Detección del Contorno de la Lengua sobre Imágenes de Ultrasonido del Plano Sagital del Tracto Vocal Utilizando Técnicas de Aprendizaje Profundo

Jorge Enrique Morantes Bohórquez

Trabajo de Grado para Optar al Título de Ingeniero Electrónico

Director
Franklin Alexander Sepúlveda Sepúlveda
Doctor en Ingeniería

Universidad Industrial de Santander
Facultad de Ingenierías Fisicomecánicas
Escuela de Ingenierías Eléctrica, Electrónica y de Telecomunicaciones
Bucaramanga
2021

Dedicatoria

A mi madre Ruth Bohórquez por ser el pilar de mi vida, por estar siempre para mí, por su constante afecto y su paciencia. Gracias por brindarme palabras de aliento desde el inicio hasta el final de mi carrera.

A mi padre Jorge Morantes por su apoyo y confianza en mí.

A mi hermana María Paula, por acompañarme durante todos estos años de vida. Gracias por tantas risas y una que otra pelea.

A mis compañeros y amigos, con los cuales compartí muchas experiencias, horas de estudio y algunas horas de ocio.

A Natalia, mi compañera de vida, junto a quien me desvele en más de una ocasión y cuyo apoyo incondicional fue imprescindible para lograr este objetivo.

Agradecimientos

Al profesor Franklin Sepúlveda, por guiarme durante todo este proceso. Gracias por su colaboración, su predisposición para ayudarme y por sus consejos y correcciones.

A todos los profesores que me formaron en múltiples disciplinas, permitiéndome adquirir una base de conocimientos sólida para desempeñarme profesionalmente.

A la Universidad Industrial de Santander, por no solo formarme como ingeniero sino como ciudadano integral.

A la Escuela de Ingenierías Eléctrica, Electrónica y de Telecomunicaciones, por acogerme y poner todos sus recursos a mi disposición.

A todas las personas que fueron parte de mi proceso de aprendizaje, a los laboratoristas, auxiliares y tutores por el tiempo que dedican constantemente.

Tabla de Contenido

Pág.

Introducción 9

1. Objetivos 11

1.1 Objetivo General 11

1.2 Objetivos Específicos 11

2. Marco Teórico 12

2.1 Estado del Arte 12

2.1.1 Contornos Activos 12

2.1.2 Grafos 13

2.1.3 Modelos de Apariencia Activa 14

2.2.1 Método 15

2.2.2 Base de Datos 18

2.2.3 Resultados 20

3. Conclusiones 24

4. Recomendaciones 25

Referencias Bibliográficas 26

Lista de Tablas

Pág.

Tabla 1. MSD y desviación estándar promedio para diferentes umbrales usando el modelo pre-entrenado 21

Tabla 2. MSD y desviación estándar promedio para diferentes umbrales usando el modelo re-entrenado 21

Tabla 3. MSD y desviación estándar promedio para diferentes umbrales usando el modelo entrenado desde cero 22

Lista de Figuras

Pág.

Figura 1. Funcionamiento de los contornos activos..... 12

Figura 2. Representación de la secuencia de fotogramas mediante un grafo 14

Figura 3. Demostración de la operación de los modelos de apariencia activa 15

Figura 4. Arquitectura Dense U-Net..... 17

Figura 5. Emparejamiento de histograma 18

Figura 6. Imagen de ultrasonido junto a su respectiva máscara 19

Figura 7. MSD entre dos secuencias..... 20

Figura 8. Contornos correctamente generados por cada uno de los modelos..... 23

Figura 9. Contornos erróneamente extraídos..... 23

Resumen

Título: Detección del Contorno de la Lengua sobre Imágenes de Ultrasonido del Plano Sagital del Tracto Vocal Utilizando Técnicas de Aprendizaje Profundo[*]

Autor: Jorge Enrique Morantes Bohórquez[**]

Palabras Clave: Contorno de la lengua, imágenes de ultrasonido, aprendizaje profundo, redes neuronales convolucionales.

Descripción: Las imágenes de la lengua por ultrasonido proporcionan un medio no invasivo para evaluar la posición y el movimiento de la lengua. Sin embargo, la presencia de ruido y el bajo contraste afectan la usabilidad de estas imágenes. En consecuencia, extraer los contornos de la lengua a partir de estas imágenes sigue siendo una tarea no trivial que implica procedimientos manuales costosos y propensos a errores. En fonética lingüística y clínica, la detección de la lengua es el primer paso en el análisis de imágenes ecográficas. Esto tiene múltiples aplicaciones médicas, como el tratamiento de trastornos del sonido del habla, la comparación de la producción del habla sana y la deficiente, el entrenamiento y rehabilitación de un segundo idioma, entre otras. Históricamente se han presentado diversos métodos orientados a la automatización de esta tarea, como los contornos activos, los métodos basados en grafos y las redes neuronales. En este trabajo se plantea la implementación y evaluación de un algoritmo, basado en técnicas de aprendizaje profundo, capaz de extraer el contorno de la lengua sobre secuencias de imágenes de ultrasonido de forma automática.

[*] Trabajo de Grado

[**] Facultad de Ingenierías Fisicomecánicas. Escuela de Ingenierías Eléctrica, Electrónica y de Telecomunicaciones. Director: Franklin Alexander Sepúlveda Sepúlveda. Doctor en Ingeniería.

Abstract

Title: Tongue Contour Detection in Ultrasound Images of the Sagittal Plane of the Vocal Tract Using Deep Learning Techniques[*]

Author: Jorge Enrique Morantes Bohórquez[**]

Key Words: Tongue contour, Ultrasound images, Deep learning, Convolutional neural networks.

Description: Ultrasound imaging provides a non-invasive means of evaluating the position and movement of the tongue. However, the presence of noise and low contrast affect the usability of these images. Consequently, extracting the contours of the tongue from these images remains a non-trivial task involving costly and error-prone manual procedures. In linguistic and clinical phonetics, tongue tracking is the first step in ultrasound image analysis. This has multiple medical applications, such as the treatment of speech sound disorders, the comparison of healthy and poor speech production, second language training and rehabilitation, among others. Historically, various methods aimed at automating this task have been presented, such as active contours, graph-based methods, and neural networks. This work proposes the implementation and evaluation of an algorithm, based on deep learning techniques, capable of extracting the contour of the tongue from ultrasound image sequences automatically.

[*] Degree Work

[**] Faculty of Physical Mechanical Engineering. School of Electrical Engineering, Electronics and Telecommunications. Director: Franklin Alexander Sepúlveda Sepúlveda. Doctor of Engineering

Introducción

La captura de imágenes de la lengua por ultrasonido proporciona un medio no invasivo para evaluar la posición y el movimiento de la lengua. Sin embargo, la presencia de ruido y el bajo contraste afectan la usabilidad de estas imágenes. En consecuencia, extraer los contornos de la lengua a partir de imágenes de ultrasonido sigue siendo una tarea no trivial, costosa en tiempo y en recursos humanos. En fonética lingüística y clínica, la detección de la lengua es el primer paso en el análisis de imágenes ecográficas. Esto tiene múltiples aplicaciones médicas, como el tratamiento de trastornos del sonido del habla, la comparación de la producción del habla sana y la deficiente (Laporte & Ménard, 2017), el entrenamiento y rehabilitación de un segundo idioma (Gick, Bernhardt, Bacsfalvi, & Wilson, 2008), la investigación sobre deglución de alimentos (Ohkubo & Scobbie, 2018), modelado de lengua en 3D (Chen, y otros, 2018), entre otros.

Históricamente se han propuesto varios métodos para la detección del contorno de la lengua de forma semiautomática o automática, destacando los contornos activos (*Snakes*) (Akgul, Kambhamettu, & Stone, 1999; Li, Kambhamettu, & Stone, 2005; Xu, y otros, 2016), métodos basados en grafos (Tang & Hamarneh, 2010) y el uso de redes neuronales (Fasel & Berry, 2010; Jaumard-Hakoun, Xu, Roussel-Ragot, Dreyfus, & Denby, 2016). No obstante, tanto *Snakes* como los métodos basados en grafos permanecen operando de forma semiautomática, requiriendo intervención manual para entregar resultados apropiados o involucrando técnicas más complejas con el objetivo de orientar el algoritmo hacia una segmentación más automática.

Las redes neuronales convolucionales (RNC) se han convertido en el método de predilecto para aplicaciones de visión por computador en los últimos años (Zhou, Greenspan, & Shen, 2017). Han demostrado un rendimiento sobresaliente en tareas de clasificación de imágenes, así como en

la detección, reconocimiento y seguimiento de objetos. En este documento, se realiza la implementación de un algoritmo basado en RNC para la extracción automática del contorno de la lengua utilizando la arquitectura Dense U-Net.

1. Objetivos

1.1 Objetivo General

Desarrollar y evaluar el desempeño de un algoritmo basado en técnicas de aprendizaje profundo para la detección del contorno superior de la lengua sobre secuencias de imágenes de ultrasonido.

1.2 Objetivos Específicos

- Realizar una etapa de preprocesamiento a las imágenes de la base de datos (mejoramiento de contraste y selección de la región de interés).
- Implementar una red neuronal profunda cuya entrada corresponda a una imagen de ultrasonido del plano sagital del tracto vocal y su salida sea un conjunto de coordenadas espaciales (x, y) con los puntos que mejor describen el contorno de la lengua en dicha imagen.
- Evaluar los resultados obtenidos con el algoritmo teniendo como referencia los contornos extraídos de forma manual

2. Marco Teórico

2.1 Estado del Arte

A continuación, se describen los principales métodos propuestos para realizar la detección del contorno de la lengua,

2.1.1 Contornos Activos

Este método se fundamenta en la configuración del conjunto de puntos que mejor describen el contorno de la lengua en el primer fotograma. Estos servirán como referencia para el siguiente fotograma, procurando que la curva se actualice en busca de regiones con cambios grandes de intensidad. Los nuevos puntos obtenidos ayudan a generar el contorno para el siguiente fotograma, esto proceso será llevado a cabo consecutivamente logrando etiquetar toda la secuencia.

Figura 1

Funcionamiento de los contornos activos

Nota. El grafico muestra el proceso de configuración de un contorno inicial, y como este se actualiza cuando se realiza el cambio de fotograma.

La actualización de los puntos se basa en encontrar los contornos que minimizan un funcional de energía, descrito por la siguiente expresión:

$$E_{snake}(V,\ S,\ I) = \sum_{i=1}^{n} \left(\alpha E_{int}(v_i,\ S) + \beta E_{ext}(v_i, I)\right) \quad (1)$$

Esta función considera dos energías: la energía interna, encargada de mantener la regularidad en la curva mediante componentes de fluidez, equidistancia y semejanza respecto a la referencia (ecuación 2) y la energía externa que está dada por el gradiente de la imagen, y es la que permite guiar a los puntos de la curva hacia la región donde haya un mayor cambio de intensidad (ecuación 3).

$$E_{int}(v_i,\, S) = [\lambda_1 \;\; \lambda_2 \;\; \lambda_3] \begin{bmatrix} E_{smo}(v_i) \\ E_{dist}(v_i) \\ E_{sim}(v_i,\, S) \end{bmatrix} \tag{2}$$

$$E_{ext}(v_i,\, I) = -|\nabla I(v_i)| \tag{3}$$

En estas expresiones *S* representa el contorno de referencia, *I* es el fotograma en cuestión, *V* representa el conjunto de puntos que describe la curva α y β son factores de ponderación. También es importante mencionar que la suma de $\lambda_1,\ \lambda_2$ y λ_3 es igual a 1.

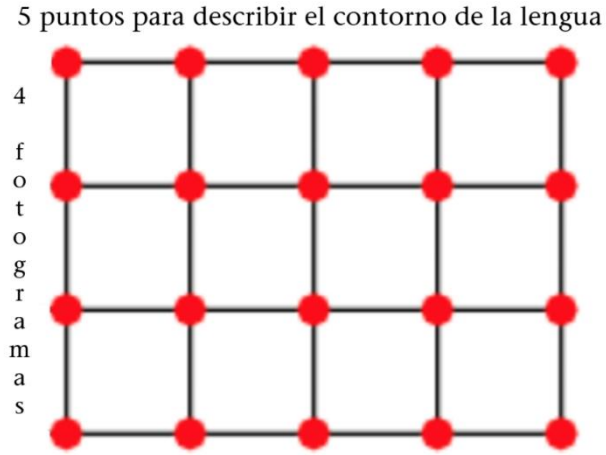
2.1.2 Grafos

La idea de esta técnica se encuentra en representar el conjunto total de puntos que describen el contorno de la lengua como vértices de un grafo tipo cuadricula. En este grafo cada fila corresponde a los n puntos de la curva para un fotograma en específico enlazados por aristas denominadas espaciales, y en cada columna se establece la unión entre un mismo punto en dos fotogramas diferentes mediante aristas temporales. Lo que se busca es etiquetar cada vértice con un vector de desplazamiento $\vec{D}$ que describa la diferencia espacial entre el iésimo punto en un fotograma y ese mismo punto en el contorno inicial.

La correcta asignación de un vector de desplazamiento a un vértice se medirá mediante la energía de datos, la cual alienta a cada punto a unirse a los bordes de la imagen, y la energía de regularización cuyo fin es asegurar que cada contorno permanezca suave y continuo y que los contornos de los fotogramas adyacentes se muevan de manera coherente.

Figura 2

Representación de la secuencia de fotogramas mediante un grafo



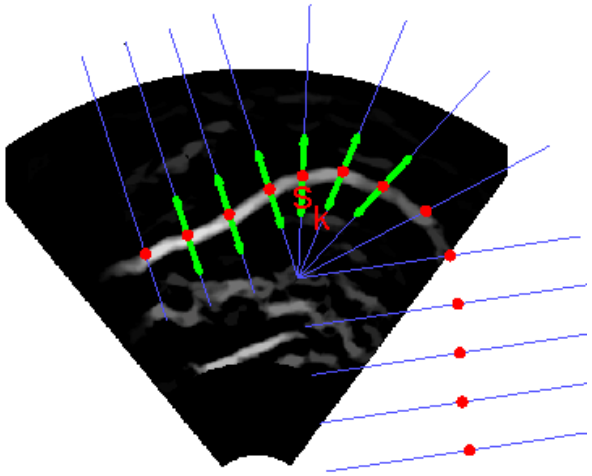
Nota. El gráfico representa el grafo de un contorno descrito por 5 puntos (número de columnas), en 4 fotogramas diferentes (número de filas).

2.1.3 Modelos de Apariencia Activa

Este método fue planteado en 2009, proponiendo una técnica novedosa y efectiva. Se basa en el trazado de líneas a lo largo del tracto vocal, para la posterior búsqueda de intersecciones de con puntos con una textura específica, en este caso la textura de la lengua, típicamente correspondiente a pixeles de una alta intensidad. Si bien la detección que se logra es precisa, los alcances de este método fueron limitados, debido a que requería información adicional, como videos de rayos X recolectados en simultaneo con el ultrasonido y estimación de la textura deseada por parte de expertos. De forma que esta propuesta resulta prácticamente obsoleta a día de hoy.

Figura 3

Demostración de la operación de los modelos de apariencia activa



Nota. Trazado de líneas a lo largo del tracto vocal. Los puntos rojos señalan las intersecciones evidentes o derivadas con pixeles de alta intensidad, correspondientes al contorno de la lengua.

2.2 Marco Referencial

El método más reciente para desempeñar la tarea de detección son las redes neuronales, una poderosa herramienta que ha probado ser de gran utilidad en múltiples labores y que con la constante mejora de las unidades de procesamiento adquiere cada vez mayor relevancia. Su funcionamiento se puede aproximar al del cerebro humano, donde por medio de un conjunto de conocimientos previos, reglas y datos se realiza un gran procesamiento con el objetivo de generar una respuesta adecuada ante determinado estimulo.

2.2.1 Método

En la literatura es posible encontrar información de múltiples tipos de redes neuronales, las seleccionadas para el desarrollo de este trabajo se denominan redes neuronales convolucionales. Este tipo de redes se ha convertido en el método predilecto para aplicaciones de visión por

computador en los últimos años, incluyendo tareas de clasificación, segmentación y generación de imágenes. El área específica en la cual se enfoca este trabajo es la de segmentación.

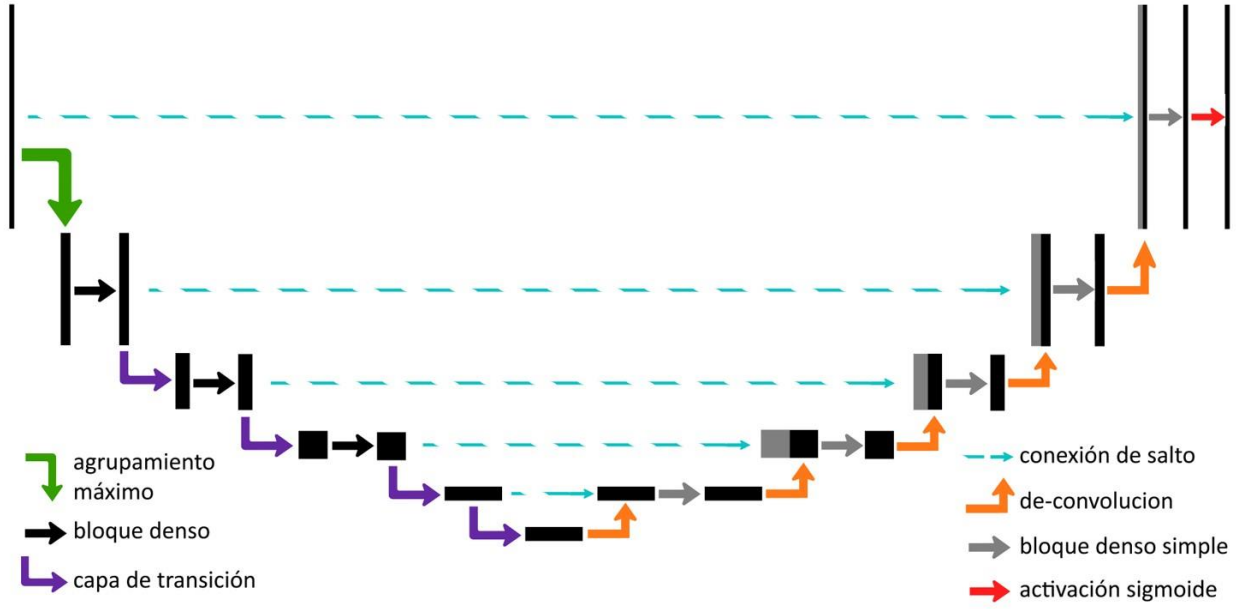
Las redes neuronales convolucionales operan de manera similar a la corteza visual del ojo humano, identificando distintas características en las imágenes que hacen posible la detección de objetos y la extracción de información valiosa. Estas redes contienen varias capas ocultas especializadas y jerarquizadas. Es decir que las primeras capas pueden detectar puntos, líneas, figuras geométricas y se van especializando hasta llegar a capas más profundas capaces de reconocer formas complejas como un rostro o una silueta.

2.2.1.1 Arquitectura de la Red. La arquitectura utilizada en este trabajo recibe el nombre Dense U-Net. Esta red inicialmente realiza la compresión de las imágenes de entrada mediante operaciones de agrupamiento y convolución, y posteriormente recupera el tamaño original de las imágenes a través de una ruta de sobremuestreo. La razón por la cual se desea recuperar el tamaño original de las imágenes es para poder realizar una clasificación de cada uno de los pixeles de acuerdo a la información que contengan, a esto se le llama segmentación.

Esta arquitectura se basa en la combinación de U-Net y Dense-Net. De U-Net se toma la estructura previamente descrita, y las conexiones de salto, operaciones que favorecen la recuperación del tamaño original de las imágenes en la ruta de subida. Por su parte DenseNet aporta el concepto de bloques densos o densamente conectados, los cuales están conformados de normalización por lotes, unidades de rectificación y convoluciones repetidas, en donde la salida de una capa se concatena en la entrada de la siguiente. El objetivo de esta concatenación es mitigar el desvanecimiento de gradiente, un problema común en redes neuronales profundas que imposibilita realizar el ajuste de los pesos de la red y por lo tanto se interrumpe el proceso de entrenamiento.

Figura 4

Arquitectura Dense U-Net



Nota. Red neuronal convolucional tipo reloj de arena, con una ruta de submuestreo y una de sobremuestreo simétricas.

2.2.1.2 **Función de pérdida.** El objetivo de esta función es evaluar el rendimiento de la red durante el entrenamiento y con base a la variación de su valor estimar cual es la versión del modelo que mejores resultados alcanza. La función empleada en este trabajo fue la función de perdida compuesta. Esta función consiste en la suma ponderada del coeficiente de similitud de Dice, el cual sirve para establecer una comparación entre dos áreas basándose en la superposición existente entre las mismas y la función de entropía cruzada estándar que evalúa la precisión en la asignación de probabilidad a cada uno de los pixeles. El valor del factor de ponderación λ (lambda) fue 5 durante la implementación.

$$L_{Compuesta} = L_{Dice} + \lambda * L_{Entropia} \quad (4)$$

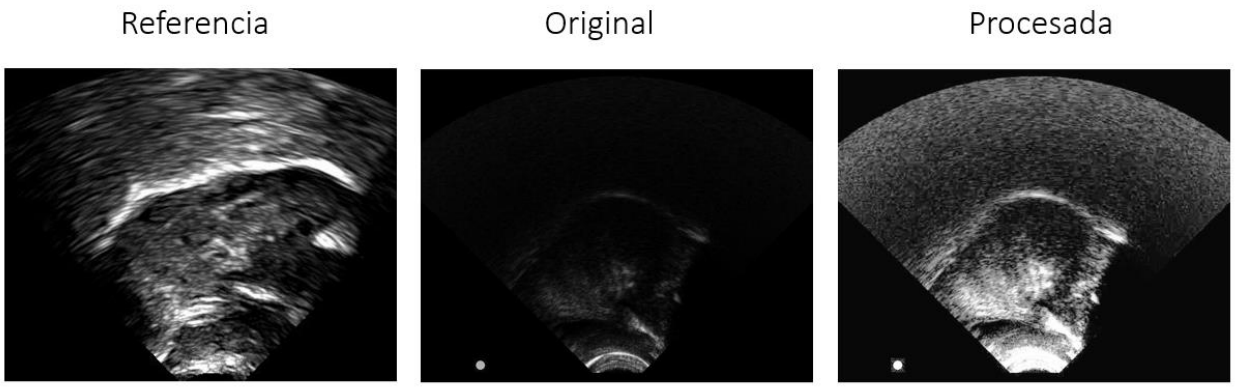
2.2.2 Base de Datos

Se contaba con una extensa base de datos conformada por más de 100000 imágenes de ultrasonido, extraídas de la grabación del ultrasonido de 30 frases cortas que fueron adquiridas por medio de un micrófono cardioide direccional. Esta base de datos incluye datos de 17 hablantes (8 hombres y 9 mujeres) de la región de Santander en Colombia, quienes informaron no tener ninguna patología del habla (Castillo, Rubio, Porras, Contreras-Ortiz, & Sepúlveda, 2019). Adicionalmente se contaba con el etiquetado del contorno de la lengua en cada fotograma realizado inicialmente por el software EdgeTrak, con posteriores correcciones manuales.

2.2.2.1 **Pre-procesamiento.** Tras realizar algunas pruebas preliminares se determinó un problema de bajo contraste en las imágenes de la base de datos, con la mayoría de pixeles concentrados en tonos muy oscuros. Para corregirlo se empleó una técnica denominada emparejamiento de histograma tomando como referencia el histograma de una imagen de ultrasonido de alta calidad.

Figura 5

Emparejamiento de histograma

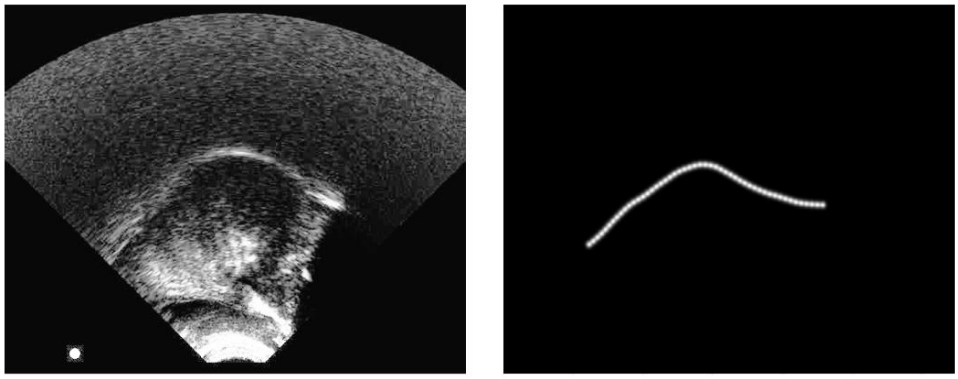


Nota. La imagen tomada como referencia fue extraída del repositorio de MTracker. Se puede apreciar un contorno mucho más visible tras el emparejamiento de los histogramas.

2.2.2 **Máscaras.** Cada contorno se describe por medio de un conjunto de 50 coordenadas (x,y), las cuales son utilizadas para la generación de máscaras o mapas de calor, las cuales serán empleadas como salida para el entrenamiento de la red.

Figura 6

Imagen de ultrasonido junto a su respectiva máscara



Nota. Se muestra a la izquierda un fotograma previamente procesado y a la derecha el contorno de referencia tomado de la base de datos.

Para la generación de esta mascar se toma cada punto como el centro de una distribución gaussiana. De esta manera, a los píxeles más cercanos a las coordenadas reales del contorno de la lengua se les asignan mayores probabilidades, mientras que todos los demás píxeles disminuyen gradualmente a medida que se alejan del contorno. Durante la implementación, la desviación estándar predeterminada fue establecida en 3.

$$I(x,y) \propto \sum_{i=1}^{n=50} exp\left(-\frac{(x-x_i)^2+(y-y_i)^2}{2\sigma^2}\right) \quad (5)$$

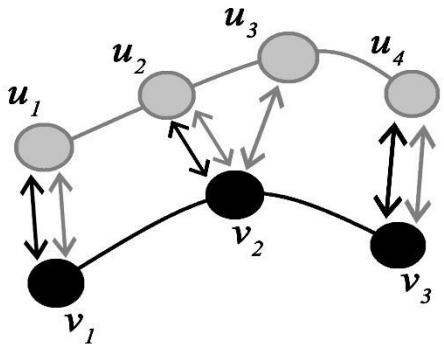
2.2.3 Resultados

Una vez fueron generados los modelos se procedió a evaluar los resultados obtenidos para el conjunto de prueba, comprendido por más de 4000 imágenes de ultrasonido.

2.2.3.1 Métrica de Evaluación. La medida seleccionada para realizar la comparación entre los contornos obtenidos en cada modelo y los contornos de referencia fue la Suma Media de la Distancia (*Mean Sum of Distance*), abreviada MSD por sus siglas en ingles. La MSD entre dos secuencias se puede calcular como la distancia promedio entre un punto dado y el punto más cercano perteneciente a la otra secuencia. Una gran ventaja de esta medida es que es aplicable incluso si el número de puntos que conforman las curvas no es el mismo. Su valor optimo es cero y podría interpretarse como una superposición total de los puntos.

Figura 7

MSD entre dos secuencias



Nota. Comparación entre dos curvas usando la MSD, incluso si el número de puntos de las curvas no es el mismo.

Cada modelo fue puesto prueba para diferentes umbrales μ, cuyo objetivo era filtrar pixeles con probabilidad por debajo de este valor. De esta manera se establece una probabilidad mínima para que un pixel sea potencialmente parte del contorno. A continuación, se presentan los resultados obtenidos por cada uno de los modelos para diferentes umbrales.

Tabla 1.

MSD y desviación estándar promedio para diferentes umbrales usando el modelo pre-entrenado

μ_{th}	**MSD (desviación estándar)**	**Porcentaje de detectados [%]**
0.9	6.163 (3.649)	64.238
0.7	3.540 (3.265)	82.793
0.5	**3.497 (3.471)**	**87.148**
0.3	3.591 (3.772)	90.861
0.1	3.968 (4.488)	96.294
0.01	4.370 (4.091)	99.752

Nota: Esta configuración logra la mejor MSD de todas, sin embargo, el porcentaje de detección no alcanza el 100% y la desviación estándar tiene un valor elevado.

Tabla 2.

MSD y desviación estándar promedio para diferentes umbrales usando el modelo re-entrenado

μ_{th}	**MSD (desviación estándar)**	**Porcentaje de detectados [%]**
0.9	**4.778 (2.704)**	**100**
0.7	4.854 (2.743)	100
0.5	4.939 (2.771)	100
0.3	4.993 (2.798)	100
0.1	5.006 (2.652)	100
0.01	11.01 (101.86)	100

Nota: Los resultados obtenidos con este modelo usando un umbral de 0.9 son buenos, detectando la totalidad de los contornos. No obstante, su MSD y su desviación estándar no son sobresalientes.

Tabla 3.

MSD y desviación estándar promedio para diferentes umbrales usando el modelo entrenado desde cero

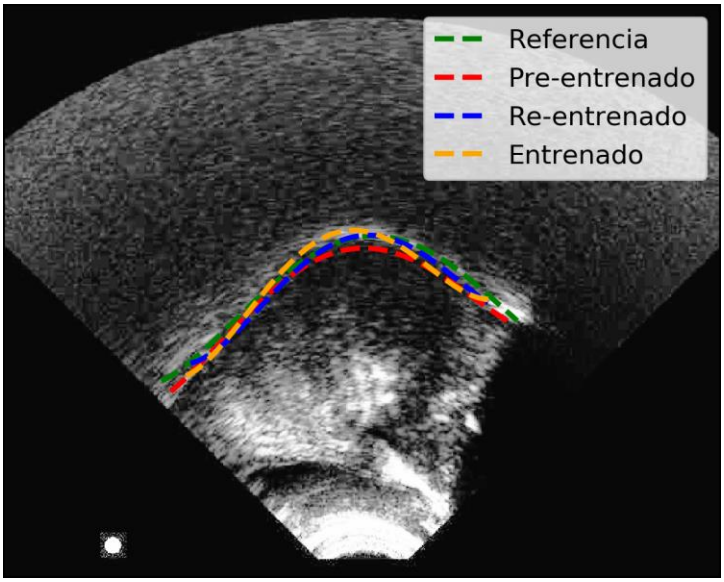
μ_{th}	**MSD (desviación estándar)**	**Porcentaje de detectados [%]**
0.9	3.978 (2.21)	100
0.7	3.895 (2.109)	100
0.5	3.838 (2.083)	100
0.3	**3.645 (1.975)**	**100**
0.1	4.054 (2.14)	100
0.01	5.125 (2.476)	100

Nota: La detección lograda por este modelo es la de mejor desempeño global, con una MSD cercana a la mínima, un porcentaje de detección del 100% y la mejor desviación estándar.

2.1.2.1 Discusión. Con base en la información presentada en las Tabla 1, se descarta el modelo pre-entrenado, debido a su bajo porcentaje de detección y su alta desviación estándar, factores que indican que este modelo es capaz de extraer los contornos de forma muy precisa en ciertos fotogramas, pero en otros la estimación obtenida es muy pobre o no llega a hacerse. Por otra parte, tras analizar la Tabla 2 se determina la desestimación del modelo re-entrenado, al no destacar en ninguna métrica. Con base en lo anterior y apoyado con los registros presentados en la Tabla 3 se justifica la selección del modelo entrenado desde cero como el de mejor desempeño, pues sus resultados son sobresalientes en todos los ítems puestos en consideración.

Figura 8

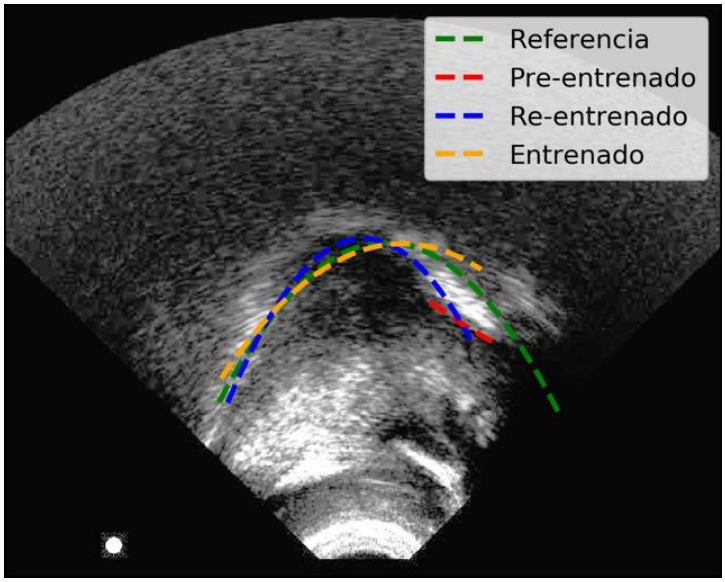
Contornos correctamente generados por cada uno de los modelos



Nota. Contornos exitosamente generados por los tres modelos empleados.

Figura 9

Contornos erróneamente extraídos



Nota. Contornos detectados de forma errónea (modelo pre-entrenado) o parcialmente correcta.

3. Conclusiones

- Los resultados demuestran la validez en el uso de las redes neuronales para el desempeño de esta tarea de forma automática, rápida y precisa.
- El modelo entrenado desde cero alcanza los mejores resultados globales, logrando un 100% de detección, una baja MSD y la menor desviación estándar.
- Una gran ventaja de este algoritmo es que permite obtener una estimación en cada fotograma de forma independiente.
- La extracción automática del contorno presentada en este trabajo puede potencialmente facilitar las anotaciones manuales que requieren mucho tiempo en la investigación fonética y clínica.

4. Recomendaciones

Una de las principales razones por la que en muchos casos se dificulta llevar a cabo trabajos o proyectos relacionados con redes neuronales o computación de alto rendimiento es por la no disponibilidad de equipos con altas prestaciones, o con las prestaciones mínimas para ejecutar los algoritmos. Con base en mi experiencia recomiendo ampliamente el uso de entornos de ejecución en la nube, tales como Jupyter o Google Colab, siendo este último el cual emplee para el desarrollo de este trabajo. Este servicio permite el uso de unidades de procesamiento grafico (GPU) por periodos de alrededor de 12 horas diarias de forma totalmente gratuita. Los tiempos de ejecución con relación a los obtenidos en mi computadora portátil con un procesador AMD A4-6210 disminuyeron a la décima parte, donde la detección de los contornos de forma local lograba una velocidad de 1 fotograma por segundo, mientras en la nube se alcanzaban a etiquetar hasta 12 fotogramas por segundo.

Referencias Bibliográficas

Akgul, Y. S., Kambhamettu, C., & Stone, M. (1999). Automatic extraction and tracking of the tongue contours. *IEEE Transactions on Medical Imaging, 18*, 1035-1045. doi:10.1109/42.811315

Castillo, M., Rubio, F., Porras, D., Contreras-Ortiz, S. H., & Sepúlveda, A. (4 de 2019). A small vocabulary database of ultrasound image sequences of vocal tract dynamics. *2019 XXII Symposium on Image, Signal Processing and Artificial Vision (STSIVA)*, (págs. 1-5). doi:10.1109/STSIVA.2019.8730224

Chen, S., Zheng, Y., Wu, C., Sheng, G., Roussel, P., & Denby, B. (2018). Direct, Near Real Time Animation of a 3D Tongue Model Using Non-Invasive Ultrasound Images. *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, (págs. 4994-4998). doi:10.1109/ICASSP.2018.8462096

Fasel, I., & Berry, J. (2010). Deep Belief Networks for Real-Time Extraction of Tongue Contours from Ultrasound During Speech. *2010 20th International Conference on Pattern Recognition*, (págs. 1493-1496). doi:10.1109/ICPR.2010.369

Gick, B., Bernhardt, B., Bacsfalvi, P., & Wilson, I. (1 de 2008). Ultrasound imaging applications in second language acquisition. doi:10.1075/sibil.36.15gic

Jaumard-Hakoun, A., Xu, K., Roussel-Ragot, P., Dreyfus, G., & Denby, B. (2016). Tongue contour extraction from ultrasound images based on deep neural network. *CoRR, abs/1605.05912*. Obtenido de http://arxiv.org/abs/1605.05912

Laporte, C., & Ménard, L. (12 de 2017). Multi-hypothesis tracking of the tongue surface in ultrasound video recordings of normal and impaired speech. *Medical Image Analysis, 44*. doi:10.1016/j.media.2017.12.003

Li, M., Kambhamettu, C., & Stone, M. (2005). Automatic contour tracking in ultrasound images. *Clinical Linguistics & Phonetics, 19*, 545-554. doi:10.1080/02699200500113616

Ohkubo, M., & Scobbie, J. (2018). Tongue Shape Dynamics in Swallowing Using Sagittal Ultrasound. *Dysphagia, 34*, 112-118.

Tang, L., & Hamarneh, G. (2010). Graph-based tracking of the tongue contour in ultrasound sequences with adaptive temporal regularization. *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*, (págs. 154-161). doi:10.1109/CVPRW.2010.5543597

Xu, K., Yang, Y., Stone, M., Jaumard-Hakoun, A., Leboullenger, C., Dreyfus, G., . . . Denby, B. (2016). Robust contour tracking in ultrasound tongue image sequences. *Clinical Linguistics & Phonetics, 30*, 313-327.

Zhou, S. K., Greenspan, H., & Shen, D. (2017). En S. K. Zhou, H. Greenspan, & D. Shen (Edits.), *Deep Learning for Medical Image Analysis* (pág. xxi). Academic Press. doi:https://doi.org/10.1016/B978-0-12-810408-8.00031-6