q1 en 2017), vol. 2017, no. 1, p. 22, 2017.

2015 Daniel Erro, Agustín 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.

PUBLICACIONES EN CONGRESOS

2019 Luis Serrano, Sneha Raman, David Tavárez, Eva Navas, Inma Hernáez, *"Parallel vs. Non-parallel Voice Conversion for Esophageal Speech"*, In Proceedings of INTERSPEECH 2019, Graz, Austria, aceptado en proceso de publicación.

2018 Luis Serrano, David Tavárez, Xabier Sarasola, Sneha Raman, Ibon Saratxaga, Eva Navas, Inma Hernáez *"LSTM based voice conversion for laryngectomees"*, In Proceedings of IberSPEECH 2018, Barcelona, Spain, pp. 122-126, 2018

2018 Sneha Raman, Inma Hernáez, Eva Navas, Luis Serrano, *"Listening to Laryngectomees: A study of Intelligibility and Self-reported Listening Effort of Spanish Oesophageal Speech"*, In Proceedings of IberSPEECH 201, Barcelona, Spain, pp. 107-111, 2018

2018 Xabier Sarasola, Eva Navas, David Tavárez, Luis Serrano, Ibon Saratxaga, *"Speech and monophonic singing segmentation using pitch parameters"*, In Proeedings of IberSPEECH 2018, Barcelona, Spain, pp. 147-151, 2018

2018 Igor Odriozola, Inma Hernáez, Eva Navas, Luis Serrano, *"The observation likelihood of silence: analysis and prospects for VAD applications"*, In Proceedings of IberSPEECH 2018, Barcelona, Spain, pp. 50-54, 2018

2017 David Tavárez, Xabier Sarasola, Agustín 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, Agustín 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, Agustín 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

2016 Daniel Erro, Inma Hernáez, Luis Serrano, Ibon Sratxaga, Eva Navas, *"Objective Comparison of Four GMM-Based Methods for PMA-to-Speech Conversion"*, In Advances in Speech and Language Technologies for Iberian Languages, Lisboa, Portugal, pp. 24-32, 2016

2016 Luis Serrano, David Tavárez, Igor Odriozola, Inma Hernáez, Ibon Saratxaga, *Aholab system for Albayzin 2016 Search-on-Speech Evaluation*, In Proceedings of IberSPEECH 2016, Lisboa, Portugal, pp. 23-25, 2016

2014 Igor Odriozola, Luis Serrano, Inma Hernáez, Eva Navas, *"The AhoSR Automatic Speech Recognition System"*, In Advances in Speech and Language Technologies for Iberian Languages, Lecture Notes in Computer Science, vol. 8854, pp. 279–288. Las Palmas de Gran Canaria, Spain, 2014

7.3 Participación en campañas de evaluación

ALBAYZIN

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

ALBAYZIN (SPOKEN TERM DETECTION):

2016 *Second best STD system in the Albayzin search on speech evaluation*, In IberSPEECH 2016, Lisboa, Portugal.

VOICE CONVERSION CHALLENGE:

2016 *GMM system submited to the* 1^{st} *Voice Conversion Challenge*.

2018 *LSTM system submited to the Hub task of the* 2^{nd} *Voice Conversion Challenge*.

Bibliografía

[1] Abad, Alberto, Rodriguez-Fuentes, Luis Javier, Penagarikano, Mikel, Varona, Amparo, & Bordel, Germán. 2013. On the calibration and fusion of heterogeneous spoken term detection systems. *Pages 20–24 of: Interspeech*. 86

[2] Abe, Masanobu, Nakamura, Satoshi, Shikano, Kiyohiro, & Kuwabara, Hisao. 1990. Voice conversion through vector quantization. *Journal of the Acoustical Society of Japan (E)*, **11**(2), 71–76. 20

[3] Aguilar, Gualberto, Pérez-Meana, Héctor, Nakano-Miyatake, Mariko, & Becerril-Mendoza, Héctor. 2004. Speech enhancement of voice produced by an electronic larynx. *Pages iii–37 of: Circuits and Systems, 2004. MWSCAS'04. The 2004 47th Midwest Symposium on*, vol. 3. IEEE. 19, 25

[4] Aguilar-Torres, Gualberto, Nakano-Miyatake, Mariko, & Perez-Meana, Hector. 2006. Enhancement and restoration of alaryngeal speech signals. *Pages 31–31 of: Electronics, Communications and Computers, 2006. CONIELECOMP 2006. 16th International Conference on*. IEEE. 19, 25

[5] Alku, Paavo. 1992. Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering. *Speech communication*, **11**(2-3), 109–118. 56

[6] Arslan, Levent M. 1999. Speaker Transformation Algorithm Using Segmental Codebooks (STASC) 1. *Speech Communication*, **28**(3), 211–226. 20

[7] Benisty, Hadas, & Malah, David. 2011. Voice conversion using GMM with enhanced global variance. *In: Twelfth Annual Conference of the International Speech Communication Association*. 20

[8] Bonafonte, Antonio, Kain, Alexander, Santen, Jan van, & Duxans, Helenca. 2004. Including dynamic and phonetic information in voice conversion systems. *In: Eighth International Conference on Spoken Language Processing*. 20

[9] Brown, Dale H, Hilgers, Frans JM, Irish, Jonathan C, & Balm, Alfons JM. 2003. Postlaryngectomy voice rehabilitation: state of the art at the millennium. *World journal of surgery*, **27**(7), 824–831. 4

[10] Brownlee, Jason. 2017. *Long Short-Term Memory Networks With Python. Develop Sequence Prediction Models With Deep Learning*. Machine Learning Mastery. 131

[11] Can, Doğan, & Saraclar, Murat. 2011. Lattice indexing for spoken term detection. *IEEE Transactions on Audio, Speech, and Language Processing*, **19**(8), 2338–2347. 87

[12] Cervera, Teresa, Miralles, José L, & González-Àlvarez, Julio. 2001. Acoustical analysis of Spanish vowels produced by laryngectomized subjects. *Journal of speech, language, and hearing research*, **44**(5), 988–996. 14, 25, 62

[13] Cheah, Lam A., Bai, Jie, Gonzalez, Jose A., Ell, Stephen R., Gilbert, James M., Moore, Roger K., & Green, Phil D. 2015. A User-centric Design of Permanent Magnetic Articulography based Assistive Speech Technology. *Pages 109–116 of: Proc. Biosignals*. 126

[14] Chen, Guoguo, Yilmaz, Oguz, Trmal, Jan, Povey, Daniel, & Khudanpur, Sanjeev. 2013. Using proxies for OOV keywords in the keyword search task. *Pages 416–421 of: Automatic Speech Recognition and Understanding (ASRU), 2013 IEEE Workshop on*. IEEE. 88

[15] Chen, Ling-Hui, Ling, Zhen-Hua, Liu, Li-Juan, & Dai, Li-Rong. 2014. Voice conversion using deep neural networks with layer-wise generative training. *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, **22**(12), 1859–1872. 20

[16] Cuenca, MH, Barrio, M, & Páez, A. 2006. Evaluación acústica y análisis prosódico de la voz esofágica. *In: Lingüística clínica y neuropsicología cognitiva. Actas del Primer Congreso Nacional de Lingüística Clínica*, vol. 2. 14, 25

[17] Cullinan, Walter L, Brown, Catherine S, & Blalock, P David. 1986. Ratings of intelligibility of esophageal and tracheoesophageal speech. *Journal of communication disorders*, **19**(3), 185–195. 104, 106

[18] de Cerio Canduela, Pedro Díaz, González, Ismael Arán, Durban, Rafael Barberá, Suárez, Alexander Sistiaga, Secall, Marc Tobed, Arias, Pablo L Parente, *et al.* 2018. Rehabilitación del paciente laringectomizado. Recomendaciones de la Sociedad Española de Otorrinolaringología y Cirugía de Cabeza y Cuello. *Acta Otorrinolaringológica Española*. 7

[19] Del Pozo, Arantza, & Young, Steve. 2006. Continuous tracheoesophageal speech repair. *Pages 1–5 of: Signal Processing Conference, 2006 14th European*. Citeseer. 19, 26

[20] Del Pozo, Arantza, & Young, Steve. 2008. Repairing tracheoesophageal speech duration. *Pages 187–190 of: Proc Speech Prosody*. Citeseer. 19, 26

[21] Deller Jr, JR, Liu, MS, Ferrier, LJ, & Robichaud, P. 1993. The Whitaker database of dysarthric (cerebral palsy) speech. *The Journal of the Acoustical Society of America*, **93**(6), 3516–3518. 23

[22] Desai, Srinivas, Black, Alan W, Yegnanarayana, B, & Prahallad, Kishore. 2010. Spectral mapping using artificial neural networks for voice conversion. *IEEE Transactions on Audio, Speech, and Language Processing*, **18**(5), 954–964. 20

[23] Doi, Hironori. 2013. *Augmented speech production beyond physical constraints using statistical voice conversion – Alaryngeal speech enhancement and singing voice quality control*. Ph.D. thesis, Nara Institute of Science and Technology. 21

[24] Doi, Hironori, Nakamura, Keigo, Toda, Tomoki, Saruwatari, Hiroshi, & Shikano, Kiyohiro. 2009. Enhancement of esophageal speech using statistical voice conversion. 20

[25] Doi, Hironori, Nakamura, Keigo, Toda, Tomoki, Saruwatari, Hiroshi, & Shikano, Kiyohiro. 2010a. Esophageal speech enhancement based on statistical voice conversion with Gaussian mixture models. *IEICE TRANSACTIONS on Information and Systems*, **93**(9), 2472–2482. 21

[26] Doi, Hironori, Nakamura, Keigo, Toda, Tomoki, Saruwatari, Hiroshi, & Shikano, Kiyohiro. 2010b. Speaking-Aid Systems Based on One-to-Many Eigenvoice Conversion for Total Laryngectomees. 21

[27] Doi, Hironori, Nakamura, Keigo, Toda, Tomoki, Saruwatari, Hiroshi, & Shikano, Kiyohiro. 2010c. Statistical approach to enhancing esophageal speech based on Gaussian mixture models. *Pages 4250–4253 of: Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*. IEEE. 20

[28] Doi, Hironori, Nakamura, Keigo, Toda, Tomoki, Saruwatari, Hiroshi, & Shikano, Kiyohiro. 2011. An evaluation of alaryngeal speech enhancement methods based on voice conversion techniques. *Pages 5136–5139 of: Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*. IEEE. 21

[29] Drugman, Thomas, & Alwan, Abeer. 2011. Joint robust voicing detection and pitch estimation based on residual harmonics. *In: Twelfth Annual Conference of the International Speech Communication Association*. 57

[30] Drugman, Thomas, Rijckaert, Myriam, Janssens, Claire, & Remacle, Marc. 2015. Tracheoesophageal speech: A dedicated objective acoustic assessment. *Computer Speech & Language*, **30**(1), 16–31. 15, 24, 72

[31] Erro, D., Navas, E., & Hernaez, I. 2013. Parametric voice conversion based on bilinear frequency warping plus amplitude scaling. *IEEE Transactions on Audio, Speech and Language Processing*, **21**(3), 556 – 566. 20

[32] Erro, Daniel. 2016. Two-band radial postfiltering in cepstral domain with application to speech synthesis. *IEEE Signal Processing Letters*, **23**(2), 202–206. 136

[33] Erro, Daniel, Navas, Eva, & Hernáez, Inma. 2012. Iterative MMSE estimation of vocal tract length normalization factors for voice transformation. *In: Thirteenth Annual Conference of the International Speech Communication Association*. 119

[34] Erro, Daniel, Sainz, Iñaki, Navas, Eva, & Hernaez, Inma. 2014a. Harmonics plus noise model based vocoder for statistical parametric speech synthesis. *IEEE Journal of Selected Topics in Signal Processing*, **8**(2), 184–194. 63, 119, 127, 133, 150, 157

[35] Erro, Daniel, Hernáez, Inma, Navas, Eva, Alonso, Agustın, Arzelus, Haritz, Jauk, Igor, Hy, Nguyen Quy, Magarinos, Carmen, Pérez-Ramón, Rubén, Sulır, Martin, *et al.* 2014b. ZureTTS: online platform for obtaining personalized synthetic voices. *Proceedings of eNTERFACE*, 1178–1193. 34

[36] Erro, Daniel, Alonso, Agustín, Serrano, Luis, Tavarez, David, Odriozola, Igor, Sarasola, Xabier, del Blanco, Eder, Sánchez, Jon, Saratxaga, Ibon, Navas, Eva, *et al.* 2016a. ML Parameter Generation with a Reformulated MGE Training Criterion-Participation in the Voice Conversion Challenge 2016. *Pages 1662–1666 of: INTERSPEECH*. 118, 129, 132, 136, 138, 141, 143, 147

[37] Erro, Daniel, Hernaez, Inma, Serrano, Luis, Saratxaga, Ibon, & Navas, Eva. 2016b. Objective Comparison of Four GMM-Based Methods for PMA-to-Speech Conversion. *Pages 24–32 of: International Conference on Advances in Speech and Language Technologies for Iberian Languages*. Springer. 147, 182

[38] Eye, Massachusetts, & Infirmary, Ear. 1994. *Voice disorders database, version. 1.03 (cd-rom)*. 22

[39] Fairbanks, G. 1960. The rainbow passage. *Voice and articulation drillbook*, **2**. 22

[40] Fiscus, J, Ajot, Jerome, & Doddington, George. 2006. The spoken term detection (STD) 2006 evaluation plan. *NIST USA, Sep*. 86

[41] Fiscus, Jonathan G, Ajot, Jerome, Garofolo, John S, & Doddingtion, George. 2007. Results of the 2006 spoken term detection evaluation. *Pages 51–57 of: Proc. sigir*, vol. 7. 91

[42] Fuchs, Anna Katharina, Morales-Cordovilla, Juan Andres, & Hagmüller, Martin. 2013. ASR for electro-laryngeal speech. *Pages 234–238 of: ASRU*. 17, 24

[43] García-León, Francisco Javier, García-Estepa, Raúl, Romero-Tabares, Antonio, & Borrachina, Jaime Gómez-Millán. 2017. Tratamiento del cáncer de laringe avanzado y calidad de vida. Revisión sistemática. *Acta Otorrinolaringológica Española*, **68**(4), 212–219. 7

[44] Gonzalez, Jose A., Cheah, Lam A., Gilbert, James M., Bai, Jie, Ell, Stephen R., Green, Phil D., & Moore, Roger K. 2016. A silent speech system based on permanent magnet articulography and direct synthesis. *Computer Speech & Language*, **39**, 67–87. 126

[45] Gonzalez, Jose A, Cheah, Lam A, Gomez, Angel M, Green, Phil D, Gilbert, James M, Ell, Stephen R, Moore, Roger K, & Holdsworth, Ed. 2017. Direct speech reconstruction from articulatory sensor data by machine learning. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, **25**(12), 2362–2374. 159

[46] Guimarães, Isabel. 2007. A ciência e a arte da voz humana. *Alcoitão, Escola Superior de Saúde de Alcoitão*. 59

[47] Haderlein, Tino, Steidl, Stefan, Nöth, Elmar, Rosanowski, Frank, & Schuster, Maria. 2004. Automatic recognition and evaluation of tracheoesophageal speech. *Pages 331–338 of: International Conference on Text, Speech and Dialogue*. Springer. 16, 24

[48] Hadjitodorov, Stefan, & Mitev, Petar. 2002. A computer system for acoustic analysis of pathological voices and laryngeal diseases screening. *Medical Engineering and Physics*, **24**(6), 419–429. 53

[49] Helander, Elina, Virtanen, Tuomas, Nurminen, Jani, & Gabbouj, Moncef. 2010. Voice conversion using partial least squares regression. *IEEE Transactions on Audio, Speech, and Language Processing*, **18**(5), 912–921. 20

[50] Hochreiter, Sepp. 1998. The vanishing gradient problem during learning recurrent neural nets and problem solutions. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, **6**(02), 107–116. 131

[51] Hochreiter, Sepp, & Schmidhuber, Jürgen. 1997. Long short-term memory. *Neural computation*, **9**(8), 1735–1780. 131

[52] Hofe, Robin, Ell, Stephen R., Fagan, Michael J., Gilbert, James M., Green, Phil D., Moore, Roger K., & Rybchenko, Sergey I. 2011. Speech Synthesis Parameter Generation for the Assistive Silent Speech Interface MVOCA. *Pages 3009–3012 of: Proc. Interspeech*. 126

[53] Hogikyan, Norman D, & Sethuraman, Girish. 1999. Validation of an instrument to measure voice-related quality of life (V-RQOL). *Journal of voice*, **13**(4), 557–569. 14

[54] Huang, Andy, Falk, Tiago H, Chan, Wai-Yip, Parsa, Vijay, & Doyle, Philip. 2009. Reference-free automatic quality assessment of tracheoesophageal speech. *Pages 6210–6213 of: Engineering in Medicine and Biology Society, 2009. EMBC 2009. Annual International Conference of the IEEE*. IEEE. 15, 25

[55] Ishaq, Rizwan, & Zapirain, Begona Garcia. 2013. Esophageal Speech enhancement using modified voicing source. *Pages 000210–000214 of: Signal Processing and Information Technology (ISSPIT), 2013 IEEE International Symposium on*. IEEE. 19, 25

[56] Ishaq, Rizwan, Gowda, Dhanananjaya, Alku, Paavo, & Zapirain, Begonya Garcia. 2015. Vowel enhancement in early stage spanish esophageal speech using natural glottal flow pulse and vocal tract frequency warping. *Pages 55–59*

of: Proceedings of SLPAT 2015: 6th Workshop on Speech and Language Processing for Assistive Technologies. 56

[57] Jacobson, Barbara H, Johnson, Alex, Grywalski, Cynthia, Silbergleit, Alice, Jacobson, Gary, Benninger, Michael S, & Newman, Craig W. 1997. The voice handicap index (VHI): development and validation. *American Journal of Speech-Language Pathology*, **6**(3), 66–70. 14

[58] Kain, Alexander, & Macon, Michael W. 1998. Spectral voice conversion for text-to-speech synthesis. *Pages 285–288 of: Acoustics, Speech and Signal Processing, 1998. Proceedings of the 1998 IEEE International Conference on*, vol. 1. IEEE. 20, 114, 115

[59] Katsurada, Kouichi, Miura, Seiichi, Seng, Kheang, Iribe, Yurie, & Nitta, Tsuneo. 2013. Acceleration of spoken term detection using a suffix array by assigning optimal threshold values to sub-keywords. *Pages 11–14 of: INTERSPEECH*. 86

[60] Kawahara, Hideki. 2006. STRAIGHT, exploitation of the other aspect of VOCODER: Perceptually isomorphic decomposition of speech sounds. *Acoustical science and technology*, **27**(6), 349–353. 165

[61] Kim, Heejin, Hasegawa-Johnson, Mark, Perlman, Adrienne, Gunderson, Jon, Huang, Thomas S, Watkin, Kenneth, & Frame, Simone. 2008. Dysarthric speech database for universal access research. *In: Ninth Annual Conference of the International Speech Communication Association*. 23

[62] Koehn, Philipp. 2005. Europarl: A parallel corpus for statistical machine translation. *Pages 79–86 of: MT summit*, vol. 5. 84

[63] Kominek, John, & Black, Alan W. 2004. The CMU Arctic speech databases. *Pages 223–224 of: Proc. 5th ISCA Speech Synthesis Workshop*. 127

[64] Koutsogiannaki, Maria, Petkov, Petko N, & Stylianou, Yannis. 2015. Intelligibility enhancement of casual speech for reverberant environments inspired by clear speech properties. *In: Sixteenth Annual Conference of the International Speech Communication Association*. 136

[65] Lachhab, Othman, Di Martino, Joseph, Elhaj, Elhassane Ibn, & Hammouch, Ahmed. 2015. A preliminary study on improving the recognition of esophageal speech using a hybrid system based on statistical voice conversion. *SpringerPlus*, **4**(1), 644. 17, 24

[66] Lee, Chung-Han, Wu, Chung-Hsien, & Guo, Jun-Cheng. 2010. Pronunciation variation generation for spontaneous speech synthesis using state-based voice transformation. *Pages 4826–4829 of: Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*. IEEE. 20

[67] Levenshtein, Vladimir I. 1966. Binary codes capable of correcting deletions, insertions, and reversals. *Pages 707–710 of: Soviet physics doklady*, vol. 10. 89

[68] Ling, Zhen-Hua, Kang, Shi-Yin, Zen, Heiga, Senior, Andrew, Schuster, Mike, Qian, Xiao-Jun, Meng, Helen M, & Deng, Li. 2015. Deep learning for acoustic modeling in parametric speech generation: A systematic review of existing techniques and future trends. *IEEE Signal Processing Magazine*, **32**(3), 35–52. 20

[69] Lorenzo-Trueba, Jaime, Yamagishi, Junichi, Toda, Tomoki, Saito, Daisuke, Villavicencio, Fernando, Kinnunen, Tomi, & Ling, Zhenhua. 2018a. The voice conversion challenge 2018: Promoting development of parallel and nonparallel methods. *arXiv preprint arXiv:1804.04262*. 138

[70] Lorenzo-Trueba, Jaime, Yamagishi, Junichi, Toda, Tomoki, Saito, Daisuke, Villavicencio, Fernando, Kinnunen, Tomi, & Ling, Zhenhua. 2018b. The Voice Conversion Challenge 2018: Promoting Development of Parallel and Nonparallel Methods. *In: Submitted to Odyssey:The Speaker and Language Recognition Workshop*. 141

[71] MacQueen, James, *et al.* 1967. Some methods for classification and analysis of multivariate observations. *Pages 281–297 of: Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, vol. 1. Oakland, CA, USA. 98

[72] Maier, Andreas, Haderlein, Tino, Eysholdt, Ulrich, Rosanowski, Frank, Batliner, Anton, Schuster, Maria, & Nöth, Elmar. 2009. PEAKS–A system for the automatic evaluation of voice and speech disorders. *Speech Communication*, **51**(5), 425–437. 16, 24

[73] Mantilla, Alfredo, Pérez-Meana, Héctor, Mata, Daniel, Angeles, Carlos, Alvarado, Jorge, & Cabrera, Laura. 2006. Recognition of vowel segments in spanish esophageal speech using hidden markov models. *Pages 115–120 of: Computing, 2006. CIC'06. 15th International Conference on*. IEEE. 17, 25

[74] Mantilla-Caeiros, A, Nakano-Miyatake, Mariko, & Perez-Meana, H. 2010. A pattern recognition based esophageal speech enhancement system. *Journal of applied research and technology*, **8**(1), 56–70. 17

[75] Mashimo, Mikiko, Toda, Tomoki, Shikano, Kiyohiro, & Campbell, Nick. 2001. Evaluation of cross-language voice conversion based on GMM and STRAIGHT. 173

[76] McAuliffe, Michael, Socolof, Michaela, Mihuc, Sarah, Wagner, Michael, & Sonderegger, Morgan. 2017. Montreal Forced Aligner: trainable text-speech alignment using Kaldi. *In: Proceedings of interspeech*. 47

[77] McDonald, Rob, Parsa, Vijay, Doyle, Philip, & Chen, Guo. 2008a. On the prediction of speech quality ratings of tracheoesophageal speech using an auditory model. *Pages 4517–4520 of: Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*. IEEE. 15, 25

[78] McDonald, Rob, Parsa, Vijay, & Doyle, Phillip. 2008b. Prediction of the quality ratings of tracheoespohageal speech using adaptive time-frequency representations. *Pages 001715–001718 of: Electrical and Computer Engineering, 2008. CCECE 2008. Canadian Conference on*. IEEE. 15, 25

[79] McGarrigle, Ronan, Munro, Kevin J, Dawes, Piers, Stewart, Andrew J, Moore, David R, Barry, Johanna G, & Amitay, Sygal. 2014. Listening effort and fatigue: What exactly are we measuring? A British Society of Audiology Cognition in Hearing Special Interest Group 'white paper'. *International journal of audiology*, **53**(7), 433–440. 104

[80] Menendez-Pidal, Xavier, Polikoff, James B, Peters, Shirley M, Leonzio, Jennie E, & Bunnell, H Timothy. 1996. The Nemours database of dysarthric speech. *Pages 1962–1965 of: Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on*, vol. 3. IEEE. 22

[81] Mohammadi, Seyed Hamidreza, & Kain, Alexander. 2014. Voice conversion using deep neural networks with speaker-independent pre-training. *Pages 19–23 of: Spoken Language Technology Workshop (SLT), 2014 IEEE*. IEEE. 20

[82] Mohammadi, Seyed Hamidreza, & Kain, Alexander. 2017. An overview of voice conversion systems. *Speech Communication*, **88**, 65–82. 20

[83] Moreno Sandoval, Antonio. 2008. *Corpus MAVIR, Laboratorio de Lingüística Informática de la UAM*. `http://www.lllf.uam.es/ESP/CorpusMavir.html`. [Online; ultimo acceso 03-Junio-2008]. 84

[84] Moukarbel, Roger V, Doyle, Philip C, Yoo, John H, Franklin, Jason H, Day, Adam MB, & Fung, Kevin. 2011. Voice-related quality of life (V-RQOL) outcomes in laryngectomees. *Head & neck*, **33**(1), 31–36. 14

[85] Narendranath, M, Murthy, Hema A, Rajendran, S, & Yegnanarayana, B. 1995. Transformation of formants for voice conversion using artificial neural networks. *Speech communication*, **16**(2), 207–216. 20

[86] Norouzian, Atta, & Rose, Richard. 2014. An approach for efficient open vocabulary spoken term detection. *Speech Communication*, **57**, 50–62. 86

[87] Pietruch, Rafal, & Grzanka, Antoni. 2010a. Combining acoustic and visual modalities in vowel recognition system for laryngectomees. *Pages 175–179 of: Neural Network Applications in Electrical Engineering (NEUREL), 2010 10th Symposium on*. IEEE. 17, 25

[88] Pietruch, Rafal W, & Grzanka, Antoni D. 2010b. Vowel recognition of patients after total laryngectomy using mel frequency cepstral coefficients and mouth contour. *Journal of Automatic Control*, **20**(1), 33–38. 17, 25

[89] Povey, Daniel, Ghoshal, Arnab, Boulianne, Gilles, Burget, Lukas, Glembek, Ondrej, Goel, Nagendra, Hannemann, Mirko, Motlicek, Petr, Qian, Yanmin, Schwarz, Petr, *et al.* 2011. The Kaldi speech recognition toolkit. *In: IEEE 2011 workshop on automatic speech recognition and understanding*. IEEE Signal Processing Society. 44, 82

[90] Pützer, Manfred, & Barry, William J. *Saarbrücken Voice Database, Institute of Phonetics, Univ. of Saarland*. `http://www.stimmdatenbank.coli.uni-saarland.de/`. 22

[91] Raman, Sneha, Hernaez, Inma, Navas, Eva, & Serrano, Luis. 2018. Listening to Laryngectomees: A study of Intelligibility and Self-reported Listening Effort of Spanish Oesophageal Speech. *Pages 107–111 of: Proc. IberSPEECH 2018*. 104, 112

[92] Rath, Shakti P, Povey, Daniel, Veselỳ, Karel, & Cernockỳ, Jan. 2013. Improved feature processing for deep neural networks. *Pages 109–113 of: Interspeech*. 82

[93] Rodrigo, Juan P, López, Fernando, Llorente, José L, Álvarez-Marcos, César, & Suárez, Carlos. 2015. Resultados de la laringectomía total en carcinoma localmente avanzado de laringe en la era de la organopreservación. *Acta Otorrinolaringológica Española*, **66**(3), 132–138. 7

[94] Rosso, Marinela, Širić, Ljiljana, Tićac, Robert, Starčević, Radan, Šegec, Igor, & Kraljik, Nikola. 2013. Perceptual evaluation of alaryngeal speech. *Collegium antropologicum*, **36**(2), 115–118. 14

[95] Rudzicz, Frank, Namasivayam, Aravind Kumar, & Wolff, Talya. 2012. The TORGO database of acoustic and articulatory speech from speakers with dysarthria. *Language Resources and Evaluation*, **46**(4), 523–541. 23

[96] Sainz, Iñaki, Erro, Daniel, Navas, Eva, Hernáez, Inma, Sánchez, Jon, & Saratxaga, Ibon. 2012a. Aholab speech synthesizer for albayzin 2012 speech synthesis evaluation. *Proc. Iberspeech*, 645–652. 88

[97] Sainz, Iñaki, Erro, Daniel, Navas, Eva, Hernáez, Inma, Sanchez, Jon, Saratxaga, Ibon, & Odriozola, Igor. 2012b. Versatile Speech Databases for High Quality Synthesis for Basque. *Pages 3308–3312 of: LREC*. Citeseer. 170

[98] Saltürk, Z, Arslanoğlu, A, Özdemir, E, Yıldırım, G, Aydoğdu, İ, Kumral, TL, Berkiten, G, Atar, Y, & Uyar, Y. 2016. How do voice restoration methods affect the psychological status of patients after total laryngectomy? *Hno*, **64**(3), 163–168. 14

[99] Saz, Oscar, Lleida, Eduardo, Vaquero, Carlos, & Rodríguez, William Ricardo. 2010. The Alborada-I3A Corpus of Disordered Speech. *In: LREC*. Citeseer. 23

[100] Schoeffler, Michael, Stöter, Fabian-Robert, Edler, Bernd, & Herre, Jürgen. 2015. Towards the next generation of web-based experiments: A case study assessing basic audio quality following the ITU-R recommendation BS. 1534 (MUSHRA). *Pages 1–6 of: 1st Web Audio Conference*. 142

[101] Schuster, Maria, Noth, Elmar, Haderlein, Tino, Steidl, Stefan, Batliner, Anton, & Rosanowski, Frank. 2005. Can you understand him? Let's look at his word accuracy-automatic evaluation of tracheoesophageal speech. *Pages I–61 of: Acoustics, Speech, and Signal Processing, 2005. Proceedings.(ICASSP'05). IEEE International Conference on*, vol. 1. IEEE. 16, 24

[102] Schuster, Maria, Haderlein, Tino, Nöth, Elmar, Lohscheller, Jörg, Eysholdt, Ulrich, & Rosanowski, Frank. 2006. Intelligibility of laryngectomees' substitute speech: automatic speech recognition and subjective rating. *European Archives of Oto-Rhino-Laryngology and Head & Neck*, **263**(2), 188–193. 16, 24

[103] Serrano, Luis, Raman, Sneha, Tavárez, David, Navas, Eva, & Hernaez, Inma. Parallel vs. Non-parallel Voice Conversion for Esophageal Speech. *In: aceptado en Interspeech 2019, a la espera de publicación*. 178

[104] Serrano, Luis, Tavárez, David, Odriozola, Igor, Hernaez, Inma, & Saratxaga, Ibon. 2016. Aholab system for Albayzin 2016 Search-on-Speech Evaluation. *In: Proceedings of IberSPEECH 2016, Lisbon, Portugal, November 23-25, 2016*. 86, 111

[105] Serrano, Luis, Tavarez, David, Sarasola, Xabier, Raman, Sneha, Saratxaga, Ibon, Navas, Eva, & Hernaez, Inma. 2018. LSTM based voice conversion for laryngectomees. *Proc. IberSPEECH 2018*, 122–126. 157, 178

[106] Stolcke, Andreas. 2002. SRILM-an extensible language modeling toolkit. *In: Seventh international conference on spoken language processing*. 85

[107] Stylianou, Yannis, Cappé, Olivier, & Moulines, Eric. 1998. Continuous probabilistic transform for voice conversion. *IEEE Transactions on speech and audio processing*, **6**(2), 131–142. 20, 115

[108] Sun, Lifa, Kang, Shiyin, Li, Kun, & Meng, Helen. 2015. Voice conversion using deep bidirectional long short-term memory based recurrent neural networks. *Pages 4869–4873 of: Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE. 20

[109] Sun, Lifa, Li, Kun, Wang, Hao, Kang, Shiyin, & Meng, Helen. 2016. Phonetic posteriorgrams for many-to-one voice conversion without parallel data training. *Pages 1–6 of: 2016 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE. 165, 173

[110] Tejedor, Javier, Toledano, Doroteo T, Wang, Dong, King, Simon, & Colás, José. 2014. Feature analysis for discriminative confidence estimation in spoken term detection. *Computer Speech & Language*, **28**(5), 1083–1114. 86

[111] Tejedor, Javier, Toledano, Doroteo T, Rodriguez-Fuentes, LJ, Penagarikano, M, Varona, A, Diez, M, & Bordel, G. 2016a. The ALBAYZIN 2016 search on speech evaluation plan. *Proc. IberSPEECH*. 84, 86

[112] Tejedor, Javier, Toledano, Doroteo T, Lopez-Otero, Paula, Docio-Fernandez, Laura, & Garcia-Mateo, Carmen. 2016b. Comparison of ALBAYZIN query-by-example spoken term detection 2012 and 2014 evaluations. *EURASIP Journal on Audio, Speech, and Music Processing*, **2016**(1), 1. 86

[113] Tejedor, Javier, Toledano, Doroteo T, Lopez-Otero, Paula, Docio-Fernandez, Laura, Serrano, Luis, Hernaez, Inma, Coucheiro-Limeres, Alejandro, Ferreiros, Javier, Olcoz, Julia, & Llombart, Jorge. 2017. ALBAYZIN 2016 spoken term

detection evaluation: an international open competitive evaluation in Spanish. *EURASIP Journal on Audio, Speech, and Music Processing*, **2017**(1), 22. 93, 111, 181

[114] Tejedor, Javier, Toledano, Doroteo T, Lopez-Otero, Paula, Docio-Fernandez, Laura, Proença, Jorge, Perdigão, Fernando, García-Granada, Fernando, Sanchis, Emilio, Pompili, Anna, & Abad, Alberto. 2018. ALBAYZIN Query-by-example Spoken Term Detection 2016 evaluation. *EURASIP Journal on Audio, Speech, and Music Processing*, **2018**(1), 2. 86

[115] Tiple, Cristina, Matu, Silviu, Dinescu, Florina Veronica, Muresan, Rodica, Soflau, Radu, Drugan, Tudor, Giurgiu, Mircea, Stan, Adriana, David, Daniel, & Chirila, Magdalena. 2015. Voice-related quality of life results in laryngectomies with today's speech options and expectations from the next generation of vocal assistive technologies. *Pages 1–4 of: E-Health and Bioengineering Conference (EHB), 2015*. IEEE. 14

[116] Toda, Tomoki, Black, Alan W, & Tokuda, Keiichi. 2007. Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory. *IEEE Transactions on Audio, Speech, and Language Processing*, **15**(8), 2222–2235. 20, 116, 118, 128, 158

[117] Toda, Tomoki, Black, Alan W., & Tokuda, Keiichi. 2008. Statistical mapping between articulatory movements and acoustic spectrum using a Gaussian mixture model. *Speech Commun.*, **50**(3), 215–227. 128

[118] Tokuda, Keiichi, Masuko, Takashi, Miyazaki, Noboru, & Kobayashi, Takao. 2002. Multi-space probability distribution HMM. *IEICE Trans. Inf. Syst.*, **E85-D**(3), 455–464. 116

[119] van As-Brooks, Corina J, Koopmans-van Beinum, Florien J, Pols, Louis CW, & Hilgers, Frans JM. 2006. Acoustic signal typing for evaluation of voice quality in tracheoesophageal speech. *Journal of Voice*, **20**(3), 355–368. 14, 25

[120] Van Den Oord, Aäron, Dieleman, Sander, Zen, Heiga, Simonyan, Karen, Vinyals, Oriol, Graves, Alex, Kalchbrenner, Nal, Senior, Andrew, & Kavukcuoglu, Koray. 2016. Wavenet: A generative model for raw audio. *CoRR abs/1609.03499*. 184

[121] van der Molen, Lisette, van Rossum, Maya A, Ackerstaff, Annemieke H, Smeele, Ludi E, Rasch, Coen RN, & Hilgers, Frans JM. 2009. Pretreatment organ function in patients with advanced head and neck cancer: clinical outcome measures and patients'views. *BMC Ear, Nose and Throat Disorders*, **9**(1), 10. 23

[122] van Son, RJJH, Jacobi, Irene, Hilgers, Frans JM, *et al.* 2010. Manipulating treacheoesophageal speech. *Pages 274–277 of: Interspeech*. 19, 24

[123] Wang, Dong. 2010. *Out-of-vocabulary spoken term detection*. Ph.D. thesis, The University of Edinburgh. 86

[124] Wszołek, Wiesław, Modrzejewski, Maciej, & Przysiężny, Monika. 2014. Acoustic analysis of esophageal speech in patients after total laryngectomy. *Archives of Acoustics*, **32**(4 (S)), 151–158. 14, 25

[125] Xu, Ning, Tang, Yibing, Bao, Jingyi, Jiang, Aiming, Liu, Xiaofeng, & Yang, Zhen. 2014. Voice conversion based on Gaussian processes by coherent and asymmetric training with limited training data. *Speech Communication*, **58**, 124–138. 20

[126] Ye, Hui, & Young, Steve. 2006. Quality-enhanced voice morphing using maximum likelihood transformations. *IEEE Transactions on Audio, Speech, and Language Processing*, **14**(4), 1301–1312. 20, 115

[127] Young, S, Evermann, G, Gales, M, Hain, T, Kershaw, D, Liu, X, Moore, G, Odell, J, Ollason, D, Povey, D, *et al.* 2006. The HTK book (v3. 4). *Cambridge University*. 91

[128] Zen, Heiga, Nankaku, Yoshihiko, & Tokuda, Keiichi. 2011. Continuous stochastic feature mapping based on trajectory HMMs. *IEEE Transactions on Audio, Speech, and Language Processing*, **19**(2), 417–430. 20

[129] Zhang, Xiaohui, Trmal, Jan, Povey, Daniel, & Khudanpur, Sanjeev. 2014. Improving deep neural network acoustic models using generalized maxout networks. *Pages 215–219 of: 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 170

ANEXO

A

Corpus grabado

A continuación se muestra el texto que aparece en pantalla a los locutores.

El texto de las 100 frases en castellano del corpus Zure TTS son las siguientes:

0001 El subcampeonato europeo de Isaac Viciosa cierra el círculo.
0002 Ya están rogando a Dios que no llueva esta tarde, aunque por otro lado, lo necesitan.
0003 Durán Lleida coincidió con Rodríguez Zapatero en el llamamiento a la unidad.
0004 Todo ello mezclado con logias mitad místicas, mitad militares, y una densa atmósfera ocultista.
0005 De Filadelfia vino el grupo Judíos por una Paz Justa.
0006 Ello hacía intuir un duelo en toda regla.
0007 Nacho García Vega, ex cantante de "Nacha Pop", dará la bienvenida mañana a "Los Secretos".
0008 Deja mucha buena obra hecha, pero me rehuye el balance.
0009 Hoy jueves, dieciocho de julio de dos mil trece.
0010 La cuestión más acuciante es qué va a pasar con Mas y su hasta ahora incuestionable liderazgo.

0011 Ver el reto y la química entre la nena hipnótica y la gran dama supone un lujo.
0012 El otro que está, dos semáforos más allá, con sus pañuelitos de papel.
0013 ¿Cree usted que él no se ocupa de otra cosa que de proteger a mozuelas como usted?
0014 ¿Ha podido reunirse con Chávez tras su elección?
0015 Quizá ustedes no lo advirtieron, por eso lo refiero ahora.
0016 Los chiíes iraquíes que han vuelto a Irak después de pasar veinte años en Irán, no imitarán el modelo iraní.
0017 Es un test de autoevaluación que puede administrarse en grupo o individualmente.
0018 Hay indicios claros, pero todavía hay llanuras para afirmar con rotundidad.
0019 Maman un fútbol de balonazo, choque y puño cerrado y lo escupen al rival de turno.
0020 Abdulá Abdulá, ministro de Exteriores, va más allá.
0021 Como aquéllas de Belmonte que hicieron crujir los fundamentos científicos y antitaurinos de Eugenio Noel.
0022 En cuanto a su utilidad, a él le sirve para ahorrarse siquiatras.
0023 Da igual; no importa de dónde extraiga uno la emoción.
0024 Sólo el chileno Mark González ponía una pizca de orgullo.
0025 Sin embargo, según Ustinov, no influyeron directamente en el hundimiento del submarino.
0026 ¿Pero usted ya conoce por dentro el mundo del cine?
0027 Hay voces que ya hablan de indulto, ¿sería factible?
0028 Reconoció que en su barrio se mueven los Ñetas aunque nunca se ha interesado por este grupo.
0029 No ofrecer mensajes de ayuda durante el proceso de conexión que puedan guiar a usuarios no autorizados.
0030 Lo que no cree nadie aquí es que cacen a Sadam con vida.
0031 Unos días de euforia y meses de atonía.
0032 Así lo afirma Antonio Gómez Rufo, quien cruza en su novela una historia épica y otra intimista.
0033 Blasco Ibáñez hizo alguna vez la misma cosa.

0034 No hay chinos mendigando en las calles, y prefieren cuidarse entre ellos.
0035 Junio es, desde hace ya veinticinco años, el mes de la comunidad galesa y estadounidense.
0036 En Vietnam hay una Iglesia católica muy viva, sobre todo en el antiguo sur.
0037 Me niegas el auxilio que yo, pobre industrial, vengo a pedirte.
0038 Hay mucha pantalla repitiendo las imágenes de Encarna y Toñi, como un videoclip de lujo e incluso lujuria.
0039 En el equipo vasco hay un hombre que va feliz por el vestíbulo de su hotel.
0040 El Athletic firme hasta entonces, con un Gurpegui muy activo y un Yeste inspirado, se resquebrajó.
0041 Así ocurre con Teherán pero también con Trípoli, Bagdad, Belgrado y un largo etcétera.
0042 En el hueco de ella había un banco de madera y paja y un reloj de pie.
0043 El balet flamenco Raíces, dirigido por Miguel Fuente, llega a su fin hoy.
0044 El equipo médico habitual volverá a reunirse y quizá opte por una nueva intervención quirúrgica.
0045 Por otro, un subproducto sólido llamado orujo y del que se obtenía un aceite de orujo perfectamente sano.
0046 El Fórum de dos mil cuatro ya ha encontrado apoyo en la sociedad civil.
0047 El andaluz lleva dentro un caballero, y tiene una alta estima de sí mismo.
0048 Las cifras de la Agencia Antidroga tuvieron ayer un ejemplo en carne y hueso.
0049 El empeño puntúa en un club rojiblanco en busca de notoriedad.
0050 Allí leí relatos de Onetti, artículos de Onetti, y vi la fotografía de su rostro atónito y despiadado.
0051 Todavía estoy escuchando los aullidos de esas bestias siguiendo nuestro rastro.
0052 Tenía la voz alborotada y la amistad ruidosa.
0053 Hace dos años le dijimos que no nos había emocionado nada su intervención, y usted fue muy comprensivo.
0054 Su mensaje no llega o llega con cuentagotas, y además, huele a antiguo.
0055 Venus ganó el primer set por la vía rápida y Serena se ofendió.

0056 Apliqué el oído a esta rayita y percibí un murmullo.
0057 Acabamos de inaugurar otro módulo libre de drogas en otra unidad en Galicia y vamos a hacerlo.
0058 Gastó todo el agua, incluyendo el agua de las lluvias.
0059 ¿Usted llama razonable a un convenio que convalida una defraudación?
0060 Si el club no hubiera cambiado, ¿se hubiera ido?
0061 ¿Un diagnóstico de la situación a esta hora de reloj?
0062 Goliat estuvo a punto de engullir a David.
0063 El taxi se detuvo ante la reja del Club Grau; el motor siguió rugiendo y humeando.
0064 La chusma está en el centro y por eso me he ido a vivir al centro.
0065 Tal vez fue hace siglos, o acaso hace tan sólo unas décadas.
0066 Se ha hecho especialmente famosa en los clubes del circuito gaitero londinense.
0067 A Jesús Bonilla le siguió Dafne Fernández, la chica Upa que tiene el récord de asistencia a saraos.
0068 Yo no fui a buscarte a tu celda ni soy yo quién te he hecho venir hoy a mi casa.
0069 Los niños le han hecho un lifting eterno, y a cambio, él nunca les ha fallado.
0070 Su adjudicación a Luis Roldán fue irregular y nunca se regularizó.
0071 No hay receta para aliviar la sed del mal de Teseo, que envía a su hijo Hipólito al destierro y a la muerte.
0072 Al llegar yo donde está esa persona, ¿cómo acredito mi calidad de embajador?
0073 Aún no sabemos qué fue a hacer a Taiwan.
0074 El pueblo noruego rechazó vía referéndum la adhesión.
0075 El buque efectuó ayer una maniobra de zigzag durante una hora.
0076 El nacionalismo hindú recorrió al ya septuagenario líder para aumentar su popularidad en la región.
0077 Con este álbum llegará seguro al número uno de ventas.
0078 Mi aldea estaba a la orilla de un riachuelo como éste.
0079 A reglón seguido, Pinochet confiesa sin resquemor su enraizada fobia anti marxista.

0080 Fui yo, por consejo del señor Regueiro Souza.

0081 Fui en un Seat seiscientos, y como ella tenía allí otro, nos volvimos cada uno en el nuestro.

0082 El pazo Bayón ha sido el buque insignia de Laureano Oubiña.

0083 Margot guardó un silencio embarazoso y después habló con un hilo de voz.

0084 ¿Felipe González puede ser llevado ante los tribunales y condenado por actos terroristas?

0085 Llegó a la puerta antes que Bella, quien maldecía en voz baja que hubiera llegado.

0086 Se ha muerto Celia Cruz sin volver a Cuba, aunque Cuba era ella y en ella vivirá para siempre.

0087 Si hay un problema o al chico del subfusil le tiembla el dedo, vamos a tener un problema.

0088 Hoy no juega al golf y el traje es azul cielo y oro.

0089 Así ocurrió con Sudán y Afganistán, y así figuraba en el guión de Irak.

0090 Obviamente, hay mucha diferencia entre reducir e impedir.

0091 Los fines de semana no se distinguen del resto de los días cuando no se tiene trabajo ni sitio a donde ir.

0092 Occidente y el islam son dos miedos que se acechan.

0093 Su otro éxito fue una película, La hora de la araña, que se aupó al noveno puesto.

0094 El club ya ha hecho una oferta al jugador y le pide una respuesta rápida.

0095 El diputado por Teruel era un sesentón, alto, enjuto, y de rostro huesudo, ceñudo, y totalmente afeitado.

0096 Manuel Rey fue otro de los jefes del Cesid que salió damnificado por la reunión de octubre.

0097 Núñez ya tiene a su hijo predilecto en casa.

0098 ¿Qué diferencia hay entre el caucho y la hevea?

0099 ¿Por qué la voluntad torera de Rivera Ordóñez no alcanza los niveles de la inteligencia y la eficacia?

0100 ¿Una fiesta en Florida Park con glamur?

El texto de las 100 frases en euskera del corpus Zure TTS son las siguientes:

0001 Bera ere ez da, ez ametsetako, ez filmetako neska.

0002 Kapak alde batera uzten dira, eta nire kuttun berdea ere ziur aski zintzilik dago.

0003 Amona gaixo hura eta bilobak bizi ziren etxe hark bazuen, beraz, ukuilua.

0004 Kontseilariak txosten bat nahi du, eta oraintxe nahi du.

0005 Espainiako barne sailak espetxe kupoa ez transferitzeko erabakiaz ere hitz egin du.

0006 Ilargia gaur ez da ilunari ihes egiteko leiho estua, ez da malda berde gozo bat taupada urrunak batzen.

0007 Badu obsesio bat mintzaira honek, eta ez da mihiekin, oinekin baino.

0008 Kontzeptua zeinu osoari dagokionez, diferentzia eta oposizio kontzeptuak garatu zituen.

0009 Bagdadeko auzo horretan bi milioi pertsona bizi dira, gehienak xiitak, eta oso pobrea da.

0010 Aitzolek Juan Ignazio Uranga pertsonaia aipatzen du, bertsolaritzaz ari den batean.

0011 Anttonik Jexuxi muxu eman dio kopetan eta gela utzi dute.

0012 Andoni Arriagak hiru hilabete egin ditu Medellinen.

0013 Airekoan ehun eta lau lagun zihoazen eta istripu honetatik inor salbu irten denik ez da uste.

0014 Harri guztiekin moldatzen da ongi, eta oso gutxitan egiten du huts.

0015 Behar baino segundo bat lehenago atzera itzuli izan banintz, neure buruaz ezer ez nuke jakingo.

0016 Formek formulei utziko diete lekua, hieroglifikoek alfabetoko letrei utzi dieten moduan.

0017 Soinu uhinek diafragma mugitzen dute, eta horrekin batera bobina.

0018 Arartekoak ez du dokumentazio hori jaso honako txosten hau ixteko epera artean.

0019 Arbuiatu eginen dut neuretzat sagaratu dudan etxe hau.

0020 Euskal jolasez blai du izena, eta bertan, euskal jolas desberdinak egingo dira dinamizatzaileekin.

0021 Luis Fernandezek ez zuela jarraituko erabaki bezain pronto, Rojo fitxatzea pentsatu genuen.
0022 Jakin badakite, operazioak benetako funts juridikorik ez zuela.
0023 Eusko jaurlaritzako osasun saileko koordinatzailea da CID.
0024 Gladiadore txikia eta samurai txikia deitzen diote.
0025 Elsa ez da ondo egon joan den hilean Adolf txikia jaio zenetik.
0026 Amnistiaren aldeko batzordeek Maite Perez omenduko dute asteartean Bilbon.
0027 Ur uherrak bere onera ekartzeko, aurki bilera egingo dute Jose Maria Aznar eta Jose Luis Rodriguez Zapaterok.
0028 Bi eta lau minutu artean iraungo du eraztun eklipseak Europan.
0029 Pertzepziozko eskemen existentzian oinarritzen da ikuslearen pertzepzio ahalmen hau.
0030 Taberna euli eta eltxoz josirik dagoela dirudi argi fluoreszenteek sortzen duten soinu elektrikoa dela eta.
0031 Aitzin etaparik gabe hasi da aurtengo tourra, erlojupeko zailarekin.
0032 Aske utzi dute Afganistanen duela hiru hilabete bahitutako ingeniaria.
0033 Juan Guzman Tapia epailea Augusto Pinochet Txileko diktadore ohia epaitzeagatik egin zen nazioartean ezaguna.
0034 Udaltzaingoaren patruila bat biktima oraindik bizirik zela heldu da istripu lekura.
0035 Ez dakit zenbat iraun zuen euforia hark.
0036 Bertsio ofiziala ontzat eman eta sumarioa itxi egin zuen Rojas epaileak.
0037 Joxe Juan Gonzalez Txabarri Gipuzkoako diputatu nagusiak ere Iñaki Uria aske utz dezatela eskatu du.
0038 Epaileak aske utzi du Jose Luis Elkoro, epaia irmoa izan artean.
0039 Bi itzulpen ofizial daude, bat frantsesez eta bestea ingelesez.
0040 Zoaz jende horren bila eta bataia ezazu ume hori.
0041 Norbait lan mahaira gehiegitxo hurbiltzen zitzaionean, eskuak poltsikora, eta kitto.
0042 Kattalin ere ohi baino gogotsuago ari da klaseen prestatzen.
0043 Zenbat jende da halere, joko hau aurrera eraman ahal izateko beharrezkoa?
0044 Itsas txakur fraidea babestea eta zaintzea ekintza ona ala txarra da?

0045 Nire desberdintasun ñimiñoa nortasun iturri bihurtzen da homogeneoegia dirudien mundu honetan.
0046 Poxpolin aparta da urrunean, ea distantzia hurbiletan are hobeto moldatzen den.
0047 Gauez imajinatu dugu, lurreratu nahi duten nabigatzaileei foku indartsuekin ingurua argiztatuz.
0048 Hauxe ei duk euren plazer lurtar bakarra.
0049 Ez dakit nola sinetsi ahal izan dudan horrelako ele huts saltsa horretan.
0050 Iñigo Garciak lortu du lehenengo domina.
0051 Keinu oro ez da obra artistikoa, interjekzio oro hitz egitea ez den bezala.
0052 Txosten hau idazteko unean, oraindik adjudikatu barik dago.
0053 Izan ere, auzia atxiloketa agindu zuen epaileari bueltatu diote, Cesar Floresi.
0054 Horiek ziren Jesus Arzallus, Mixel Xalbador, Txomin Ezponda eta Imanol Lazkano.
0055 Udal batek obra publiko bat legez kontra adjudikatzea erreklamatu da.
0056 Zentzu garrantzitsu batean marxismoa erlijioa da.
0057 Llorentek gol abagune argia izan du oinetan, baina aurrelaria horretan ez da fin ibili.
0058 Horri buruz zerbait gehiago jakin nahi duenak web toki hau bisita dezake hiru.
0059 Zein da horrenbeste golf zelai egitearen arrazoia?
0060 Nahi al duzu naftaren efektua morroi honekin probatzea?
0061 Ez dugu nahi hau ghetto bat bihur dadin.
0062 Dena den, sailkapen nagusian citroen etxeko pilotua da jaun eta jabe.
0063 Kremlinek negozio gizon handien aurka hasitako eraso judizialak sutan jarri ditu oligarkak.
0064 Ohean makurtu eta atso biei zerbait xuxurlatu zien.
0065 Bai gizon atsegina, hiztun goxoa eta adiskidetsua, Felix.
0066 Pepponek bere jarduna eten eta solemneki jaso zuen hatz bat.
0067 Informazio hutsa eman beharrean, erabiltzailea web gunearen partaide bihurtu behar genuen.
0068 Ez du balio inork hau txarto edo ondo dagoela esateak.

0069 Hain zuzen, pub batean goxo goxo geundela, barran zeuden bi tipo oilartu eta elkarri bultzaka hasi ziren.

0070 Guardia zibil piloa dabil, udaletxeko plazan lau jeep daude eta hobeto duzu trenbide ondotik joan.

0071 Bikaina da hori, Jerusalem gure hiriburua eta hiru erlijiorena.

0072 Hainbeste su eginda, txotx bakar bat ere ez zuan geratu hurrengo egunerako.

0073 Oraintxe Asiako hondamendia da albiste, tsunami dela eta ez dela.

0074 Babcockeko langileek bertan behera utzi dute greba.

0075 Oinarrizko curriculum diseinuen kultur egokitzapen eza.

0076 Kinka hau hartzen du abiapuntutzat Jim idazle iparramerikarraren ihesa eleberriak.

0077 Bolada hori apurtzeko asmoarekin joango dira gaur deabru gorriak Bridge futbol zelaira.

0078 Bakoitzari muxu eta hitz goxo bat paratu ondoren sartu da ohean, Patriciaren gona jostera.

0079 Beren haur ttipi eta arra haur ttipiez ongi inguratuak.

0080 Tronbosiaz hitz egingo dute forum saioan.

0081 Erasoen jomugan jarraitzen dute erromes xiitek.

0082 Jende gehienak prozesu logiko huts hustzat dauka ikerkuntza, eta jarduera hotz eta zorroztzat.

0083 Higuina diet aljebrari, geometriari eta aritmetikari.

0084 Auzo txiroetan, pub merkeetan, sentimendu hitsetan ehizatzen du epifania.

0085 ñirñir egingo digute begiek, hunkiturik eta liluraturik.

0086 Hau da anai Rufinori dedikaturiko zenbait ataletako lehena.

0087 Tipo bitxia da Adam, ume herabetiaren moduan dantzan.

0088 Mesopotamiara ihesi joan eta bertan morroi egon zen Jakob, emakume baten truke.

0089 Jexux Mari Irazu irratia aditzea gauza ederra da.

0090 Bizkaian lau golf klub daude, konpetentzia handia?

0091 Patxi Perez da nagusi, zeinak honezkero Euskal Herriko plaza gehientsuenak ezagutuko baititu.

0092 Daniel Redondok esan zuen urmael zaharrak oroitzen zuena baino txikiagoa ematen zuela.
0093 Ttottek ez zuen ulertzen orduan gertatu zitzaiona.
0094 Ur jarioa berez xurgatzen zuen lurrak.
0095 Xabier Etxaniz Rojo eta Aitor Arana Luzuriaga izan dira euskarazko alorretan irabazleak.
0096 Epaitegiek Lapatx sozietate anonimo ilegaltzat jo ondoren, zein asmo dago?
0097 Liburu hau idatzi aurretik zein neurritan, nola ezagutzen zenuen Bilintx?
0098 Getxon ospatuko da euskal graffitien bigarren lehiaketa uztailaren hamaseian.
0099 Gorka Urbiolak deklaratu du aurrena, eta Aitziber Perez Blanco hasi da gero.
0100 Atturri eta Bidasoaren artean apneak jota dago euskara, bizi edo ito kinka larrian.

Los términos a grabar son las siguientes:

0101 BALANCE
0102 COMUNIDAD
0103 EMPEÑO
0104 MANIOBRA
0105 EMBARAZOSO
0106 PROBLEMA
0107 PREDILECTO
0108 LIBRE
0109 AGUA
0110 ELLA
0111 AYUNTAMIENTO
0112 MURCIÉLAGO
0113 ACUÍFERO
0114 ACEITUNO

Por ultimo se pide que sostengan la pronunciación de las 5 vocales del castellano cuatro veces:

v01 AEIOU
v02 AEIOU
v03 AEIOU
v04 AEIOU

ANEXO

B

Formulario de consentimiento informado

B. FORMULARIO DE CONSENTIMIENTO INFORMADO

FORMULARIO DE CONSENTIMIENTO INFORMADO

TÍTULO DEL PROYECTO: Adquisición de una base de datos de voces de personas laringuectomizadas**.** RESTORE.

Investigador responsable: Inmaculada Hernáez Rioja.

Código:

Yo (*Nombre y Apellidos del Participante*).....

declaro que he leído la Hoja de Información al Participante, de la que se me ha entregado una copia. Se me han explicado las características del estudio, así como los derechos que puedo ejercitar, y las previsiones sobre el tratamiento de datos. He recibido suficiente información sobre el estudio.

Sé que se mantendrá en secreto mi identidad y que se identificarán mis datos con un sistema de codificación (anonimización de los datos).

Soy libre de revocar mi consentimiento en cualquier momento y por cualquier motivo, sin tener que dar explicaciones.

- ☐ Doy mi consentimiento para que se utilicen mis datos como parte de este estudio. Consiento en participar voluntariamente.
- ☐ Autorizo a que mis datos anonimizados puedan en el futuro ser utilizados por el grupo de investigación responsable de este proyecto en otros proyectos de la misma línea de investigación.
- ☐ Autorizo a que mis datos anonimizados puedan en el futuro ser cedidos a otros grupos de investigación cuyos países tengan una legislación de protección de datos equiparable a la legislación española

Fecha:…… Firma del participante:……

Constato que he explicado las características del estudio y las condiciones de conservación que se aplicarán a los datos conservados.

Nombre del Investigador: Inmaculada Hernáez Rioja

Fecha:…… Firma Investigador:……..

RESTORE Adquisición de una base de datos de voces de personas laringuectomizadas

Figura B.1: Formulario de consentimiento informado

ANEXO

C

Características de cada locutor

A continuación se muestra una tabla con los datos de los locutores para cada sesión.

C. CARACTERÍSTICAS DE CADA LOCUTOR

Sesión	Fase	Genero	Edad	Tiempo desde la operación	Frases grabadas
01M3	3.ª	Hombre	62 años 1 mes	9 años 3 meses	100
02M3	3.ª	Hombre	75 años 5 meses	8 años 1 mes	100
03M3	3.ª	Hombre	56 años 3 meses	1 año 4 meses	100
04M3	3.ª	Hombre	59 años 4 meses	1 año 7 meses	100
05M3	3.ª	Hombre	64 años 9 meses	4 años 2 meses	100
06M3	3.ª	Hombre	71 años 4 meses	5 años 2 meses	100
07M3	3.ª	Hombre	57 años 6 meses	1 año 2 meses	100
08M3	3.ª	Hombre	63 años 10 meses	17 años 1 mes	100
09M3	3.ª	Hombre	60 años 6 meses	11 meses	33
09MT	TE	Hombre	60 años 6 meses	11 meses	100
10M3	3.ª	Hombre	76 años 8 meses	10 meses	100
11F3	3.ª	Mujer	70 años 7 meses	3 años	100
12M3	3.ª	Hombre	59 años 2 meses	1 año 6 meses	100
13M2	2.ª	Hombre	66 años 11 meses	1 año	91
14M2	2.ª	Hombre	72 años 5 meses	1 año 11 meses	100
15F2	2.ª	Mujer	60 años 11 meses	1 año 4 meses	33
15F3	3.ª	Mujer	61 años 1 mes	1 año 7 meses	100
16M2	2.ª	Hombre	66 años 1 mes	1 año 6 meses	33
16M3	3.ª	Hombre	66 años 4 meses	1 año 10 meses	100
17M3	3.ª	Hombre	70 años 6 meses	5 años	100
18M3	3.ª	Hombre	60 años 1 mes	2 años 2 meses	100
19M3	3.ª	Hombre	79 años 11 meses	32 años 7 meses	100
20M3	3.ª	Hombre	73 años 7 meses	19 años 2 meses	100
21M3	3.ª	Hombre	82 años 5 meses	3 años 7 meses	100
22M3	3.ª	Hombre	60 años 8 meses	1 año 7 meses	100
23M3	3.ª	Hombre	65 años 2 meses	10 años 1 mes	100
24M3	3.ª	Hombre	60 años 11 meses	9 años 7 meses	100
25F3	3.ª	Mujer	59 años 3 meses	11 años 11 meses	100
26M3	3.ª	Hombre	55 años 2 meses	4 años 8 meses	100
27M3	3.ª	Hombre	65 años 6 meses	15 años 1 mes	100
28F3	3.ª	Mujer	59 años 2 meses	9 años 11 meses	100
29M3	3.ª	Hombre	51 años 4 meses	2 años 11 meses	100
30M3	3.ª	Hombre	60 años 6 meses	Desconocido	100
31M3	3.ª	Hombre	72 años	4 años 2 meses	100
32M3	3.ª	Hombre	77 años 5 meses	2 años 2 meses	100

ANEXO

D

Caracterización acústica de los locutores

Se ha calculado la frecuencia media y la desviación estándar sobre vocales sostenidas para todos los locutores de la base de datos, y también para un locutor sano (H1). Los resultados están recogidos en la tabla D.1.

Así mismo, se han calculado sobre la parte estable de la grabación de vocales sostenidas los distintos valores de jitter para todos los locutores esofágicos. También se ha incluido un locutor de voz sana (S1) para poder comparar los valores. Los resultados se muestran en la tabla D.2

Por último, también se ha calculado el shimmer para todos los hablantes esofágicos y para un hablante de voz sana (S1). El cálculo se realiza sobre la parte estable de las vocales sostenidas, y los valores resultantes aparecen en la tabla D.3.

Tabla D.1: f_0 media y desviación estándar para un locutor de voz sana (S1) y de todos los locutores esofágicos calculada con tres métodos distintos sobre vocales sostenidas.

Locutor	Autocor.		PSIAIF + Autocor.		PSIAIF + SRH	
	Media (Hz)	STD	Media (Hz)	STD	Media (Hz)	STD
S1	110.29	1.15	110.23	1.34	109.51	8.28
01	206.98	139.95	55.28	6.74	55.54	6.92
02	311.26	87.35	39.97	12.44	49.84	11.40
03	188.51	125.85	52.74	20.15	57.39	16.83
04	371.92	96.08	44.65	17.75	51.43	15.42
05	278.93	179.43	51.75	16.31	60.48	14.86
06	418.55	94.81	44.94	12.00	45.29	11.75
07	244.76	158.13	51.41	18.52	53.91	19.27
08	327.58	107.72	45.18	12.79	48.69	10.80
09	226.73	94.96	47.74	19.14	52.03	16.71
10	110.78	89.19	59.41	15.25	61.78	15.12
11	100.37	93.43	59.56	19.88	62.49	16.04
12	228.36	173.25	44.60	14.13	53.70	17.39
13	367.10	101.19	38.05	14.27	47.68	11.18
14	165.62	135.30	55.23	16.35	54.09	18.64
15	341.79	74.27	35.59	12.54	39.90	13.93
16	127.61	34.83	61.28	12.66	52.05	17.30
17	159.11	108.69	60.06	13.72	49.72	14.99
18	318.04	108.40	46.24	16.02	48.92	16.09
19	356.61	128.24	47.57	8.19	48.57	7.34
20	193.87	156.97	52.29	15.05	50.50	15.71
21	412.71	90.44	51.79	16.32	47.24	15.98
22	290.50	159.11	53.09	13.29	53.97	15.17
23	187.47	82.04	53.35	12.29	56.47	11.00
24	288.18	129.33	52.38	7.33	53.98	5.93
25	101.22	78.82	56.34	16.28	65.61	16.34
26	156.24	42.55	59.51	15.10	60.73	16.04
27	234.36	151.59	45.41	18.44	54.31	15.87
28	248.56	123.22	49.66	14.29	55.07	7.74
29	318.05	105.23	36.91	12.42	46.06	13.22
30	127.90	107.61	59.45	14.01	60.26	14.59
31	162.22	108.16	56.42	8.58	50.71	15.71
32	353.63	100.79	46.48	20.08	50.22	16.62

Tabla D.2: Jitter calculado sobre vocales sostenidas para un locutor de voz sana (S1) y para los 32 locutores esofágicos mediante tres métodos distintos.

Locutor	Autocor.			PSIAIF + Autocor.			PSIAIF + SRH		
	jitta(μs)	jitt(%)	rap(%)	jitta(μs)	jitt(%)	rap(%)	jitta(μs)	jitt(%)	rap(%)
S1	7.98	0.09	0.02	45.01	0.50	0.24	333.79	3.62	2.40
01	213.64	2.35	1.47	155.13	0.84	0.32	261.31	1.43	0.87
02	134.36	3.31	2.01	617.51	2.23	1.26	1355.06	6.32	3.94
03	425.14	5.63	3.53	687.67	2.95	1.75	3084.79	15.67	10.04
04	75.90	2.63	1.43	814.09	3.12	1.77	3137.55	14.46	9.43
05	321.84	4.55	2.88	602.62	2.74	1.54	1352.05	7.53	4.71
06	180.80	6.33	4.05	734.38	3.04	1.69	2600.01	10.83	6.82
07	451.00	6.40	4.11	1095.69	4.76	2.90	3207.78	14.75	9.16
08	111.42	2.91	1.78	546.34	2.24	1.31	1501.77	6.94	4.53
09	336.82	5.79	3.63	937.14	3.72	2.29	2872.72	13.20	8.39
10	345.18	2.88	1.77	618.50	3.38	1.89	1073.42	6.16	3.87
11	491.68	3.84	2.35	777.08	3.85	2.29	1789.07	10.19	6.36
12	246.29	3.04	1.86	574.55	2.33	1.35	2293.67	11.01	7.00
13	170.05	4.84	3.02	779.91	2.59	1.49	1956.79	8.73	5.54
14	571.08	5.94	3.80	876.22	4.36	2.69	3548.94	16.45	10.37
15	63.80	2.07	1.14	643.50	2.06	1.27	5352.89	19.05	12.08
16	412.33	4.82	2.77	823.77	4.79	2.77	3203.63	14.39	8.81
17	272.45	3.28	2.01	557.29	3.11	1.82	3754.08	16.76	10.70
18	224.89	5.52	3.43	1007.26	4.03	2.44	3950.77	17.07	10.76
19	152.09	4.06	2.54	277.90	1.27	0.65	531.43	2.51	1.59
20	327.96	3.84	2.30	544.92	2.52	1.43	3473.44	15.45	9.87
21	126.40	4.40	2.72	985.20	4.51	2.78	4223.72	17.56	11.25
22	365.62	6.38	4.01	659.24	3.23	1.84	2245.82	11.03	6.86
23	96.14	1.50	0.78	417.58	2.08	1.14	1440.23	7.79	5.04
24	57.17	1.08	0.58	144.87	0.74	0.30	412.74	2.20	1.39
25	373.42	3.15	1.90	554.41	2.88	1.66	1142.46	6.92	4.48
26	107.97	1.60	0.86	512.99	2.78	1.58	2120.86	11.71	7.40
27	295.68	4.25	2.63	714.95	2.69	1.56	2956.70	14.37	9.13
28	277.81	4.25	2.65	303.55	1.34	0.71	310.72	1.68	1.04
29	206.08	4.81	3.04	678.80	2.23	1.31	1753.49	7.34	4.66
30	291.10	2.45	1.46	820.80	4.43	2.58	1389.03	7.66	4.69
31	229.83	2.96	1.77	436.17	2.40	1.37	4033.58	17.58	11.30
32	51.19	1.68	0.94	1359.60	5.10	3.17	3851.84	16.89	10.45

Tabla D.3: Jitter calculado sobre vocales sostenidas para un locutor de voz sana (S1) y para los 32 locutores esofágicos mediante tres métodos distintos.

Locutor	Shim (%)	ShdB (dB)	Num. Tramas de vocal
S1	2.22	0.19	401
01	14.78	1.22	247
02	28.21	2.04	133
03	19.15	1.61	154
04	39.09	3.50	50
05	24.59	2.32	140
06	35.90	3.55	61
07	26.65	2.23	145
08	16.34	1.50	246
09	22.79	2.07	113
10	19.01	1.61	94
11	22.53	2.00	96
12	26.73	2.23	248
13	30.44	2.68	68
14	29.39	2.81	84
15	24.47	2.87	77
16	27.75	2.81	49
17	26.02	2.26	117
18	26.48	2.30	193
19	13.32	1.12	155
20	25.98	2.43	218
21	43.33	3.56	66
22	23.55	2.12	145
23	15.93	1.41	170
24	14.34	1.12	305
25	19.66	1.72	124
26	17.26	1.57	117
27	23.26	2.08	252
28	17.16	1.44	130
29	25.69	1.94	279
30	23.49	2.17	77
31	25.07	2.49	54
32	32.87	2.96	73

ANEXO

E

Tasa de error de los locutores esofágicos

Utilizando el reconocedor expuesto en 4.3 se han evaluado todas las sesiones de los locutores esofágicos que contienen las 100 frases en castellano. Los resultados se muestran en la tabla E.1. En la primera columna se muestran los resultados utilizando el diccionario general y con un modelo de lenguaje de trigramas. En la segunda columna se muestran las tasas de error con el diccionario limitado al vocabulario de las 100 frases del corpus ZureTTS y un modelo de lenguaje basado en unigramas equiprobables. En la tercera columna están los resultados para el diccionario reducido y el modelo de lenguaje de unigramas, pero los modelos acústicos del ASR están entrenados con las propias voces esofágicas (esf.).

E. TASA DE ERROR DE LOS LOCUTORES ESOFÁGICOS

Tabla E.1: WER (%) para las sesiones de locutores esofágicas. Las dos primeras columnas muestran los resultados con los modelos acústicos de voz sana y la tercera con voces esofágicas.

Sesión	WER(%) ASR estándar	WER(%) ASR lexicón reducido	WER(%) ASR lexicón reducido (esf.)
14M2	100.08	93.11	103.33
06M3	95.16	96.21	59.50
21M3	93.26	93.94	76.68
32M3	92.43	91.14	80.85
07M3	92.20	91.37	69.57
15F3	90.31	91.75	61.85
16M3	90.08	90.39	62.00
10M3	88.80	82.29	48.68
31M3	88.19	81.23	54.50
04M3	86.83	74.34	40.05
05M3	86.45	77.97	46.48
26M3	85.31	80.17	68.13
23M3	83.88	76.23	41.48
30M3	83.50	71.84	37.32
18M3	83.12	74.41	57.38
11F3	80.24	71.61	44.51
20M3	72.67	63.89	43.07
29M3	70.40	56.25	32.78
17M3	70.10	61.54	39.52
08M3	67.98	56.7	38.99
12M3	66.84	55.26	34.44
02M3	65.93	56.25	32.32
19M3	65.48	54.13	34.11
24M3	64.27	55.94	41.86
22M3	63.97	55.26	30.96
01M3	62.38	51.02	37.17
03M3	61.39	51.85	25.36
28F3	60.79	48.68	29.45
25F3	56.78	43.38	22.94
09MT	52.23	37.70	28.77

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 2019 under the supervision of Inma Hernáez at the University of the Basque Country.

Bilbao, July 2019

This dissertation was finished writing in Bilbao on Thursday 4th July, 2019