DESARROLLO DE UN SISTEMA DE VISUALIZACIÓN DE LA FORMA DEL TRACTO VOCAL SOBRE SECUENCIAS DE IMÁGENES DE RESONANCIA MAGNÉTICA

ANDRÉS MAURICIO CRISTANCHO JULIAO

UNIVERSIDAD INDUSTRIAL DE SANTANDER
FACULTAD DE INGENIERÍAS FÍSICO-MECÁNICAS
ESCUELA DE INGENIERÍA ELÉCTRICA, ELECTRÓNICA Y TELECOMUNICACIONES
BUCARAMANGA
2015

DESARROLLO DE UN SISTEMA DE VISUALIZACIÓN DE LA FORMA DEL TRACTO VOCAL SOBRE SECUENCIAS DE IMÁGENES DE RESONANCIA MAGNÉTICA

ANDRÉS MAURICIO CRISTANCHO JULIAO

Trabajo de grado para optar al título de Ingeniero Electrónico

Director
FRANKLIN ALEXANDER SEPÚLVEDA SEPÚLVEDA
Ingeniero Electrónico, Ph.D

UNIVERSIDAD INDUSTRIAL DE SANTANDER
FACULTAD DE INGENIERÍAS FÍSICO-MECÁNICAS
ESCUELA DE INGENIERÍA ELÉCTRICA, ELECTRÓNICA Y TELECOMUNICACIONES
BUCARAMANGA
2015

CONTENIDO

	Pág.
INTRODUCCIÓN	14
1. JUSTIFICACIÓN	17
2. PLANTEAMIENTO DEL PROBLEMA	18
3. OBJETIVOS	20
3.1 OBJETIVO GENERAL	20
3.2 OBJETIVOS ESPECÍFICOS	20
4. ESTADO DEL ARTE	21
5. MÉTODO	23
5.1 BASE DE DATOS	23
5.1.1 Descripción de la base de datos	23
5.2 SEGMENTACION BASADA EN MODELOS DE APARIENCIA ACTIVA (AAM), MÉTODO 1	25
5.1.2 Modelado de la forma	27
5.1.3 Modelado de la apariencia	30
5.2 SEGMENTACIÓN ROBUSTA DE LOS LÍMITES DE LOS TEJIDOS DE LA VÍA AÉREA, MÉTODO 2	31
5.3 CRITERIOS DE COMPARACIÓN	34
5.4 ANÁLISIS ACÚSTICO-ARTICULATORIO	35
5.5 REPRESENTACIÓN DE LA SEÑAL ACÚSTICA	36
5.5.1 Espectrograma	36

5.5.2 Formantes 36

6. RESULTADOS 38
6.1 RESULTADOS DEL MÉTODO 1 38
6.2 RESULTADOS DEL MÉTODO 2 44
6.3 ANALISIS DE RESULTADOS 48
6.4 SISTEMA DE VISUALIZACIÓN 51

7. CONCLUSIONES 53

8. TRABAJOS FUTUROS 54

REFERENCIAS BIBLIOGRÁFICAS 55
BIBLIOGRAFIA 59

ANEXOS 63

LISTA DE FIGURAS

Pág.

Figura 1. Muestra de una imagen MRI con las estructuras anatómicas etiquetadas. 15
Figura 2. Imágenes del medio sagital del tracto vocal para cada uno de los diez hablantes que actualmente están en la base de datos; (a) imágenes MRI de la configuración oral, (b)imágenes MRI de la configuración nasal. 25
Figura 3. Segmentaciones deseadas para las configuraciones con el velo cerrado (izquierda) y el velo abierto (derecha). 26
Figura 4. Imagen MRI con la línea guía y los puntos anatómicos etiquetados para la elaboración de la rejilla. 32
Figura 5. Imagen MRI con el proceso de elaboración de rejilla terminado. 32
Figura 6. Imagen de la interfaz gráfica que se emplea como herramienta para la realización de la segmentación manual. 35
Figura 7. Imágenes MRI con 26 puntos referencia mostrados (rojo) y puntos secundarios (verdes) para la elaboración de modelos. 39
Figura 8. Región de análisis velar. 40
Figura 9. Ejemplos de segmentaciones empleando el método 1. (a) segmentaciones de la configuración nasal. (b) segmentaciones de la configuración oral. 42
Figura 10. Diagramas de cajas del coeficiente DSC de las segmentaciones ejecutadas con el método 1, (a) con inicialización manual, (b) con inicialización automática. 43
Figura 11. Ejemplos del procedimiento de elaboración de la rejilla. 45
Figura 12. Ejemplos de segmentaciones realizadas con el método 2. 46
Figura 13. Diagrama de cajas del coeficiente DSC de las segmentaciones ejecutadas con el método 2. 47

Figura 14. Variabilidad en los resultados de la convergencia del modelo según su ubicación inicial. (a) Inicialización cercana a la laringe y la cavidad nasal, (b) resultado de la inicialización cercana a la laringe y la cavidad nasal, (c) inicialización cercana a los labios, (d) resultado de la Inicialización cercana a los labios. 49
Figura 15. Falencias típicas de la segmentación del método 2. 50
Figura 16. Diagramas de cajas del coeficiente DSC donde se comparan las segmentaciones realizadas con el método 1 (rojo) y el método 2 (azul). 50
Figura 17. Sistema de visualización. 52

LISTA DE TABLAS

Pág.

Tabla 1. Promedio y desviación estándar del coeficiente DSC de la segmentación realizada con el método 1 para cada hablante. 44

Tabla 2. Promedio y desviación estándar del coeficiente DSC de la segmentación realizada con el método 2 para cada hablante. 47

Tabla 3. Promedio y desviación estándar del coeficiente DSC de la segmentación realizada con cada método. 51

LISTA DE ANEXOS

	Pág.
ANEXO A. SELECCIÓN DE IMÁGENES	63

ABREVIATURAS Y ACRÓNIMOS

AAM	Modelo de apariencia activa (Active Appearance Model)
ASM	Modelo de forma activa (Active Shape Model)
DHT	Transformada de Hough dinámica (Dynamic Hough Transform)
DSC	Coeficiente de similitud de Dice (Dice similarity coefficient)
EMA	Articulógrafo electromagnético (Electromagnetic Articulometry)
MRI	Imágenes de resonancia magnética (Magnetic Resonance Imaging)
PCA	Análisis de componentes principales (Principal Component Analysis)
RF	Radio frecuencia (Radio Frequency)
RM	Resonancia magnética
RT-MRI	Imágenes de resonancia magnética en tiempo real (Real-Time Magnetic Resonance Imaging)

RESUMEN

TÍTULO: DESARROLLO DE UN SISTEMA DE VISUALIZACIÓN DE LA FORMA DEL TRACTO VOCAL SOBRE SECUENCIAS DE IMÁGENES DE RESONANCIA MAGNÉTICA[*]

AUTOR: ANDRÉS MAURICIO CRISTANCHO JULIAO[**]

PALABRAS CLAVE: Imágenes de MRI; Imágenes MRI en tiempo real; Tracto vocal; Segmentación de imágenes; Modelos de forma activa (ASM); Modelos de apariencia activa (AAM); Producción del habla.

DESCRIPCIÓN:

Actualmente, las imágenes de resonancia magnética (MRI) son ampliamente utilizadas en el estudio de los articuladores del tracto vocal. La segmentación de estas imágenes y de videos MRI es difícil debido a la alta variabilidad de los contornos del tracto vocal durante el proceso de producción del habla. Además, la segmentación de vídeos MRI es una tarea laboriosa debido al gran número de cuadros que forman estos videos, por tanto, se requieren métodos que minimicen la intervención del usuario.

El presente trabajo se centra en el desarrollo de un sistema de visualización de secuencias de imágenes de resonancia magnética que describe las formas del tracto vocal. Para llevar a cabo esta tarea, dos métodos de segmentación recientemente desarrollados son analizados y comparados; donde, el Coeficiente de Similitud de Dice (DSC) se utiliza como criterio de comparación de rendimiento. El primer método se basa en Modelos de Apariencia Activa (AAM). El segundo método realiza una estimación robusta de la trayectoria de la vía aérea entre las paredes del tracto vocal utilizando el algoritmo de Viterbi. Luego, se construye una rejilla del tracto vocal. Los límites de los tejidos de las vías respiratorias se encuentran para cada línea de la rejilla mediante la búsqueda del píxel más cercano de mayor intensidad.

Después, con el fin de evaluar el estado de las técnicas de los dos métodos, se aplica la prueba t-student sobre los valores DSC. Se encontró que no existen diferencias estadísticamente significativas entre los resultados de los dos métodos analizados en este trabajo. Además, se desarrolló un sistema de visualización, donde se muestran los contornos del tracto vocal y la acústica del habla. En particular, se pueden observar el oscilograma, el espectrograma y los formantes. Por lo tanto, se podría llevar a cabo un análisis acústico-articulatorio.

[*] Proyecto de Grado

[**] Facultad de Ingenierías Físico-mecánicas. Escuela de Ingeniería Eléctrica, Electrónica y Telecomunicaciones. Director: Dr. Ing. Franklin Alexander Sepúlveda Sepúlveda.

SUMMARY

TITLE: DEVELOPMENT OF A SYSTEM FOR VIEWING VOCAL TRACT SHAPE ON SEQUENCE MAGNETIC RESONANCE IMAGING [*]

AUTHOR: ANDRÉS MAURICIO CRISTANCHO JULIAO[**]

KEYWORDS: Magnetic Resonance Imaging; Real-Time Magnetic Resonance Imaging; Vocal Tract; Image segmentation; Active Shape Model (ASM); Active Appearance Model (AAM); Speech production.

DESCRIPTION:

Currently, magnetic resonance imaging (MRI) are widely used in the study of vocal tract articulators. The segmentation of these MRI images and videos is difficult due to the high variability of vocal tract contours during speech production process. In addition, MRI video segmentation is a high intensive labour task because of the large number of frames forming these videos, so, methods where the user intervention is minimized are required.

Present work is focused on the development of a display system of MRI image sequences describing the vocal tract shapes. In order to accomplish this task, two segmentation methods recently developed are analyzed and compared; where, the Dice Similarity Coefficient (DSC) value is used as the criterion of performance comparison. The first method is based on Active Appearance Models (AAM). The second method performs robust estimation of airway path between the vocal tract walls by using the Viterbi algorithm. Then, a grid line of the vocal tract is constructed. The tissue-airway boundaries are found for each grid line by searching the closest pixel of higher intensity.

Afterwards, in order to evaluate the two state of art methods, the t-student test is applied on DSC values. It was found that no statistically significant differences exist between the performances of the two methods analyzed in present work. Furthermore, it is developed a visualization system where vocal tract contours and speech acoustics are shown. In particular, the oscillogram, the spectrogram and the formants can be observed. Thus, acoustis-articulatory analysis could been carried out.

[*] Grade Works

[**] Facultad de Ingenierías Físico-mecánicas. Escuela de Ingeniería Eléctrica, Electrónica y Telecomunicaciones. Director: Dr. Ing. Franklin Alexander Sepúlveda Sepúlveda.

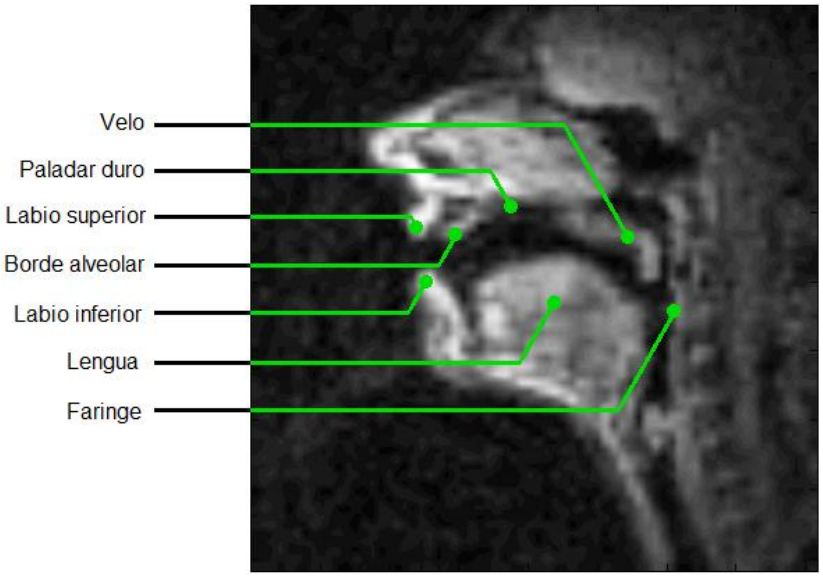
INTRODUCCIÓN

Los seres humanos desde el principio de los tiempos han creado diversas formas de comunicación, una de ellas es la comunicación oral. La comunicación oral es la manifestación de nuestros pensamientos a través de la palabra hablada con fines comunicativos. Hoy en día la voz es la principal manera de comunicación entre los humanos; además de contener información sobre el mensaje que se desea transmitir, también contiene información sobre las emociones y estados fisiológicos del hablante. En consecuencia, se han estudiado los mecanismos de producción de la voz humana y se han creado sistemas de simulación y reconocimiento de voz [1].

Para la producción de la voz humana son necesarios tres elementos fundamentales. Primero, un cuerpo elástico el cual vibra y está formado por dos membranas situadas en la garganta denominadas cuerdas vocales. Segundo, un medio mecánico, cabe destacar que en este caso es el aire. Y tercero, la caja de resonancia, la cual está formada por parte de la garganta y por la boca.

Hablar un sólo idioma limita la habilidad de una persona para comunicarse efectivamente tanto en reuniones como en otros lugares. Hoy en día, así como es de fundamental hablar correctamente también lo es saber bien más de un idioma [2]. Además, poder comunicarse en una segunda lengua es de importancia con fines comerciales y de trabajo.

Figura 1. Muestra de una imagen MRI con las estructuras anatómicas etiquetadas.



El poder visualizar *los labios, dientes, paladar duro, velo del paladar y mandíbula,* definidos como *los articuladores* es de utilidad en varios campos: fonoaudiología, fonética, en el desarrollo de sistemas de comunicación alternativos y en sistemas de aprendizaje de la pronunciación de un segundo idioma. En la figura 1 se observa una imagen MRI de la cabeza, en la cual se etiquetan las estructuras anatómicas del tracto vocal.

La visualización anteriormente mencionada, se puede realizar mediante videos en los cuales se observa la técnica de resonancia magnética en tiempo real. Esta herramienta se ha desarrollado para diversas aplicaciones como lo son la adquisición de imágenes cardiacas e imágenes abdominales [3].

Este proyecto de grado se encuentra centrado en el desarrollo de un sistema que permita visualizar la configuración de los articuladores del habla, construido a partir de datos articulatorios provenientes de videos de resonancia magnética en tiempo real. Lo que se busca es que la presente aplicación siente bases para que

a futuro se pueda desarrollar un sistema que permita apoyar a los estudiantes de un idioma extranjero para dominar la articulación de los sonidos del habla que no existen en su lengua madre y a su vez ayudar en la terapia de los trastornos del habla como una técnica de estimulación visual [4].

1. JUSTIFICACIÓN

En la actualidad existe una gran preocupación por el aprendizaje del idioma inglés y es que el aprender una segunda lengua diferente a la lengua nativa de una persona, requiere de ciertos métodos y conocimientos de aprendizaje en cuanto a la gramática y fonética del idioma. Generalmente el idioma original o nativo se aprende de manera natural. Pero aprender un segundo idioma es más complejo [5]. El aprendizaje mismo no es simplemente una cuestión de inteligencia y aptitud sino de actitud y motivación también. Ésta ha sido un área de interés para profesores de lenguas extranjeras y psicólogos.

De la misma manera, pacientes con trastornos del habla como la apraxia pueden beneficiarse de la percepción visual de la dinámica de los movimientos articulatorios en lugar de ver imágenes estáticas (es decir, posiciones de destino consonánticos) [4]. El sistema de visualización permitiría que dichos pacientes, y personas que estén en el proceso de aprendizaje del idioma inglés, observen la segmentación del tracto vocal, con la cual detallarían lo movimientos vocálicos y consonánticos durante el proceso de producción del habla, asimismo, permitirá la realización del análisis acústico-articulatorio de la producción del habla.

La Universidad Industrial de Santander, específicamente el Grupo de Investigación en Control, Electrónica, Modelado y Simulación (CEMOS) como parte de su trabajo investigativo en el análisis y estudio del desarrollo de sistemas de visualización de imágenes MRI, centra sus actividades en el desarrollo de trabajos de grado de pregrado y posgrado teniendo como uno de sus ejes temáticos esta rama de la ingeniería electrónica. Por lo tanto, este proyecto de grado servirá como complemento y soporte para los estudios que se realicen basados en este tema que en la actualidad son de gran interés en el modelado y simulación.

2. PLANTEAMIENTO DEL PROBLEMA

En [6] se muestra que los sistemas de animación (*talking heads*) ayudan al mejoramiento del aprendizaje de un segundo idioma. Ya que el uso de tal información como medio de entrenamiento en la pronunciación de un segundo lenguaje podría ser benéfico y ya ha empezado a extenderse.

Sin embargo se requiere que el sistema sea confiable, por lo tanto se prefiere contar con información proveniente de mediciones reales acerca del funcionamiento del tracto vocal. Para ello se pueden utilizar dispositivos tales como articulógrafos electromagnéticos (EMA), sistema de toma de imágenes de resonancia magnética (MRI), rayos X de energía reducida (XRMB) y sistemas de ultrasonido, entre otros. Dentro de los cuales, los más viables por sus características corresponde a los datos EMA y MRI [7].

Las imágenes MRI poseen mejor resolución espacial con respecto a los sistemas EMA; pero los sistemas EMA poseen mayor resolución temporal [7]. Los sistemas EMA no distorsionan la señal de voz, mientras que el sistema MRI sí lo hace. Sin embargo, la distorsión está un nivel que aún permite escuchar los mensajes contenidos en la voz. Lo ideal consiste en tener un sistema que fusione ambos tipos de información, con lo cual se mezclarían las ventajas de ambos. Sin embargo, primero se debe desarrollar un sistema que permita extraer la información del contorno del tracto vocal y el movimiento de puntos cruciales de articulación para luego poder fusionar la información MRI con la información EMA.

En el presente trabajo se plantea la comparación de dos métodos de segmentación de imágenes de resonancia magnética en tiempo real, teniendo como criterio de comparación el coeficiente de similitud de Dice (DSC, *Dice*

Similarity Coefficient); Adicionalmente, se desarrolla un sistema de visualización que permite mostrar los resultados de la segmentación del tracto vocal y los patrones de respuesta acústica de la voz, con lo cual, se facilitaría el análisis acústico-articulatorio del proceso de producción del habla.

3. OBJETIVOS

3.1 OBJETIVO GENERAL

Desarrollar un sistema de segmentación de secuencias de imágenes de resonancia magnética del tracto vocal que permita la visualización y el análisis acústico-articulatorio del proceso de producción del habla.

3.2 OBJETIVOS ESPECÍFICOS

- Seleccionar un algoritmo de segmentación para secuencias de imágenes de resonancia magnética (MRI) mediante el cual se pueda obtener los contornos que describen la forma del tracto vocal.

- Desarrollar un sistema de visualización 2D (dos dimensiones) el cual permitirá la observación de la forma del tracto vocal, de la misma manera tendrá la facultad de posibilitar el análisis acústico-articulatorio de la producción del habla.

4. ESTADO DEL ARTE

La resonancia magnética es una herramienta útil e importante al momento de proporcionar imágenes para el análisis de la anatomía del tracto vocal, además es una técnica influyente en la investigación de la producción del habla. Las imágenes MRI brindan un alto contraste de los tejidos blandos, por lo cual hace que se empleen típicamente en el análisis de cada uno de los articuladores durante la producción del habla. Lo anterior contribuye al desarrollo de posibles aplicaciones de diagnóstico y/o modelado del tracto vocal [8].

Diversos documentos se han realizado con lo relacionado respecto al estudio de la producción del habla. Similarmente, se encuentra gran cantidad de documentación relacionada con el análisis y segmentación del tracto vocal empleando diversas técnicas. A continuación se presentan algunas técnicas de los documentos mencionados.

En [9] los autores realizan la combinación de la transformada de Hough dinámica (DHT, *Dynamic Hough Transform*) - algoritmo empleado en reconocimiento de patrones de una imagen - con la elaboración de modelos de forma activa (ASM, *Active Shape Model*). Dichos modelos fueron formados con un conjunto de 39 imágenes donde se representan diferentes configuraciones en la forma de la lengua. La combinación mencionada se realiza para realizar la segmentación automática y un seguimiento de manera confiable a la dinámica de la lengua en secuencias de imágenes. Una desventaja de este método es la alta complejidad computacional que requiere.

En [10] se describe un método para la segmentación de imágenes mediante la representación en el dominio de la frecuencia. Este algoritmo fue desarrollado

para segmentar grandes secuencias de imágenes de resonancia magnética del corte sagital con el fin de extraer el contorno del tracto vocal en tiempo real. Cada imagen de la secuencia se procesa de manera independiente, además, realiza una identificación explícita de los diferentes articuladores basado en un modelo geométrico a priori, cuyo ajuste a los datos es optimizado jerárquicamente empleando un procedimiento de gradiente descendente.

En [11] se propone un marco variacional, basado en la evolución de la curva resultante de la segmentación del contorno de la lengua en imágenes de resonancia magnética de corte sagital. Elaboran un modelo de análisis de componentes principales (PCA, *Principal Component Analysis*) de los contornos de la lengua para las diferentes configuraciones de un hablante de referencia, y lo utilizan como forma a priori. Los parámetros de la curva de representación son entonces manipulados para minimizar una función objetivo.

En [12] se realiza la segmentación del tracto vocal de imágenes de resonancia magnética en tiempo real mediante la elaboración de modelos de forma activa orientados. Igualmente, propone un método con el cual se establecen automáticamente los puntos de referencia para formar los modelos, pero no garantiza la ubicación de puntos de referencia importantes como en el labio superior, esto trae como consecuencia una variabilidad en la posición de dichos puntos [13].

5. MÉTODO

La mayor dificultad del presente trabajo corresponde a la adecuada selección de un método de segmentación que permita extraer el contorno del tracto vocal de secuencias de imágenes MRI del tracto vocal, las cuales harán parte del sistema de visualización. A continuación se presentan dos métodos de segmentación del tracto vocal, los cuales corresponden a dos de los métodos más recientemente desarrollados, uno de ellos data del año 2014 y el otro es del 2015. Estos métodos se comparan mediante un criterio de evaluación típicamente utilizado en este tipo de imágenes. También se hará una breve descripción de la base de datos USC-TIMIT y por último se presentaran los criterios de comparación.

5.1 BASE DE DATOS

En esta sección se hará una breve exposición de los datos usados en el presente trabajo. Los datos provienen de la base de datos USC-TIMIT la cual contiene videos de resonancia magnética a una rata de 23.18 cuadros/s [7]. Durante el proceso de adquisición de datos, las frases de estímulo fueron presentados en texto grande en una pantalla de retroproyección donde los hablantes podían leer desde dentro de la cavidad del escáner sin necesidad de mover la cabeza. Información adicional relacionada con el hardware y el sistema de adquisición de datos se puede consultar en [7].

5.1.1 Descripción de la base de datos Las imágenes de resonancia magnética en tiempo real (rtMRI) son una importante herramienta emergente para la investigación del habla ya que proporciona información dinámica del plano sagital de la vía aérea superior del hablante, o cualquier otro plano de exploración de

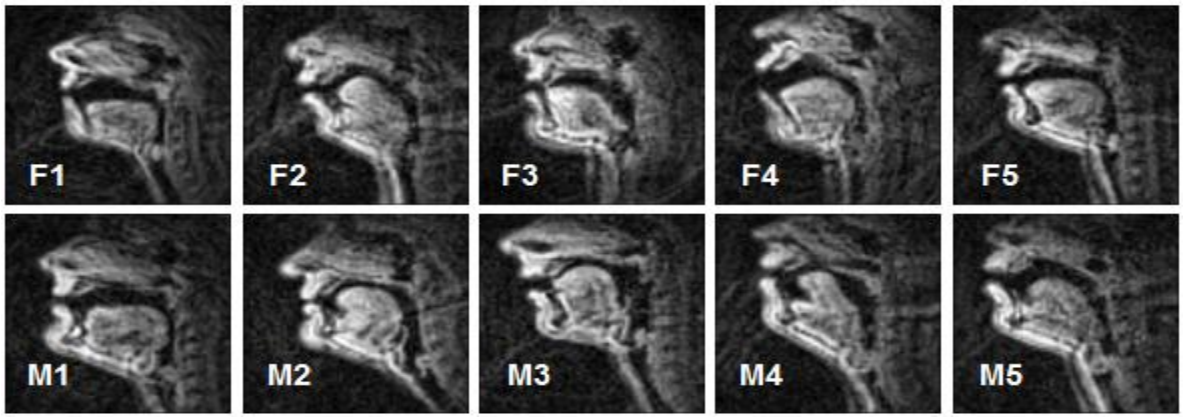
interés. El plano medio sagital de las rtMRI capta el movimiento de la lengua, los labios, la mandíbula y el velo del paladar; además de mostrar la faringe y la laringe, las cuales son regiones que no se puede visualizar fácilmente con otras técnicas de medición de articulación del habla. Las frecuencias de muestreo son actualmente más bajas que las utilizadas en articulometria electromagnética y en rayos X de energía reducida (XRMB). La herramienta rtMRI ofrece información dinámica sobre la conformación del tracto vocal y coordinación articulatoria global [7].

USC-TIMIT es una extensa base de datos que contiene datos de la producción del habla, desarrollada para complementar los recursos existentes a disposición de la comunidad de investigación del habla y con la intención de ser refinada y aumentada continuamente. La base de datos incluye actualmente datos de imágenes de resonancia magnética en tiempo real de cinco hombres y cinco mujeres hablantes de Inglés Americano. Los datos del articulógrafo electromagnético también se han recogido en la actualidad de cuatro de estos hablantes. Las dos modalidades se registraron en dos sesiones independientes, mientras que los sujetos producen el mismo corpus de 460 frases utilizada previamente en la base de datos MOCHA-TIMIT **[7]**.

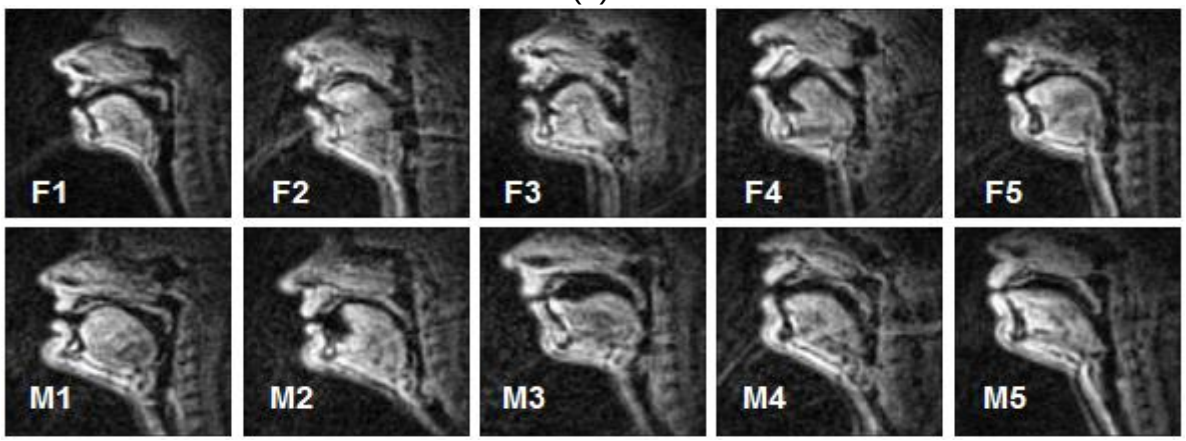
Este conjunto de frases están diseñados para obtener todos los fonemas de Inglés Americano en una amplia gama de contextos prosódicos y fonológicos. Además de proporcionar una muestra fonológicamente integral de inglés, este corpus fue elegido para proporcionar un recurso adicional para los investigadores que ya han hecho uso de la base de datos MOCHA-TIMIT. En ambos casos se registró la señal de audio y sincronizado con los datos articulatorios [7].

En la figura 2 se muestran imágenes MRI del medio sagital del tracto vocal para cada uno de los diez hablantes que actualmente están en la base de datos.

Figura 2. Imágenes del medio sagital del tracto vocal para cada uno de los diez hablantes que actualmente están en la base de datos; (a) imágenes MRI de la configuración oral, (b)imágenes MRI de la configuración nasal.



(a)



(b)

5.2 SEGMENTACION BASADA EN MODELOS DE APARIENCIA ACTIVA (AAM), MÉTODO 1

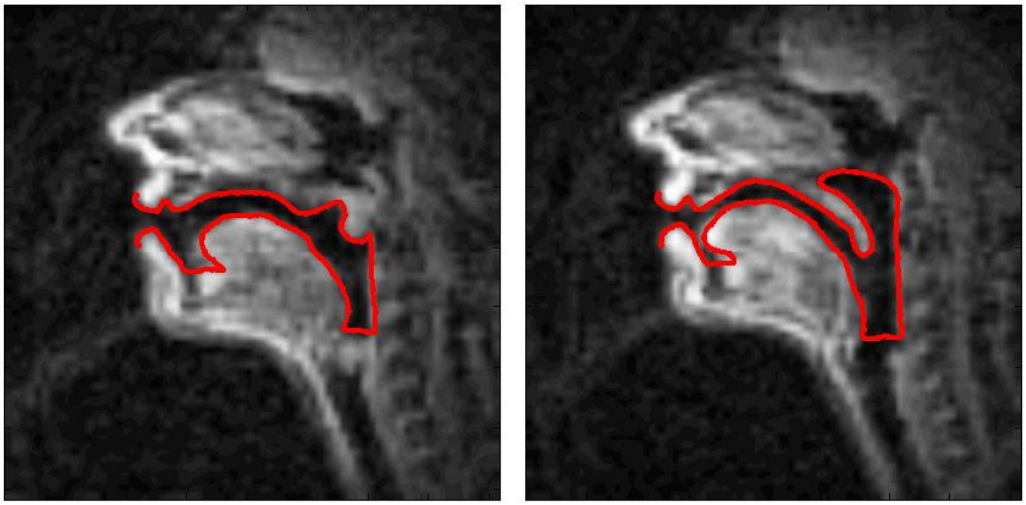
En [13] se propone un método basado en Modelos de Apariencia Activa (AAM por sus siglas en ingles) para la segmentación automática de imágenes de resonancia magnética del tracto vocal. Estas imágenes provienen de videos MRI del perfil sagital de la cabeza. El método AAM se mejora con la elaboración de Modelos de Forma Activa (ASM por sus siglas en ingles).

Tradicionalmente, cuando se utiliza el método de ASM se requiere de una alta complejidad computacional [9]. De la misma manera, cuando se emplean modelos de forma/apariencia activa se presenta un entorno de aplicación limitado [14]. Así mismo, en [12] se elabora una segmentación del tracto vocal usando ASM, para ello establecen un método automático para asignar los puntos de referencia, esto resulta en la falta de puntos de referencia importantes [13].

En contraste con lo anterior, este método realiza una segmentación más rápida y el modelo de convergencia se mejora en comparación con un enfoque tradicional. Este método también tiene en cuenta un esquema de multi-resolución propuesto por [15].

El método de AAM permite la segmentación automática de toda la base de datos usando imágenes etiquetadas previamente de forma manual a modo de insumo para el proceso de entrenamiento.

Figura 3. Segmentaciones deseadas para las configuraciones con el velo cerrado (izquierda) y el velo abierto (derecha).



Se prefiere el uso de dos modelos, uno con el velo del paladar abierto y otro con el velo cerrado, debido a consideraciones de tipo práctico, ya que la diferencia entre los contornos de una y otra configuración es notablemente diferente, ver figura 3. Como resultado se obtiene un modelo para cada configuración. Por lo tanto, es necesario definir un criterio para decidir cuál de los dos modelos se debe aplicar a cada cuadro.

Para entrar un poco en contexto, se hace una concisa explicación de los aspectos más relevantes de los modelos ASM y AAM. El lector puede acceder a información complementaria al respecto en [16], [17] y [18].

5.1.2 Modelado de la forma En este paso es importante elegir puntos de referencia adecuados [18], tales como esquinas notables en los bordes del contorno. Para el proceso es necesario elegir un conjunto de $\boldsymbol{N}$ imágenes de entrenamiento. Luego, estas imágenes se etiquetan manualmente con $\boldsymbol{n}$ puntos de referencia. El vector $\boldsymbol{x_i}$ que contiene los puntos de referencia etiquetados para la forma $\boldsymbol{i}$ es:

$$\boldsymbol{x_i} = \{(\boldsymbol{x_{i1}}, \boldsymbol{y_{i1}}), (\boldsymbol{x_{i2}}, \boldsymbol{y_{i2}}), \dots, (\boldsymbol{x_{in}}, \boldsymbol{y_{in}})\}^{\boldsymbol{T}} \quad (1)$$

Donde $\boldsymbol{i} = \boldsymbol{1}, \boldsymbol{2}, \dots, \boldsymbol{N}$.

Antes de iniciar el análisis estadístico, es de suma importancia realizar una alineación de las diferentes formas. Esto se hace reduciendo al mínimo la distancia entre cada forma y la forma promedio, es decir, minimizando la siguiente expresión:

$$D = \sum |\boldsymbol{x_i} - \bar{\boldsymbol{x}}|^2 \quad (2)$$

Terminado el proceso de alineación, se procede con el análisis estadístico.

Mediante el Análisis de Componentes Principales (PCA por sus siglas en inglés) es posible expresar la forma de un nuevo objeto [18] de la siguiente manera:

Se calcula la forma promedio,

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{x_i} \tag{3}$$

Luego, se calcula la covarianza de los datos,

$$S = \frac{1}{N-1}\sum_{i=1}^{N} (x_i - \bar{x})\,(x_i - \bar{x})^T \tag{4}$$

Después se calculan los vectores propios también conocidos como modos de variación, $\boldsymbol{P_i}$, y los valores propios correspondientes, $\boldsymbol{\lambda_i}$, de la matriz de covarianza, $\boldsymbol{P} = (\boldsymbol{p_1}|\boldsymbol{p_2}|\ldots|\boldsymbol{p_t})$, a continuación se puede aproximar la forma de un nuevo objeto, como sigue:

$$x \approx \bar{x} + Pb \tag{5}$$

Donde

$$b = (b_1, b_2, \ldots, b_t)^T \tag{6}$$

Es un vector de parámetros de peso asociado a cada modo de variación, dicho de otra manera, al variar $\boldsymbol{b}$ se hace variar la forma. Si se limita al parámetro $\boldsymbol{b_i}$ a

$\pm 3\sqrt{\lambda_i}$, se garantiza que la forma generada es similar a la forma del conjunto de entrenamiento [18].

Cada valor propio da la varianza de los datos alrededor del promedio en la dirección del vector propio correspondiente. Seguidamente se calcula la varianza total $V_T = \sum_t \lambda_i$.

Finalmente se elige el primer valor de t de manera tal que,

$$\sum_{i=1}^{t} \lambda_i \geq f_v V_T \tag{7}$$

Donde f_v define la proporción de la variación total que se desea explicar, tradicionalmente varia de 90% a 99,5% [13].

Por otro lado, la convergencia del modelo a los nuevos puntos, se realiza normalmente mediante una transformación euclidiana que define la posición, (X_t, Y_t), la orientación, θ, y la escala, s, del modelo de la imagen [18]. Según lo anteriormente descrito, los puntos del modelo en la imagen, X, se representan por medio de:

$$X = T_{X_t, Y_t, s, \theta}(\bar{x} + Pb) \tag{8}$$

Donde la función $T_{X_t, Y_t, s, \theta}$ realiza una rotación por θ, un escalamiento por s, y una traslación por (X_t, Y_t). Por ejemplo, si se aplica a un solo punto (x, y) seria de la siguiente manera:

$$T_{X_t, Y_t, s, \theta}\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} X_t \\ Y_t \end{pmatrix} + \begin{pmatrix} s \cdot cos\theta & -s \cdot sin\theta \\ s \cdot sin\theta & s \cdot cos\theta \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix} \quad (9)$$

Así mismo si se quiere encontrar los parámetros para la mejor posición, en cuanto a traslación, escalamiento y rotación, de forma que coincida el modelo X a un nuevo conjunto de puntos, Y, se debe minimizar la suma de las distancias cuadradas entre los puntos del modelo y la nueva imagen, esto es equivalente a minimizar la siguiente expresión [18]:

$$\left|Y - T_{X_t, Y_t, s, \boldsymbol{\theta}}(\bar{x} + Pb)\right|^2 \quad (10)$$

5.1.3 Modelado de la apariencia Para ejecutar el modelado de la apariencia se realiza la combinación del modelo de variación de la forma con el modelo de variación de la textura, entiéndase como textura al patrón de intensidades de una imagen [17]. Análogamente al modelado de la forma, el modelado de la apariencia también requiere del Análisis de Componentes Principales para expresar la apariencia de un nuevo objeto, de esta manera el modelo de apariencia se puede expresar como [19]:

$$g \approx \bar{g} + P_g b_g \quad (11)$$

Donde $\bar{\boldsymbol{g}}$ es la textura promedio, $\boldsymbol{P_g}$ describe la matriz de vectores propios o modos de variación de textura del conjunto de entrenamiento y $\boldsymbol{b_g}$ es el vector de parámetros de peso asociado a cada modo de variación.

5.2 SEGMENTACIÓN ROBUSTA DE LOS LÍMITES DE LOS TEJIDOS DE LA VÍA AÉREA, MÉTODO 2

Por otro lado se analizó el método exhibido en [20], donde se presenta un algoritmo para la segmentación robusta de los límites del tejido de la vía aérea registradas en imágenes de resonancia magnética en tiempo real. Dicho documento utiliza el algoritmo de Viterbi para encontrar la ruta de la vía aérea, la ubicación de los labios, y la parte superior de la laringe. Los límites de los tejidos del tracto vocal se encuentran para cada línea de la rejilla por medio de la búsqueda del pixel más cercano de mayor intensidad.

Este método realiza un pre-procesamiento de las imágenes MRI usando una aproximación basada en análisis de multi-resolución con el fin de minimizar los efectos del ruido. Igualmente requiere la elaboración de una rejilla, la cual ayudará a la detección de los labios y la parte superior de la laringe; además esta rejilla permitirá la búsqueda de la ruta optima de la vía aérea para la posterior segmentación de los límites del tracto vocal.

La clave de la segmentación de los límites del tracto vocal es encontrar una ruta de la vía aérea precisa [20]. Para ello, se dibuja manualmente una línea guía curva desde los labios hasta la faringe, de la misma manera es necesario situar manualmente tres puntos anatómicos ubicados en el centro de los labios, otro en el punto más alto del paladar y por ultimo uno en la laringe, ver figura 4. Lo anterior se realiza con el fin de la posterior elaboración de una rejilla, cuyo centro es la línea dibujada manualmente, como se muestra en la figura 5.

Figura 4. Imagen MRI con la línea guía y los puntos anatómicos etiquetados para la elaboración de la rejilla.

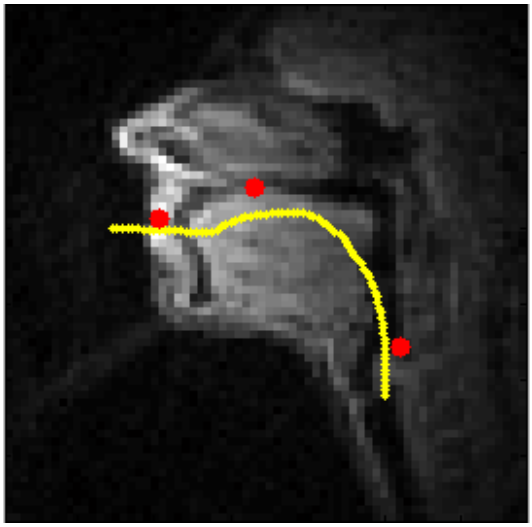
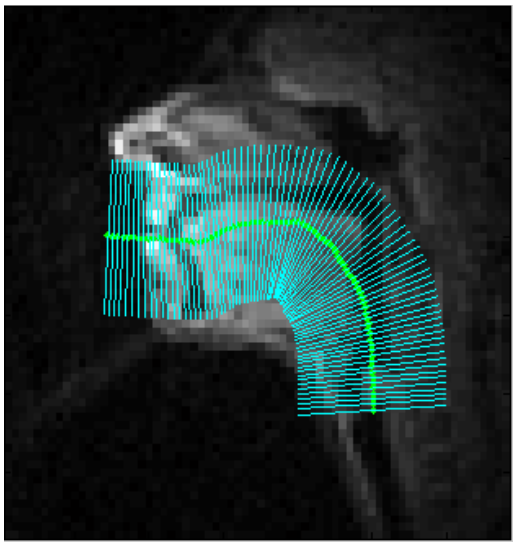


Figura 5. Imagen MRI con el proceso de elaboración de rejilla terminado.



Con el fin de determinar la ruta del aire más probable se utiliza un procedimiento de optimización de la ruta más corta. Se asume que $\boldsymbol{q_t}$ es un estado de la instancia $\boldsymbol{t}$. $\boldsymbol{N}$ indica el número de estados. $\boldsymbol{S^T_{q_i,q_j}}$ denota el índice de transición de $\boldsymbol{q_i}$ a $\boldsymbol{q_j}$. $\boldsymbol{S^L_{q_i}}$ es el índice de probabilidad (de la observación) de $\boldsymbol{q_i}$. $\boldsymbol{P_i}$ es el índice a priori de $\boldsymbol{q_i}$. $\boldsymbol{K}$ es el número de instancias. $\boldsymbol{Q}$ denota la secuencia de estados

$\boldsymbol{q_1, q_2, \ldots, q_K}$, un estado por cada instancia. El índice objetivo $\boldsymbol{\mathcal{J}}$ de $\boldsymbol{Q}$ se define como:

$$\mathcal{J} = \left(P_1 S^L_{q_1} + w S^T_{q_2,q_1}\right) + \left(\sum_{u=2}^{K-1} S^L_{q_u} + w S^T_{q_{u+1},q_u}\right) \tag{12}$$

Donde $\boldsymbol{w}$ es un factor de peso para $\boldsymbol{S^T_{q_i,q_j}}$. La secuencia óptima $\boldsymbol{Q}^*$ es obtenida mediante la búsqueda de $\boldsymbol{Q}$ asociado al minimo de $\boldsymbol{\mathcal{J}}$:

$$Q^* = \underset{[q_1,q_2,\ldots,q_K]}{\arg\min} \ \mathcal{J} \tag{13}$$

Utilizando el algoritmo de Viterbi se determinan las rutas óptimas de la vía aérea que pasan por todas las líneas de la rejilla mediante la búsqueda de las rutas con el mínimo índice $\mathcal{J}$. Cada línea de la rejilla corresponde a una instancia y cada punto posible en una línea corresponde a un estado. En este caso, $\boldsymbol{q_i}$ corresponde al i-esimo estado, donde $\boldsymbol{q_{N/2}}$ se encuentra en la mitad de la línea de la rejilla. $\boldsymbol{S^T_{x,y}}$ es la distancia euclidiana entre los estados $\boldsymbol{x}$ y $\boldsymbol{y}$ ubicados en sus respectivas líneas adyacentes de la rejilla; es decir, ocupará un único estado en cada línea de la rejilla. $\boldsymbol{S^L_x}$ es la intensidad de los pixeles del estado $\boldsymbol{x}$. Entonces, la ruta óptima de la vía aérea se encuentra minimizando el índice de posibles estados tal como se muestra en la ecuación 13.

Luego de haber encontrado la ruta óptima de la vía aérea, se elabora otra rejilla cuyo centro es la ruta óptima encontrada. La nueva rejilla empieza en la parte superior de la laringe y termina en el borde anterior de los labios, es decir, la rejilla empieza a variar lentamente desde la parte superior de la laringe hasta terminar en el borde anterior de los labios. Para que este proceso sea exitoso el método asume que el hablante está mirando hacia la parte izquierda de la imagen.

Posteriormente, se procede con la localización de los labios, los límites de los tejidos del tracto vocal y la laringe. Los límites de los tejidos de las vía aérea se encuentran para cada línea de la rejilla por medio de la búsqueda del pixel más cercano de mayor intensidad. Después de tener los límites estimados de los tejidos de la vía aérea, se realiza un suavizado por medio de un procedimiento de regresión local usando mínimos cuadrados ponderados.

5.3 CRITERIOS DE COMPARACIÓN

En [13] se explica que los modelos iniciales obtenidos de un hablante específico se comparan utilizando el Coeficiente de similitud Dice.

El Coeficiente de similitud Dice (DSC) se puede expresar como:

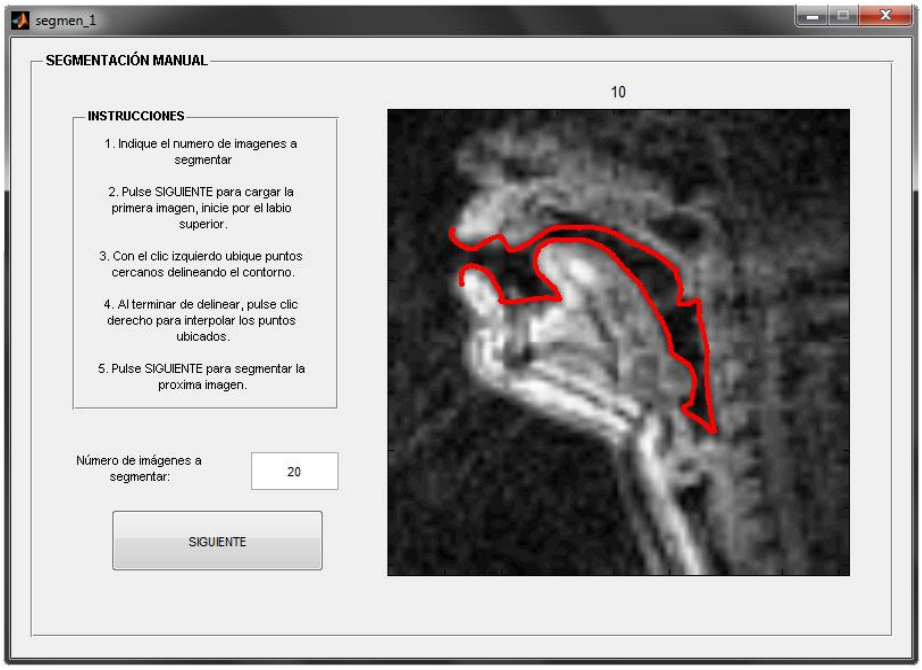
$$DSC = \frac{2|A \cap B|}{|A| + |B|} \tag{14}$$

Donde A y B son las segmentaciones que se desean comparar. Las medidas de DSC hacen referencia a la cantidad de superposición entre las dos segmentaciones normalizadas por las áreas sumadas: por lo tanto, es uno cuando las regiones contenidas dentro de ambos contornos coinciden y cero cuando están completamente diferentes. La literatura sugiere que un valor de DSC > 0,7 representa un excelente solapamiento [21].

Aunque los métodos de segmentación se pueden evaluar usando el índice de Williams, el uso de este índice se ajusta mejor a aquellos casos en los que se tienen varios observadores que segmentan las imágenes.

Para la validación se utilizará una metodología similar a la empleada en [13], esta metodología consiste en realizar una comparación entre la segmentación realizada con cada uno de los métodos y la segmentación realizada de manera manual. La segmentación manual se realizará a un grupo de 20 imágenes por cada hablante, 10 por cada modelo, para un total de 200 imágenes segmentadas manualmente por parte del autor. En la figura 6 se muestra la interfaz gráfica desarrollada para el proceso de segmentación manual.

Figura 6. **Imagen de la interfaz gráfica que se emplea como herramienta para la realización de la segmentación manual.**



5.4 ANÁLISIS ACÚSTICO-ARTICULATORIO

El sistema de visualización permitirá realizar el análisis acústico-articulatorio del proceso de producción del habla, principalmente, mediante la observación del espectrograma de la señal de voz, ya que la representación en el dominio de la

frecuencia facilita una mejor comprensión de la señal, lo cual permite una mejor identificación de sonidos fonéticos. Igualmente, el espectrograma muestra las frecuencias de los primeros cuatro formantes. Estos por su parte, sirven como una herramienta importante para la percepción de vocales [22]. El espectrograma empleado en el sistema de visualización corresponde al desarrollado en [23] y está disponible bajo petición.

5.5 REPRESENTACIÓN DE LA SEÑAL ACÚSTICA

La señal de voz transmite varios tipos de información. La facilidad de interpretar la información de una señal depende de su representación; Por lo tanto, la elección de la representación es de gran importancia. Por ejemplo, la información inferida por los seres humanos de la forma de onda visualizada es escasa y limitada; sin embargo, una gran cantidad de información se convierte fácilmente accesible al transformar la señal de voz en el dominio espectral [22]. De hecho, la lectura de espectrogramas se ha utilizado por los expertos para inferir la información fonética [24].

5.5.1 Espectrograma Desde la invención del espectrógrafo, el espectrograma ha sido la forma más utilizada de visualización para el análisis habla. Sin duda, un espectrograma de voz a veces introduce distorsiones en las estructuras acústicas del habla y con frecuencia no proporciona información adecuada sobre ciertas señales lingüísticamente pertinentes, como el estrés y la entonación. Sin embargo, un espectrograma de voz da una buena descripción de las señales acústicas segmentarias de discurso, y ha sido una herramienta muy valiosa para el desarrollo de nuestra comprensión de los procesos de producción del habla [24].

5.5.2 Formantes Las frecuencias de resonancia del tracto vocal son de fundamental importancia en la producción del habla. Dichas frecuencias son las

frecuencias naturales, o frecuencias propias, de la ruta del aire en el tracto vocal desde la glotis hasta los labios, además la ruta del aire tiene principalmente la forma de la lengua, la mandíbula y otros articuladores [25]. Debido a que tales resonancias del tracto vocal se definen como características de un sistema físico, están obligadas a existir en algunos valores de frecuencia en todo momento, incluso cuando la boca está completamente cerrada emiten una señal acústica débil o no medible [22].

A diferencia de las frecuencias de resonancia, los formantes se definen en el dominio acústico. Los formantes están asociados con picos o prominencias en el espectro de potencia suavizado de la señal acústica del habla, en otras palabras, están relacionados con máximos locales en la amplitud del espectro y no se debe a las propiedades relacionadas con la fuente de espectro [22]. Teniendo en cuenta la definición acústica, los formantes desaparecerían durante el cierre consonántico completo [25].

Los patrones de los formantes sirven de herramienta para la indicación primaria para la percepción de vocales. Cuando las vocales se han sintetizado utilizando frecuencias formantes estimadas del habla natural, los resultados han sido generalmente satisfactorios. Por otro lado, se han desarrollado experimentos basados en sintetizadores articulatorios que muestran la relación entre formantes y representaciones del tracto vocal [22].

6. RESULTADOS

En esta sección se describe el proceso para la obtención de los resultados con cada uno de los métodos mencionados. Igualmente se realiza el análisis de los resultados logrados y finalmente se toma la decisión pertinente. Los códigos para la implementación del métodos 1 y el método 2 están disponibles en [26] y [27] respectivamente.

6.1 RESULTADOS DEL MÉTODO 1

El proceso de segmentación empleando este método inicia con la elección de imágenes de entrenamiento para la elaboración de los modelos. Para ello, se establecen algunos criterios [13]:

- Incluir imágenes de todos y cada uno de los hablantes
- Elegir imágenes con las diferentes posibles configuraciones del tracto vocal
- El número de imágenes seleccionada no debe ser innecesariamente grande

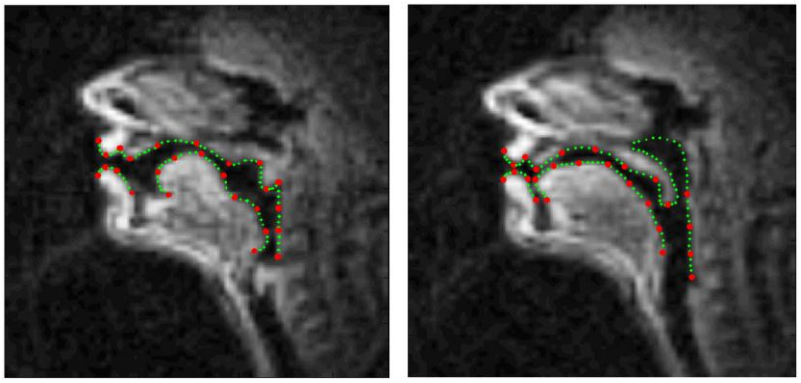
Con el fin de crear modelos apropiados, para elaboración del conjunto de imágenes de entrenamiento, además de cumplir con los criterios anteriormente descritos, se deben elegir imágenes representativos como aquellas donde el velo se encuentre completamente abierto, además, no se tuvieron en cuenta imágenes donde los labios estaban completamente cerrados; las anteriores consideraciones se hacen con el fin de generar mejores resultados [13]. La formación del conjunto de imágenes de entrenamiento se realizó con 12 imágenes por cada hablante, de las cuales 7 pertenecen al modelo nasal y 5 al modelo oral; por consiguiente, el conjunto de imágenes de entrenamiento está conformado por 120 imágenes en total, 70 para el modelo nasal y 50 para el modelo oral.

Para la realización de los modelos oral y nasal, se utilizaron 26 puntos de referencia en cada modelo de tal manera que permitieran un completo cubrimiento del tracto vocal. Dichos puntos se ubicaron de la siguiente manera [13]:

- 3 puntos en el labio superior
- 1 en el borde alveolar
- 2 en el paladar
- 3 en el velo
- 4 en la faringe
- 1 en la raíz de la lengua
- 6 puntos se ubicaron de manera equidistante a lo largo del dorso y la espalda de la lengua
- 1 en la punta de la lengua
- 2 puntos en el frenillo lingual
- 3 puntos en el labio inferior

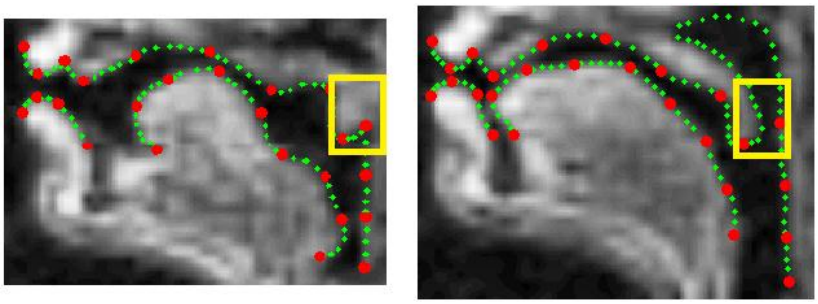
El procedimiento usado para la ubicación manual de los puntos de referencia también permite la colocación de puntos secundarios, como se observa en la figura 7, con el fin de guiar la interpolación entre los puntos de referencia.

Figura 7. **Imágenes MRI con 26 puntos referencia mostrados (rojo) y puntos secundarios (verdes) para la elaboración de modelos.**



Como consecuencia de la elaboración de dos modelos, es necesario establecer un criterio por medio del cual se elija el modelo que se debe aplicar a cada cuadro. Dicho criterio es la región ubicada entre los puntos de referencia de la punta del velo y la parte superior de la faringe. En la figura 8 se muestra la región de análisis velar en la cual se calcula la intensidad media y se compara con un umbral especificado para cada hablante, si la intensidad media de la región velar es menor al umbral señalado se aplica el modelo nasal, en contraparte si la intensidad media de dicha región es mayor al umbral establecido se designa el modelo oral.

Figura 8. **Región de análisis velar.**



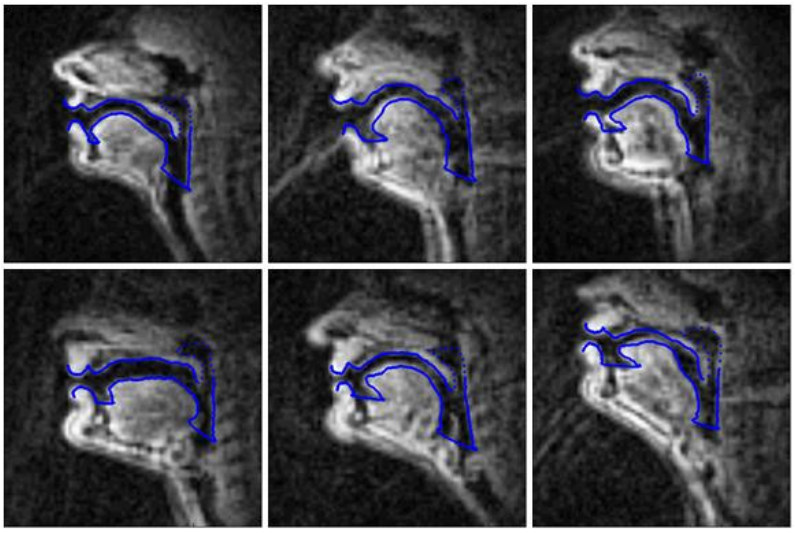
Una vez se han creado los modelos y se ha establecido el criterio para seleccionar el modelo adecuado, se puede proceder con la segmentación. Para este propósito, la inicialización de este método requiere la ubicación del modelo promedio calculado sobre el cuadro que se va a segmentar. Teniendo en cuenta lo anterior, es importante constituir la manera como se realizara la aplicación de los modelos.

Primero, se decidió realizar la aplicación del modelo de manera manual a cada una de las imágenes que hacen parte del conjunto de imágenes de evaluación. Para definir si la segmentación es adecuada, se debe determinar si el modelo convergió con el tamaño del tracto vocal del hablante así como también en zonas importantes como los labios y la cavidad nasal [13].

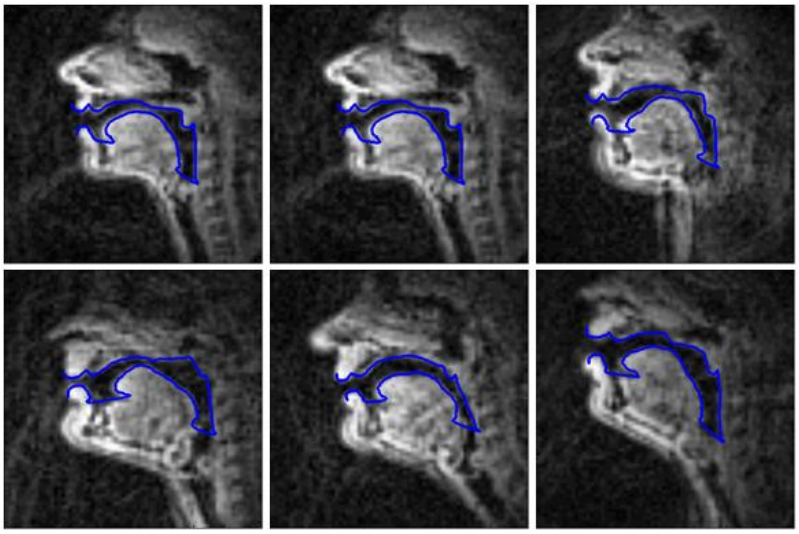
Durante este proceso se detalló que el método presenta una alta sensibilidad en lo concerniente a la ubicación inicial del modelo promedio calculado. Asimismo, se observó que las segmentaciones realizadas a las imágenes de los hablantes masculinos no convergían de manera adecuada; por esta razón se hace necesario la elaboración de modelos exclusivos para dichos hablantes.

En la figura 9 se muestran algunos resultados de las segmentaciones realizadas con el presente método. Igualmente, en la figura 10 se ilustra el diagrama de cajas para los resultados obtenidos de la comparación entre la segmentación del presente método, empleando la inicialización automática y la segmentación manual.

Figura 9. Ejemplos de segmentaciones empleando el método 1. (a) segmentaciones de la configuración nasal. (b) segmentaciones de la configuración oral.

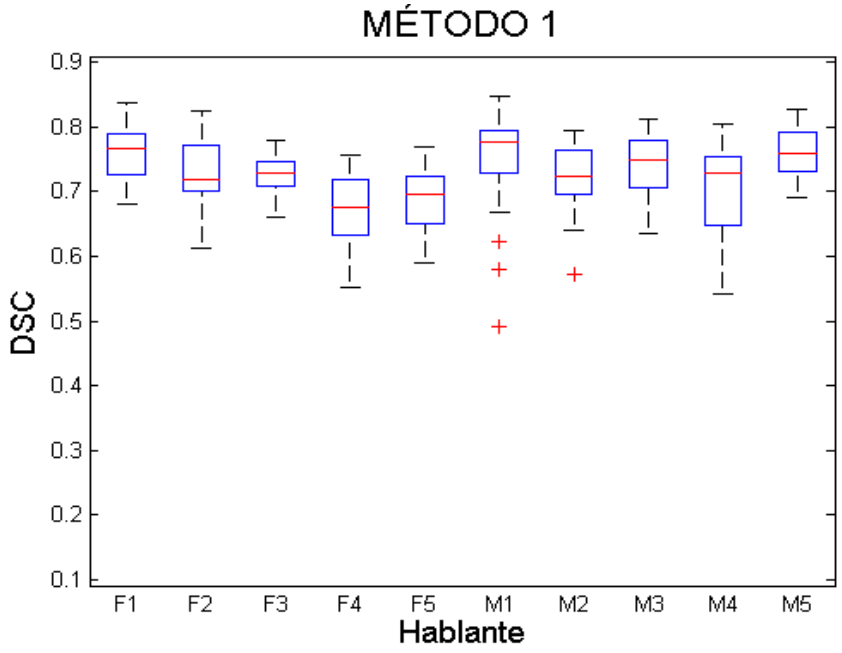


(a)

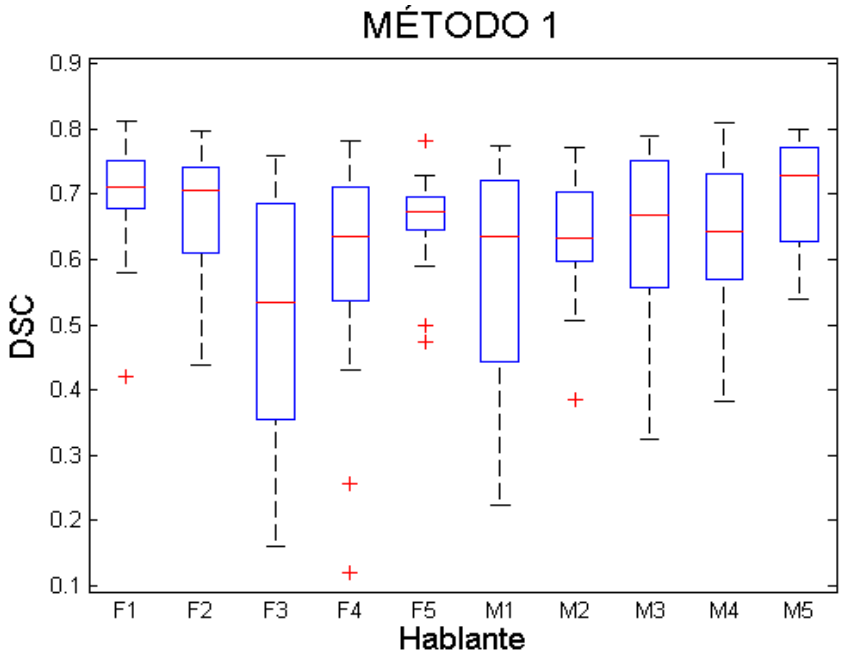


(b)

Figura 10. Diagramas de cajas del coeficiente DSC de las segmentaciones ejecutadas con el método 1, (a) con inicialización manual, (b) con inicialización automática.



(a)



(b)

Para efectos prácticos, se plantea una inicialización de un modo automático. Dicha inicialización se ejecuta de manera tal que el punto de ubicación del modelo promedio calculado corresponda a la media de los puntos obtenidos en la inicialización manual. Sin embargo, debido a la alta sensibilidad del método no se obtienen buenos resultados, como lo refleja el diagrama de cajas del coeficiente DSC en la figura 10(b).

En la Tabla 1 se detalla el promedio y la desviación estándar del coeficiente DSC obtenido para cada hablante.

Tabla 1. Promedio y desviación estándar del coeficiente DSC de la segmentación realizada con el método 1 para cada hablante.

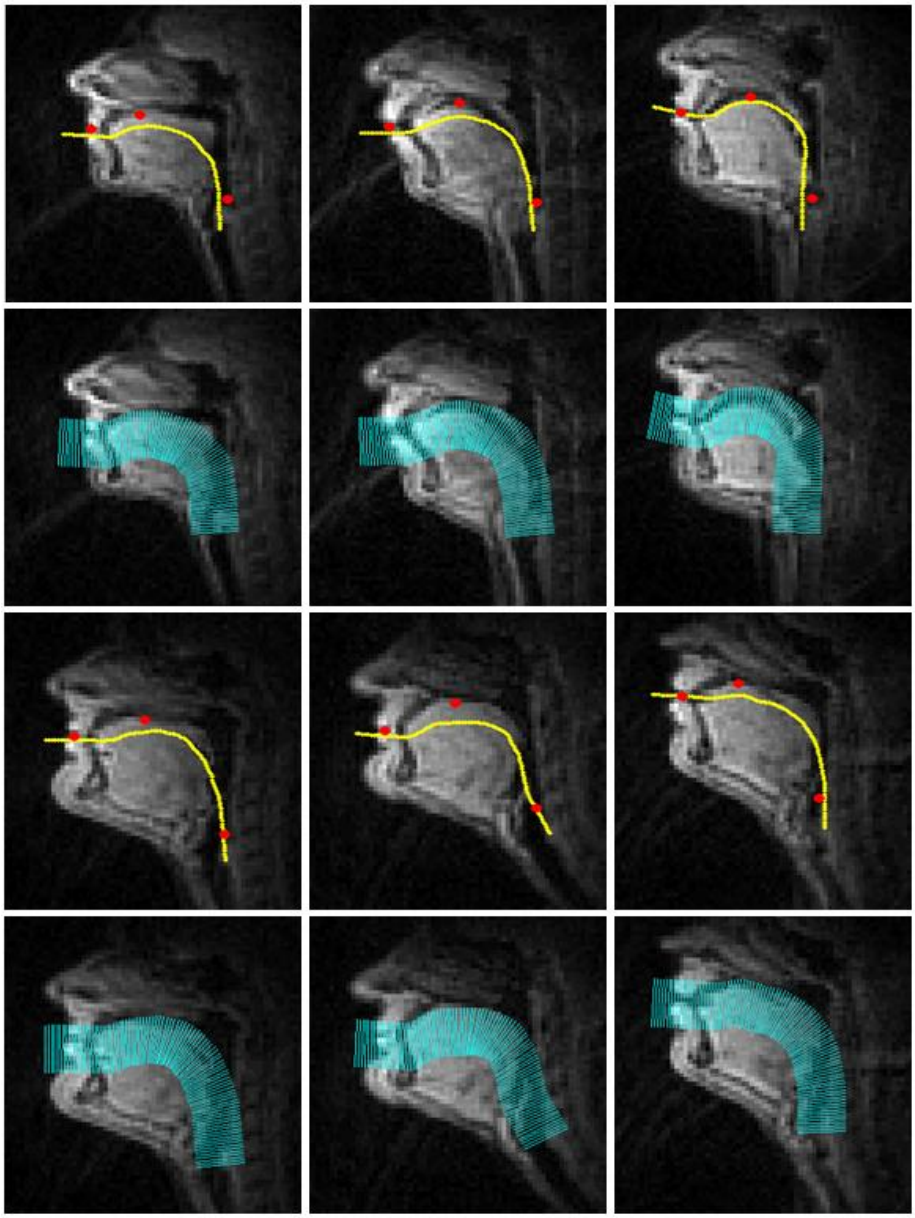
HABLANTE	PROMEDIO	DESVIACIÓN ESTÁNDAR
F1	0.7677	0.0401
F2	0.7196	0.0575
F3	0.7289	0.0350
F4	0.6753	0.0561
F5	0.6974	0.0510
M1	0.7765	0.0890
M2	0.7227	0.0540
M3	0.7503	0.0530
M4	0.7298	0.0782
M5	0.7587	0.0400

6.2 RESULTADOS DEL MÉTODO 2

Como se describió en la sección 5.3, el proceso de segmentación de este método requiere la elaboración de una rejilla. Para lograrlo se necesita esbozar una línea curva desde los labios hasta la faringe, así mismo es necesario situar

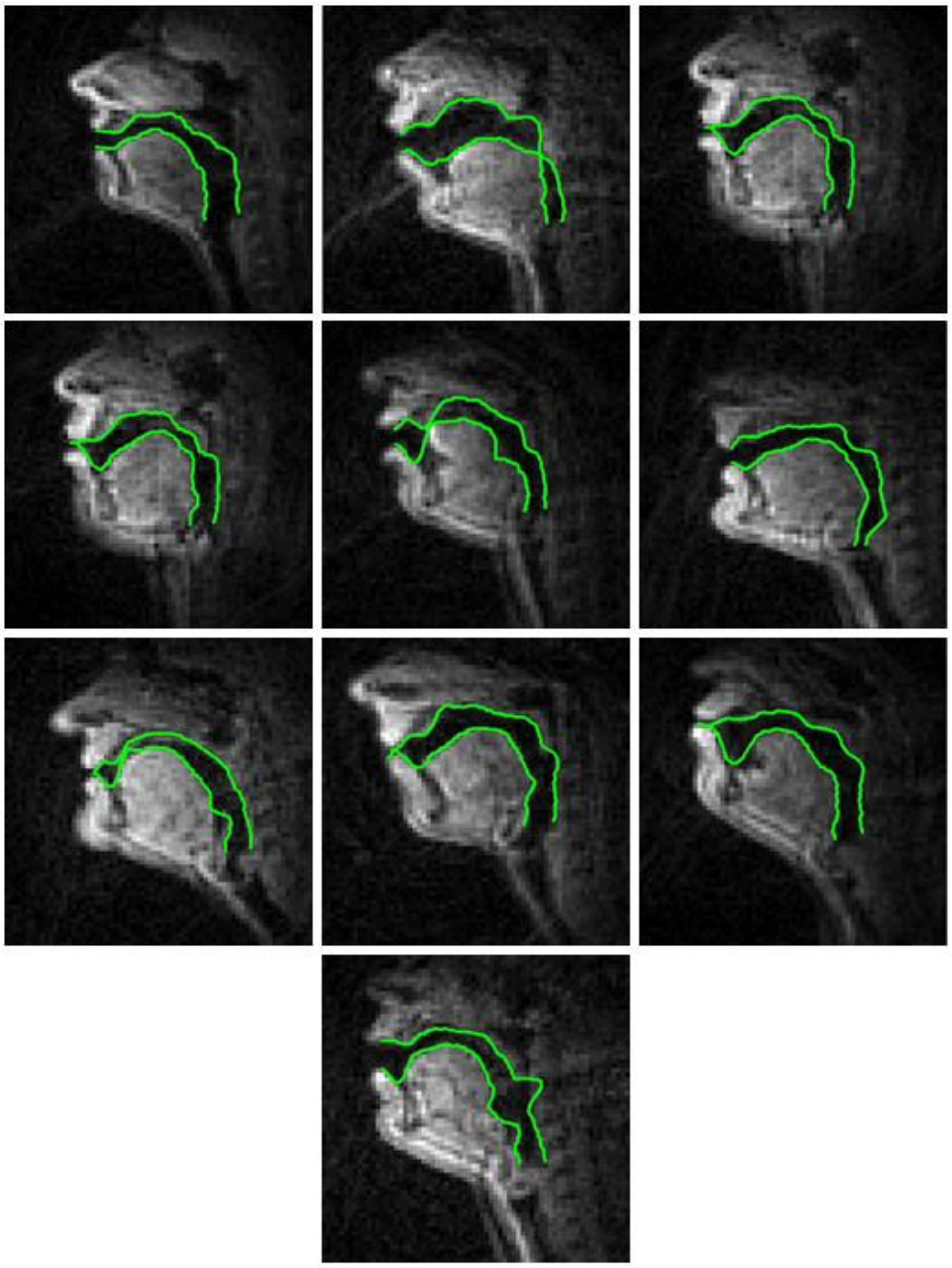
manualmente tres puntos anatómicos ubicados en el centro de los labios, otro en el punto más alto del paladar y por ultimo uno en la laringe. En la figura 11 se muestran algunos ejemplos de lo mencionado anteriormente. El proceso descrito se ejecuta para cada hablante, debido a la alta variabilidad en cuanto a ubicación y tamaño del tracto vocal de cada uno de ellos.

Figura 11. **Ejemplos del procedimiento de elaboración de la rejilla.**



Terminado el proceso de elaboración de la nueva rejilla, se procede con la segmentación. En este punto es importante mencionar que el presente método, a diferencia del método 1, no tiene en cuenta la cavidad nasal. En la figura 12 se ilustran algunas imágenes segmentadas con el método 2.

Figura 12. **Ejemplos de segmentaciones realizadas con el método 2.**



En la figura 13 se observa el diagrama de cajas para los resultados obtenidos de la comparación entre la segmentación del método 2 y la segmentación manual. En la Tabla 2 se detalla el promedio y la desviación estándar del coeficiente DSC obtenido de las segmentaciones realizadas con el método 2 para cada hablante.

Figura 13. Diagrama de cajas del coeficiente DSC de las segmentaciones ejecutadas con el método 2.

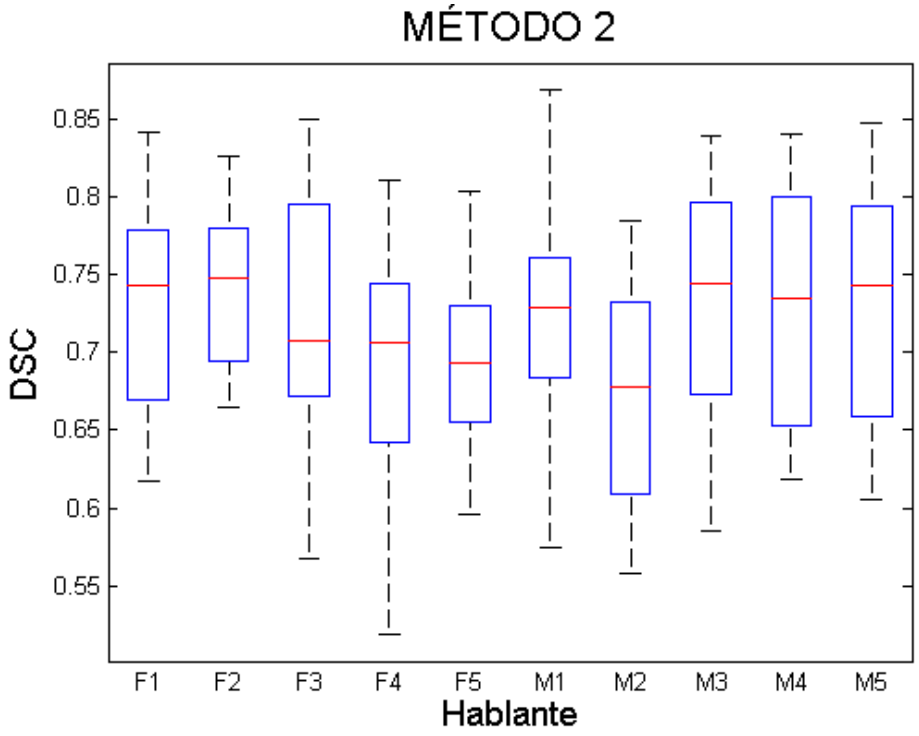


Tabla 2. Promedio y desviación estándar del coeficiente DSC de la segmentación realizada con el método 2 para cada hablante.

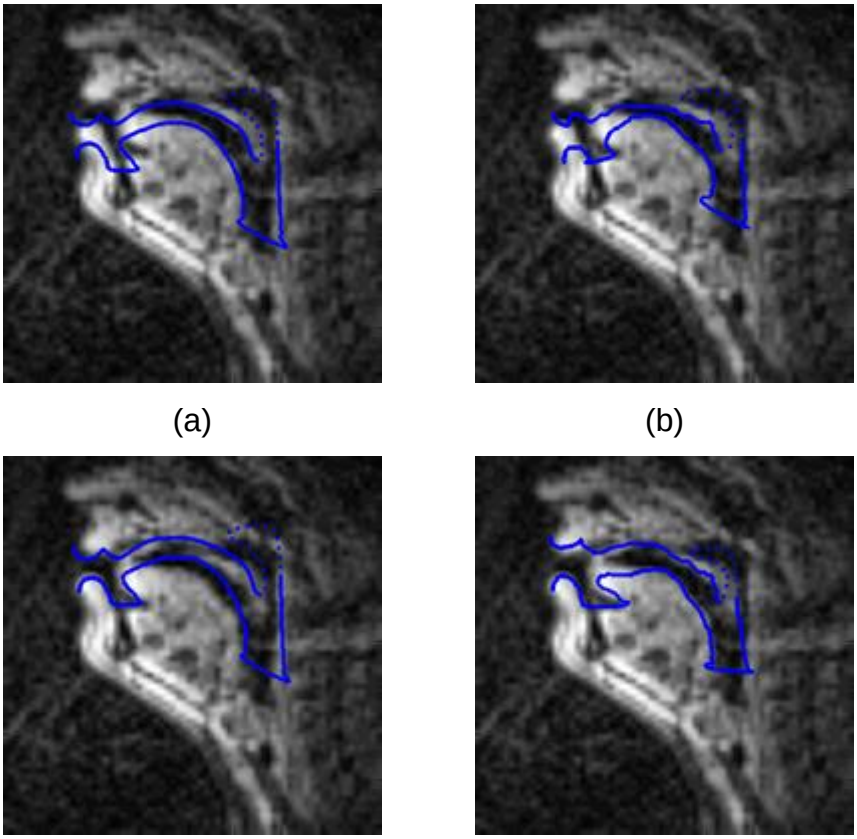
HABLANTE	PROMEDIO	DESVIACIÓN ESTÁNDAR
F1	0.7425	0.0656
F2	0.7474	0.0497
F3	0.7069	0.0802
F4	0.7056	0.0767
F5	0.6932	0.0531
M1	0.7286	0.0794

HABLANTE	PROMEDIO	DESVIACIÓN ESTÁNDAR
M2	0.6778	0.0696
M3	0.7446	0.0724
M4	0.7343	0.0768
M5	0.7435	0.0790

6.3 ANALISIS DE RESULTADOS

Las segmentaciones realizadas con el método 1 en su mayoría presentan buenos resultados, como se muestra en la figura 9, 10(a) y la tabla 1. Sin embargo, se observa la alta sensibilidad que tiene el método en lo concerniente a la ubicación del modelo promedio en hablantes cuyo contorno del tracto vocal variaba de cierta manera en la forma y el tamaño del modelo promedio. Por ejemplo, si la forma del tracto vocal del cuadro que se va a segmentar tiene cierta inclinación en comparación al modelo promedio y el modelo se ubica más cerca de los labios que de la laringe y la cavidad nasal, en el proceso de convergencia el modelo sufre una deformación y al final del proceso se observará una convergencia aceptable del modelo en la zona de los labios y en el mejor de los casos, una convergencia poco aceptable en la parte de la laringe al igual que en la cavidad nasal; igualmente si la ubicación del modelo promedio se realiza cercana a la cavidad nasal y la laringe, el modelo no convergerá en los labios. Lo anterior se observa en la figura 14.

Figura 14. Variabilidad en los resultados de la convergencia del modelo según su ubicación inicial. (a) Inicialización cercana a la laringe y la cavidad nasal, (b) resultado de la inicialización cercana a la laringe y la cavidad nasal, (c) inicialización cercana a los labios, (d) resultado de la Inicialización cercana a los labios.

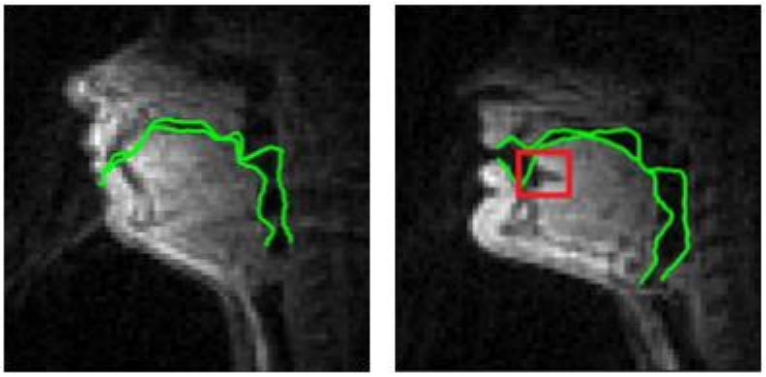


(a) (b)

(c) (d)

Por otro lado, en los resultados del método 2 se observa que, aunque no tiene en cuenta la cavidad nasal, las segmentaciones presentan una buena discriminación del tracto vocal; como se puede apreciar en la figura 12. Sin embargo, la delineación en las imágenes donde los labios están completamente cerrados presenta un pequeño desvío, como se puede ver en la figura 15. Igualmente, debido a la naturaleza del método se observa que la parte inmediatamente debajo de la punta de la legua no es segmentada en los cuadros donde la punta de la lengua se levanta. En figura 15 muestra lo anteriormente mencionado.

Figura 15. Falencias típicas de la segmentación del método 2.



Como se observa en la figura 16 y en la tabla 3, los resultados obtenidos con los dos métodos analizados son muy similares. Para analizar la relación entre los resultados obtenidos por cada método se realiza una prueba t. Luego de realizar la prueba t, se concluye que no existe evidencia en contra de que las medias de las dos muestras sean iguales, o lo que es lo mismo, no se han encontrado diferencias estadísticamente significativas.

Figura 16. Diagramas de cajas del coeficiente DSC donde se comparan las segmentaciones realizadas con el método 1 (rojo) y el método 2 (azul).

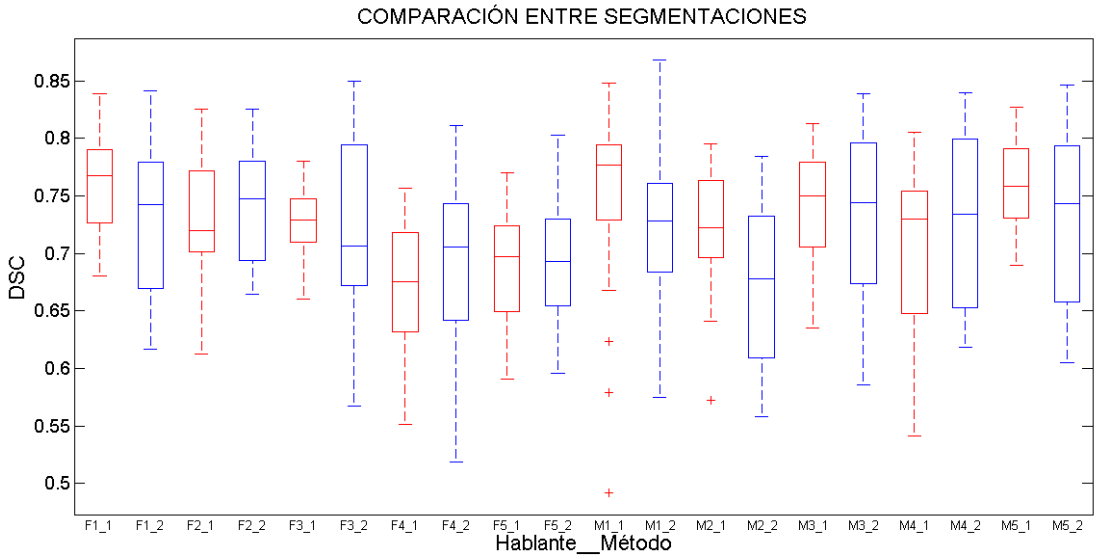


Tabla 3. Promedio y desviación estándar del coeficiente DSC de la segmentación realizada con cada método.

	Promedio	Desviación estándar
Método 1	0.7307	0.0626
Método 2	0.7186	0.0725

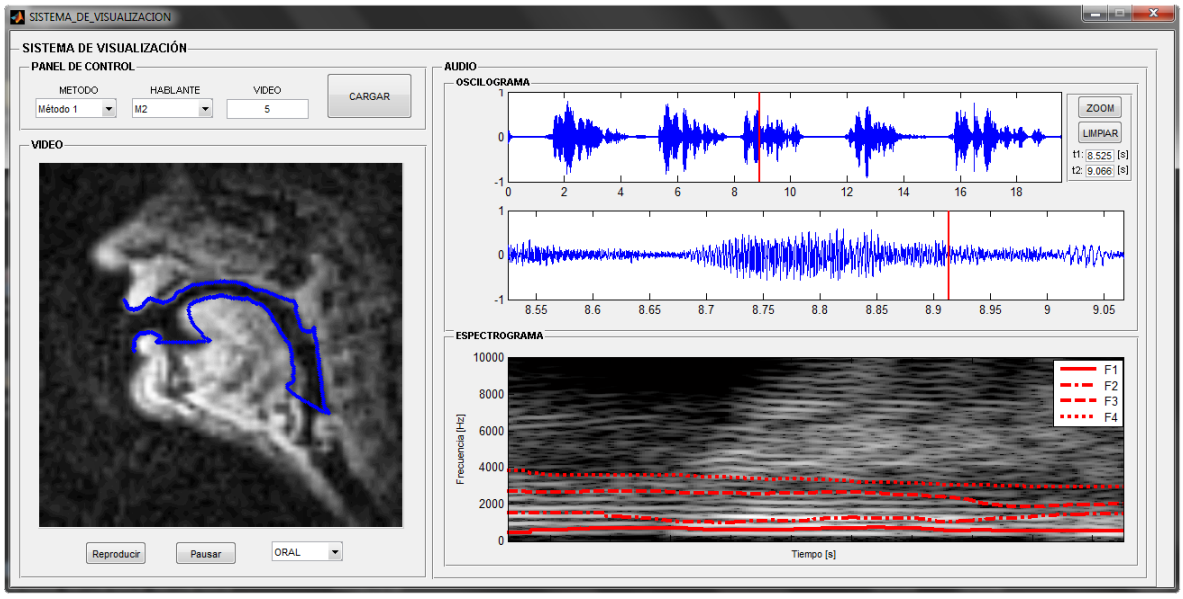
Teniendo en cuenta los resultados obtenidos se decide beneficiar al sistema de visualización con las ventajas que posee cada método, es decir, el sistema de visualización le permitirá al usuario elegir con cuál de los dos métodos decide realizar el análisis.

El método 1 permitirá al usuario analizar detalladamente la ubicación de los articuladores en un cuadro específico. Por otro lado, el método 2 le brindará la facilidad de observar la reproducción del video, previamente seleccionado, donde se contempla la segmentación del tracto vocal.

6.4 SISTEMA DE VISUALIZACIÓN

Como se muestra en la figura 17, el sistema de visualización permite al usuario seleccionar el método, el hablante y el video del cual desea hacer el análisis. Igualmente, proporciona la observación del oscilograma y espectrograma de la señal de audio correspondiente al video escogido. Además, brinda la posibilidad de hacer un acercamiento sobre el oscilograma para un mejor análisis. De la misma manera, el espectrograma muestra la frecuencia de los cuatro primeros formantes.

Figura 17. Sistema de visualización.



7. CONCLUSIONES

- Los resultados obtenidos durante el proceso de segmentación muestran una gran similitud en el coeficiente DSC obtenido con los dos métodos analizados. Si bien el promedio del coeficiente DSC del método 1 es superior al promedio del coeficiente DSC del método 2, la prueba t realizada permite concluir que no se encontraron diferencias estadísticamente significativas. Como consecuencia de lo anterior, se decide implementar los dos métodos en el sistema de visualización.

- En el presente trabajo se desarrolló un sistema de visualización del tracto vocal junto con algunas mediciones espectrales de la señal acústica del habla. Sin embargo, las señales acústicas están contaminadas con ruido provocado por el sistema de adquisición de imágenes MRI, lo cual afecta mediciones tales como el espectrograma.

- Se elaboró un sistema de segmentación de secuencias de imágenes de resonancia magnética del tracto vocal, el cual permite realizar la visualización y análisis acústico-articulatorio del proceso de producción del habla. Para el proceso de segmentación se emplean los dos métodos presentados; por otro lado, el sistema de visualización permite la observación de herramientas importantes en el análisis de la producción del habla como los son el oscilograma y el espectrograma; este último muestra la frecuencia de los cuatros primeros formantes.

8. TRABAJOS FUTUROS

Como consecuencia de la variabilidad en que existe cuanto a la forma, tamaño e inclinación del tracto vocal entre los hablantes pertenecientes a la base de datos USC-TIMIT, se sugiere para trabajos relacionados con la segmentación empleando el método 1, la elaboración de dos modelos para cada hablante, un modelo oral y un modelo nasal. De esta manera se mejorarían los resultados del proceso de segmentación con AAM.

Por otro lado, debido al ruido presente en el audio de los videos de resonancia magnética en tiempo real y sabiendo que la base de datos USC-TIMIT está conformada por videos de imágenes MRI en tiempo real y datos EMA, se recomienda como trabajo futuro el aprovechamiento de la ventaja que tienen los sistemas EMA (no distorsionan la señal de voz) para realizar la concatenación de los videos MRI con el audio obtenido de estos datos.

REFERENCIAS BIBLIOGRÁFICAS

[1] C. Duque, M. Morales. “Caracterización de voz empleando análisis tiempo-frecuencia aplicada al reconocimiento de emociones”. Proyecto de grado, Universidad Tecnológica de Pereira, 2007.

[2] N. Cavassa, “La importancia de hablar diferentes idiomas”, 2008. [Online]. Available: http://importanciadelosidiomas.blogspot.com/. [Accessed: 21-Jan-2015].

[3] S. Narayanan, K. Nayak, S. Lee, A. Sethy, D. Byrd. “An approach to real-time magnetic resonance imaging for speech production”. Journal of the acoustical society of America, 115(4), pp. 1771-1776, 2004.

[4] B. Kröger, J. Gotto, S. Albert, C. Neuschaefer-Rube. "A visual articulatory model and its application to therapy of speech disorders: a pilot study". ZAS Papers in Linguistics (ZASPiL), vol. 40, pp. 79-94, 2005.

[5] D. Wilen, B. Días, NASP Resources. “Aprendizaje de un segundo idioma: Información para los padres”. [Online], Available: http://www.nasponline.org/resources/translations/secondlanguage_spanish.aspx. [Accessed: 21-Jan-2015]

[6] X. Wang, T. Hueber, P. Badin. “On the use of an articulatory talking head for second language pronunciation training: the case of Chinese learners of French”. 10th International Seminar on Speech Production, ISSP10, 2014.

[7] S. Narayanan, A Toutios. “Real-time magnetic resonance imaging and electromagnetic articulography database for speech production research (TC)”. Journal of the acoustical society of America, 136, pp. 1307-1311, 2014.

[8] A. Scott, M. Wylezinska, M. Birch, M. Miquel. “Speech MRI: Morphology and function”. Physica Medica, vol. 30, no. 6, pp. 604-618, 2014.

[9] M. Avila-García, J. Carter, R. Damper. "Extracting tongue shape dynamics from magnetic resonance image sequences". Transactions on Engineering, Computing and Technology, vol. 2, pp. 288-291, 2004.

[10] E. Bresch, S. Narayanan. “Region Segmentation in the Frequency Domain Applied to Upper Airway Real-Time Magnetic Resonance Images”. IEEE Transactions on Medical Imaging, vol. 28, no. 3, pp. 323-338, 2009.

[11] T. Peng, E. Kerrien, M. Berger. "A shape-based framework to segmentation of tongue contours from MRI data". Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on. IEEE, pp. 662-665, 2010.

[12] Z. Raeesy, S. Rueda, J. Udupa, J. Coleman. "Automatic segmentation of vocal tract MR images". Biomedical Imaging (ISBI), 2013 IEEE 10th International Symposium on. IEEE, pp. 1328-1331, 2013.

[13] S. Silva, A. Teixeira. “Unsupervised segmentation of the vocal tract from real-time MRI sequences”. Computer Speech & Language, Vol. 33, no. 1, pp. 25-46, 2015.

[14] M. Vasconcelos, S. Ventura, D. Freitas, J. Tavares. “Towards the automatic study of the vocal tract from magnetic resonance images”. Journal of Voice, Vol. 25, no. 6, pp. 732-742, 2011.

[15] T. Cootes, C. Taylor, A. Lanitis. “Active shape models: evaluation of a multi-resolution method for improving image search”. In: Proc. BritishMachine Vision Conference, pp. 327-336, 1994.

[16] T. Cootes, C. Taylor, D. Cooper, J. Graham. “Active shape models – their training and application”. Comput. Vision Image Underst. 61, pp. 38-59, 1995.

[17] T. Cootes, G. Edwards, C. Taylor. “Active appearance models”. In Pattern Analysis and Machine Intelligence, IEEE Transactions on , vol. 23, no.6, pp. 681-685, 2001.

[18] Appears as Chapter 7: ”Model-Based Methods in Analysis of Biomedical Images” in ”Image Processing and Analysis”, Ed.R.Baldock and J.Graham,Oxford University Press, 2000, pp. 223-248.

[19] T. Cootes, G. Edwards, C. Taylor. “Active appearance models”. In: Proc. European Conference on Computer Vision, pp. 484-498, 1998.

[20] J. Kim, N. Kumar, S. Lee, S. Narayanan. “Enhanced airway-tissue boundary segmentation for real-time magnetic resonance imaging data". In Proceedings of 10-th International Seminar on Speech Production (ISSP), 2014.

[21] A. Zijdenbos, B. Dawant, R. Margolin, A. Palmer. "Morphometric analysis of white matter lesions in MR images: method and validation". In Medical Imaging, IEEE Transactions on, vol. 13, no. 4, pp. 716-724, 1994.

[22] A. Sepulveda. “Estimation of articulatory parameters from the acoustic speech signal”. Tesis de doctorado, Universidad Nacional de Colombia-sede Manizales, 2012.

[23] K. Mustafa, I. Bruce. "Robust formant tracking for continuous speech with speaker variability“, in Audio, Speech, and Language Processing, IEEE Transactions on, vol.14, no.2, pp.435-444, 2006.

[24] V. Zue, R. Cole. "Experiments on spectrogram reading". In Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '79, vol. 4, pp.116-119, 1979.

[25] L. Deng, A. Acero, I. Bazzi. "Tracking vocal tract resonances using a quantized nonlinear function embeddedin a temporal constraint". In Audio, Speech, and Language Processing, IEEE Transactions on, vol. 14, no. 2, pp.425-434, 2006.

[26] Mathworks, “Active Shape Model (ASM) and Active Appearance Model (AAM)”. [Online]. Available: http://www.mathworks.com/matlabcentral/fileexchange/46443-rtmri-seg-v3-zip?s_tid=srchtitle. [Accessed: 03-Jun-2015].

[27] Mathworks, “rtmri_seg_v3.zip”. [Online]. Available: http://www.mathworks.com/matlabcentral/fileexchange/46443-rtmri-seg-v3-zip?s_tid=srchtitle. [Accessed: 21-Jun-2015].

BIBLIOGRAFIA

A. SCOTT, M. WYLEZINSKA, M. BIRCH, M. MIQUEL. “Speech MRI: Morphology and function”. Physica Medica, vol. 30, no. 6, pp. 604-618, 2014.

A. SEPULVEDA. “Estimation of articulatory parameters from the acoustic speech signal”. Tesis de doctorado, Universidad Nacional de Colombia-sede Manizales, 2012.

A. ZIJDENBOS, B. DAWANT, R. MARGOLIN, A. PALMER. "Morphometric analysis of white matter lesions in MR images: method and validation". In Medical Imaging, IEEE Transactions on, vol. 13, no. 4, pp. 716-724, 1994.

APPEARS AS CHAPTER 7: ”Model-Based Methods in Analysis of Biomedical Images” in ”Image Processing and Analysis”, Ed.R.Baldock and J.Graham,Oxford University Press, 2000, pp. 223-248.

B. KRÖGER, J. GOTTO, S. ALBERT, C. NEUSCHAEFER-RUBE. "A visual articulatory model and its application to therapy of speech disorders: a pilot study". ZAS Papers in Linguistics (ZASPiL), vol. 40, pp. 79-94, 2005.

C. DUQUE, M. MORALES. “Caracterización de voz empleando análisis tiempo-frecuencia aplicada al reconocimiento de emociones”. Proyecto de grado, Universidad Tecnológica de Pereira, 2007.

D. WILEN, B. DÍAS, NASP Resources. “Aprendizaje de un segundo idioma: Información para los padres”. [Online], Available:

http://www.nasponline.org/resources/translations/secondlanguage_spanish.aspx. [Accessed: 21-Jan-2015]

E. BRESCH, S. NARAYANAN. “Region Segmentation in the Frequency Domain Applied to Upper Airway Real-Time Magnetic Resonance Images”. IEEE Transactions on Medical Imaging, vol. 28, no. 3, pp. 323-338, 2009.

J. KIM, N. KUMAR, S. LEE, S. NARAYANAN. “Enhanced airway-tissue boundary segmentation for real-time magnetic resonance imaging data". In Proceedings of 10-th International Seminar on Speech Production (ISSP), 2014.

K. MUSTAFA, I. BRUCE. "Robust formant tracking for continuous speech with speaker variability“, in Audio, Speech, and Language Processing, IEEE Transactions on, vol.14, no.2, pp.435-444, 2006.

L. DENG, A. ACERO, I. BAZZI. "Tracking vocal tract resonances using a quantized nonlinear function embeddedin a temporal constraint". In Audio, Speech, and Language Processing, IEEE Transactions on, vol. 14, no. 2, pp.425-434, 2006.

M. AVILA-GARCÍA, J. CARTER, R. DAMPER. "Extracting tongue shape dynamics from magnetic resonance image sequences". Transactions on Engineering, Computing and Technology, vol. 2, pp. 288-291, 2004.

M. VASCONCELOS, S. VENTURA, D. FREITAS, J. TAVARES. “Towards the automatic study of the vocal tract from magnetic resonance images”. Journal of Voice, Vol. 25, no. 6, pp. 732-742, 2011.

MATHWORKS, “Active Shape Model (ASM) and Active Appearance Model (AAM)”. [Online]. Available:

http://www.mathworks.com/matlabcentral/fileexchange/46443-rtmri-seg-v3-zip?s_tid=srchtitle. [Accessed: 03-Jun-2015].

MATHWORKS, “rtmri_seg_v3.zip”. [Online]. Available: http://www.mathworks.com/matlabcentral/fileexchange/46443-rtmri-seg-v3-zip?s_tid=srchtitle. [Accessed: 21-Jun-2015].

N. CAVASSA, “La importancia de hablar diferentes idiomas”, 2008. [Online]. Available: http://importanciadelosidiomas.blogspot.com/. [Accessed: 21-Jan-2015].

S. NARAYANAN, A TOUTIOS. “Real-time magnetic resonance imaging and electromagnetic articulography database for speech production research (TC)”. Journal of the acoustical society of America, 136, pp. 1307-1311, 2014.

S. NARAYANAN, K. NAYAK, S. LEE, A. SETHY, D. BYRD. “An approach to real-time magnetic resonance imaging for speech production”. Journal of the acoustical society of America, 115(4), pp. 1771-1776, 2004.

S. SILVA, A. TEIXEIRA. “Unsupervised segmentation of the vocal tract from real-time MRI sequences”. Computer Speech & Language, Vol. 33, no. 1, pp. 25-46, 2015.

T. COOTES, C. TAYLOR, A. LANITIS. “Active shape models: evaluation of a multi-resolution method for improving image search”. In: Proc. BritishMachine Vision Conference, pp. 327-336, 1994.

T. COOTES, C. TAYLOR, D. COOPER, J. GRAHAM. “Active shape models – their training and application”. Comput. Vision Image Underst. 61, pp. 38-59, 1995.

T. COOTES, G. EDWARDS, C. TAYLOR. “Active appearance models”. In Pattern Analysis and Machine Intelligence, IEEE Transactions on , vol. 23, no.6, pp. 681-685, 2001.

T. COOTES, G. EDWARDS, C. TAYLOR. “Active appearance models”. In: Proc. European Conference on Computer Vision, pp. 484-498, 1998.

T. PENG, E. KERRIEN, M. BERGER. "A shape-based framework to segmentation of tongue contours from MRI data". Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on. IEEE, pp. 662-665, 2010.

V. ZUE, R. COLE. "Experiments on spectrogram reading". In Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '79, vol. 4, pp.116-119, 1979.

X. WANG, T. HUEBER, P. BADIN. “On the use of an articulatory talking head for second language pronunciation training: the case of Chinese learners of French”. 10th International Seminar on Speech Production, ISSP10, 2014.

Z. RAEESY, S. RUEDA, J. UDUPA, J. COLEMAN. "Automatic segmentation of vocal tract MR images". Biomedical Imaging (ISBI), 2013 IEEE 10th International Symposium on. IEEE, pp. 1328-1331, 2013.

ANEXOS

ANEXO A. SELECCIÓN DE IMÁGENES

Para cada hablante se escogieron imágenes de diferentes videos, de modo que de un mismo hablante se eligiera una sola imagen por video.

B.1. IMÁGENES PARA ENTRENAMIENTO DE MODELOS

Para el entrenamiento de los modelos se emplearon 12 imágenes por cada hablante, 5 para el modelo oral y 7 para el modelo nasal.

Tabla 1. Relación de imágenes seleccionadas del hablante F1 para el entrenamiento de los modelos.

HABLANTE F1				
Modelo	**Imagen**	**Video**	**Cuadro**	**Tiempo del video [s]**
ORAL	1	3	431	18,5936
	2	29	511	22,0449
	3	47	481	20,7506
	4	61	171	7,3770
	5	78	301	12,9853
NASAL	1	8	541	23,3391
	2	16	421	18,1622
	3	33	391	16,8680
	4	45	381	16,4366
	5	51	41	1,7688
	6	87	321	13,8481
	7	90	51	2,2002

Tabla 2. Relación de imágenes seleccionadas del hablante F2 para el entrenamiento de los modelos.

HABLANTE F2				
Modelo	**Imagen**	**Video**	**Cuadro**	**Tiempo del video [s]**
ORAL	1	9	111	4,7886
	2	18	221	9,5341
	3	35	451	19,4564
	4	56	201	8,6713
	5	73	151	6,5142
NASAL	1	11	51	2,2002
	2	21	481	20,7506
	3	37	541	23,3391
	4	49	81	3,4944
	5	62	31	1,3374
	6	78	321	13,8481
	7	88	451	19,4564

Tabla 3. Relación de imágenes seleccionadas del hablante F3 para el entrenamiento de los modelos.

HABLANTE F3				
Modelo	**Imagen**	**Video**	**Cuadro**	**Tiempo del video [s]**
ORAL	1	7	491	21,1821
	2	23	141	6,0828
	3	41	281	12,1225
	4	52	551	23,7705
	5	74	421	18,1622
NASAL	1	15	331	14,2796
	2	21	341	14,7110
	3	32	561	24,2019
	4	49	41	1,7688
	5	60	351	15,1424
	6	83	231	9,9655
	7	92	121	5,2200

Tabla 4. Relación de imágenes seleccionadas del hablante F4 para el entrenamiento de los modelos.

HABLANTE F4				
Modelo	**Imagen**	**Video**	**Cuadro**	**Tiempo del video [s]**
ORAL	1	2	221	9,5341
	2	14	61	2,6316
	3	25	281	12,1225
	4	53	81	3,4944
	5	79	361	15,5738
NASAL	1	11	361	15,5738
	2	28	61	2,6316
	3	37	401	17,2994
	4	55	201	8,6713
	5	69	51	2,2002
	6	88	161	6,9456
	7	91	491	21,1821

Tabla 5. Relación de imágenes seleccionadas del hablante F5 para el entrenamiento de los modelos.

HABLANTE F5				
Modelo	**Imagen**	**Video**	**Cuadro**	**Tiempo del video [s]**
ORAL	1	1	211	9,1027
	2	38	281	12,1225
	3	55	431	18,5936
	4	71	131	5,6514
	5	84	41	1,7688
NASAL	1	9	121	5,2200
	2	22	291	12,5539
	3	43	391	16,8680
	4	58	81	3,4944
	5	67	301	12,9853
	6	76	191	8,2399
	7	82	21	0,9060

Tabla 6. Relación de imágenes seleccionadas del hablante M1 para el entrenamiento de los modelos.

HABLANTE M1				
Modelo	**Imagen**	**Video**	**Cuadro**	**Tiempo del video [s]**
ORAL	1	4	141	6,0828
	2	22	421	18,1622
	3	51	461	19,8878
	4	86	301	12,9853
	5	92	171	7,3770
NASAL	1	18	421	18,1622
	2	30	491	21,1821
	3	45	41	1,7688
	4	59	541	23,3391
	5	61	661	28,5160
	6	74	391	16,8680
	7	85	31	1,3374

Tabla 7. Relación de imágenes seleccionadas del hablante M2 para el entrenamiento de los modelos.

HABLANTE M2				
Modelo	**Imagen**	**Video**	**Cuadro**	**Tiempo del video [s]**
ORAL	1	3	421	18,1622
	2	14	361	15,5738
	3	31	311	13,4167
	4	59	141	6,0828
	5	70	411	17,7308
NASAL	1	19	51	2,2002
	2	38	261	11,2597
	3	41	141	6,0828
	4	54	521	22,4763
	5	67	381	16,4366
	6	75	241	10,3969
	7	82	441	19,0250

Tabla 8. Relación de imágenes seleccionadas del hablante M3 para el entrenamiento de los modelos.

HABLANTE M3				
Modelo	**Imagen**	**Video**	**Cuadro**	**Tiempo del video [s]**
ORAL	1	9	151	6,5142
	2	15	181	7,8085
	3	47	331	14,2796
	4	68	471	20,3192
	5	83	211	9,1027
NASAL	1	11	451	19,4564
	2	35	121	5,2200
	3	40	501	21,6135
	4	63	31	1,3374
	5	79	521	22,4763
	6	87	141	6,0828
	7	90	551	23,7705

Tabla 9. Relación de imágenes seleccionadas del hablante M4 para el entrenamiento de los modelos.

HABLANTE M4				
Modelo	**Imagen**	**Video**	**Cuadro**	**Tiempo del video [s]**
ORAL	1	5	401	17,2994
	2	23	131	5,6514
	3	52	391	16,8680
	4	70	261	11,2597
	5	88	31	1,3374
NASAL	1	14	111	4,7886
	2	25	231	9,9655
	3	34	431	18,5936
	4	47	221	9,5341
	5	56	501	21,6135
	6	71	131	5,6514
	7	80	541	23,3391

Tabla 10. Relación de imágenes seleccionadas del hablante M5 para el entrenamiento de los modelos.

HABLANTE M5				
Modelo	**Imagen**	**Video**	**Cuadro**	**Tiempo del video [s]**
ORAL	1	15	71	3,0630
	2	31	501	21,6135
	3	59	531	22,9077
	4	64	401	17,2994
	5	92	291	12,5539
NASAL	1	1	131	5,6514
	2	27	391	16,8680
	3	38	271	11,6911
	4	43	241	10,3969
	5	69	41	1,7688
	6	76	311	13,4167
	7	83	61	2,6316

B.2. IMÁGENES PARA SEGMENTACIÓN MANUAL

La segmentación manual se realizó sobre 20 imágenes por cada hablante, 10 imágenes de la configuración oral y 10 imágenes de la configuración nasal; para un conjunto de 200 imágenes en total.

Tabla 11. Relación de imágenes seleccionadas del hablante F1 para segmentación manual.

HABLANTE F1				
Modelo	Imagen	Video	Cuadro	Tiempo del video [s]
ORAL	1	1	61	2,6316
	2	10	551	23,7705
	3	22	291	12,5539
	4	35	51	2,2002
	5	49	431	18,5936
	6	57	621	26,7903
	7	60	531	22,9077
	8	71	401	17,2994
	9	83	361	15,5738

HABLANTE F1				
Modelo	Imagen	Video	Cuadro	Tiempo del video [s]
	10	92	341	14,7110
NASAL	1	6	441	19,0250
	2	13	251	10,8283
	3	25	221	9,5341
	4	39	541	23,3391
	5	43	241	10,3969
	6	58	331	14,2796
	7	64	681	29,3788
	8	72	41	1,7688
	9	85	391	16,8680
	10	91	431	18,5936

Tabla 12. Relación de imágenes seleccionadas del hablante F2 para segmentación manual.

HABLANTE F2				
Modelo	**Imagen**	**Video**	**Cuadro**	**Tiempo del video [s]**
ORAL	1	2	241	10,3969
	2	10	561	24,2019
	3	27	341	14,7110
	4	33	301	12,9853
	5	41	481	20,7506
	6	54	161	6,9456
	7	65	51	2,2002
	8	72	81	3,4944
	9	83	251	10,8283
	10	90	91	3,9258
NASAL	1	7	471	20,3192
	2	13	141	6,0828
	3	19	371	16,0052
	4	24	41	1,7688
	5	31	391	16,8680
	6	45	301	12,9853
	7	58	111	4,7886
	8	67	441	19,0250
	9	77	81	3,4944
	10	89	331	14,2796

Tabla 13. Relación de imágenes seleccionadas del hablante F3 para segmentación manual.

HABLANTE F3				
Modelo	**Imagen**	**Video**	**Cuadro**	**Tiempo del video [s]**
ORAL	1	5	391	16,8680
	2	12	201	8,6713
	3	26	331	14,2796
	4	38	161	6,9456
	5	46	281	12,1225
	6	51	41	1,7688
	7	59	181	7,8085
	8	63	421	18,1622
	9	72	71	3,0630
	10	85	451	19,4564
NASAL	1	9	131	5,6514
	2	17	341	14,7110
	3	29	31	1,3374
	4	34	311	13,4167
	5	42	421	18,1622
	6	49	41	1,7688
	7	55	441	19,0250
	8	66	621	26,7903
	9	73	191	8,2399
	10	81	61	2,6316

Tabla 14. Relación de imágenes seleccionadas del hablante F4 para segmentación manual.

HABLANTE F4				
Modelo	**Imagen**	**Video**	**Cuadro**	**Tiempo del video [s]**
ORAL	1	8	241	10,3969
	2	17	461	19,8878
	3	22	261	11,2597
	4	35	131	5,6514
	5	41	41	1,7688
	6	48	81	3,4944
	7	59	401	17,2994
	8	60	571	24,6333
	9	73	341	14,7110

HABLANTE F4				
Modelo	Imagen	Video	Cuadro	Tiempo del video [s]
	10	82	201	8,6713
NASAL	1	3	391	16,8680
	2	12	431	18,5936
	3	29	191	8,2399
	4	38	31	1,3374
	5	45	261	11,2597
	6	57	71	3,0630
	7	61	351	15,1424
	8	76	281	12,1225
	9	84	81	3,4944
	10	87	141	6,0828

Tabla 15. Relación de imágenes seleccionadas del hablante F5 para segmentación manual.

HABLANTE F5				
Modelo	Imagen	Video	Cuadro	Tiempo del video [s]
ORAL	1	4	351	15,1424
	2	14	431	18,5936
	3	28	241	10,3969
	4	32	501	21,6135
	5	40	381	16,4366
	6	51	201	8,6713
	7	65	31	1,3374
	8	72	61	2,6316
	9	86	491	21,1821
	10	91	91	3,9258
NASAL	1	6	21	0,9060
	2	15	311	13,4167
	3	19	111	4,7886
	4	27	411	17,7308
	5	34	181	7,8085
	6	48	321	13,8481
	7	53	271	11,6911
	8	66	161	6,9456
	9	79	451	19,4564
	10	87	51	2,2002

Tabla 16. Relación de imágenes seleccionadas del hablante M1 para segmentación manual.

HABLANTE M1				
Modelo	**Imagen**	**Video**	**Cuadro**	**Tiempo del video [s]**
ORAL	1	1	201	8,6713
	2	29	561	24,2019
	3	37	481	20,7506
	4	40	111	4,7886
	5	52	91	3,9258
	6	63	651	28,0846
	7	75	211	9,1027
	8	81	261	11,2597
	9	88	301	12,9853
	10	92	451	19,4564
NASAL	1	9	41	1,7688
	2	15	71	3,0630
	3	24	301	12,9853
	4	36	171	7,3770
	5	43	81	3,4944
	6	57	651	28,0846
	7	54	531	22,9077
	8	65	211	9,1027
	9	79	361	15,5738
	10	89	121	5,2200

Tabla 17. Relación de imágenes seleccionadas del hablante M2 para segmentación manual.

HABLANTE M2				
Modelo	**Imagen**	**Video**	**Cuadro**	**Tiempo del video [s]**
ORAL	1	5	421	18,1622
	2	12	391	16,8680
	3	27	291	12,5539
	4	33	171	7,3770
	5	49	231	9,9655
	6	51	541	23,3391
	7	64	261	11,2597
	8	77	511	22,0449
	9	85	481	20,7506

HABLANTE M2				
Modelo	**Imagen**	**Video**	**Cuadro**	**Tiempo del video [s]**
	10	91	591	25,4961
NASAL	1	4	121	5,2200
	2	17	451	19,4564
	3	22	71	3,0630
	4	26	511	22,0449
	5	30	561	24,2019
	6	43	241	10,3969
	7	55	371	16,0052
	8	69	31	1,3374
	9	72	41	1,7688
	10	88	151	6,5142

Tabla 18. Relación de imágenes seleccionadas del hablante M3 para segmentación manual.

HABLANTE M3				
Modelo	**Imagen**	**Video**	**Cuadro**	**Tiempo del video [s]**
ORAL	1	8	471	20,3192
	2	11	411	17,7308
	3	27	281	12,1225
	4	33	61	2,6316
	5	46	241	10,3969
	6	52	531	22,9077
	7	59	311	13,4167
	8	64	691	29,8102
	9	75	351	15,1424
	10	80	171	7,3770
NASAL	1	19	111	4,7886
	2	23	331	14,2796
	3	28	241	10,3969
	4	31	421	18,1622
	5	42	231	9,9655
	6	54	171	7,3770
	7	65	381	16,4366
	8	77	531	22,9077
	9	81	81	3,4944
	10	85	541	23,3391

Tabla 19. Relación de imágenes seleccionadas del hablante M4 para segmentación manual.

HABLANTE M4				
Modelo	**Imagen**	**Video**	**Cuadro**	**Tiempo del video [s]**
ORAL	1	2	401	17,2994
	2	13	371	16,0052
	3	29	151	6,5142
	4	36	281	12,1225
	5	41	351	15,1424
	6	54	491	21,1821
	7	65	201	8,6713
	8	68	91	3,9258
	9	72	51	2,2002
	10	81	511	22,0449
NASAL	1	12	431	18,5936
	2	24	171	7,3770
	3	33	291	12,5539
	4	38	41	1,7688
	5	40	321	13,8481
	6	57	531	22,9077
	7	61	631	27,2217
	8	75	461	19,8878
	9	84	21	0,9060
	10	90	311	13,4167

Tabla 20. Relación de imágenes seleccionadas del hablante M5 para segmentación manual.

HABLANTE M5				
Modelo	**Imagen**	**Video**	**Cuadro**	**Tiempo del video [s]**
ORAL	1	7	401	17,2994
	2	13	281	12,1225
	3	20	331	14,2796
	4	35	161	6,9456
	5	45	81	3,4944
	6	51	571	24,6333
	7	67	111	4,7886
	8	73	481	20,7506
	9	74	241	10,3969

HABLANTE M5				
Modelo	**Imagen**	**Video**	**Cuadro**	**Tiempo del video [s]**
	10	86	501	21,6135
NASAL	1	3	311	13,4167
	2	18	461	19,8878
	3	22	171	7,3770
	4	29	401	17,2994
	5	33	251	10,8283
	6	45	21	0,9060
	7	54	531	22,9077
	8	61	381	16,4366
	9	78	11	0,4745
	10	87	141	6,0828