Biblioteca Digital
F C E N - U B A

BIBLIOTECA CENTRAL LUIS F LELOIR
BIBLIOTECA CENTRAL LELOIR
FACULTAD DE CIENCIAS EXACTAS Y NATURALES UBA

Tesis Doctoral

Modelado del sistema vocal humano y su aplicación a estudios de percepción y producción de habla

Assaneo, María Florencia

2014-09-09

Este documento forma parte de la colección de tesis doctorales y de maestría de la Biblioteca Central Dr. Luis Federico Leloir, disponible en digital.bl.fcen.uba.ar. Su utilización debe ser acompañada por la cita bibliográfica con reconocimiento de la fuente.

This document is part of the doctoral theses collection of the Central Library Dr. Luis Federico Leloir, available in digital.bl.fcen.uba.ar. It should be used accompanied by the corresponding citation acknowledging the source.

Cita tipo APA:
Assaneo, María Florencia. (2014-09-09). Modelado del sistema vocal humano y su aplicación a estudios de percepción y producción de habla. Facultad de Ciencias Exactas y Naturales. Universidad de Buenos Aires.

Cita tipo Chicago:
Assaneo, María Florencia. "Modelado del sistema vocal humano y su aplicación a estudios de percepción y producción de habla". Facultad de Ciencias Exactas y Naturales. Universidad de Buenos Aires. 2014-09-09.

EXACTAS UBA
Facultad de Ciencias Exactas y Naturales

UBA
Universidad de Buenos Aires

Dirección: Biblioteca Central Dr. Luis F. Leloir, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires. Intendente Güiraldes 2160 - C1428EGA - Tel. (++54 +11) 4789-9293

Contacto: digital@bl.fcen.uba.ar

UNIVERSIDAD DE BUENOS AIRES

Facultad de Ciencias Exactas y Naturales
Departamento de Física

Modelado del sistema vocal humano y su aplicación a estudios de percepción y producción de habla.

Tesis presentada para optar al título de Doctor de la Universidad de Buenos Aires en el área Ciencias Físicas

María Florencia Assaneo

Director del trabajo: Dr. Marcos Alberto Trevisan
Consejero de estudios: Dr. Gabriel Mindlin
Lugar de trabajo: Laboratorio de Sistemas Dinámicos, Instituto de Física de Buenos Aires - Departamento de Física, FCEyN, UBA.

Buenos Aires
9 de Septiembre del 2014

Modelado del sistema vocal humano y su aplicación a estudios de percepción y producción de habla.

María Florencia Assaneo

Buenos Aires 2014

Resumen

Desde el punto de vista biológico el proceso del habla puede separarse en dos etapas moduladas entre sí: la producción y la percepción. En este trabajo nos ocupamos de ambas, concentrándonos especialmente en la primera.

El sistema vocal humano está formado por dos grandes bloques: las cuerdas vocales y el tracto vocal. Las cuerdas vocales constituyen la fuente acústica, determinando la entonación del discurso, mientras que el contenido fonético (los sonidos propios de la lengua) es definido por la dinámica del tracto vocal. En esta tesis presentamos un modelo completo de producción vocal, incluyendo el estudio dinámico de un modelo detallado de cuerdas vocales y su adaptación a un modelo de baja dimensión del tracto vocal.

Para evaluar la calidad de la voz sintetizada con el modelo, utilizamos una combinación de test perceptuales y de resonancia magnética funcional, cuyos resultados muestran que la voz sintética es indistinguible de segmentos de voz real. Los sintetizadores basados en la física de la producción de voz permiten además el estudio de la percepción de voz controlando parámetros biológicos. En particular, en este trabajo mostramos que la identidad de la voz está codificada en términos de las dimensiones relativas entre las cuerdas vocales y el tracto vocal.

Usamos este modelo de voz verificado experimentalmente para responder preguntas de la biolingüística y la biomimética. En primer lugar, investigamos el rol de la física del aparato vocal en la formación de las onomatopeyas. A pesar de considerarse palabras vinculadas directamente con la imitación, es difícil establecer qué se preserva acústicamente entre los sonidos y sus onomatopeyas. Utilizamos el modelo vocal para mostrar que las configuraciones del tracto vocal que producen los sonidos más parecidos a los originales corresponden a consonantes co-articuladas. Estos pares vocal-consonante se corresponden, además, con las sílabas más estables de las onomatopeyas en distintos idiomas, sugiriendo un mecanismo por el cual la imitación vocal

permite asociar sonidos simples a estructuras de habla más complejas.

Por otra parte, nos preguntamos cuál es la dimensionalidad del espacio motor que gobierna la producción de habla. Para abordar este problema diseñamos un dispositivo experimental que permite monitorear tres puntos de la cavidad oral durante el discurso. Con esta herramienta, logramos una descripción discreta para las coordenadas motoras de las vocales y consonantes oclusivas del español, mostrando además la viabilidad de controlar el modelo de producción vocal con variables anatómicas para la síntesis de voz en tiempo real a partir de los gestos anatómicos producidos durante el habla.

Palabras clave
producción y percepción de voz - biolinguística - modelado matemático - dinámica nolineal - resonancia magnética funcional

Abstract

Modeling of the human vocal system and its application to studies of speech perception and production.

From a biological point of view the ability of speaking can be split in two intermodulated processes: production and perception. In this work we investigated both of them from a physical perspective, focusing on the first one.

The physical process associated with the production of voice rely on the vocal anatomy, composed of two main blocks: the vocal folds and the vocal tract. The folds are the acoustic source that specify the intonation of the speech, while the phonetic content is determined by the vocal tract dynamics. In this thesis we developed a complete model of voice production, we studied the different dynamic regimes of a detailed mathematical model of the folds, and adjusted it to a low dimensional model of the tract. This model allows to synthesize voice by controlling physical parameters of the vocal system.

In order to evaluate the quality of the synthetic voices, we carried out a combination of perceptual and fMRI tests, showing that synthetic voices are indistinguishable from real ones. Such an articulatory synthesizer, based on the physics processes involved in the voice production, allows to study the perceptual effects of precise variations in the anatomical parameters. We used it to show that the voice identity is encoded in the relative dimensions of the tract and the folds.

Using this validated model, we addressed two specific questions. First, we investigated the role of imitation within the generation of onomatopoeias. Despite it is widely know that onomatopoeias are based on imitation, it remains unclear which are the acoustic features shared between the sounds and their onomatopoeias. Using our vocal model we show that co-articulated

consonants are the sounds that best fit the original noises. This pairs of vowel-consonant also are the more stable syllables within the onomatopoeias across languages, suggesting a mechanism through which vocal imitation associates simple sounds with more complex speech structures.

We also inquire about the dimension of the vocal motor space controlling the production of speech, in order to study this problem we designed an experimental device that allows monitoring 3 points of the upper vocal tract while speaking. Making use of this novel tool, we reach a discrete description for the motor coordinates of Spanish vowels and occlusive consonants. This results show the plausibility to control the vocal model with direct anatomical measures, synthesizing speech in real time from simple motor gestures produced during the vocalization.

Keywords
speech perception ans poduction - biolinguistcs - mathemathical modelling - nonlinear dynamics - fMRI

A Marta y Guillermo por esto,
a Luciana Biazutti por aquello.

Índice general

Capítulo 1

Introducción

Son muchas las especies que utilizan su capacidad de producir sonidos para comunicarse, ya sea para transmitir una señal de alarma o como medio de cortejo. Sin embargo, el ser humano es el único que posee la capacidad de hablar, entendida como la codificación de un concepto abstracto en un conjunto de instrucciones articulatorias, que devienen en una concatenación controlada de sonidos. Esta facultad descansa en una anatomía con la complejidad necesaria para producir un repertorio variado de sonidos, una estructura neuronal capaz de ejercer un control fino sobre la periferia, y capacidades cognitivas más elevadas. A lo largo de esta tesis abordamos distintas etapas de este proceso, evidenciando las ventajas que presenta contar con una descripción matemática apropiada de la anatomía vocal.

El sistema vocal humano está formado por dos bloques: las cuerdas vocales y el tracto vocal, esquematizados en la figura 1.1. Las cuerdas vocales son un par de membranas capaces de oscilar modulando el flujo de aire proveniente de los pulmones. Más detalladamente, son dos membranas de forma cónica, constituidas por repliegues de tejido mucoso y elástico. Uno de sus márgenes se encuentra adherido a la pared laríngea. El otro, libre, permite que se forme una rendija entre las dos membranas, denominada glotis, por la que circula el aire haciéndolas vibrar. Los bordes que delimitan la glotis se encuentran engrosados formando el ligamento vocal que se extiende desde

el cartílago tiróideo hasta el aritenoides. Contienen además al músculo vocal que se encarga de acortar o relajar las cuerdas controlando propiedades acústicas de la voz durante el discurso. En su posición de reposo las cuerdas se encuentran alejadas, deben acercarse para que la fonación sea posible. La apertura o cierre de la glotis, responsable del comienzo o finalización de las vocalizaciones, se produce gracias a los músculos cricoaritenoideos (laterales y mediales) que se encargan de mover lateral o medialmente los cartílagos que sostienen las membranas. Aparte, las perturbaciones en la presión producidas en la glotis son inyectadas en el tracto vocal, un conjunto de cavidades por el que deben propagarse hasta alcanzar el exterior. El tracto vocal está formado por la cavidad laríngea, la faringe, la cavidad oral y la nasal (figura 1.1). El tracto contiene articuladores que le permiten variar notablemente su configuración. Estos son mandíbula, labios, parte blanda del paladar (velo) y lengua. Por ejemplo, subiendo el velo se bloquea la entrada de aire a la cavidad nasal restringiendo el tracto a un número menor de cavidades, o redondeando y estirando los labios podemos aumentar su longitud.

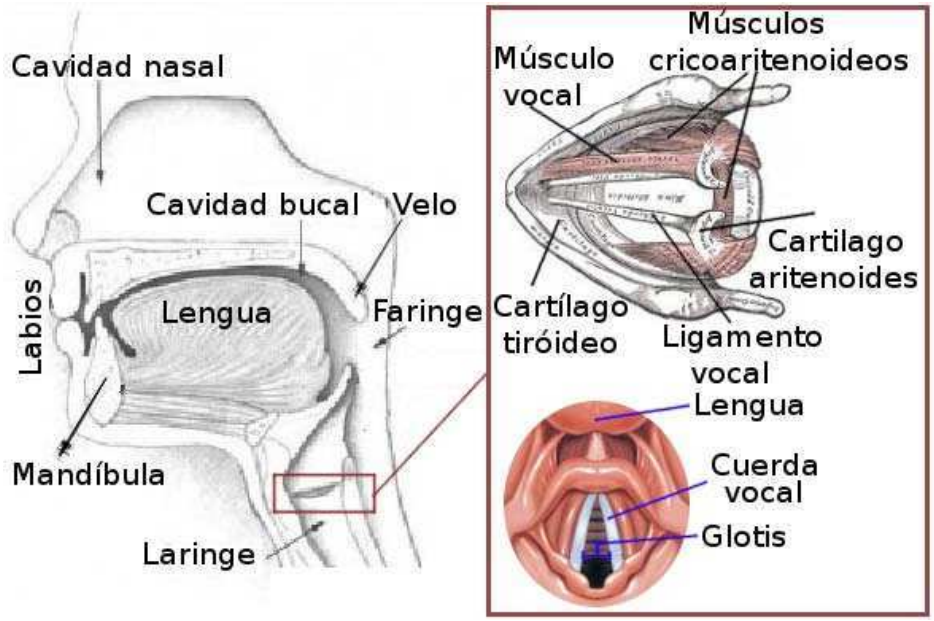

Figura 1.1: Anatomía del sistema vocal humano. Esquema del tracto vocal y de las cuerdas vocales (recuadro rosado.)

Veamos ahora como es la interacción de estos dos bloques. La presencia del tracto tiene dos efectos sobre las cuerdas: de acoplamiento, modificando su dinámica, y de filtrado del sonido emergente. El primero propone que la dinámica de las cuerdas es afectada por los rebotes en el tracto de las perturbaciones de presión generadas en tiempos posteriores [1]. Sin embargo, se sabe que durante el habla este efecto es despreciable [2, 3], y el tracto actúa tan solo como un filtro para la onda sonora generada por las cuerdas, lo que se conoce como teoría *fuente-filtro* [4, 5]. Según esta teoría, en el espacio de las frecuencias el desacople acústico resulta en que el espectro de un sonido voceado sea la multiplicación entre el espectro de la fuente (las cuerdas) y el del filtro (el tracto), la figura 1.2 explica este proceso.

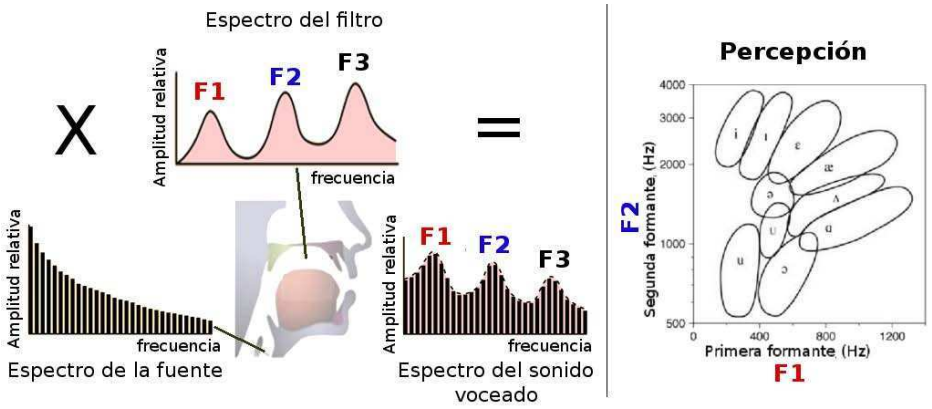

Figura 1.2: Panel derecho: Teoría fuente filtro, multiplicando el espectro de la fuente (cuerdas vocales) por el del filtro (tracto vocal) se obtiene el espectro del sonido resultante. Panel izquierdo: Regiones en el espacio (F_1, F_2) que caracterizan a las distintas vocales del inglés. Vocalizaciones de distintos sujetos incluyendo niños, mujeres y hombres son percibidas como distintas vocales de acuerdo con los valores de la primera y segunda formante [6].

Datos experimentales muestran que las oscilaciones de las cuerdas generan señales de alto contenido espectral, cuya densidad de potencia decae según: $P_s(f)\alpha f^{-1}$ [2]. Estas perturbaciones viajan a través del tracto vocal, el cual puede pensarse como un tubo no uniforme, de aproximadamente 17,5cm, caracterizado por una función de transferencia $P_t(f)$ con máximos en las fre-

cuencias resonantes F_i, comúnmente denominadas formantes. De esta forma, la densidad de potencia del espectro del sonido a la salida del tracto viene dada por: $P_s(f)P_t(f)$. La capacidad de transmitir información descansa en la plasticidad del tracto, ya que la identidad fonémica viene dada,mayormente, por $P_t(f)$. Por ejemplo, se sabe que las vocales son percibidas y clasificadas según las dos primeras resonancias del tracto [2, 6, 7], sus dos primeras formantes, como se muestra en la figura 1.2.

La anatomía vocal permite generar dos tipos de sonidos: voceados y no voceados. Los primeros, se producen a partir de las vibraciones de las cuerdas vocales. Los segundos, gracias a la posibilidad de modificar la forma del tracto vocal, generando una constricción u oclusión donde el flujo de aire se vuelve turbulento. Es precisamente esta turbulencia la que actúa de fuente sonora para los sonidos no voceados. Si bien el espacio de los sonidos de la voz es continuo, no lo es el de los que componen el habla, denominados fonemas (unidad acústica básica capaz de distinguir significado). Existen dos grandes familias de fonemas: las vocales y las consonantes. En las vocales las cuerdas están activas, el tracto no presenta constricciones y actúa como una guía de ondas para el sonido. Las distintas formas que puede adoptar el tracto modifican las resonancias de las cavidades que lo componen, y son estas resonancias las que fijan la identidad de la vocal. En el caso de las consonantes, las cuerdas pueden, o no, estar activas y presentan una constricción en algún lugar del tracto. Para identificar cada fonema, los fonetistas construyeron un alfabeto internacional (IPA del inglés *International Phonetic Alphabet*), otorgándole un símbolo a cada sonido con contenido fonémico en alguna lengua y caracterizándolo según su anatomía. Las vocales quedan descriptas por la posición de la lengua, apertura de la mandíbula y redondeado de los labios (figura 1.3). Las consonantes según si son sordas o voceadas (cuerdas inactivas y activas respectivamente), por la posición y por el grado de la constricción (figura 1.3). Las columnas, representan el sitio donde se genera la constricción. Las filas el grado de la misma, que determina a la vez el *tipo de fuente* sonora, por ejemplo: las fricativas presentan constricción capaz de generar una turbulencia que actúa como una fuente ruidosa soste-

nida en el tiempo, mientras que en las plosivas se produce una oclusión y su posterior liberación, lo que genera un silencio previo a un aumento abrupto en la intensidad.

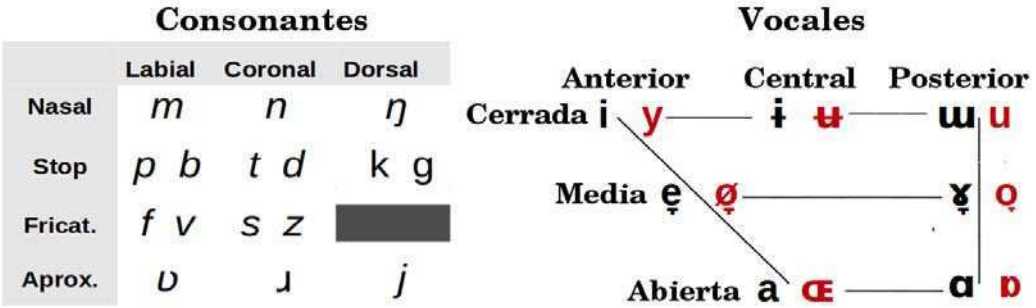

Figura 1.3: Representación de fonemas según el alfabeto fonético internacional. Consonantes: Los sitios que contienen dos fonemas corresponden a consonantes sordas (izquierda) y voceadas (derecha). Las columnas representan el sitio de la constricción, mientras que las filas están relacionadas con el tipo de fuente o número y grado de constricción. Nasales: son voceadas, el velo está *abierto* y el sonido se transmite por la cavidad nasal. Fricativas: la fuente es una turbulencia generada en una constricción. Oclusivas o plosivas: las caracteriza una oclusión y su posterior liberación que genera un crecimiento abrupto en la presión acústica (conocido como *ataque*) y una posterior turbulencia. Aproximantes: Presentan una constricción en el tracto pero no lo suficientemente estrecha como para generar una turbulencia audible, son voceadas. Vocales: pueden ser descriptas en un espacio bidimensional, donde cada sitio tiene dos fonemas, según si se pronuncian, o no, con los labios redondeados (en rojo, redondeados). La dimensión vertical se corresponde con la posición de la lengua respecto del paladar. Mientras que la horizontal describe la ubicación de la lengua en el eje dorsoventral (con respecto a la región posterior de la boca).

En esta tesis estudiamos el modelado matemático de la anatomía vocal, para investigar como repercuten las restricciones de la anatomía en las distintas etapas del proceso de producción y percepción del habla. Desde este enfoque investigamos la percepción de la voz, la presencia de imitación dentro del lenguaje y los gestos motores que caracterizan los distintos fonemas. Estos distintos puntos se encuentran detallados a lo largo de este trabajo de la siguiente manera:

En el capítulo 2 describimos un sintetizador articulatorio, a través del modelado matemático del sistema vocal que permite generar habla a partir de variaciones en distintos parámetros anatómicos. Para esto, describimos y estudiamos distintos modelos para los dos bloques que constituyen el sistema vocal: las cuerdas y el tracto. Investigamos como repercute en la síntesis de sonido el nivel de detalle en el modelado, arribando a un sintetizador de baja dimensión, que presenta un buen compromiso entre la descripción del modelo y la calidad de la síntesis. Los resultados de este capítulo se pueden encontrar publicados en [8, 9]

En el capítulo 3 validamos nuestro modelo, comparando la percepción de las voces sintéticas con la de voces reales. En la literatura hay antecedentes de tests sobre habla sintética, acotados a su comprensibilidad, no investigan la *calidad vocal*. Nosotros mostramos que nuestras síntesis son indistinguibles de voces reales, tanto a nivel comportamental como en la actividad cerebral que generan. Por otro lado, típicamente la percepción de la voz se estudia realizando variaciones acústicas arbitrarias sobre grabaciones experimentales. En contraste con este enfoque clásico, una vez testeado nuestro sintetizador, lo utilizamos para explorar la codificación neuronal de la voz a partir de variaciones de parámetros anatómicos. Mostrando que contar con un sintetizador articulatorio testeado, constituye una herramienta novedosa y relevante para el estudio de la percepción de la voz en un espacio fisiológico en vez de acústico.

Las reglas de formación de las palabras han sido clásicamente, estudiadas por la lingüística. Sin embargo, recientemente se ha comenzado a estudiar el alcance de los mecanismos básicos de la biología, como la imitación, en este campo [10]. Los objetos naturales para estudiar los efectos de la imitación en el habla son las onomatopeyas, que transforman sonidos de la naturaleza en palabras, a través de la imitación. Sin embargo, las diferencias acústicas entre los sonidos originales y sus onomatopeyas son notables. En el capítulo 4 estudiamos este problema combinando el modelado físico del sistema vocal

con experimentos perceptuales, investigando cómo la física de la producción vocal puede modular el proceso mimético. Esta investigación se encuentra publicada en [11].

Como mencionamos anteriormente, los sonidos que componen el habla, a diferencia de los de la voz en general, son discretos. El espacio fonémico está constituido por un número finito de elementos, como se muestra en la figura 1.3. Por otro lado, trabajos recientes muestran que la codificación neuronal que controla la ejecución de los distintos fonemas es también discreta. Es decir, que existen poblaciones neuronales que se activan selectivamente a la articulación de distintos fonemas [12, 13]. Si bien estos dos extremos en el proceso del habla son discretos, las configuraciones de tracto vocal ocurren en un espacio continuo. Nuestra hipótesis, siguiendo a Goldstein et. al. [14, 15], es que esta información discreta se encuentra codificada en el tracto vocal. En el capítulo 5 estudiamos la dinámica de los articuladores del tracto vocal superior durante el discurso, buscando recuperar la información discreta a partir de la dinámica continua del tracto. Para esto, diseñamos un dispositivo experimental que permite monitorear el movimiento de la mandíbula, labios y lengua durante el discurso. Encontramos una descripción para las vocales y las consonantes oclusivas del español en un espacio discreto de coordenadas motoras de estos articuladores. Parte de este trabajo ha sido publicado en [16].

Capítulo 2

Modelado

El desarrollo y estudio de un modelo detallado del sistema fonador humano es un problema que ha sido abordado desde distintas áreas de la ciencia, debido a las diversas aplicaciones que presenta. Algunas de ellas son: aplicaciones bioprostéticas [17], mejoras al diagnóstico de patologías [18] y el desarrollo de un sintetizador de habla articulatorio de calidad [19]. Debido a esto, a lo largo de las últimas décadas, una serie de investigaciones se concentraron en lograr una descripción matemática adecuada de las cuerdas vocales y del tracto vocal, los dos grandes bloques que forman el aparato vocal. En este capítulo describimos y estudiamos algunas de ellas.

Como mencionamos en el capítulo anterior, las cuerdas vocales son un par de membranas alojadas en la laringe capaces de oscilar generando sonido, mientras que el tracto es la cavidad que abarca desde la salida de la laringe hasta los labios. Durante el habla, es válida la teoría *fuente-filtro* [4, 5], lo que permite modelar cada uno de estos bloques constitutivos por separado.

2.1. Fuente: Cuerdas vocales.

Recapitulando un poco la historia del modelado de las cuerdas, el trabajo fundacional es el de Ishizaka y Flanagan [20] del año 1971, donde aproximan cada cuerda vocal como un sistema de dos osciladores amortiguados aco-

plados, sentando las bases de lo que hoy se conoce como *el modelo de dos masas*. Este modelo se basa en datos experimentales [21, 22] que muestra que las cuerdas no se mueven como un bloque, sino que existe una onda transversal propagándose en el tejido. Esto produce un desfasaje en el desplazamiento de los márgenes superior e inferior de las membranas, y es el modelo de dos masas el sistema de osciladores más simple que puede dar cuenta de este efecto. Debido a la falta de herramientas computacionales que permita su tratamiento, aquel primer modelo cuadri-dimensional permaneció inexplorado por algunas décadas. Los primeros estudios analíticos del problema se concentraron en una aproximación denominada *el modelo de flameo*, una aproximación bidimensional que se basa en asumir la onda transversal propagándose a lo largo de las cuerdas [4, 23]. Este modelo simple reproduce las características esenciales de la producción de voz como las oscilaciones auto-sostenidas de las cuerdas y el perfil de los pulsos glotales, es por eso que los primeros estudios se realizaron sobre esta simplificación del problema.

En los últimos años, con el desarrollo de nuevas herramientas computacionales, el modelo cuadri-dimensional de dos masas volvió a cobrar protagonismo. Incluyendo en el mismo: una descripción más detallada del flujo glotal, que permite un tratamiento apropiado de la fuerza hidrodinámica [24], y las colisiones de las cuerdas vocales [25, 26].

2.1.1. Modelo de flameo

Este modelo simplificado permite dar cuenta de una transferencia neta de la energía cinética del flujo de aire al movimiento de las cuerdas [4]. Las membranas que forman las cuerdas presentan dos modos de vibración: una onda que se propaga transversalmente hacia arriba, y una oscilación lateral de su punto medio en torno a la posición de equilibrio. Datos obtenidos mediante laringoscopía [2] sugieren que durante la fonación estos modos se coordinan, de manera de que exista una ganancia neta de energía en cada

ciclo, proveniente del flujo de aire. El perfil es divergente cuando las cuerdas se acercan y convergente cuando se alejan, como muestra en el panel superior de la figura 2.1. La presión intraglotal es diferente en las distintas partes del ciclo, lo que permite la ganancia de energía. Si bien la ley de Bernoulli no es válida durante todo el proceso, sirve para predecir la desigualdad en las presiones: la presión disminuye en un tubo que se angosta y aumenta en uno que se ensancha. Como la presión a la salida es cero (desacople con el tracto) la presión intraglotal resulta mayor en un perfil que en el otro.

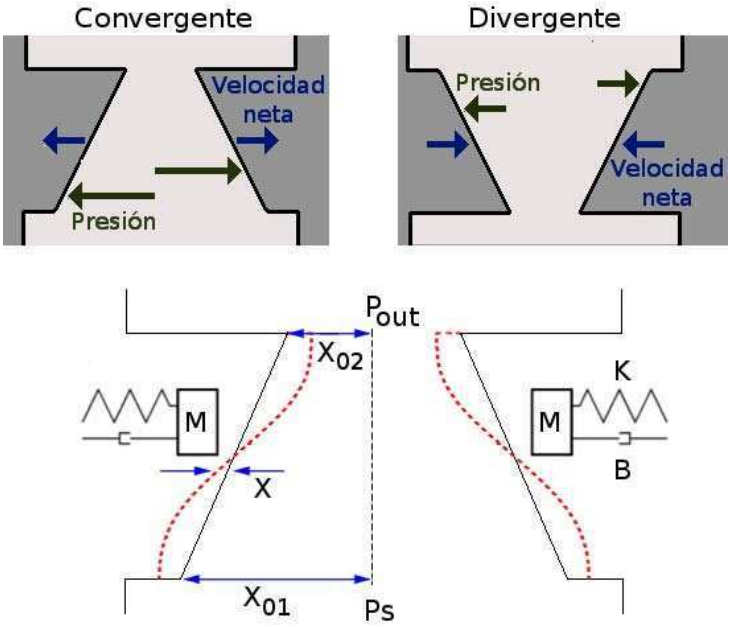

Figura 2.1: Esquema del modelo de flameo. Panel superior: Perfil convergente y divergente de la glotis, en azul la dirección de la velocidad del desplazamiento horizontal de las cuerdas, y en verde se representa la intensidad de la presión intraglotal, en cada etapa del ciclo. Panel inferior: Aproximación de flameo, cada membrana es aproximada por una oscilador amortiguado de masa m, acoplado a la pared de la laringe por K, con una disipación B. El desplazamiento del punto medio de las cuerdas, medido desde la posición de equilibrio, viene dado por x, mientras que x_{01} y x_{02} representan la posición del margen inferior y superior, respectivamente, para el perfil de equilibrio de la glotis. P_s representa la presión subglotal, y P_{out} la correspondiente a la entrada del tracto.

La forma matemática de escribir una onda transversal propagándose a lo largo de las cuerdas, es planteando que el desplazamiento del margen superior está retrasado 2τ con respecto al inferior. Entonces, la sección transversal del área a la entrada y salida de la glotis (a_1 y a_2) puede ser aproximada por:

$$\begin{cases} a_1 = 2l_g(x_0 + x + \tau\dot{x}) \\ a_2 = 2l_g(x_0 + x - \tau\dot{x}) \end{cases} \tag{2.1}$$

donde x es el desplazamiento del punto medio de las cuerdas a partir del equilibrio, x_0, y τ es la mitad del tiempo que le toma a la onda transversal viajar del margen inferior al superior, y l_g el largo de las membranas en la dirección perpendicular al plano de la figura 2.1. De esta forma se reduce la dimensión del problema, ya que alcanza con estudiar la dinámica del punto medio de las cuerdas. El punto medio se modela como un oscilador amortiguado de masa efectiva m [4, 5], como se muestra en la figura 2.1, y su ecuación de movimiento es:

$$m\ddot{x} = -k\,x - \beta\,\dot{x} + a_g p_s \frac{\Delta + 2\tau\dot{x}}{a_{01} + x + \tau\dot{x}}, \tag{2.2}$$

donde $k = k(x) = k_1 + k_2x^2$ y $\beta = \beta(x, \dot{x}) = \beta_1 + \beta_2x^2 + \beta_3\dot{x}^2$ representan las características restitutiva y disipativa del tejido, respectivamente. Utilizando una versión modificada fenomenológicamente de las ecuaciones de Bernoulli [4], la presión puede escribirse según: $p_g = p_s(1 - a_2/a_1)$ con p_s la presión subglotal. Esta corrección, conjunto la ecuación 2.1, dan origen al último término que representa la presión intraglotal, siendo a_g el área de las cuerdas.

Finalmente, de acuerdo con [1,27], las perturbaciones de presión acústica que se propagan por el tracto se relacionan con el movimiento de las membranas según: $p_i(x, \dot{x}) = \sqrt{p_s/\rho}\,x$, siendo ρ la densidad del aire.

2.1.2. Modelo de dos masas

En esta sección estudiamos el modelo completo de dos masas descripto por Lucero y Koening en [28]. Elegimos este modelo porque presenta un buen compromiso entre simplicidad matemática y nivel de descripción de los fenómenos físicos presentes en la dinámica de las cuerdas vocales.

En el modelo de dos masas cada una de las cuerdas vocales se representa con dos osciladores amortiguados acoplados, según se muestra en la figura 2.2.

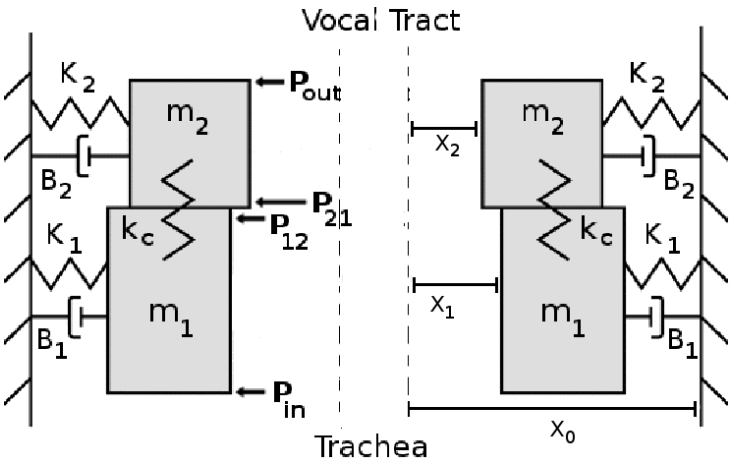

Figura 2.2: Esquema del modelo de dos masas: Cada cuerda es representada por las masas m_1 y m_2 acopladas entre sí mediante una fuerza elástica k_c y a las paredes de la laringe por K_1 y K_2, con sus respectivas pérdidas B_1 y B_2. El desplazamiento de cada masa a partir de la posición de equilibrio x_0 viene dada por x_1 y x_2. P representa la presión aerodinámica actuando en los distintos puntos de las cuerdas.

Adoptando esta descripción, si se asume simetría con respecto al plano sagital, las masas de la izquierda y de la derecha son idénticas y las ecuaciones de movimiento de cada una viene dada por:

$$\begin{aligned} \dot{x_i} &= y_i \\ \dot{y_i} &= \frac{1}{m_i}\left[f_i - K_i(x_i) - B_i(x_i, y_i) - k_c(x_i - x_j)\right] \end{aligned} \tag{2.3}$$

donde $i, j = 1$ o 2 para la masa inferior y superior respectivamente, K y

B son funciones que representan las características restitutivas y disipativas del tejido de las cuerdas, m la masa, k_c el coeficiente de acoplamiento, y f la fuerza hidrodinámica ejercida por la presión intraglotal sobre las paredes de las membranas. El desplazamiento horizontal de las masas a partir del reposo, x_0, es representado por x.

Las formas funcionales de B y K son las descriptas en [20,28], modificadas para que sean derivables:

$$K_i(x_i) = k_i x_i(1 + 100{x_i}^2) + \Theta\left(\frac{x_i + x_0}{x_0}\right) 3k_i(x_i + x_0)[1 + 500(x_i + x_0)^2] \tag{2.4}$$

$$B_i(x_i) = \left[1 + \Theta\left(\frac{x_i + x_0}{x_0}\right)\frac{1}{\epsilon_i}\right] r_i(1 + 850{x_i}^2)y_i \tag{2.5}$$

$$\text{con } \Theta(x) = \begin{cases} 0 & \text{si } x \leq 0 \\ \frac{x^2}{8\ 10^{-4}+x^2} & \text{si } x > 0 \end{cases} \tag{2.6}$$

donde $r_i = 2\epsilon_i\sqrt{k_i m_i}$, y ϵ_i el coeficiente de amortiguamiento. Los términos con la función escalón Θ modelan el incremento en la restitución y la disipación durante la colisión de las cuerdas.

Para escribir la fuerza hidrodinámica (f) que se ejerce sobre las cuerdas, necesitamos conocer la presión en los distintos puntos de la glotis. De acuerdo con datos experimentales, se adoptan para modelar la presión las ecuaciones de Bernoulli, incluyendo pérdidas por viscosidad y el *modelo de capa límite*. Este modelo, descripto en [17,24,28], asume un flujo de aire unidimensional, cuasiestacionario e incompresible desde la traquea hasta un *punto de separación* donde el flujo se separa de la superficie del tejido, disipando energía en forma de turbulencia. Experimentalmente, se muestra que el sitio donde

se genera esta turbulencia depende del *grado de divergencia* del perfil de la glotis [29] (entendiendo que el perfil es divergente si $x_1 < x_2$). En el modelo el *punto de separación* se desplaza desde la salida de la glotis a la interfase entre las dos masas m_1 y m_2 cuando el perfil glotal supera el nivel de divergencia: $a_2 > k_s a_1$.

De esta forma, las ecuaciones para la presión en distintos puntos de la glotis, de acuerdo con la nomenclatura de la figura 2.2, vienen dadas por:

$$P_{in} = P_s + \frac{\rho u_g^2}{2a_1^2}, \tag{2.7}$$

$$P_{12} = P_{in} - \frac{12\mu u_g d_1 l_g^2}{a_1^3}, \tag{2.8}$$

$$P_{21} = \begin{cases} \frac{12\mu u_g d_2 l_g^2}{a_2^3} + P_{out} & \text{si } a_2 > k_s a_1 \\ 0 & \text{si } a_2 \leq k_s a_1, \end{cases} \tag{2.9}$$

$$P_{out} = 0. \tag{2.10}$$

donde P_s representa la presión subglotal (controlada activamente por los pulmones), l_g es el ancho de las cuerdas (dimensión en el plano normal a la figura 2.2), d_1 y d_2 representan el largo de la masa inferior y superior respectivamente, la sección transversal de la glotis es $a_i = 2l_g(x_i + x_0)$, μ y ρ son los coeficientes de viscosidad y densidad del aire, u_g es el flujo de aire dentro de la glotis, y $k_s = 1{,}2$ un coeficiente hallado experimentalmente [19]. Además, se asume que no hay perdidas a la entrada de la glotis (ecuación 2.7), y presión cero a la entrada del tracto vocal (ecuación 2.10).

Finalmente, según [20,28,30] la fuerza hidrodinámica actuando sobre cada masa puede escribirse como:

$$f_1 = \begin{cases} d_1 l_g P_s & \text{si } x_1 \leq -x_0 \text{ o } x_2 \leq -x_0 \\ \frac{P_{in}+P_{12}}{2} & \text{en otro caso} \end{cases} \tag{2.11}$$

$$f_2 = \begin{cases} d_2 l_g P_s & \text{si } x_1 > -x_0 \text{ y } x_2 \leq -x_0 \\ 0 & \text{si } x_1 \leq -x_0 \\ \frac{P_{21}+P_{out}}{2} & \text{en otro caso} \end{cases} \tag{2.12}$$

Estas ecuaciones incluyen la apertura, clausura parcial y clausura total de la glotis, es decir que modelan el choque de las membranas.

Diagrama de bifurcaciones

Si bien este modelo fue utilizado en distintos trabajos [19, 28, 31], para generar pulsos glotales compatibles con datos experimentales, no han sido estudiados, hasta el momento, los distintos regímenes a los que el sistema puede acceder. En esta sección, realizamos un diagrama de bifurcaciones explorando los distintos regímenes de oscilación de las cuerdas, variando parámetros relevantes dentro de rangos fisiológicamente compatibles. Los resultados que se muestran a continuación se encuentran publicados en [9].

Lo primero que se debe determinar, a la hora de construir un diagrama de bifurcaciones, es el espacio de los parámetros adecuado para estudiar el problema. En este caso, es de interés analizar la dinámica de las cuerdas en términos de la presión subglotal y la tensión de las cuerdas, ya que son parámetros que pueden variarse activamente durante la producción de habla. Como mencionamos en el capítulo 1, la presión es controlada por el flujo de aire proveniente de los pulmones, y la tensión por la acción del músculo vocal. Modificar la tensión en las cuerdas significa alterar tanto su elasticidad como su masa efectiva. Por esto, la actividad muscular es modelada por un factor Q [20] que escala propiedades mecánicas de las cuerdas según: $k_c = Qk_{c0}$, $k_i = Qk_{i0}$ y $m_i = \frac{m_{i0}}{Q}$.

Realizamos un diagrama de bifurcaciones del modelo descripto anteriormente en el espacio $(Q; P_s)$, moviéndonos en un rango que incluye los parámetros de fonación normal: $\sim (1; 800)$. De esta forma investigamos los distintos

regímenes a los que puede acceder la dinámica del sistema, en un rango de parámetros fisiológicamente relevante.

En la figura 2.3 identificamos 5 zonas, que quedan delimitadas por las

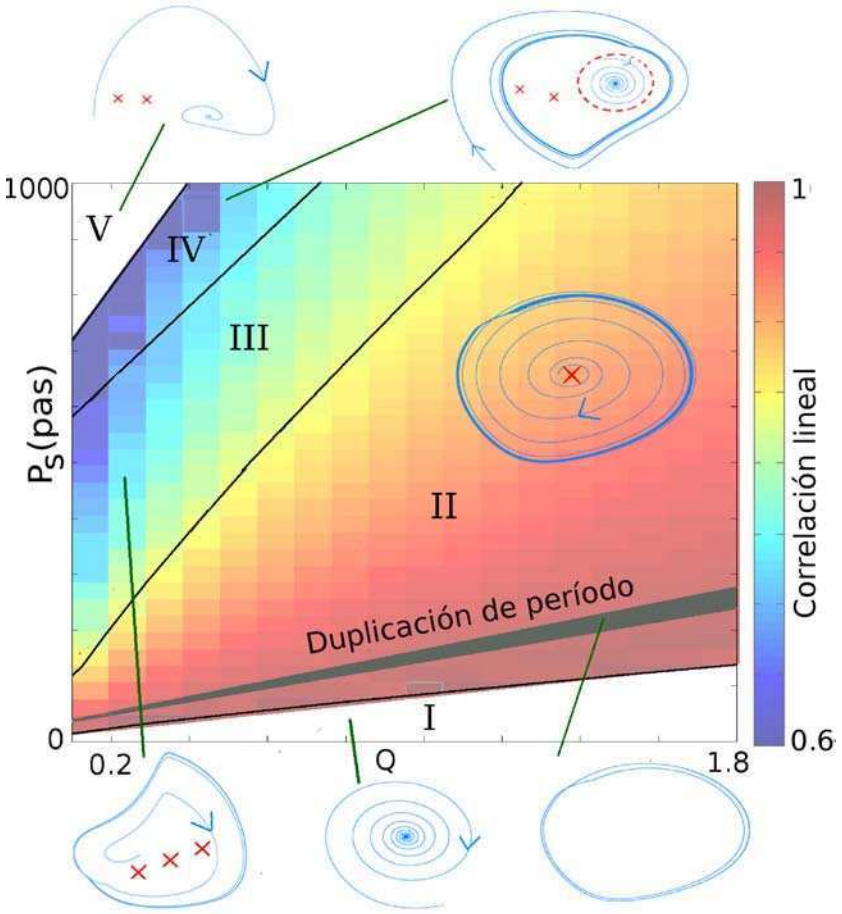

Figura 2.3: Diagrama de bifurcaciones en el plano de presión subglotal versus tensión de las cuerdas (P_s, Q), y proyecciones bidimensionales del flujo en el plano (v_1,x_1). Las cruces rojas representan puntos fijos inestables y las líneas punteadas ciclos límites inestables. La fonación normal ocurre en $(Q, P_s) \sim (1, 800)$. El código de colores representa el valor de la correlación lineal entre $(x_1 - x_2)$ y $(y_1 + y_2)$, yendo de rojo oscuro para $R = 1$ a azul oscuro $R = 0{,}6$. Este diagrama se confeccionó utilizando el programa *AUTO continuation software* [32]. Los parámetros del modelo se fijaron en: $m_1 = 0{,}125g$, $m_2 = 0{,}025g$, $k_{10} = 80N/m$, $k_{20} = 8N/m$, $k_c = 25N/m$, $\epsilon_1 = 0{,}1$, $\epsilon_2 = 0{,}6$, $l_g = 1{,}4cm$, $d_1 = 0{,}25cm$, $d_2 = 0{,}05cm$ y $x_0 = 0{,}02cm$, de acuerdo con [28].

bifurcaciones que sufre el sistema. Para valores bajos de presión (región I) el sistema presenta un punto fijo estable, y no hay oscilaciones de las cuerdas. En la región II el punto fijo cambia su estabilidad y aparece un ciclo límite estable. Si bien no se alcanza a ver en la figura 2.3, en el límite de estas dos regiones ocurren tres bifurcaciones en un rango acotado de valores de presión, como se detalla en el panel izquierdo figura 2.4.

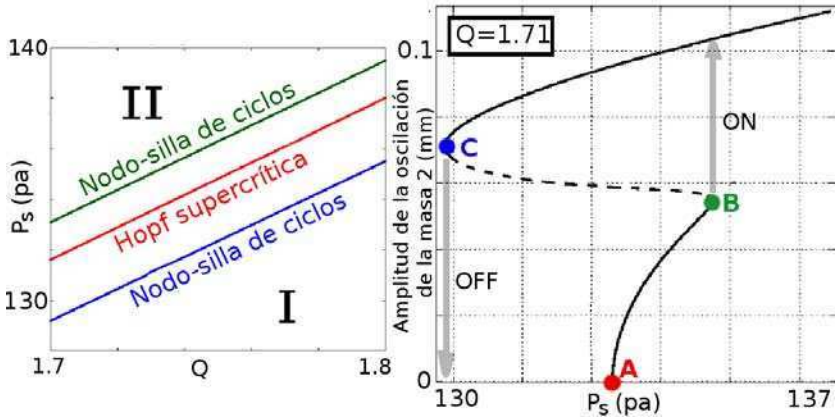

Figura 2.4: Fenómeno de histéresis presente en el proceso de aparición y desaparición de las oscilaciones. Panel izquierdo: zoom del limite entre las regiones I y II. Las línea azul y la verde representan bifurcaciones del tipo pliegue de ciclos límites (bifurcación de nodo-silla en el mapa), mientras que la roja una Hopf supercrítica. Panel izquierdo: amplitud de la oscilación de la masa superior (x_2) en función de la presión subglotal P_s, para un valor de $Q = 1{,}71$. La continuación de las soluciones periódicas fue realizada con el paquete de software AUTO [32].

El panel derecho de la figura 2.4 describe el comportamiento del sistema para un valor fijo de Q. Más precisamente, muestra como varía la amplitud de la oscilación de la masa superior, x_2, con la presión. En el punto **A** se genera un ciclo límite en una bifurcación de Hopf supercrítica [33]. La amplitud de la oscilación aumenta con la presión hasta el punto **B**, donde el ciclo estable se aniquila con uno inestable en una bifurcación nodo-silla de ciclos y la amplitud de la oscilación salta a la rama superior. Una vez en esta rama, si la presión decrece, la oscilación continúa hasta el punto **C** (correspondiente a

un valor de P_s menor que el del punto **A**), donde el sistema retorna abruptamente al reposo. Este fenómeno de histéresis presente en el arranque y el final de las oscilaciones fue reportado experimentalmente [7].
La existencia de la rama **AB** depende de la viscosidad del aire, μ. Al disminuir el valor de μ los puntos **A** y **B** se aproximan hasta que colisionan para $\mu = 0$, recuperando el resultado reportado en [25, 28]. En esos trabajos no se incluyen pérdidas por viscosidad, y las oscilaciones se originan como una combinación de una bifurcación de Hopf subcrítica y un pliegue de ciclos. Por otro lado, la rama **BC** depende del punto de separación donde el flujo se vuelve turbulento. Más específicamente, al aumentar k_s, el punto de separación se mueve hacia la entrada de la glotis y los puntos **C** y **B** se acercan hasta colapsar. En este caso, las oscilaciones nacen en una bifurcación de Hopf supercrítica y desaparece el fenómeno de histéresis, al igual que sucede en la aproximación de flameo [34].

Volviendo al diagrama de bifurcaciones de la figura 2.3, las regiones II y III están separadas por una bifurcación del tipo repulsor-silla. Si bien esta bifurcación no representa un cambio cualitativo en la dinámica del sistema (ya que no se generan ni desaparecen puntos fijos estables), su efecto puede ser relevante si se quiere estudiar el mecanismo completo de producción vocal.
Repasemos rápidamente como se generan los sonidos voceados: las modulaciones en el flujo de aire producidas por las oscilaciones de las cuerdas son inyectadas en el tracto vocal. Luego, la perturbación en el flujo se propaga a lo largo del tracto que actúa como un filtro para la señal original, enfatizando las frecuencias de la fuente cercanas a las resonancias propias del tracto. Los sonidos voceados son percibidos y clasificados de acuerdo con estas resonancias, que son las que determina su entidad fonémica [2]. Por lo tanto, un aspecto clave en la generación de voz es que la fuente sea capaz de emitir una señal espectralmente rica, siendo así capaz de relevar con precisión las resonancias del tracto vocal.
Curiosamente, la fonación normal ocurre en una región cercana a la bifurcación repulsor-silla. Si bien, como se dijo, no se altera el régimen dinámico de las cuerdas, sí se altera la forma de las oscilaciones. Observamos que el

ciclo límite se acerca a la variedad estable del punto silla, como se muestra en la figura 2.5. A pesar de que la deformación es suave y ocurre en una región limitada alrededor de la interfaz entre las regiones I y II, podría ser un mecanismo para aumentar el contenido espectral de la fuente, un aspecto clave en la producción de voz. Para cuantificar este resultado utilizamos el índice de contenido espectral, un indicador de la riqueza espectral de una señal, definido por: $\frac{\sum_{i=0}^{\infty}|S(f_i)|f_i}{\sum_{i=0}^{\infty}|S(f_i)|f_0}$, donde $S(f)$ es el espectro complejo y f_0 es la frecuencia fundamental. En el panel superior de la figura 2.5, mostramos la evolución de este índice para $x_1(t)$ en función de P_s para un valor fijo de Q. Se puede ver que, efectivamente, el valor del índice aumenta cuando el sistema se acerca a la bifurcación repulsor-silla. Con este modelo, el flujo glotal para valores de fonación normal, $(1; 800)$, tiene la forma que se muestra en el panel derecho de la figura 2.5.

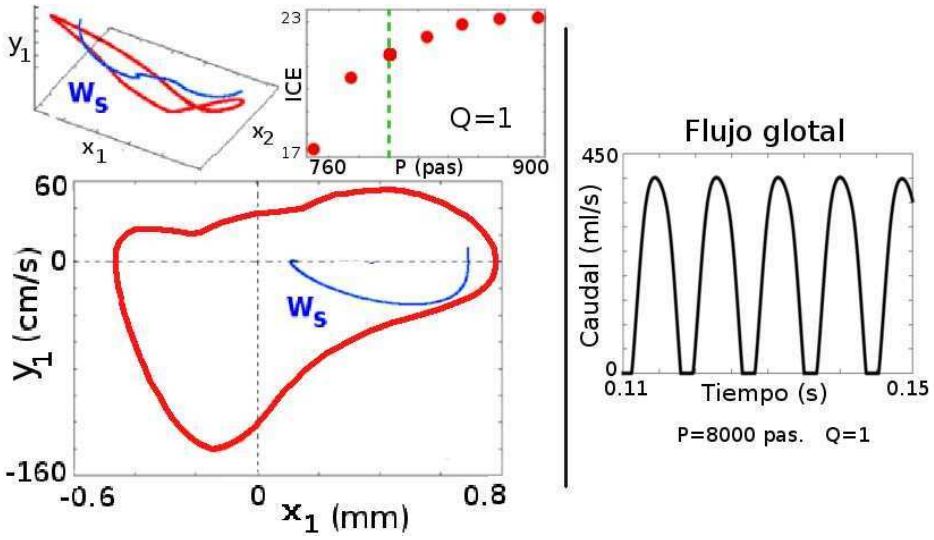

Figura 2.5: Contenido espectral de las oscilaciones para el régimen de habla normal. Proyección del ciclo límite en $(x_1; v_1)$ (rojo) conjunto la variedad estable del punto silla (azul). Los parámetros se fijaron en valores consistentes con condiciones normales de fonación, $(Q, P_s) = (1, 850)$ contenidos en la región III. Panel superior izquierdo: proyección en el espacio tridimensional (y_1, x_1,x_2). Panel superior derecho: Índice de contenido espectral para $x_1(t)$ como función de P_s para un valor fijo de $Q = 1$. Se indica en verde el valor donde ocurre la bifurcación repulsor-silla. En el panel derecho se muestra el flujo glotal para valores de fonación normal: $(P_s; Q) = (800; 1)$

Volviendo a la figura 2.3, el pasaje de la región III a la IV ocurre cuando uno de los puntos inestables generados en la bifurcación repulsor-silla cambia su estabilidad y se genera un ciclo límite inestable, a través de una bifurcación de Hopf supercrítica [33]. Finalmente, al ingresar en la región V colisionan y desaparecen los ciclos con estabilidades opuestas en un pliegue de ciclos, en esta región no existen oscilaciones en el sistema.

Además de estudiar la dinámica del sistema, investigamos las diferencias entre las soluciones del modelo completo y las correspondientes a la aproximación del *modelo de flameo* descripta en la sección 2.1. La figura 2.3 muestra un mapa de colores que cuantifica la diferencia entre las soluciones de estos dos modelos. La ecuación que vincula las áreas en los extremos de las cuerdas en el modelo de flameo, 2.1, se puede reescribir como: $(x_1 - x_2) = \tau(y_1 + y_2)$, de forma que las variables: $(x_1 - x_2)$ y $(y_1 + y_2)$, guardan una relación lineal. Teniendo esto en cuenta, para cuantificar la diferencia entre los modelos, calculamos el coeficiente de correlación lineal entre $(x_1 - x_2)$ y $(y_1 + y_2)$, para las soluciones del modelo de dos masas. El mapa de colores de la figura 2.3 corresponde al valor de este coeficiente de correlación R. Encontramos que la correlación disminuye al aumentar P_s o disminuyendo Q, mientras que en la zona de fonación normal la aproximación de flameo es relativamente buena con $R \sim 0{,}8$. Este mapa lo realizamos para distintos valores de x_0 viendo que R aumenta medida que lo hace x_0, es decir que las soluciones del modelo completo se aproximan a las de la aproximación de flameo. Resultado esperado si tenemos en cuenta que el efecto de colisión no está incluido en el modelo de flameo, y los choques disminuyen al aumentar la separación entre cuerdas. Variar x_0 tiene un sentido fisiológico, ya que la separación media de las cuerdas puede controlarse activamente mediante la actividad de los músculos cricoaritenóideos.

Curvas de isofrecuencia

La frecuencia fundamental de la oscilación de las cuerdas, f_0, conocida como *pitch*, es una de las propiedades perceptuales más relevantes del habla.

Contiene información de la identidad del hablante, por ejemplo: típicamente los valores de pitch femeninos se mueven en una rango de 110 a 160 Hz, mientras que los masculinos lo hacen en uno más bajo que va de los 70 a los 120 Hz, aproximadamente. Además, la capacidad de variar activamente valor de f_0 durante el discurso permite producir distintas entonaciones.
A pesar de ser una característica importante de la voz, existen muy pocos registros experimentales sobre el control activo del pitch a través de la actividad de los músculos laríngeos y la presión subglotal. Esto se debe a la dificultad experimental de realizar estas mediciones simultáneamente. Sin embargo, existe un trabajo [35] donde encuentran que, estando el músculo vocal inactivo, un incremento en la la presión subglotal resulta en un aumento en el valor del pitch.

Reproduciendo de forma teórica las condiciones de este trabajo, estudiamos la dependencia de f_0 con P_s dejando fijo Q, parámetro relacionado con la actividad del músculo vocal. Analizamos el comportamiento de pitch para distintas representaciones de la fuerza restitutiva, explorando cuál resulta acertada al comparar los resultados teóricos con los experimentales.
Típicamente, existen dos descripciones para las características restitutivas del tejido de las cuerdas vocales: la cúbica [20, 28] (ver ecuación 2.4) y la lineal [25,30] ($K_i(x_i) = k_i x_i + \Theta(\frac{x_i+x_0}{x_0})3k_i(x_i+x_0)$). En la figura 2.6 se muestran los resultados obtenidos para las curvas de isofrecuencia en ambos casos, para valores de los parámetros dentro del rango de fonación normal.
Si bien al comienzo de las oscilaciones las curvas $f_0(P_s)$ no se ven afectadas por el tipo de restitución, la diferencia se hace evidente al aumentar P_s. La pendiente resulta positiva cuando la restitución es cúbica y negativa para el caso lineal. Al comparar este resultado con la evidencia experimental anteriormente mencionada [35], la restitución cúbica resulta ser la descripción acertada para modelar las cualidades elásticas del tejido.

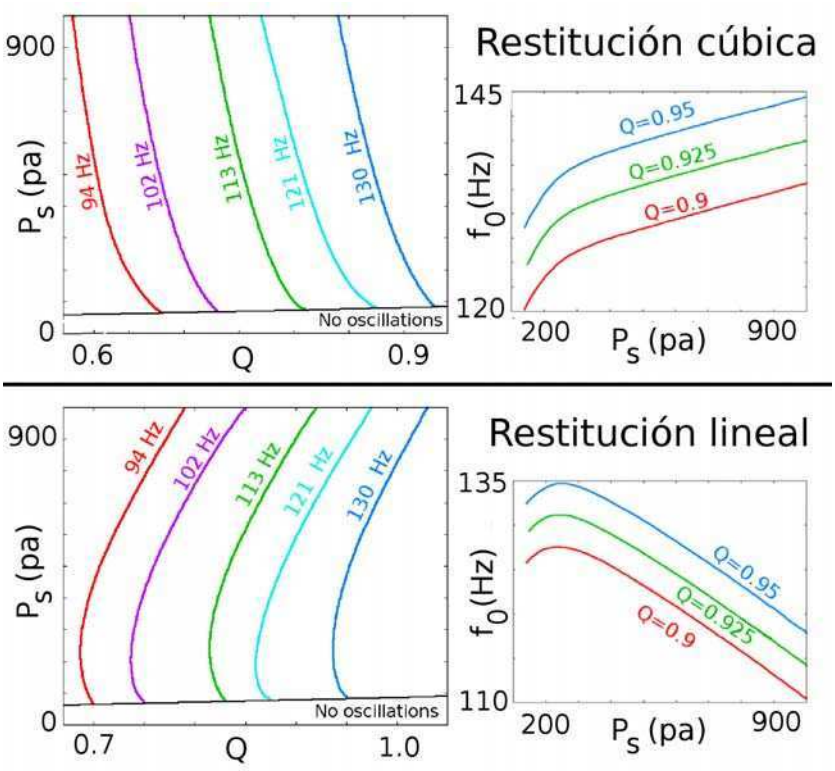

Figura 2.6: Dependencia del pitch con la tensión en las cuerdas y la presión subglotal. Paneles izquierdos: curvas de isofrecuencias en el plano (Q,P_s). Paneles derechos: Curvas $f_0(P_s)$ para Q=0.9, Q=0.925 y Q=0.95. Los gráficos superiores corresponden al modelo completo descripto en la sección 2.1.2, con una restitución cúbica según la ecuación 2.4. Los inferiores se obtuvieron alterando el modelo usando una restitución lineal dada por: $K_i(x_i) = k_i x_i + \Theta(\frac{x_i+x_0}{x_0})3k_i(x_i + x_0)$.

2.2. Filtro: Tracto vocal

Como mencionamos en el capítulo 1, el rol que desempeña el tracto vocal queda determinado por el tipo de fonema. Para las vocales, actúa simplemente como un filtro para las perturbaciones de presión acústica generadas por las cuerdas. Mientras que en las consonantes, no solo actúa de filtro sino también de fuente, en estos fonemas existe una constricción en el tracto donde se genera una turbulencia que contribuye como fuente sonora. Debido a las dificultades que presenta el correcto modelado de fenómenos que incluyen turbulencia, los primeros modelados se desarrollaron para configuraciones de

tracto vocal correspondientes a vocales.
En este caso, el problema se reduce a resolver la ecuación de propagación de una onda de presión acústica en un tubo de sección variable. En el caso de que la sección del tubo sea pequeña comparada con la longitud de onda, esta ecuación viene dada por [36]:

$$\frac{1}{S(x)}\frac{\partial}{\partial x}(S(x)\frac{\partial p}{\partial x}) = \frac{1}{c^2}\frac{\partial^2 p}{\partial t^2} \tag{2.13}$$

donde x es la posición en el tubo, t el tiempo, S la sección y c la velocidad del sonido. Para evitar el costo computacional de resolver esta ecuación diferencial en derivadas parciales, se aproxima el tubo de sección variable por distintos segmentos de sección constante, como se muestra en la figura 2.7.

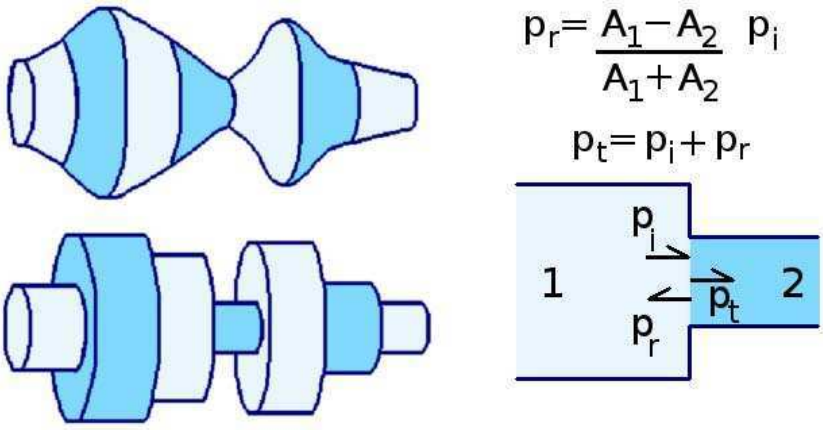

Figura 2.7: Panel izquierdo: Contorno de un tracto vocal y su discretización en 7 segmentos de sección constante. Panel derecho: Interfaz entre dos tubos, con: p_i onda de presión acústica incidente, p_r reflejada, p_t transmitida, A_1 área del tubo anterior, A_2 área del tubo posterior.

En este caso la onda sonora se separa en dos en cada interfaz entre los tubos n y $n \pm 1$, una onda reflejada y una transmitida, de acuerdo con la relación entre las áreas, según se describe en el panel izquierdo de la figura 2.7.

Los primeros modelados de tracto consistieron en sistemas de tres tubos [2], donde el primero representaba la faringe, el segundo la boca y el tercero los labios. Estos modelos permitían reproducir la posición de las dos primeras resonancias del tracto de varias vocales. Con el aumento del poder de cálculo, estos modelos se fueron complejizando: aumentando el número de tubos e incluyendo distintos fenómenos físicos de pérdida de energía. La obtención de imágenes por resonancia magnética de la disposición de tracto para los distintos fonemas permitió un estudio más detallado [37] del problema, y en los últimos años se desarrollaron descripciones dinámicas del tracto que permiten integrar consonantes y sintetizar habla [38].

Típicamente, en la literatura existen dos formas de resolver el problema de una onda sonora propagándose en un conjunto de tubos concatenados: planteando condiciones de contorno para la onda de presión acústica en cada interfaz (*modelo de reflexión de ondas*), o mediante el análogo eléctrico. Este último modela cada tubo como un elemento en una línea eléctrica con su capacitancia, inductancia y resistencia correspondientes [39] al largo y ancho que lo definen. A lo largo de esta tesis trabajamos con modelos de tracto con distinto nivel de complejidad, utilizando siempre el *modelo de reflexión de ondas* para su resolución.

2.2.1. Modelo simple de vocales y fricativas

Existen en la literatura, varios modelos para el tracto vocal [2, 37, 40] basados en la teoría de la propagación del sonido en un tubo delgado [36]. Todos ellos aproximan el tracto por una concatenación de un número variable de tubos uniformes, yendo desde 2 [2] hasta 44 [37], según el nivel de detalle con el que se quiera describir el perfil del tracto. Estos modelos asumen dos ondas en cada tubo: una que avanza, p_{if}, y una retrocede, p_{ib}; y planteando las condiciones de contorno en cada una de las interfaces, se obtiene la onda a la salida del último tubo. Más precisamente, para un sistema formado por N tubos, las ondas de presión que se propagan en cada uno de ellos vienen

dadas por el siguiente sistema de ecuaciones con *delay*:

$$\begin{aligned} p_{1f}(t) &= p_v(t) + r_{1,0}p_{1b}(t-\tau_1), \\ p_{1b}(t) &= r_{1,2}p_{1f}(t-\tau_1) + t_{2,1}p_{2b}(t-\tau_2), \\ p_{2f}(t) &= t_{1,2}p_{1f}(t-\tau_1) + r_{2,1}p_{2b}(t-\tau_2), \\ &... \\ p_{Nf}(t) &= r_{N,N-1}p_{Nf}(t-\tau_N) + t_{N-1,N}p_{(N-1)f}(t-\tau_{N-1}), \\ p_{Nb}(t) &= r_{N,out}p_{Nf}(t-\tau_N) \end{aligned} \tag{2.14}$$

donde $\tau_i = l_i/c$ es el tiempo que tarda el sonido en recorrer una longitud l_i (el largo del tubo i), $r_{i,i\pm1} = a_i/a_{i\pm1}$ y $t_{i,i\pm1} = 1 - r_{i,i\pm1}$ son los coeficientes de reflexión y de transmisión respectivamente, $r_{N,out}$ es el coeficientes de reflexión con la atmósfera y $p_v(t)$ es la presión acústica emergente de las cuerdas vocales. El sonido a la salida del tracto es proporcional a p_{Nf}.
Esta descripción es suficiente para modelar vocales. Fijando un modelo para las cuerdas, ya sea tipo dos masas o de flameo, se obtiene $p_v(t)$. Luego, se determinan los parámetros geométricos de cada tubo y mediante las ecuaciones 2.15 se obtiene la presión a la salida del último tubo, que permite sintetizar los sonidos vocálicos.
El número de tubos que componen el modelo se fija buscando un buen compromiso entre costo computacional y calidad de la síntesis. Para el caso de las vocales, alcanza con tres tubos para poder distinguirlas, ya que un modelo con $N = 3$ alcanza para generar funciones de transferencia, cuyas dos primeras resonancias coincidan con formantes de las 5 vocales del español. Una discretización más fina es necesaria si se quiere ajustar un número mayor de formantes, modificando el timbre de la vocalización.

El desafío consiste ahora en integrar consonantes al modelo. A diferencia de las vocales, la separación fuente-filtro ya no es válida y no existe en la literatura una descripción precisa de la fuente sonora. En esta sección nos concentramos en el modelado de un tipo de consonantes: las fricativas. Algunos ejemplos de estas consonantes son: [*f*, *s*, $\int$, *x*] que suenan según la letra en

negrita de cada palabra: **f**ino, **s**al, **y**o y **j**amón. Para este tipo de fonemas, la fuente de sonido es una turbulencia producida por el pasaje de aire a través de una constricción en el tracto, lograda acercando dos articuladores, como por ejemplo: el labio inferior sobre los dientes superiores, en el caso de $[f]$; la parte posterior de la lengua contra el paladar blando, en el caso de $[x]$; o el lado de la lengua contra los molares, en el caso de $[\int]$.
Distintos estudios realizados sobre este tipo de sonidos [3, 41] muestran que pueden modelarse como una fuente de ruido coloreado ubicada a la salida de la constricción. Inspirados por estos resultados, modelamos la presión acústica que actúa como fuente para este tipo de fonemas (p_u) como un oscilador forzado con ruido blanco ($n(t)$):

$$\ddot{p}_u = -\kappa p_u - \beta \dot{p}_u + n(t) \tag{2.15}$$

La amplitud de un oscilador forzado depende de la frecuencia del forzante, forzando al oscilador con ruido blanco obtenemos una señal ruidosa, con un espectro que presenta un pico del cual podemos variar su ancho y su posición moviendo los parámetros κ y β, panel inferior de la figura 2.8.
Para integrar las fricativas a nuestro modelo, cuando el área de uno o más tubos es menor que un dado umbral, se ubica a la salida del tubo más angosto una fuente extra de sonido del tipo p_u. De esta manera, si la constricción está en el tubo $i-1$, las ecuaciones para las ondas 2.15 quedan todas iguales salvo: $p_{if}(t) = t_{i,i+1} b_f(t-\tau) + r_{i+1,i} d_b(t-\tau) + p_u(t)$.
Teniendo en cuenta que: la constricción de una fricativa abarca aproximadamente 1,7 cm, y que el largo de un tracto promedio es de 17 cm [3]; necesitamos que el modelo cuente, al menos, con diez segmentos ($10 \leq N$) para poder sintetizar tanto vocales como fricativas.

De esta manera obtenemos un modelo relativamente sencillo, como el esquematizado en la figura 2.8, que permite reproducir las características acústicas de vocales y fricativas.

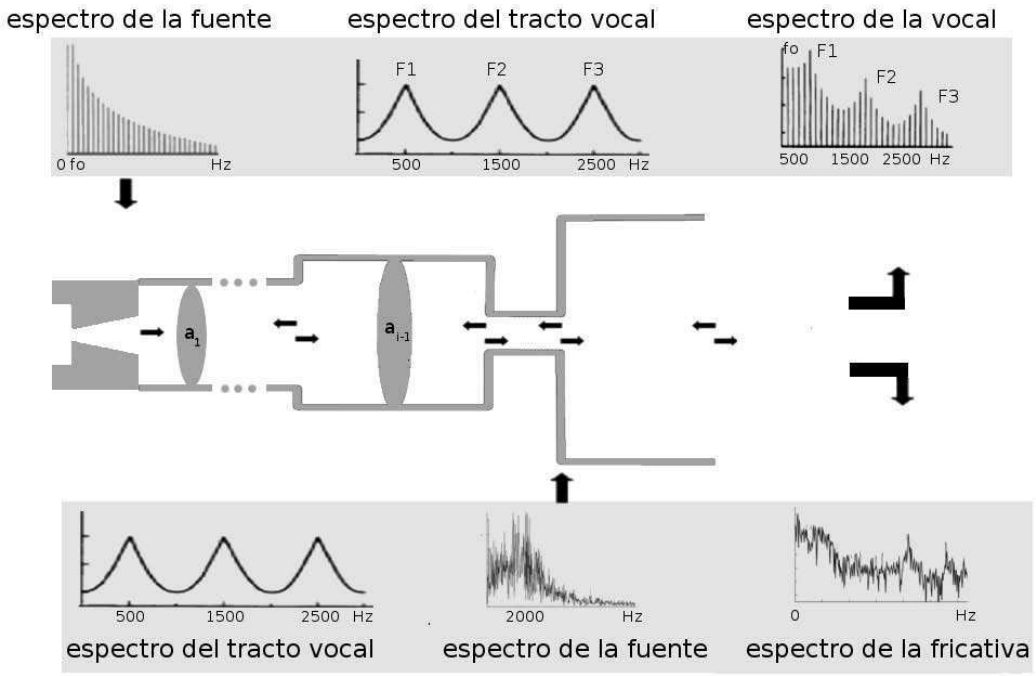

Figura 2.8: Esquema del modelo vocal. La figura central representa la concatenación de tubos que modela el tracto vocal. El panel superior, de izquierda a derecha: espectro de la fuente para los sonidos voceados con una frecuencia fundamental f_0, función de transferencia del tracto en este caso un tubo uniforme de 17.5 cm y la convolución de ambos espectros que corresponde al fonema voceado resultante. En el panel inferior: nuevamente la función de transferencia del tracto, el espectro de un ruido coloreado que caracteriza el flujo turbulento a la salida de una constricción y el espectro del sonido emergente del tracto, la fricativa.

2.2.2. ¿Cómo ir del espacio acústico al anatómico?

Hasta ahora contamos con un modelo que nos permite sintetizar distintos fonemas. Sin embargo, muchas veces nos encontramos con el problema inverso: dado un sonido queremos conocer la anatomía que lo genera. Para esto desarrollamos un algoritmo computacional que permite reconstruir la configuración de tracto correspondiente a una vocalización, a partir de sus propiedades espectrales. Utilizamos para esto un algoritmo genético, un método estocástico de búsqueda, inspirado en la evolución biológica, que

permite encontrar distintas familias de soluciones a un problema de optimización [42]. Esta herramienta permite encontrar conjuntos de áreas y largos para la ecuación 2.15 compatibles con un dado espectro experimental (*espectro objetivo*).

Primeramente, este algoritmo de búsqueda fue testeado en un modelo simple de tracto: el de un ave. Existe un interés especial en el estudio del canto de las aves debido a las similitudes entre el proceso del canto y el del habla. El sistema fonador de las aves está constituido, al igual que el humano, por unas membranas capaces de oscilar produciendo sonido (siringe) y de un conjunto de cavidades posteriores que se ocupan de filtrarlo. En el caso humano, la transmisión de información por medio de distintos fonemas es posible gracias a un tracto vocal dinámico, capaz de alterar activamente su configuración, y una fuente con un alto contenido espectral que permite un relevamiento detallado del perfil del tracto. La estrategia del canto de las aves es distinta: la siringe genera sonidos cuasi armónicos, de baja riqueza espectral, la variedad de sílabas depende de amplias variaciones en la frecuencia fundamental de la fuente. Es decir, que transmite información variando la frecuencia de oscilación de las membranas que constituyen la siringe. Debido a esto, históricamente los trabajos realizados sobre el canto de los pájaros se concentraron en el modelado de la fuente, perdiendo de vista al tracto vocal. Sin embargo, trabajos recientes revelan un control activo en la configuración de tracto vocal durante el canto en algunas aves oscinas [39, 40, 43].
El trabajo que describiremos a continuación se encuentra publicado en [8]. Este, además de servir para testear el algoritmo, permitió demostrar la existencia de dinámica en el tracto de aves suboscinas, grupo de aves un poco menos desarrolladas que las oscinas, tanto a nivel anatómico como comportamental.

Nuestro objeto de estudio fue una vocalización específica de un ave suboscina llamada Benteveo (*Pitangus Sulfuratus*). Esta vocalización estereotipada, denominada *llamada*, es emitida por el ave en situaciones de alerta, y su espectro es el que se muestra en el panel superior de la figura 2.9.

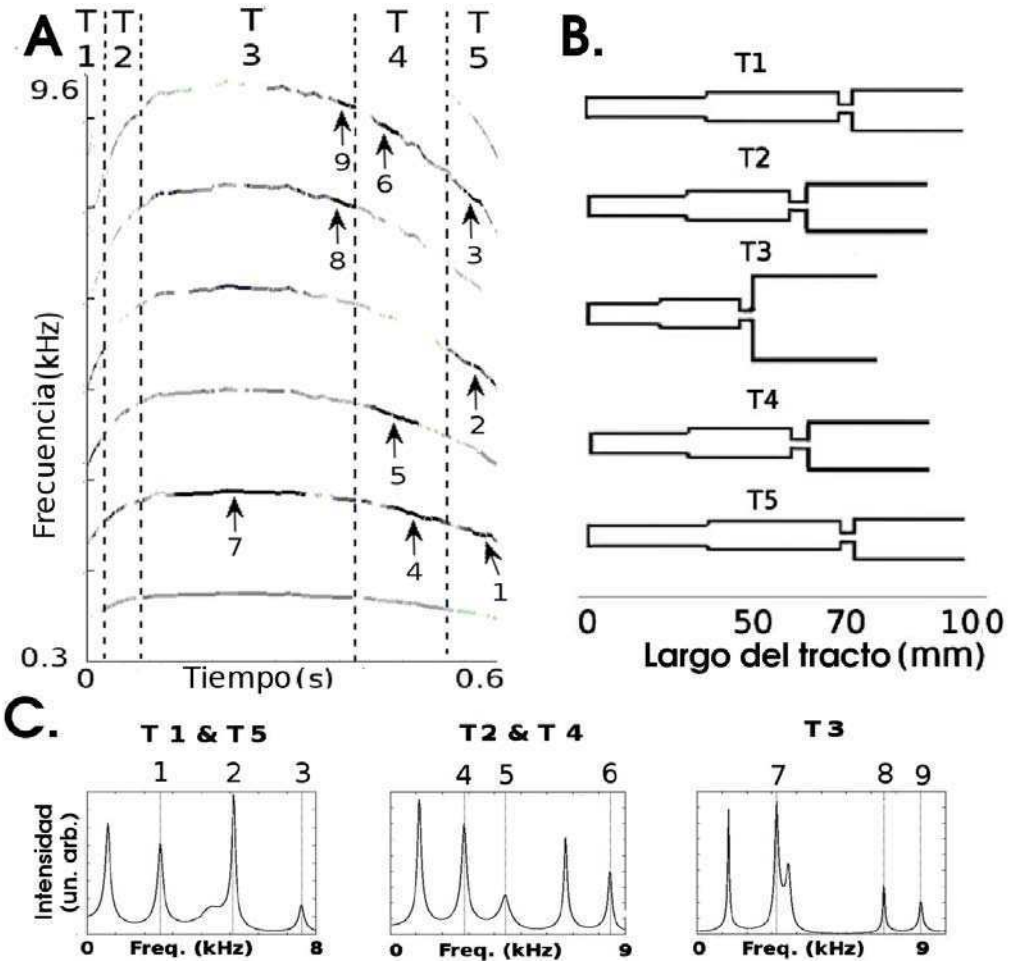

Figura 2.9: Dinámica del tracto vocal del benteveo durante la emisión de una *llamada*. **A.** Espectrograma reasignado [44, 45] de una llamada. Los números indican las frecuencias enfatizadas, las mismas se enfatizan al comienzo y al final ($T_1 = T_5$ y $T_2 = T_4$), las primeras no se alcanzan a ver en la figura. **B.** Esquema de la configuración de tracto hallada con el algoritmo para cada tiempo. Los parámetros $\{l_1; l_2; l_3; l_4 | a_1; a_2; a_3; a_4\}$ son: $T_1(T_5) \rightarrow \{35; 36; 3{,}7; 23 | 1; 2{,}1; 0{,}4; 1{,}6\}$, $T_2(T_4) \rightarrow \{34{,}3; 25{,}1; 3{,}7; 23{,}6 | 1; 2; 0{,}6; 2\}$ y $T_3 \rightarrow \{25{,}4; 24.; 3{,}7; 29{,}6 | 1; 2{,}2; 0{,}6; 3{,}6\}$. **C.** Función de transferencia para la configuración de tracto halladas con el algoritmo para cada tiempo, las resonancias coinciden con las frecuencias enfatizadas experimentales.

Si asumimos que las frecuencias en las cuales uno, o un subgrupo, de los armónicos presenta un aumento repentino en la intensidad se corresponden con resonancias del tracto vocal, podemos identificar nueve resonancias por debajo de los $10kHz$, ver figura 2.9. Realizando simulaciones numéricas para sistemas con distinto número de tubos ($2 \leq n \leq 20$) con un largo total aproximado de $90mm$ (de acuerdo con mediciones directas de la anatomía del ave), encontramos entre 3 y 5 resonancias por debajo de los $10kHz$. Este resultado indica que no es posible que un solo perfil de tracto presente todas las resonancias experimentales. Es necesaria una dinámica en el tracto que permita cambiar su forma, por ende sus resonancias, para dar cuenta de las frecuencias enfatizadas a lo largo de toda la vocalización. Los armónicos enfatizados de forma simultánea representan las resonancias de las distintas formas que adopta el tracto a lo largo de la llamada. Separando la llamada en 5 segmentos no hay más de 3 resonancias simultáneas, buscamos las configuraciones de tracto cuyas resonancias coinciden con las frecuencias enfatizadas en cada segmento (ver figura 2.9). Es decir, fijamos para el sistema de ecuaciones 2.15 $N = 4$ (en consonancia con trabajos previos [39]) y mediante el algoritmo genético, buscamos las formas de tubos que reproducen las resonancias experimentales.

Las configuración encontradas para cada tiempo pueden pensarse como perturbaciones de una única configuración, constituida por un tracto vocal inferior (formado por los dos primeros tubos de áreas similares) separado de una cavidad más ancha por una constricción (figura 2.9). Esto permite identificar los dos primeros tubos con la traquea del ave, un tubo cartilaginoso y extensible de aproximadamente $60mm$ de largo. El tercer tubo con la constricción glotal, y el último como una cavidad efectiva que incluye la cavidad oral y el pico. La evolución temporal encontrada consiste en una reducción de la longitud total y una expansión del tracto superior al comienzo (configuraciones de T_1 a T_3) y el movimiento opuesto al terminar (configuraciones de T_4 a T_5) la vocalización. Este esquema coincide con el movimiento que realiza el ave al vocalizar la llamada: al comenzar retrae su cabeza a la vez que abre el pico, mantiene este gesto y luego invierte el movimiento volviendo

a la posición inicial al terminar la llamada.

Para construir confianza en el modelo, sintetizamos llamadas con un modelo para la siringe descripto por la ecuación 2.2 y la dinámica de tracto hallada. Cabe aclarar, que trabajos previos [27] prueban que el modelo de flameo es capaz de sintetizar sonidos compatibles con los emitidos por el Benteveo, al ser alimentado por mediciones experimentales de la presión en el saco aéreo del ave fonando. Ingresando funciones suaves para la presión subglotal $p_s(t)$, proporcionales a la frecuencia fundamental, se obtienen sonidos compatibles con las grabaciones experimentales.

El proceso completo indicaría que el ave aumenta la presión en el saco aéreo, provocando un aumento en la frecuencia del sonido, al mismo tiempo que modifica las dimensiones de las cavidades del tracto vocal, haciendo coincidir el primer armónico de la fuente con la segunda resonancia del tracto. Esta estrategia para aumentar la intensidad de las vocalizaciones coincide con la descripta previamente en aves oscinas [40, 43].

2.2.3. Un modelo más detallado que mantiene la baja dimensión

Según mencionamos en el capítulo 1, el aumento en la capacidad de cálculo y la disponibilidad de imágenes por resonancia magnética, permitieron una descripción mucho más detallada del tracto vocal. Los modelos más modernos aumentan el número de tubos que componen el sistema, e incluyen pérdidas energéticas por distintos fenómenos físicos.

Los primeros trabajos de Story et.al. [46, 47] consisten en el estudio, mediante imágenes de resonancia magnética, de las configuraciones de tracto correspondientes a un mismo hablante pronunciando las 10 vocales inglesas. Toman el área del tracto en 44 puntos equiespaciados, yendo de la glotis a los labios, construyendo así lo que se conoce como funciones de área para

las vocales. Realizan un análisis de componentes principales y encuentra que los dos primeros *vectores propios* (o modos empíricos), más un tracto neutro, son suficientes para reproducir los perfiles de tracto de todas las vocales, con un 90 % de precisión. El tracto neutro es sujeto dependiente y es lo que otorga la identidad del hablante, mientras que los dos modos representan las restricciones que los distintos articuladores imponen sobre la configuración del tracto. Más precisamente, fijando en 44 el número de tubos el diámetro de cada uno de viene dado por:

$$d(i) = \Omega(i) + q_1\varphi_1(i) + q_2\varphi_2(i) \tag{2.16}$$

donde $1 \leq i \leq 44$ es el número de tubo, φ_1 y φ_2 son los dos modos espaciales, Ω el tracto neutro que depende de cada individuo, q_1 y q_2 los coeficientes de cada modo que son quienes definen la vocal. Esta descripción permite sintetizar diptongos haciendo evolucionar q_1 y q_2 en el tiempo.

Lo que nos resulta novedoso y relevante de esta investigación es que permite una reducción en la dimensionalidad del problema, basada en restricciones anatómicas. Si queremos sintetizar distintas vocales con un modelo de tracto según la ecuación 2.15, tomando una descripción equiespaciada donde $l_i = L_{tracto}/N$, tenemos $N + 1$ parámetros a determinar: $\{a_i; L\}$ con $i = 1, 2, ..., N$. Esta estrategia permite reducir la dimensión de $N + 1$ a 2, alcanza con determinar la contribución de cada modo normal para obtener el tracto de cualquier vocal.

En un trabajo posterior [37] este resultado es extendido a consonantes. Incluyendo una oclusión en distintos lugares del tracto, logran reproducir funciones de área de distintas consonantes, halladas nuevamente a partir de imágenes de resonancia magnética funcional. Más precisamente, la oclusión es modelada por una función suave entre 0 y 1, C_k, que vale 1 para todos los tubos menos en las cercanías de k , modificando la ecuación 2.16 según:

$$d(i) = [\Omega(i) + q_1\varphi_1(i) + q_2\varphi_2(i)]C_k(i) \tag{2.17}$$

variando k se obtienen las configuraciones de tracto correspondientes a las distintas consonantes oclusivas (para más detalle ver [37]).

Si bien una característica de las consonantes plosivas es lo que se llama el ataque (aumento repentino y pronunciado en la presión que se produce cuando se libera la oclusión y se genera un flujo turbulento), estudios perceptuales muestran que, aún en ausencia del ataque, las plosivas voceadas son identificadas correctamente [38]. Esto permite modelar este tipo de fonemas sin la necesidad de incluir un fuente turbulenta a la salida de la constricción. Incorporando esta descripción al sistema vocal, sintetizamos vocales y oclusivas voceadas, utilizando un modelo de cuerdas como el descripto en 2.1.2, y uno de tracto como el descripto en esta sección, incluyendo pérdidas energéticas. Para esto último se incluyen en la ecuación 2.15 términos que dan cuenta de pérdidas por viscosidad y por vibración de las paredes, para más detalle ver [48]. La síntesis se logra haciendo evolucionar q_1, q_2 en el tiempo e imponiendo sobre este sustrato vocálico la constricción correspondiente. A cada oclusiva le corresponde una función c_k centrada en distinto número de tubo, con distintos parámetros de extensión y simetría según los datos experimentales hallados en [37]. Con una función temporal $m(t)$ se *prende* y *apaga* la constricción obteniendo distintas concatenaciones de fonemas tipo vocal-consonante-vocal (vcv). La figura 2.10 aclara este procedimiento.

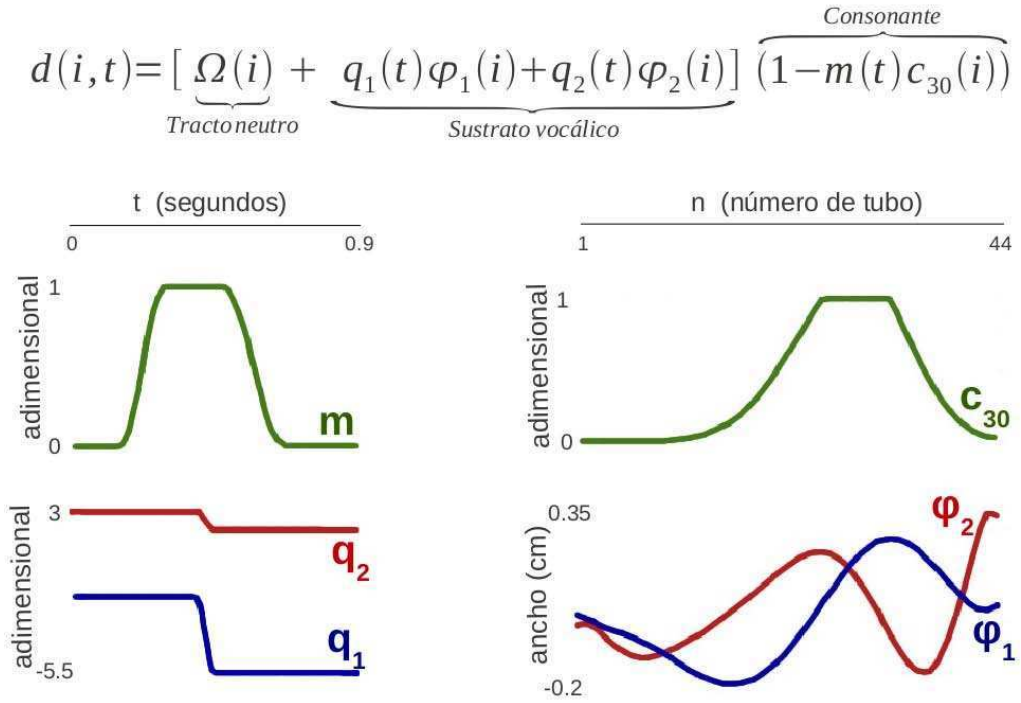

Figura 2.10: Modelo dinámico del tracto vocal para sintetizar la sílaba */egi/*. En el encabezado la ecuación que modela la evolución temporal del diámetro d para cada uno de los tubos i que conforman el tracto, con $1 \leq i \leq 44$. El primer término corresponde al tracto neutro, cuyo perfil no varía en el tiempo, es una variable sujeto dependiente. El primer factor del segundo término representa la evolución temporal para la consonante oclusiva */g/*, oclusión centrada en el tubo 30. El segundo es el estrato vocálico sobre el que se impone la constricción, determina el contexto vocálico de la sílaba. A la derecha la evolución temporal de los parámetros que definen: la consonante $m(t)$ y las vocales $(q_1; q_2)$. Para ajustar las vocales rioplatenses */e/* $\rightarrow$ */i/* utilizamos $(q1(t_0), q2(t_0)) = (-1; 3) \rightarrow (q1(t_f), q2(t_f)) = (-5{,}5; 2)$. A la izquierda las funciones espaciales para la consonante */g/* y para los modos normales vocálicos, según se describen en [37].

Capítulo 3

Explorando la percepción de la voz con un sintetizador articulatorio

Un sintetizador de voz realista presenta una herramienta con aplicaciones variadas, yendo del estudio de percepción de la voz al desarrollo tecnológico, por esto en los últimos años numerosos esfuerzos se realizaron en esta dirección. Existen tres tipos de sintetizadores: los concatenativos, que generan habla a partir de extensas bases de datos, concatenando secciones de audio experimentales para producir una nueva síntesis; los acústicos, sintetizan manipulando en el tiempo distintos parámetros acústicos como pitch, formantes o ruido [49,50]; y los articulatorios que se basan en los principios físicos de la producción de la voz [2,23]. A lo largo de esta tesis nos concentramos en sintetizadores del último grupo, haciendo uso de la ventaja que presenta frente al resto: permite estudiar las consecuencias acústicas, y por ende perceptuales, de variaciones en parámetros fisiológicos. No trivial, ya que variaciones en parámetros anatómicos pueden resultar en transformaciones no lineales en el espacio acústico.

Tradicionalmente, la percepción de la voz es estudiada en términos de parámetros acústicos [51,52]. Sin embargo, en los últimos años, varios estu-

dios apuntan a un programa del habla sensori-motor, que integra percepción y producción en un mismo proceso, proponiendo que la percepción de habla incluye procesos neuronales que reconstruyen los *gestos motores* del discurso percibido [12, 53]. En este marco, proponemos estudiar la codificación neuronal de la voz en términos de parámetros anatómicos en vez de acústicos, y un sintetizador articulatorio resulta la herramienta natural para estudiar este problema.

Primeramente, es necesario testear el modelo. Típicamente los sintetizadores se validan por su comprensibilidad, es decir, por su capacidad de transmitir información fonológica, perdiendo de vista la calidad vocal de la síntesis. En este capítulo, verificamos la *fidelidad* de un sintetizador articulatorio, cotejando sus síntesis con voces reales. Comparamos tanto la respuesta comportamental, como la activación cerebral que generan, por medio de imágenes de resonancia magnética funcional (*fMRI*). Una vez testeado el modelo, mostramos que es necesaria una relación entre las dimensiones del tracto vocal y de las cuerdas vocales para que la voz sintética sea percibida como *natural*. Este trabajo está terminado y en proceso de escritura, en colaboración con el Laboratorio de Neuropsicología y Neuroimágenes, ICM de París.

3.1. Experimento 1: Voces reales vs. sintéticas

En esta sección comparamos voces sintéticas con reales, estudiando la respuesta comportamental y cerebral frente a ambas. En las últimas décadas, varios trabajos se concentraron en el estudio del procesamiento cerebral de la voz. Por ejemplo, Belin et. al. realizaron estudios con fMRI donde muestran que en la corteza auditiva existen zonas que codifican sonidos voceados [54,55], nombradas como TVA (del inglés *temporal vocal áreas*). Estas áreas, localizadas bilateralmente a lo largo de STS (surco temporal superior),

presentan una actividad mayor frente a estímulos voceados humanos (habla, habla invertida, risas, llanto etc.) que frente a sonidos de control de distinto tipo (animales, instrumentos musicales, campanas, etc). Estos antecedentes sugieren que hay áreas cerebrales que responden a la voz, independientemente de su contenido semántico. Haciendo uso de este resultado investigamos si las mismas áreas son estimuladas por las voces sintetizadas, además de estudiar comportamentalmente su percepción.

3.1.1. Métodos

El sintetizador se construyó utilizando para las cuerdas vocales el modelo de dos masas descripto en la sección 2.1.2, y el modelo de tracto de la sección 2.2.3.

Luego, buscamos comparar la respuesta (comportamental y cerebral) generada con nuestro sintetizador con la correspondiente a voces reales, enfocándonos en la *fidelidad* acústica (cuan naturales suenan los estímulos), más allá del contenido fonológico. Para esto, utilizamos 4 tipos de estímulos. Naturalmente, uno corresponde a voces reales (Humanos) y otro a voces sintetizadas con parámetros óptimos (Óptimos), estos últimos estímulos generados buscando el mayor nivel de *realismo* posible. El tercer tipo consiste en síntesis *robóticas* pero con contenido fonológico (Robóticos), incluido para investigar la actividad neuronal relacionada exclusivamente con la *fidelidad* vocal. El último corresponde a sonidos no voceados (No Voceados), a modo de parámetro de control. Más detalladamente, cada familia de estímulos esta construida de la siguiente manera:

1. *No voceados (NV)*

   Se incluyen vocalizaciones de animales, sonidos de la naturaleza, producidos por aparatos (timbres, campanas, rings telefónicos etc.) y por instrumentos musicales. Audios extraídos de un protocolo desarrollado por Belin et. al, disponible en la web [51]. Este, consiste en distintos estímulos voceados y no voceados, que al ser escuchados por un sujeto dentro del resonador, permiten localizar las áreas vocales.

2. *Humanas (H)*

   Se grabaron dos individuos de distinto sexo, de entre 30 y 40 años, hablantes nativos de español rioplatense, pronunciando distintas vocales, diptongos y estructuras vocal-consonante-vocal, *vcv*, con distintas entonaciones. Las consonantes permitidas son las oclusivas voceadas del español */b/, /d/ y /g/*, y las vocales las cinco rioplatenses.

3. *Óptimos (O)*

   Para obtener los estímulos de esta familia, buscamos usar parámetros *naturales* para controlar el sintetizador. Se generaron vocales, diptongos y estructuras *vcv*, las mismas que usamos en la familia anterior. Para la dinámica del tracto vocal, que es lo que determina el contenido fonológico, utilizamos la ecuación 2.17. Para sintetizar vocales, se utilizaron los pares $(q_1; q_2)$ que forman configuraciones de tracto cuyas dos primeras formantes se corresponden con las vocales rioplatenses. Los diptongos consisten en transiciones lineales en el tiempo, yendo de los valores de una vocal a los de la otra. Las estructuras *vcv* se generaron superponiendo a los diptongos la oclusión correspondiente, según se explica en la sección 2.2.3.
   Además del contenido léxico, fijamos la entonación, o prosodia, de las vocalizaciones, controlada por la tensión de las cuerdas y la presión subglotal (Q,P_s) del modelo de dos masas descripto en la sección 2.1.2. El perfil temporal de estos parámetros resulta central para determinar la *naturalidad* de las voces sintetizadas.
   Algunos parametros acústicos utilizados típicamente por los fonetistas para caracterizar la voz son: el ataque, la liberación, y la rugosidad [56–58]. Los dos primeros, corresponden a características temporales definidas por el tiempo que le toma a la onda sonora alcanzar su máximo, y el que le toma apagarse, respectivamente. En nuestro modelo de cuerdas vocales, este está relacionado con la forma en que comienzan y terminan las oscilaciones. Para dar cuenta de esta característica del timbre de la voz fijamos para la presión subglotal un arranque y una finalización suave, partiendo de cero e imponiendo un ascenso y un

descenso que no dure menos de 0.05 segundos. La rugosidad de la voz es un ruido presente en la frecuencia fundamental de la voz, el pitch. Recordemos que en nuestro modelo, la frecuencia está principalmente gobernada por el parámetro Q. Le sumamos, entonces, un ruido a este parámetro de aproximadamente el 5 % de su valor, de acuerdo con valores experimentales [57].
Por último, buscamos reproducir entonaciones naturales, fijando el perfil temporal de la presión con las variaciones en la intensidad y el de Q según las variaciones en frecuencia, de distintas grabaciones experimentales realizadas en el laboratorio.

4. *Robóticos (R)*

   El contenido léxico de esta familia es idéntico al de la familia anterior, por lo cual se utiliza la misma dinámica de tracto. La diferencia se encuentra en los parámetros de la fuente, buscamos estímulos que sean percibidos como artificiales sin perder el contenido fonológico. Para esto, utilizamos valores constantes para (Q,P_s), un ascenso y un descenso abrupto para P_s (con una duración de 0.01 segundos) y eliminamos el ruido de Q. La figura 3.1 describe la diferencia entre estos estímulos y los óptimos.

Para el experimento dentro del resonador utilizamos un diseño en bloques: se presentan los bloques *randomizados* con un silencio entre ellos donde el sujeto debe realizar la tarea correspondiente. Los estímulos fueron ordenados de forma aleatoria en 9 bloques de cada familia (NV, H, O y R), cada bloque contiene aproximadamente 10 estímulos. Además, incluimos 4 bloques señuelo, estos contienen el primer 80 % de los estímulos con contenido fonológico (H, O o R) y el último 20 % corresponde a no voceados. Cada bloque dura 9 segundos y el silencio entre bloques es de 4.5 segundos. Los sujetos debían completar el experimento perceptual mientras eran escaneados. Recibieron la instrucción de escuchar y calificar cada bloque presionando uno de los 4 botones de una botonera que sostenían en la mano derecha. Más precisamente, al finalizar cada bloque debían activar un botón siguiendo la siguiente escala: Índice: Los sonidos escuchados no son humanos. Medio: Probablemente no

sean humanos. Anular: Probablemente sea humano. Meñique: Definitivamente son humanos. Además, se les indicó no responder en el caso de los señuelos, el objetivo de estos era garantizar la atención hasta el final del bloque. Todo el procedimiento se repitió para cada participante. Se escanearon 17 adultos (7 mujeres) entre 20 y 40 años de edad, todos hablantes nativos de español. Para más detalles de la adquisición de datos ver apéndice A.

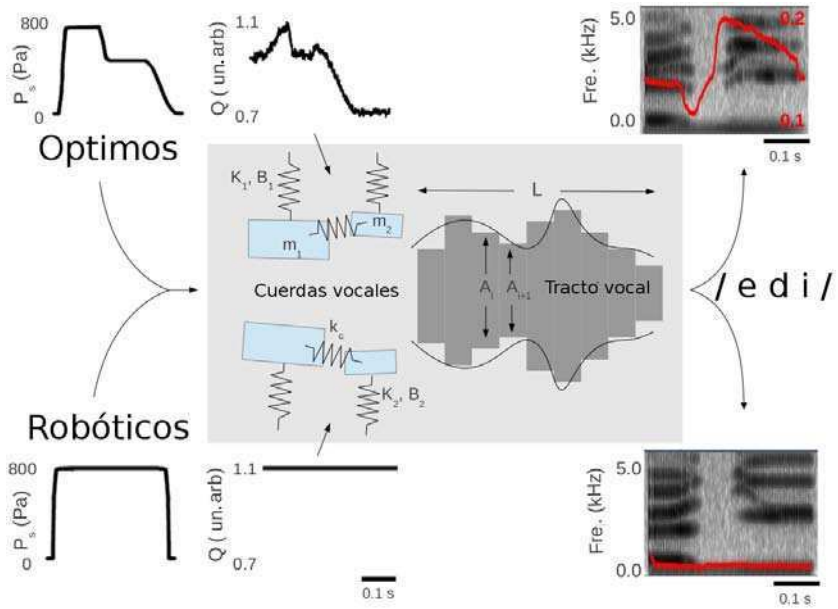

Figura 3.1: Diferencia entre estímulos óptimos y robóticos. Utilizamos distintos perfiles para la tensión en las cuerdas y la presión subglotal, parámetros que alimentan el sistema de dos masas descripto en la sección 2.1.2. La dinámica del tracto es la misma para las dos familias de estímulos, en este caso la que corresponde a sintetizar la estructura */edi/*. Mostramos a la derecha el espectrograma correspondiente a cada tipo de estímulo con el valor de pitch en rojo. Para los estímulos óptimos: $Q(t)$ y $P_s(t)$ presentan distintos valores para la primera y segunda vocal, el ascenso y el descenso de $P_s(t)$ es suave y $Q(t)$ incluye un ruido con amplitud del 5 % de su valor. Para los estímulos robóticos, en cambio, $Q(t)$ y $P_s(t)$ no varían en el tiempo, el ascenso y el descenso de $P_s(t)$ es abrupto, y $Q(t)$ no incluye ruido.

3.1.2. Resultados

Los resultados del experimento comportamental, llevado a cabo por los sujetos dentro del resonador, se muestran en el panel A de la figura 3.2. Podemos ver que los estímulos óptimos son indistinguibles de los reales.

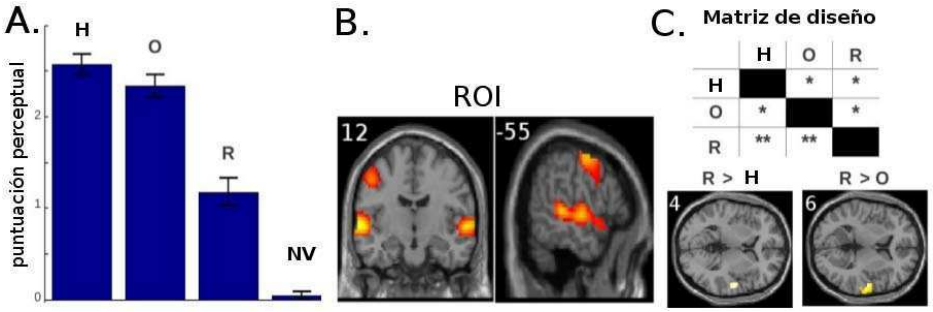

Figura 3.2: Experimento 1. **A.** Resultado del estudio comportamental, donde a cada botón se le adjudicó un valor según: Indice(0): El sonido escuchado no es humano. Medio(1): Probablemente no sea humano. Anular(2): Probablemente sea humano. Meñique(3): Definitivamente es humano. Se muestra el valor medio conjunto la desviación estándar. **B.** Vista coronal y sagital mostrando la actividad para el contraste (O+R+H-3NV) fonológico vs. no-fonológico ($p_{no\_corregido} <$ 0,005 , $nr\_voxels >$ 300). Definimos estas zonas como las regiones de interés, ROI. **C.** Matriz de diseño, cada lugar de la matriz representa un contraste. Los sitios marcados con ** indican que ese contraste presenta zonas activas para un umbral $p_{no\_corregido} <$ 0,001, mientras que * indica que no hay voxels que superen el umbral $p_{no\_corregido} <$ 0,01. Tanto el experimento comportamental como el de imágenes no muestran diferencias significativas entre las voces reales y las sintetizadas con parámetros óptimos.

Para analizar los datos de imágenes cerebrales utilizamos el programa SPM5 [59], para una descripción detallada del preprocesamiento y posterior análisis de los datos ver apéndice A. Primeramente, determinamos las regiones generales activadas por todas las familias de estímulos acústicos. Para esto realizamos 4 contrastes a nivel de grupo: NV-reposo, H-reposo, O-reposo y R-reposo. Los cuatro contrastes presentan patrones de activación similares: una amplia zona temporal y otra que se extiende del giro precentral al giro frontal inferior son activadas bilateralmente, además de una región en el giro

frontal medio que incluye áreas motoras. Las diferencias son muy tenues: los estímulos no voceados encienden regiones más posteriores del giro temporal superior, mientras que para los bloques con contenido fonológico la región frontal activada es un poco más extensa que en el caso no voceado.

Luego, estudiamos las áreas involucradas en el procesamiento fonológico, para esto buscamos las zonas estimuladas por el contraste fonológico vs no fonológico ($H + O + R > 3NV$). En este caso se activaron las tres regiones ($p_{no\_corregido} < 0{,}005$, $nr\_voxels > 300$) que se muestra en la figura 3.2 y en la tabla 3.1. El giro temporal superior, el inferior y el polo temporal son activados bilateralmente, en acuerdo con los resultados obtenidos por Belin et.al [54, 55]. La otra región activada se extiende a lo largo de los giros pre y post centrales del hemisferio izquierdo, que son áreas involucradas en la producción y la integración sensorial-motora del habla [53]. Tanto en este experimento como en el que le sigue, estudiamos la actividad cerebral correspondiente a estímulos acústicos con contenido fonológico, por esto definimos estas tres áreas como nuestra región de interés (ROI del inglés *region of interest*). Todos los análisis estadísticos que realizamos a continuación se encuentran restringidos a estas regiones.

Una vez definida nuestra región de interés, nos adentramos en el objetivo de este experimento: comparar la respuesta cerebral a voces sintéticas con la correspondiente a voces reales. Analizamos todos los contrastes posibles de dos tipos de estímulos dentro del grupo fonológico (ver la matriz de diseño en la figura 3.2). Ninguno de estos contrastes muestra regiones activadas por arriba del umbral establecido ($p < 0{,}01$ restringido a la ROI), salvo $R > H$ y $R > O$. Concluimos, entonces, que los sonidos sintetizados con parámetros *óptimos* son indistinguibles de los reales. Un área contenida en el giro temporal superior derecho, incluyendo Heschl (corteza auditiva primaria), muestra más activación para los estímulos Robóticos que para los óptimos ($p_{no\_corregido} < 0{,}001$ y $p_{corregido\_por\_cluster} = 0{,}016$). La misma tendencia la observamos en el contraste $R > H$ ($p_{no\_corregido} < 0{,}001$ y $p_{corregido\_por\_cluster} = 0{,}054$), figura 3.2 y tabla 3.1. Curiosamente, las áreas

temporales que responden a sonidos humanos descriptas en [55] son activadas por habla sintética, aún cuando esta es percibida como *no natural* (R). Mas aún, algunas subregiones de estas áreas presentan una activación mayor para este tipo de estímulos.

Anatomía	Cordenadas Talairach			T-valor	Voxels
	x	y	z		
Fonológico>No Fonológico (ROI)					
izq STG	-66	-25	4	14.6	500
	-69	-7	-2	10.8	
izq Precentral G	-51	-7	46	6.5	320
der STG	60	-19	-2	7	503
	66	-10	1	6.9	
Robóticas>Óptimos					
der STG	54	-10	4	5.9	52
Robóticas>Humanas					
der STG	54	-10	4	4	17

Tabla 3.1: Tabla de resultados para el experimento 1. Se especifica la zona anatómica, locación estereotáxica en coordenadas Talairach, T-valor y extensión en número de voxels para cada cluster. Los umbrales varían según el contraste para $Fonologico > No\_Fonologico$ (ROI) se fija un umbral de $p_{no\_corregido} < 0{,}005$, $nr\_voxels > 300$. Tanto el análisis estadístico para $R > O$ como para $R > H$ fue restricto dentro de la ROI, el primero con $p_{no\_corregido} < 0{,}001$ y $p_{corregido\_por\_cluster} < 0{,}05$ el segundo con $p_{no\_corregido} < 0{,}001$ y $p_{corregido\_por\_cluster} < 0{,}055$.

Resumiendo, utilizando perfiles adecuados para la presión y tensión en las cuerdas, las voces sintéticas son indistinguibles de las reales, tanto comportamentalmente como a partir de la actividad cerebral que suscitan. Además, las voces percibidas como artificiales provocan más actividad en las áreas de reconocimiento de voz que las reales. Este resultado es análogo al obtenido en corteza visual comparando reconocimiento de caras y caricaturas [60].

3.2. Experimento 2: Codificación de la voz a partir de parámetros anatómicos

Una vez testeado el grado de *fidelidad* del modelo vocal como sintetizador, lo utilizamos para estudiar la percepción de la voz a partir de variaciones de parámetros anatómicos. Típicamente, el procesamiento de la voz se estudia comparando las respuestas cerebrales a distintos estímulos. Estos consisten en voces reales y manipulaciones artificiales de las mismas, alterando uno o varios parámetros acústicos [61–63]. En otros casos, utilizan técnicas de *morfeo*, mezclan distintas voces para generar una nueva.

En un trabajo reciente [52], utilizando esta última técnica, estudian la codificación de la voz a partir de la existencia de una *voz patrón* para cada género. Más precisamente, promedian distintas voces femeninas o masculinas generando una voz prototípica para cada caso. Determinan, en un espacio tridimensional de parámetros acústicos (pitch, dispersión de formantes y realación armónicos-ruido), la distancia de las distintas voces a la voz patrón correspondiente a su género. Encuentran que, a mayor distancia: **a**. Comportamentalmente, las voces son identificadas como más *originales* o *características*. **b**. A nivel cerebral, la actividad en áreas, incluidas dentro de las zonas cerebrales que codifican sonidos vocales, es mayor. Este resultado depende de contar con un patrón por género. No logran reproducirlo a partir de la distancia a una voz patrón andrógina, construida promediando el total de las voces, sin distinguir género.

Nosotros nos propusimos estudiar el problema de la *identidad* de la voz a partir de parámetros anatómicos en vez de acústicos. Para esto generamos con nuestro modelo vocalizaciones en una grilla de distintos parámetros anatómicos, investigando como varía la percepción de los mismos según su ubicación en este espacio.

3.2.1. Métodos

Exploramos las consecuencias perceptuales de variar las dimensiones de los dos grandes bloques que conforman el aparato fonador: las cuerdas vocales y el tracto. Variamos el largo del tracto con un factor de escala $\lambda = \frac{L}{L_0}$, donde $L_0 = 0{,}17m$ y L es el largo total del tracto. En cuanto a las cuerdas vocales, no hay una sola magnitud que caracterice su tamaño. Es decir, variar las dimensiones de las cuerdas incluye alterar su masa, las características del tejido, su ancho y diámetro. Por esto, se utiliza un factor β [28] para rescalear los parámetros mecánicos del sistema de dos masas que modelan cada membrana, según: $k_i^{'} = \frac{k_i}{\beta}$, $m_i^{'} = \frac{m_i}{\beta^3}$, $d_i^{'} = \frac{d_i}{\beta}$ y $l_g^{'} = \frac{l_g}{\beta}$.

Los estímulos auditivos utilizados para este experimento, consistieron en vocalizaciones sintetizadas con distintas combinaciones de dimensiones de cuerdas y de tracto. Más precisamente, armamos una grilla de 49 puntos en el espacio bidimensional $(\beta; \lambda)$, correspondiente a 7 valores para β y 7 para λ como se muestra en la figura 3.3. Para cada punto de la grilla generamos 4 tipos de vocalizaciones: dos estructuras *vcv* (*/ego/* y */aba/*) cada una de ellas con dos entonaciones distintas (*pitch 1* con una duración de 0.6s y *pitch 2* de 0.9s).

Para el experimento dentro del resonador incluimos 38 silencios con una duración de 0.75s, y el diseño es *por eventos*. El total de los estímulos es presentado en orden aleatorio, con una separación de 2 segundos y se repiten 19, elegidos de forma aleatoria, debido a que la tarea a realizar es del tipo *one back*. Esto quiere decir que los participantes recibieron la instrucción de presionar un botón en caso que el sonido escuchado sea idéntico al anterior. El objetivo de la tarea es, simplemente, mantener la atención del sujeto durante el experimento. El procedimiento completo se repitió 4 veces mientras los sujetos (los mismos que participaron del experimento anterior) eran escaneados.

Por otro lado, un grupo diferente de participantes (30 adultos, 18 mujeres, con edades entre 28 y 40 años, hablantes nativos de español) completaron el

experimento comportamental. Se les presentaron los 196 estímulos, correspondientes a las 4 vocalizaciones para cada uno de los 49 puntos de la grilla, y se les indicó que puntúen cada uno de acuerdo con la siguiente escala: El sonido escuchado 0: No es voz humana. 1: Probablemente no sea voz humana. 2: Probablemente sea voz humana. 3: Definitivamente es voz humana.

3.2.2. Resultados

El resultado del experimento comportamental se muestra en el panel A de la figura 3.3. Observamos que las voces percibidas como más *naturales* se agrupan en una zona que se extiende de la esquina superior izquierda a la inferior derecha. Esto, sugiere que es necesaria cierta relación entre las dimensiones de los dos bloques que conforman el sistema vocal para que la voz sea reconocida como tal. La recta que interpola esta región viene dada por $L = -0{,}16\beta + 1{,}13$, con límites de 95 % de confianza para los respectivos coeficientes: $(-0{,}27; -0{,}05)$ y $(0{,}98; 1{,}3)$. Consistentemente, la región delimitada contiene los parámetros anatómicos típicos para hombres $(\beta; \lambda) = (1; 1)$ y mujeres $(\beta; \lambda) = (1{,}5; 0{,}9)$, según se reporta en [28]. Asimismo, las puntuaciones más altas se alcanzan en la esquina inferior derecha, presumiblemente una zona asociada a dimensiones del sistema vocal de niños. Resumiendo, las síntesis son percibidas como más naturales cuanto menor es la distancia a una línea que incluye las dimensiones anatómicas típicas, que denominamos *recta óptima*.

Luego, estudiamos las imágenes obtenidas con el resonador. A modo de control, verificamos el contraste *todos menos silencio*, fijando como regresores cada punto de la grilla. Como se esperaba, las zonas activadas son compatibles con las regiones de interés definidas en el experimento 1. Luego, guiados por los resultados comportamentales, buscamos zonas donde la activación correlacione con la distancia a la *recta óptima*. Para esto, fijamos como contraste el mapa de la figura 3.3 B y encontramos un área en el giro temporal superior derecho (ver figura 3.3 B y la tabla 3.2) donde la actividad aumenta al hacerlo la distancia entre el estímulo y la *recta óptima* ($p_{no\_corregido} < 0{,}001$,

$p_{\_corregido\_por\_cluster} > 0{,}05$). Más aún, la actividad media para cada punto de la grilla del voxel más activo de la zona, presenta un patrón similar al mapa de distancia euclídea a la *recta óptima*, figura 3.3 D.

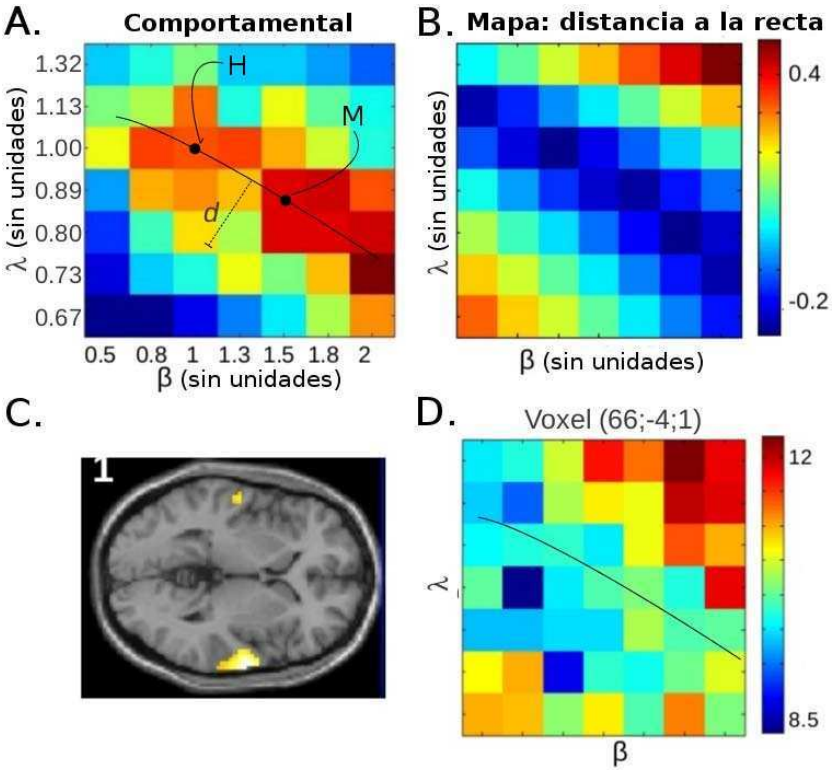

Figura 3.3: Experimento 2: codificación de la *naturalidad* de la voz. **A.** Resultado comportamental, mapa de la graduación promedio de la *naturalidad* de las vocalizaciones correspondientes a cada punto de la grilla (largo del tracto L vs. dimensiones de las cuerdas β). La escala utilizada es la misma que en el experimento 1. La línea negra ($L = -0{,}16\beta + 1{,}13$) es un ajuste por cuadrados mínimos de los puntos mejor graduados para cada β, *recta óptima*. Se indican, además, los puntos correspondientes a anatomías típicas para hombres (H) y mujeres (M) según [28]. **B.** Mapa de distancia euclídea a la *recta óptima*, restándole el valor medio. **C.** Vista transversal de las zonas que muestran una correlación ($p_{no\_corregido} < 0{,}001$, $p_{\_corregido\_por\_cluster} > 0{,}05$) entre la actividad y la distancia a la *recta óptima*, resultado de utilizar el mapa de B como contraste. **D.** Actividad media en el voxel con mayor actividad (66;-4;1) para cada punto de la grilla.

Resumiendo, encontramos que las voces son percibidas como naturales si existe cierta proporcionalidad entre las dimensiones de las cuerdas vocales y el tracto, lo que nos permite definir una recta de parámetros óptimos en este espacio anatómico. Asimismo, y en acuerdo con el resultado del primer experimento (cuanto menos natural más actividad cerebral), el estudio por imágenes revela una zona cuya actividad correlaciona con la distancia a la recta. Este resultado es compatible con el descripto por Latinus et.al. [52], la actividad aumenta al alejarnos de la recta óptima. La diferencia con aquel trabajo es que, al investigar el problema en un espacio anatómico, la codificación de la voz parece realizarse a partir de una relación entre parámetros anatómicos, en la que quedan embebidos los prototipos por género.

Por último, realizamos un análisis limitado a estímulos percibidos como naturales, cercanos a la *recta óptima*, estudiando el efecto que causa la posición en esta recta. Esto nos permite investigar la codificación de identidad, ya que distintos sitios en la recta pueden ser asociados con distintas identidades (hombres, mujeres y niños). Para esto restringimos el análisis a los puntos de la grilla graduados por arriba de la media en el experimento comportamental (máscara que se muestra en la figura 3.4) y utilizamos como regresor la posición en la *recta óptima*. El contraste utilizado se detalla en la figura 3.4 A. Dos regiones muestran actividad ($p_{no\_corregido} < 0{,}001$, $p_{\_corregido_{p}or\_cluster} > 0{,}05$): una en el giro temporal superior derecho y otra, de mayor extensión, en el izquierdo (ver figura 3.4 B y tabla 3.2). Este resultado sugiere que estas zonas estarían involucradas en el procesamiento de la identidad del hablante.

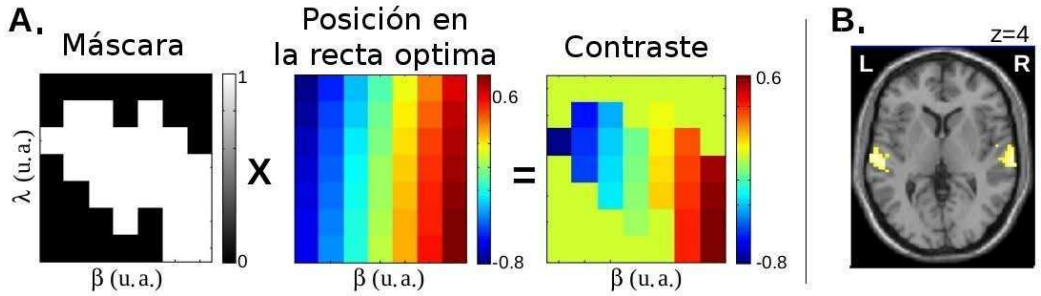

Figura 3.4: Experimento 2: codificación de identidad en el espacio anatómico. **A.** Este análisis se restringió a los sitios de la grilla puntuados por arriba de la media en el experimento comportamental (máscara). El contraste utilizado consistió en asignarle un valor a cada punto, de acuerdo con su posición en la *recta óptima* ($L = -0{,}16\beta + 1{,}13$), y un valor nulo a los puntos por fuera de la máscara. **B.** Vista transversal de las áreas activadas para el contraste descripto anteriormente, análisis restringido a la ROI y con valores: $p_{no\_corregido} < 0{,}001$ y $p_{\_corregido\_por\_cluster} > 0{,}05$.

Anatomía	Coordenadas Talairach			T-Valor	Voxels
	x	y	z		
Distancia a la recta óptima (naturalidad)					
izq STG	-57	13	1	4.2	30
der STG	66	-4	1	6.1	163
Posición en la recta óptima (identidad)					
izq STG	-60	-22	7	4.9	110
	-66	-31	13	4	
der STG	66	-19	7	4.4	90
	57	-1	-2	4.2	

Tabla 3.2: Tabla de resultados para el experimento 2. Fijando umbrales en $p_{no\_corregido} < 0{,}001$ y $p_{\_corregido\_por\_cluster} > 0{,}05$ restringidos a las regiones de interés, para los contrastes: distancia a la recta óptima, y posición en la recta óptima restringiendo el estudio a estímulos puntuados por arriba de la media. Se especifica la zona anatómica, locación estereotáxica en coordenadas Talairach, T-valor y extensión en número de voxels para cada cluster.

Capítulo 4

Cómo repercute la física vocal en la generación de estructuras del habla.

La capacidad de imitar se encuentra íntimamente ligada a la del aprendizaje, con lo cual un posible enfoque para investigar esta última es a través del estudio de pocesos miméticos. En especial, para el estudio de la imitación dentro del lenguaje, las onomatopeyas resultan el objeto natural, ya que traducen sonidos de la naturaleza a un conjunto de fonemas de la lengua. En este capítulo, estudiamos cuáles son las características acústicas que se preservan en esta transformación regida por la imitación, y como interviene la anatomía vocal en este proceso.

Definimos *imitación vocal* como la vocalización que optimiza la similitud acústica con el sonido que se busca copiar. Es decir, dentro de las restricciones que impone el sistema vocal, cuál es el sonido más parecido a un dado ruido de la naturaleza, que somos capaces de producir. Nuestro objetivo es encontrar las configuraciones de tracto correspondientes a las *imitaciones vocales* de algunos sonidos de la naturaleza, y compararlas con los perfiles de tracto de los fonemas que caracterizan las onomatopeyas que los representan. Investigando diferencias y similitudes entre estas anatomías, buscamos una

mejor comprensión de las fuerzas miméticas presentes en el lenguaje. Los resultados de este capítulo se pueden encontrar publicados en [11].

4.1. Las onomatopeyas.

Las onomatopeyas son vocalizaciones particulares que habitan el límite entre dos espacios: el de las palabras y el de los sonidos. Esta particularidad de no pertenecer estrictamente a ninguna familia presenta su contraparte a nivel neuronal: las onomatopeyas activan regiones cerebrales extensas que incluyen tanto zonas de procesado de verbos como de sonidos animales [10]. Si bien, son imitaciones de sonidos cotidianos embebidas en el espacio fonético, una comparación acústica entre sonidos y sus onomatopeyas no muestra una relación directa. Por ejemplo, mientras que los ruidos son *simples*, las onomatopeyas están formadas por una concatenación de vocales y consonantes.

En este trabajo, nos concentramos en dos pares *ruido-onomatopeya*:

1. El ruido que se produce al presionar una tecla (del *mouse*, un interruptor de luz, etc.), asociado con la onomatopeya *clik*.

2. El ruido que se genera al golpear a una puerta, asociado con la onomatopeya *toc*.

En cuanto a los ruidos, elegimos que no sean generados por un sistema vocal, de forma de extremar el esfuerzo en el proceso imitativo, y que presenten diferencias espectrales y temporales: los tipo *toc* tienen una duración aproximada de 0,1s, y un contenido espectral que decae con un perfil convexo, llegando a cero alrededor de los 5 kHz; los tipo *clik* son más cortos ($\sim$ 0,02s), con un contenido espectral cóncavo distribuido entre 0 y 6 kHz (ver figura 4.3).

En cuanto a las onomatopeyas, las tipo *clic* no presentan mucha variación fonémica a través de distintos idiomas, al ser una onomatopeya más moderna,

ligada a herramientas tecnológicas, se encuentra *globalizada*. Mientras que las tipo *toc* presentan una mayor dispersión a lo largo de distintos idiomas como se muestra en la tabla 4.1.

Lengua	**Acción**	**Onomatopeya**
Español	Golpear/tocar	to**k**
Italiano	Bussare	to**k**
Frances	Frapper	to**k**
Ingles	To knock	no**k**
Alemán	Klopfen	**k**lopf
Polaco	Pukac	pu**k**
Japonés	Takete	**k**on
Holandés	Kloppen	**k**lop
Hungaro	kopogtató	**k**op
Tailandés	kor	**k**o**k**
Bulgaro	blŭskam	chu**k**

Tabla 4.1: Onomatopeyas asociadas a la acción de golpear a la puerta para distintos idiomas. La consonante $[k]$ en un contexto vocálico $[o]$ o $[u]$ se mantiene estable a lo largo de los distintos idiomas. Otros ejemplos se encuentran disponibles en internet, por ejemplo en $http://en.wikipedia.org/wiki/Cross-linguistic_onomatopoeias$, muy pocos casos quedan por fuera de la regla.

Es interesante notar que la consonante oclusiva $[k]$ y las vocales $[o]$ y $[u]$ se mantienen estables a lo largo de los distintos idiomas. Asimismo, la consonante $[k]$ está presente, también, en la otra onomatopeya estudiada, pero con un contexto vocálico distinto, $[i]$. Este fonema estable a lo largo de los idiomas, y onomatopeyas, embebido en distintos contextos vocálicos nos da una pista de dónde podría estar codificada la imitación, según las semejanzas y diferencias acústicas de los sonidos asociados a cada tipo de onomatopeya.

4.2. Exploración del espacio fonémico

Como ya mencionamos, los sonidos que conforman el habla pueden separarse en dos grupos: *voceados* y *no voceados*. La fuente del primer grupo

son las oscilaciones de las cuerdas vocales, mientras que la del segundo es una turbulencia, consecuencia de una constricción en el tracto. En esta sección estudiamos las configuraciones de tracto correspondientes a fonemas de ambos grupos: de vocales y de fricativas. Para esto utilizamos el modelo descripto en la sección 2.2.1, fijando en 10 el número de tubos que componen el sistema, todos con longitud l ($L_{total} = 10l$) y áreas $a_1, a_2, ..., a_{10}$. Para las cuerdas vocales elegimos un modelo de tipo flameo según la sección 2.1.

Este modelo nos permite sintetizar una gran variedad de sonidos del habla a partir de un conjunto de parámetros anatómicos. Sin embargo, nosotros nos enfrentamos al problema inverso: dado un sonido rastreamos los parámetros anatómicos $\{l, a_1, ..., a_10\} \equiv \{l, A\}$, que permiten producir el sonido vocal más similar. Más precisamente, a partir de grabaciones experimentales, de diferentes hablantes, pronunciando distintos fonemas, buscamos la configuración de tracto capaz de generar estos últimos. Para realizar esta búsqueda recuperamos la estrategia detallada en la sección 2.2.2, para encontrar las configuraciones del tracto vocal de un ave, a partir del espectro de su canto. Aquí, adaptamos el algoritmo para buscar los perfiles de tracto asociados a distintos fonemas, a partir del espectro de los mismos (*espectro objetivo*). Elegimos un algoritmo genético, dado que permite una exploración eficiente de un espacio multidimensional de parámetros y nos devuelva una familia de tractos vocales compatibles con el sonido experimental. Para más detalle de cómo fue implementado este algoritmo ver apéndice B.

4.2.1. Ajuste de vocales

Primeramente, buscamos las configuraciones de tracto correspondientes a las vocales del español. Esto nos permite testear el modelo, ya que existen datos experimentales de estos perfiles. Fijamos como *espectro objetivo* para cada vocal el promedio de 10 espectros, correspondientes a distintos hablantes nativos de español rioplatense, y buscamos con el algoritmo los tractos correspondientes a cada una. Los resultados obtenidos son los que se muestran en la figura 4.1.

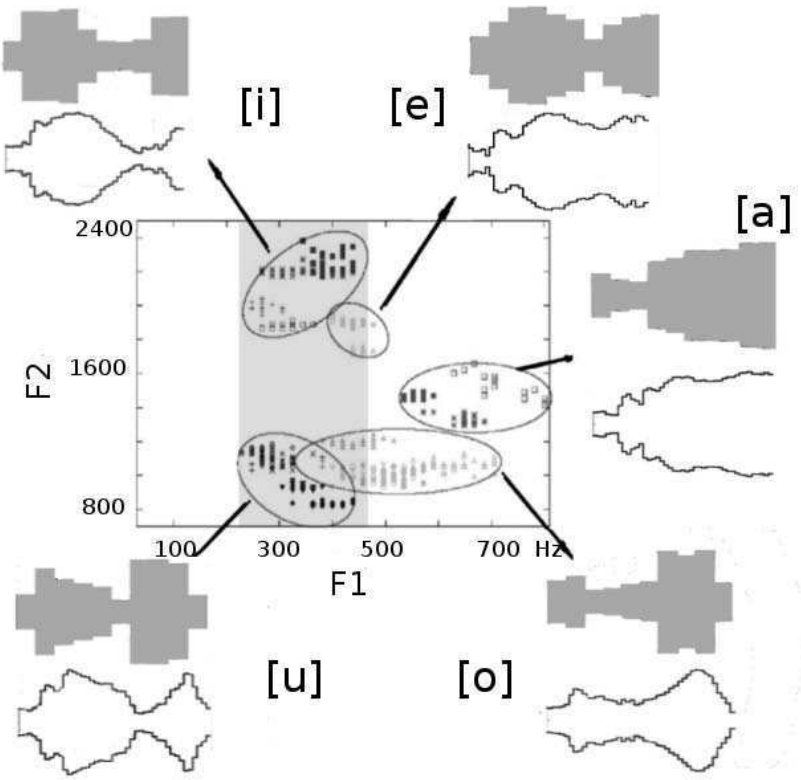

Figura 4.1: Anatomía de las vocales. Cada punto en el gráfico corresponde a una vocal ($\sim 100ms$) extraída de una base de datos, conformada por vocalizaciones de 20 hablantes nativos de español rioplatense de diferente sexo y edad. Realizamos la transformada de fourier de cada muestra, y graficamos la primera y segunda formante. Los puntos se pueden separar en 5 grupos según las 5 vocales del español. Los tractos en negro corresponden a datos anatómicos reportados en [37]. Las figuras grises representan un promedio de las diez mejores configuraciones halladas con nuestro algoritmo. Los diámetros de cada uno de los diez tubos que las componen se encuentran en el apéndice B.1.

En la figura mostramos 40 vocales producidas por 20 hablantes nativos de español rioplatense, en el espacio de las dos primeras formantes (espacio perceptualmente relevante, como mencionamos en el capítulo 1), y vemos que forman 5 grupos. Para cada grupo se presentan dos formas de tracto: Los contornos negros son datos experimentales reportados en [37], obtenidos

a partir de imágenes por resonancia magnética. En ese trabajo un hablante inglés es escaneado mientras pronuncia sostenidamente distintas vocales. Las formas grises son las obtenidas con nuestro algoritmo. Podemos ver que ambos perfiles son compatibles, teniendo en cuenta la diferencia de idiomas y de discretización (los datos de [37] corresponden a 44 tubos mientras que nosotros usamos 10).

4.2.2. Coarticulación de fricativas

Una vez validado el algoritmo genético para la búsqueda de *tractos vocálicos*, investigamos su rendimiento para consonantes. En esta sección exploramos la anatomía vocal que le corresponde a la consonante [x] en distintos contextos vocálicos. Dos preguntas naturales serían: *porqué* [x] y *porqué en distintos contextos vocálicos?*

La respuesta a la primer pregunta es: porque es muy parecida a la consonante [k], fonema estable a lo largo de las onomatopeyas que queremos estudiar. A diferencia de [x], [k] no pertenece a la familia de las fricativas, es una consonante plosiva. Este tipo de consonantes es ocasionada por una oclusión completa del tracto y su posterior liberación, produciendo un aumento abrupto en la intensidad del sonido (denominado ataque) y su posterior decaimiento. Sin embargo, la oclusión del tracto para la [k] se produce en el mismo sitio que la constricción de la [x], y mas aún la parte estable del espectro de esta plosiva es indistinguible del de la fricativa. En este trabajo excluimos el ataque de [k] y la simulamos haciendo la convolución entre su envolvente y el sonido correspondiente a simular una fricativa estable [x]. De esta manera reproducimos casi todas las características espectrales y temporales de la plosiva evitando la dificultad de simular la oclusión.

En cuanto al segundo interrogante: porque la anatomía, y la acústica, de la consonante varían según las vocales que la rodean, efecto conocido como coarticulación. Durante el habla, los gestos articulatorios son parcialmente

transmitidos de un fonema a otro. La configuración de tracto de las consonantes fricativas queda determinada por características propias y de su contexto vocálico, efecto con consecuencias perceptualmente relevantes [64].

Para estudiar este efecto grabamos 10 hablantes nativos de español pronunciando la fricativa $[x]$ en distintos contextos vocálicos, registramos dos tipos de pares vocal-consonante: $[xa, xe, xi, xo, xu]$ y $[ax, ex, ix, ox, ux]$. Luego, para estudiar el efecto de la vocal extrajimos de los audios solamente la consonante, les realizamos la transformada de fourier, promediamos 10 muestras de cada tipo y fijamos estos promedios como *espectros objetivo*. Los resultados obtenidos se muestran en la figura 4.2.

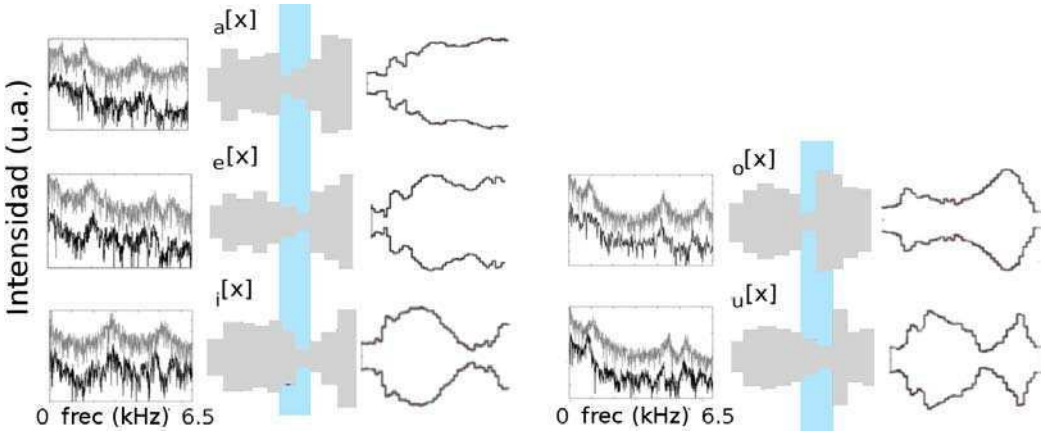

Figura 4.2: Coarticulación de fricativas. En gris la reconstrucción obtenida con nuestro algoritmo del tracto vocal correspondiente a la fricativa $[x]$ en distintos contextos vocálicos: $[a]$, $[e]$, $[i]$, $[o]$ y $[u]$. La banda celeste indica la posición de la constricción, en todos los casos cercana a la mitad del tracto (zona velar). Los datos para los diámetros de cada reconstrucción se encuentran en el apéndice B.1 En negro los datos experimentales presentados en [37] para la vocal correspondiente. Cada uno de los tractos grises, conjunto los siguientes valores para los parámetros de la fuente (ecuación 2.15) $(\sqrt{\kappa}/2\pi 10^{-3}, \beta 10^{-6})$: $[x]_a \to (3{,}5; 3{,}5)$, $[x]_e \to (3{,}8; 1{,}8)$, $[x]_i \to (3; 1{,}8)$, $[x]_o \to (1{,}55; 7{,}3)$ y $[x]_u \to (1{,}24; 6{,}2)$, generan sonidos cuyo espectro es el que se muestra en negro. Mientras que el espectro gris corresponde al sonido experimental de cada fricativa.

En la figura 4.2 mostramos los tractos obtenidos para la $[x]$ acompañada por distintas vocales, los espectros reales mas los sintéticos y nuevamente incluimos los datos de MRI para cada vocal. Como esperábamos, todos los tractos muestran una constricción a nivel velar (como resalta la marca de agua en la figura), que es justamente la signatura de la $[x]$, impuesta sobre un perfil que similar al de la vocal contigua. Si bien las consonantes efectivamente heredan la forma de la vocal vecina, no encontramos diferencias en las anatomías que dependan del orden relativo de la vocal, es decir que da lo mismo *consonante-vocal* o *vocal-consonante*. Es por eso que a lo largo del trabajo nombramos a la consonante coarticulada con una vocal v como $[x]_v$, sin importar el orden.

4.3. Anatomía de las onomatopeyas

En esta sección nos concentramos en la *imitación vocal* de los ruidos, asociados a las onomatopeyas elegidas. Buscamos los parámetros que minimizan las diferencias espectrales entre el sonido a imitar y la voz generada por el modelo. Más precisamente, generamos dos *espectros objetivo*: uno promediando 10 espectros golpes (ruidos asociados a la onomatopeya *toc*), otro 10 interruptores (ruidos asociados a la onomatopeya *clic*). Aplicando el algoritmo genético a estos *espectros objetivo* obtuvimos los tractos que se muestran en la figura 4.3. Esta figura permite comparar las anatomías correspondientes a las *imitaciones vocales* de los ruidos, con las de los fonemas más estables de las onomatopeyas que las representa.

Como esperábamos, para ambas imitaciones, el algoritmo seleccionó sonidos no voceados. Debido a la naturaleza ruidosa de los espectros a imitar se obtienen resultados mucho más satisfactorios utilizando como fuente la turbulencia a la salida de la constricción, que las oscilaciones de las cuerdas. Esto nos indica que, de existir una clave mimética, estaría embebida en la consonante $[k]$. Comparemos, entonces, las imitaciones con las coarticulaciones de la $[k]$ en las distintas onomatopeyas.

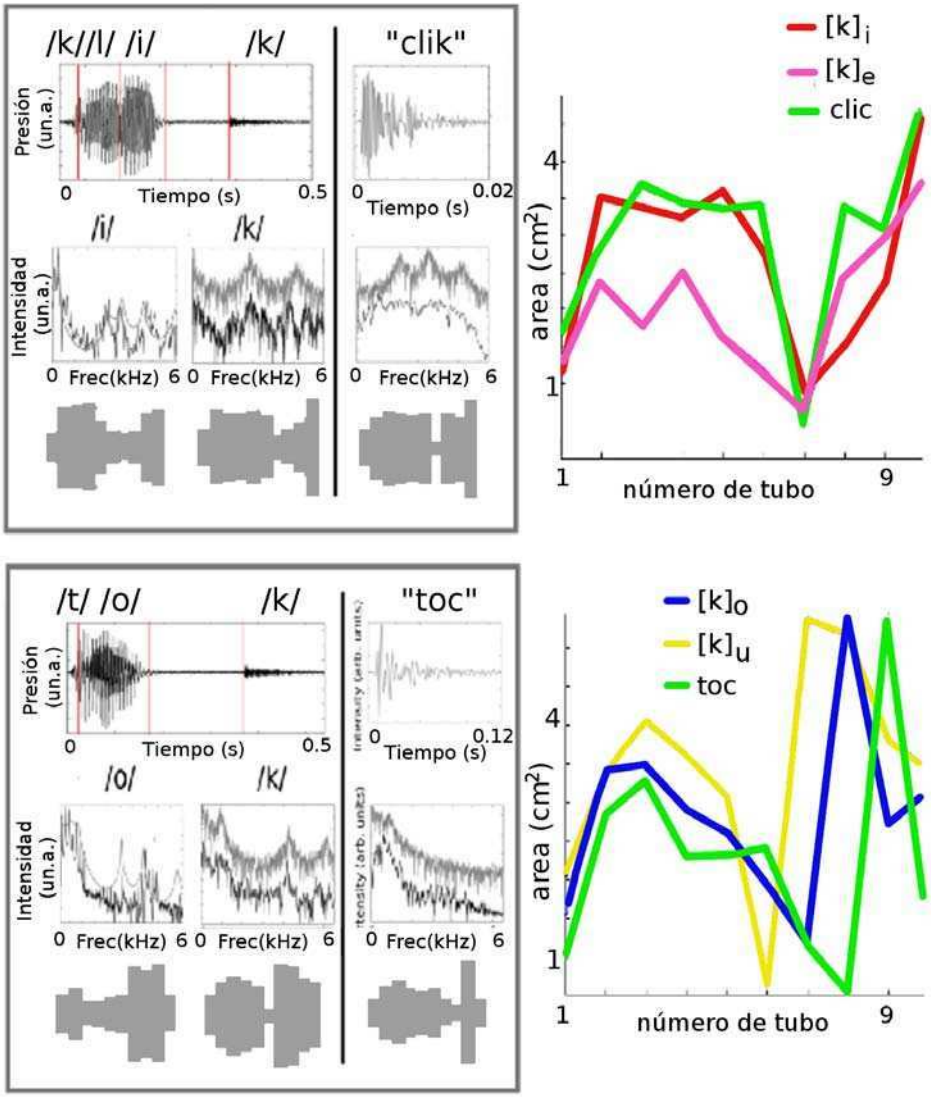

Figura 4.3: Anatomía de las onomatopeyas. Comparamos la serie temporal de la presión acústica, espectro y perfil del tracto vocal de las onomatopeyas con los sonidos que representan. Arriba los sonidos tipo click abajo tipo toc. Las funciones por fuera del recuadro muestran las funciones de área (área en función del número de tubo) para el tracto que mejor *imita* el sonido, conjunto al de las consonantes estables a lo largo de los idiomas para cada onomatopeya.

Desde un punto de vista fonético-articulatorio, la manera natural de describir el perfil del tracto vocal correspondiente a un dado fonema es según

la posición los articuladores: la apertura de la mandíbula, la posición de la lengua y el redondeamiento de los labios [3]. Esto, se puede traducir a nuestro sistema de tubos de la siguiente manera:

1. La apertura de la mandíbula y la posición de la lengua, están relacionadas con el ancho máximo y su posición en el arreglo de tubos.

2. La redondez de los labios la determina el ancho de los últimos dos tubos (por ejemplo: los perfiles de la $[o]$ y de la $[u]$, en la figura 4.1, son los únicos que tienen los labios redondeados).

Bajo esta descripción, la imitación del golpe es redondeada como $[k]_o$, mientras que la del interruptor es no redondeada al igual que $[k]_i$. Más allá de esto, existen discrepancias entre las consonantes coarticuladas de las onomatopeyas y las imitaciones. Particularmente, vemos que estas últimas presentan perfiles de tracto con variaciones más abruptas, quizás esto se deba a que nuestro modelo no impone ningún tipo de restricción anatómica. Es decir, el área de cada tubo es independiente de la del resto. Sin embargo, trabajos previos, descriptos en la sección 2.2.3, muestran que esto no es así, ya que existen restricciones anatómicas impuestas por los articuladores (mandíbula, lengua y labios).

A continuación, examinamos la plausibilidad de estas configuraciones, empleando el mismo método que Story et. al [46,65]. Recapitulando, en esos trabajos encuentran que las funciones de área, $A(x)$ (área en función de la distancia a la glotis), de varios fonemas obtenidas experimentalmente, pueden ser aproximados por $A^{PCA}(x) = \Omega(i) + q_1\varphi_1(i) + q_2\varphi_2(i)$, fijando los coeficientes q_1 y q_2 apropiados. Donde, Ω es un tracto neutro que depende de cada individuo y φ_1 y φ_2 son los dos primeros modos de un análisis de componentes principales, realizado sobre los datos experimentales. Son precisamente los dos primeros modos los que representan las restricciones impuestas por los articuladores.
Siguiendo esta idea, realizamos un análisis de componentes principales sobre las funciones de área (área en función del número de tubo) obtenidas con el algoritmo, para las vocales del español y las fricativas $[x]$ en distintos

contextos vocálicos. Asumiendo que, por pertenecer a fonemas, estas configuraciones son en efecto viables. Luego, volvimos a buscar las imitaciones, pero penalizando en nuestro algoritmo la diferencia entre el tracto propuesto y su aproximación utilizando los dos primeros componentes principales (ver apéndice B). En el panel inferior de la figura 4.4 se muestran estos resultados. Curiosamente, encontramos que las mejores imitaciones, sujetas a las *restricciones anatómicas*, se organizan en el espacio bidimensional de las dos primeras componentes $(q_1; q_2)$ de forma tal que: el tracto correspondiente a $[k]_o$ es el más cercano al que imita los golpes, mientras que el de $[k]_i$ es el más próximo al que imita al *clik*.

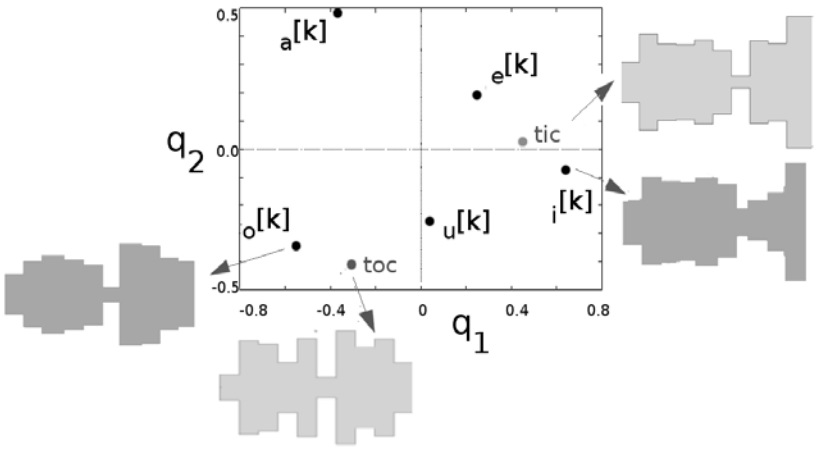

Figura 4.4: Representación de las configuraciones de tracto halladas con el algoritmo, en un espacio bidimensional, $(q_1; q_2)$, dado por los dos primeros modos de un análisis de componentes principales, realizado sobre las funciones de área para las vocales del español y las fricativas $[x]$. Las distancias entre el tracto de la mejor imitación al sonido tipo *toc* y las coarticulaciones de $[k]$ en este espacio: $d[k]_a = 0{,}9$, $d[k]_e = 0{,}82$, $d[k]_i = 1.$, $d[k]_o = 0{,}26$ y $d[k]_u = 0{,}38$. Las distancias a la imitación tipo *clik*: $d[k]_a = 0{,}95$, $d[k]_e = 0{,}26$, $d[k]_i = 0{,}21$, $d[k]_o = 1{,}07$ y $d[k]_u = 0{,}5$.

Además de trabajar en el espacio euclídeo de los coeficientes (q_1, q_2), sometimos nuestro resultados a test perceptuales. Realizamos dos tipos de experiencias, donde 20 sujetos debían puntuar un conjunto de archivos de audio,

de acuerdo a su similitud con el sonido de golpear a una puerta, o al que produce el interruptor de luz. En el primer tipo de experiencia, los audios corresponden a 5 consonantes oclusivas [k] provenientes de distintas vocales, aisladas del resto de la vocalización, y pronunciadas por un mismo sujeto masculino durante habla normal. En el segundo, corresponden a sus versiones sintéticas generadas con nuestro modelo, como se describe en la sección 4.2.2, y las síntesis correspondientes a los tractos de las mejores imitaciones de cada ruido. Los resultados perceptuales se muestran en la figura 4.5.

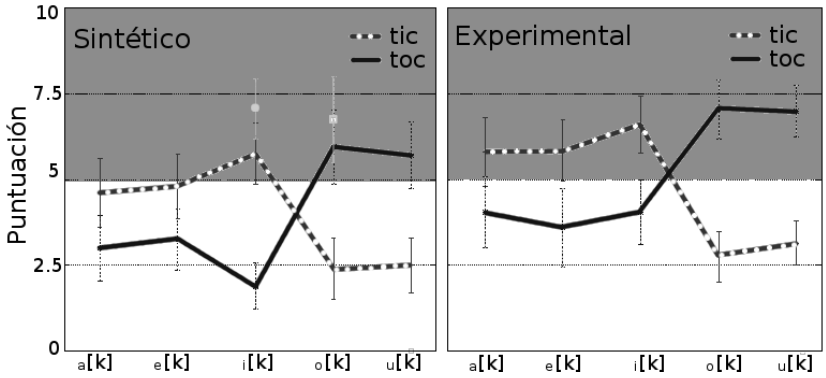

Figura 4.5: Asociación perceptual entre la consonante [k] coarticulada y sonidos tipo *toc* o *clik*. Evaluamos la similitud entre las distintas coarticulaciones de $[k]_v$ y los sonidos tipo *toc* (línea sólida) y *clik* (línea puntuada). Se les indicó a los participantes que puntuaran los sonidos entre 1 (nada de parecido) y 10 (asociación perfecta al tipo de sonido). Izquierda, graduación promedio de 20 participantes a 7 sonidos sintéticos: 5 [k] coarticuladas y las mejores imitaciones (puntos en grises), utilizamos los parámetros que se muestran en la tabla B.1. Los sonidos fueron modulados por la envolvente de una [k] experimental. Derecha, graduación promedio de 20 participantes a 5 sonidos sintéticos: consonantes oclusivas [k] aisladas, producidas en distintos contextos vocálicos. Extraídas de una grabación de habla fluida de un hablante nativo del español, de 40 años de edad y sexo masculino. Los sujetos que puntuaron los sonidos sintéticos son distintos de los que puntuaron los reales.

Los resultados para las vocalizaciones reales, panel derecho de la figura 4.5, muestran que: tomando la nota media (5) como línea de corte, hay un gru-

po de fonemas asociado a cada ruido. El grupo $\{[k]_a; [k]_e; [k]_i\}$ es asociado a ruidos tipo *clic* (Wilcoxon test $p < 4 \cdot 10^{-11}$), mientras que $\{[k]_o; [k]_u\}$ lo es a ruidos tipo *toc* (Wilcoxon test $p < 8 \cdot 10^{-11}$). Si bien no existen diferencias significativas entre las puntuaciones de los fonemas dentro de cada grupo, el fonema más asociado al *clik* es $[k]_i$, con un valor medio de $\bar{x} = 6{,}60$ ($s_{20} = 1{,}64$), mientras que para el *toc* lo es $[k]_o$, con $\bar{x} = 7{,}05$ ($s_{20} = 1{,}73$). En cuanto al experimento con los sonidos sintéticos, panel izquierdo de la figura 4.5, muestran la misma tendencia que los del caso anterior. Recordemos que nuestro modelo para sintetizar fricativas, y su posterior adaptación a oclusivas, fue diseñado para dar cuenta de las características espectrales de estas consonantes. Estos resultados sugieren que hay propiedades que son perceptualmente relevantes, y que no están consideradas en nuestro modelo. Sin embargo, las puntuaciones más altas corresponden, nuevamente, a $[k]_i$ con $\bar{x} = 5{,}75$ ($s_{20} = 1{,}77$) y $[k]_o$ con $\bar{x} = 5{,}95$ ($s_{20} = 2{,}16$). Por otro lado, los sonidos sintetizados con los tractos correspondientes a las mejores imitaciones de cada ruido, son los que reciben las puntuaciones más altas, con $\bar{x} = 7{,}05$ ($s_{20} = 1{,}76$) para *clic* y $\bar{x} = 6{,}75$ ($s_{20} = 2{,}49$) para *toc*, diferenciándose del resto de los estímulos ($p < 0{,}035$).

Estos resultados sugieren que es a través de los fonemas más estables a lo largo de distintos idiomas, que las onomatopeyas estudiadas logran *imitar* los sonidos que representan. Esta conexión parece estar tanto a nivel de producción vocal como perceptual. Desde el punto de vista de la producción, las configuraciones de tracto vocal de las consonantes coarticuladas $[k]_i$ y $[k]_o$ se encuentran próximas a las que maximizan la similitud espectral con los sonidos tipo *clik* y *toc*. Por otro lado, perceptualmente estos fonemas, aislados del resto de la vocalización, son asociados con los ruidos ligados a la onomatopeya de la que forman parte. Es importante notar que tanto la consonante como la vocal, aún si se extrae esta última, son necesarias para el proceso de mímesis, ya que es en la consonante coarticulada donde se encuentra la clave mimética

4.4. Interacción imitación-sinestesia

El resultado anterior no ofrece una descripción acabada del proceso de formación de las onomatopeyas. Como cualquier otra palabra, las onomatopeyas contienen elementos que se fueron modificando a lo largo de la historia, y exceden la pura imitación [66]. Si bien, interpretamos la elección de los fonemas más estables a lo largo de distintas lenguas como un fenómeno de imitación directa, este no explica la estructura completa de las onomatopeyas. Los fonemas que presentan variaciones permanecen inexplorados. En esta sección estudiamos la generación de onomatopeyas de una forma más global, teniendo en cuenta el total de los fonemas que las componen, y sumando al efecto imitativo el de sinestesia.

La sinestesia es un fenómeno por el cual un estímulo es percibido por un sentido que no es el *apropiado*, como por ejemplo: percibir sensaciones gustativas al escuchar una melodía [67]. Existe evidencia de que procesos relacionados con la sinestesia participan en la generación de estructuras del lenguaje. En particular, claves visuales como tamaño, forma y brillo influyen en la elección de fonemas a la hora de nombrar un objeto [68]. Un ejemplo de este fenómeno es el que se conoce como el *efecto Bouba y Kiki*, según el cual formas puntiagudas son nombradas con consonantes oclusivas y vocales *i,e*, mientras que formas redondeadas lo son con consonantes aproximantes y vocales *o,u*. Efecto que se mantiene en niños y adultos no alfabetizados [69] y a lo largo de distintas lenguas [70].

Para estudiar la interacción de este fenómeno con el de imitación, diseñamos un experimento donde hablantes de distintas lenguas deben representar, mediante una onomatopeya creada por ellos, *pseudo-onomatopeya*, un dado estímulo. En búsqueda de resultados estables, independientes de la lengua materna, nos concentramos en sujetos pertenecientes a dos grupos de distintas familias lingüísticas: hablantes nativos del francés, lengua romance de la familia indoeuropea; y hablantes nativos de una lengua japónica. El protocolo experimental se encuentra acabado, pero no la adquisición y el análisis

de los datos, en esta sección mostramos resultados parciales correspondientes al primer grupo, el de hablantes nativos de frances.

4.4.1. Métodos

Como mencionamos anteriormente, el experimento consistió en un conjunto de sujetos inventando una onomatopeya para representar distintos estímulos. El experimento se separó en tres bloques según el o los sentidos estimulados: bloque auditivo (1), visual (2) y audiovisual (2). La figura 4.6 explica este procedimiento. Las piezas constitutivas de los estímulos de todos los bloques son tres acciones básicas: deslizamiento, vibración y choque. Los estímulos auditivos son sonidos compatibles con estas acciones, cada uno en dos versiones: una de contenido espectral alto y otra de bajo. Los visuales son videos, sin audio, donde: formas puntiagudas (tipo *kiki*), o redondeadas (tipo *bouba*); grandes, o pequeñas; representan las distintas acciones. Por último, los estímulos audiovisuales corresponden a videos, con audio, de las distintas interacciones, realizadas por las distintas formas, con el sonido correspondiente a la acción en sus dos variantes de contenido espectral.

Figura 4.6: Estímulos diseñados para investigar la interacción mimesis-sinestesia. El bloque auditivo tiene 6 estímulos: 3 (acciones) X 2 (contenidos espectrales). En el bloque visual son 12: 3 (acciones) X 2 (formas) X 2 (tamaños). El audiovisual consta de 24 estímulos: 3 (acciones) X 2 (formas) X 2 (tamaños) X 2 (contenidos espectrales).

Se le indicó a un grupo de 17 hablantes nativos de francés que atendieran

a todos los bloques y que pronuncien, en su lengua materna, una *pseudo-onomatopeya* al final de cada estímulo. Luego, dos lingüistas se ocuparon de desgrabar los audios, generando un corpus de onomatopeyas trascriptas al alfabeto fonológico internacional.

4.4.2. Resultados parciales

Estudiamos en el corpus el efecto de las formas, el contenido espectral y el tipo de acción en la elección de fonemas, para los tres bloques. Cada fonema fue descripto según su ubicación en el alfabeto fonémico internacional. Como muestra la figura 1.3 del capítulo 1, esta representación distingue vocales de consonantes. Las primeras son descriptas según la posición de la lengua, apertura de la mandíbula y redondez de los labios. Mientras que las segundas, según el lugar donde se genera la constricción y el grado de la misma. Esto nos permitió buscar correlaciones entre los estímulos y las características de los fonemas elegidos para representarlos. La figura 4.7 muestra los resultados para las características que encontramos relevantes.
Analicemos primero el efecto del contenido espectral y de las formas sobre la elección de los fonemas, sin tener en cuenta a que acción corresponden:

1. Bloque auditivo, panel **A** de la figura 4.7: el contenido espectral incide sobre la elección de la vocal. Contenidos espectrales bajos son representados con vocales posteriores, mientras que contenidos altos lo son con anteriores. Este efecto está relacionado con la imitación, ya que la intensidad espectral de las vocales anteriores se centra en frecuencias más altas que la de las posteriores.

2. Bloque visual, panel **B** de la figura 4.7: recuperamos el efecto *Bouba Kiki*. Formas redondeadas correlacionan con vocales redondeadas, y hay una correlación entre la forma y el lugar de articulación de la consonante: formas puntiagudas son representadas por consonantes articuladas en zonas más posteriores (coronales o dorsales) mientras que las redondeadas prefieren labiales. La forma fija la columna a la que pertenece la consonante en el alfabeto fonémico internacional, ver figura 1.3. No se encuentra ninguna correlación entre fonemas y tamaños.

3. Bloque audio visual: desaparece el efecto sinestésico, quedando solamente el mimético. Al igual que en el bloque auditivo hay una interacción entre la elección de vocales anteriores o posteriores y el contenido espectral del sonido del estímulo.

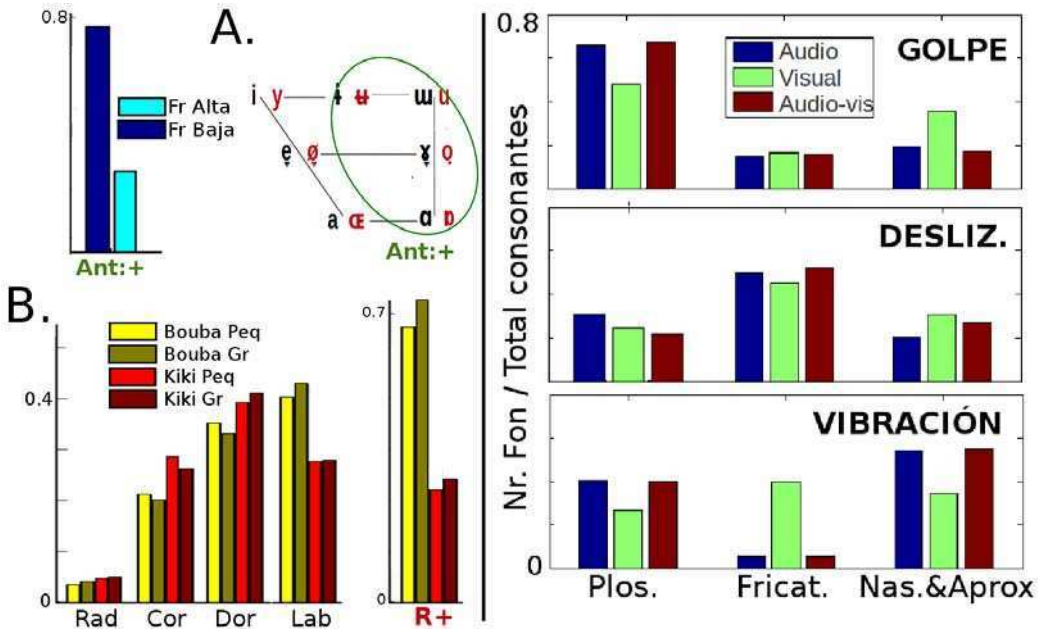

Figura 4.7: Resultados parciales, correspondientes a 17 hablantes nativos de francés. El eje vertical de los histogramas representa el número de consonantes (vocales) de un dado tipo dividido el total de consonantes (vocales) del bloque. Paneles izquierdos: **A.** Bloque auditivo: Efecto del contenido espectral del estímulo sobre la elección del fonema. Se indican en verde las vocales anteriores:+ y en rojo redondeadas:+. **B.** Bloque visual: Efecto de la forma (*Bouba/Kiki Grande/Pequeño*) en la elección de consonantes. El eje horizontal indica el sitio de articulación de la consonante, columnas de la *tabla IPA* esquematizada en la figura 1.3, en el primer gráfico, y vocales con labios redondeados en el segundo. Panel derecho: Efecto del tipo de acción sobre la elección de consonantes. En azul resultados del bloque auditivo, en verde del visual y bordo del audiovisual. Los tres sitios del eje horizontal corresponden a consonantes plosivas el primero, fricativas el segundo y el tercero integra nasales, aproximantes y róticas.

Estudiamos, luego, el efecto del tipo de accción sobre la elección de fonemas en cada bloque. Los resultados se muestran en el panel **C** de la figura

4.7. Vemos que el tipo de acción tiene un efecto sobre la elección de la consonante, independiente del tipo de bloque, salvo en la vibración resultado que discutiremos posteriormente. La acción determina el tipo de fuente de la consonante, la fila del alfabeto fonémico internacional esquematizado en la figura 1.3. Esto puede ser identificado como un efecto de imitación [71]. Los deslizamientos son representados fricativas, una fuente ruidosa sostenida en el tiempo. Las oclusivas representan golpes, ambos sonidos son caracterizados por un espectro ruidoso de aumento repentino y corta duración. Y las vibraciones eligen nasales, róticas o aproximantes sonidos voceados (con fuentes periódicas). El hecho que el bloque visual no siga la misma tendencia que el resto en los estímulos correspondientes a vibraciones lo adjudicamos a un error en la generación del estímulo. La frecuencia del movimiento no resulta adecuada para que sea percibido como una vibración.

Capítulo 5

Representación discreta de los gestos motores de vocales y consonantes oclusivas

Las unidades básicas del lenguaje son los fonemas, unidades sonoras mínimas del habla que permiten distinguir significado. Por ejemplo $[p]$ y $[d]$ son fonemas, dado que $/pan/$ y $/dan/$ transmiten distintos mensajes. Es la plasticidad del tracto vocal, su capacidad de cambiar rápida y abruptamente su perfil, lo que nos permite transmitir información a través de la concatenación de fonemas, dando origen a las palabras. Como ya mencionamos a lo largo de esta tesis, en el caso de las vocales, las oscilaciones de las cuerdas vocales actúan como fuente sonora y son los distintos perfiles de tracto lo que nos permite distinguirlas, generando distintos filtros para la fuente. Asimismo, la identidad de las consonantes viene dada por el grado y ubicación de una constricción u oclusión en el tracto vocal.

Curiosamente, si bien el espacio de las configuraciones que puede tomar el tracto vocal es continuo, no lo es el fonémico. El conjunto de sonidos del habla es finito. Más aún, existe una descripción de todos ellos a partir de propiedades anatómicas básicas: el alfabeto fonológico internacional, ya mencionado a lo largo de esta tesis y esquematizado en la figura 1.3.

Por otro lado, el avance de nuevas tecnologías, permitió en estos últimos años abordar el problema del planeamiento, o control motor de los articuladores del habla a partir de mediciones directas de actividad cerebral. Trabajos recientes muestran que tanto vocales, como consonantes, son codificadas por la actividad selectiva de algunos grupos neuronales [12, 13] que controlan la dinámica de los distintos articuladores.

Tanto al comienzo como al final del *proceso de producción* del habla la información es discreta: discreta es la actividad neuronal de las áreas que gobiernan la dinámica de los articuladores del tracto vocal, y discreto es el número de fonemas que constituyen los distintos idiomas. Sin embargo, la dinámica del tracto vocal, intermediaria entre el espacio neuronal y el fonémico, habita en un espacio continuo. En este capítulo, buscamos una medición directa de la periferia que pueda servir de paso intermedio entre las mediciones neuronales y espacio fonológico. Mas precisamente, registramos los movimientos de los articuladores del tracto durante el discurso, y los representamos en un espacio discreto que nos permite distinguir las vocales y las consonantes oclusivas del español. Parte de los resultados y procedimientos de esta sección se encuentran publicados en [16].

5.1. Dispositivo experimental

El primer desafío consistió en encontrar un método de medición que sea capaz de relevar la dinámica de distintos puntos del tracto vocal superior durante el habla, sin interferir con la articulación de los fonemas. Esta dificultad fue superada utilizando detectores de efecto hall e imanes, montados sobre una prótesis bucal odontológica como se muestra en el panel A de la figura 5.1. Estos sensores miniatura ($4mm\,x\,2mm\,x\,1mm$) detectan la intensidad de campo magnético en la dirección perpendicular a su eje, lo que permite tener una medición indirecta de la distancia y orientación entre el imán y el detector. Para una descripción más detallada del dispositivo ver apéndice C.

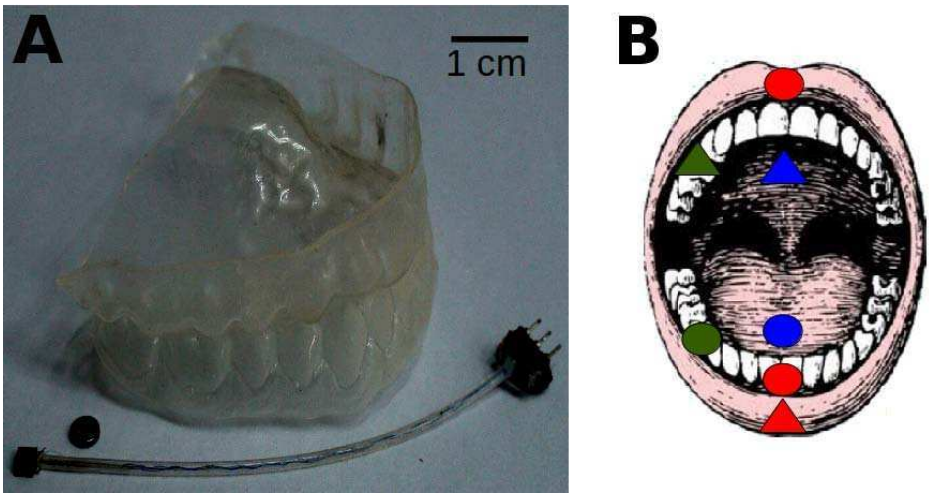

Figura 5.1: Dispositivo experimental. **A.** Se muestran los distintos elementos que conforman el equipo montado sobre el tracto bocal superior: prótesis odontológica, sensor de efecto hall e imán. Para interferir lo menos posible en la dicción los sensores son conectados a la electrónica externa de adquisición de datos, mediante cables de electrofisiología. **B.** Esquema que indica la posición de los transductores (triángulos) e imanes (círculos). En rojo: posiciones ajustadas para monitorear la clausura y proturción de los labios. Azul: ubicación que optimiza la detección de la posición de la lengua en la cavidad bucal. Verde: Par imán-detector que obtiene la apertura de la mandíbula.

Una vez escogida la herramienta de medición, buscamos determinar el número mínimo de pares emisor-detector y su ubicación adecuada, para maximizar la detección de fonemas. En este trabajo nos concentramos en un subgrupo de fonemas del español conformado por todas las vocales y consonantes oclusivas sordas de esta lengua: */k/ /t/ y /p/*. Luego de varios ensayos se escogió la disposición de imanes y sensores que muestra en el panel B de la figura 5.1, y que puede ser descripta según:

Labios Se adhirió el sensor en el centro del labio inferior y dos imanes, uno en el labio superior y otro entre los incisivos centrales inferiores sobre el protector bucal. El primer imán permite monitorear la clausura de los labios, mientras que el segundo refleja la proturción del labio inferior. La orientación de los imanes se fijo de forma tal que: la señal aumenta

al acercar los labios y al alejar el labio inferior de los incisivos inferiores (proturción).

Mandíbula En este caso, tanto el imán como el transductor se sujetaron al protector dental inferior y superior, respectivamente, entre el canino y el primer molar.

Lengua El imán se adhirió a la lengua, aproximadamente a 20 mm de la punta, mientras que el detector se encuentra en el protector sobre el paladar aproximadamente a 10 mm detrás de las paletas.

Participaron del experimento 3 sujetos (1 mujer, 2 hombres) con edades entre 29 y 40 años, todos hablantes nativos de español rioplatense. Se realizaron dos sesiones de adaptación, y siete de adquisición de datos a lo largo de dos meses. Cada sesión consiste en el registro de las señales de los tres sensores, mas el del audio de las vocalizaciones correspondientes.

5.2. Coordenadas motoras discretas

Si bien buscamos unas descripción que permita distinguir los gestos motores de distintos fonemas durante el discurso natural, primero estudiamos las vocales aisladas, y luego extendemos el espacio integrando la coarticulación de consonantes oclusivas.

5.2.1. Vocales

Comenzamos estudiando las vocales por varias razones: representan los bloques fundamentales del habla sobre el que se superponen las consonantes [37], y pueden pronunciarse de forma aislada y sostenida en el tiempo. Se realizaron, para cada uno de los sujetos, 4 sesiones a lo largo de un mes. Cada sesión consta de la pronunciación de 20 vocales aisladas, arrancando y terminando la vocalización en la posición natural de reposo (*boca cerrada*) de cada sujeto.

Al analizar las señales observamos que son del tipo activación-inactivación: estableciendo un umbral por detector, de manera que cada señal puede tomar dos valores 1 o 0 (supera o no supera el umbral), logramos diferenciar las 5 vocales del español, como se ejemplifica en el panel A de la figura 5.2.

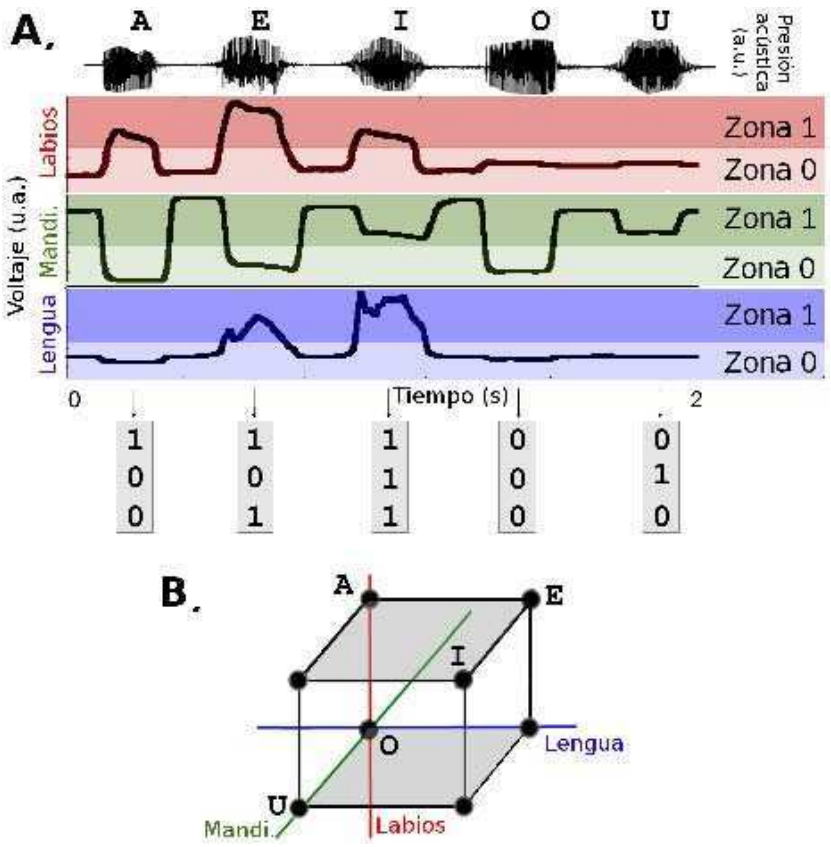

Figura 5.2: Descripción de las vocales en un espacio binario de 3 dimensiones. **A.** Un set completo de señales temporales para las 5 vocales del español: panel superior presión acústica, le siguen las señales de los distintos detectores labios en rojo, mandíbula en verde y lengua en azul. La intensidad de colores remarca el estado en el que se encuentra cada articulador. Por debajo el vector que representa cada vocal. **B.** Espacio tridimensional donde cada dimensión representa un articulador: $(Lengua; Mandibula; Labios)$. Cada articulador tiene dos estados posibles: 0 o 1. Los estados permitidos forman las aristas de un cubo, donde las vocales rioplatenses ocupan 5 de los 8 sitios disponibles.

De esta manera el espacio motor queda discretizado: a cada vocal le corresponde un vector tridimensional, donde cada dimensión corresponde a la actividad de un articulador, que puede tomar dos valores posibles 0 o 1. En este espacio tridimensional, donde cada dimensión representa el estado de un articulador (lengua, mandíbula y labios), las vocales rioplatenses ocupan 5 de los 8 estados disponibles, que forman las aristas de un cubo como el que se muestra en el panel B de la figura 5.2. Es interesante notar que, si bien existen del orden de 30 sonidos vocálicos a lo largo de las distintas lenguas, la dimensión del *cubo vocálico* es compatible con una descripción existente de una base del espacio vocálico: las vocales cardinales. Estas consisten en un subgrupo de 8 vocales utilizadas por los fonetistas como referencia, sirven para colapsar a todo el espacio de sonidos vocálicos en estos 8, elegidos con criterios tanto articulatorios como acústicos [72]. De esta forma, cada vocal puede ser descripta como una perturbación de su cardinal más cercana. Las vocales españolas mapean a distintas cardinales, sugiriendo que nuestra representación articulatoria discreta presenta una alternativa simple y novedosa de construir una *base vocálica* con 8 estados posibles.

Debido a que la descripción depende del valor en que se fije el umbral, exploramos la efectividad del método en función de este valor, y buscamos una forma de optimizarlo. Primeramente, investigamos el porcentaje de éxito en la decodificación de cada transductor en función del valor del umbral. Es decir, a cada vocal le asociamos un estado para cada articulador, como se muestran en la figura 5.2, y para distintos valores de umbral contamos para cada articulador la cantidad de vocalizaciones que alcanzan el estado correcto. En el panel izquierdo de la figura 5.3 se muestran los resultados para uno de los participantes. Encontramos que el comportamiento es similar en todos los detectores: el porcentaje de éxito aumenta hasta alcanzar valores de decodificación por arriba del 90 %, donde se mantiene para un rango de valores del umbral antes de decrecer. Los datos de los otros dos participantes muestran el mismo comportamiento. Esto indica que ciertas variaciones en las señales de los detectores no altera el reconocimiento de las vocales, cada vocal es compatible con una familia de configuraciones de tracto vocal.

Por otro lado, el rango de valores de los umbrales donde la decodificación exitosa supera el 90 % varía para los distintos sujetos. Es decir, el valor del umbral depende del hablante. Esto puede deberse a múltiples factores: diferencias relativas en los tractos vocales dependiendo, por ejemplo, del genero del hablante, diferentes hábitos de dicción o variaciones en el montaje del dispositivo. Por esto, cualquier aplicación de este método necesita comenzar con un período de calibración.

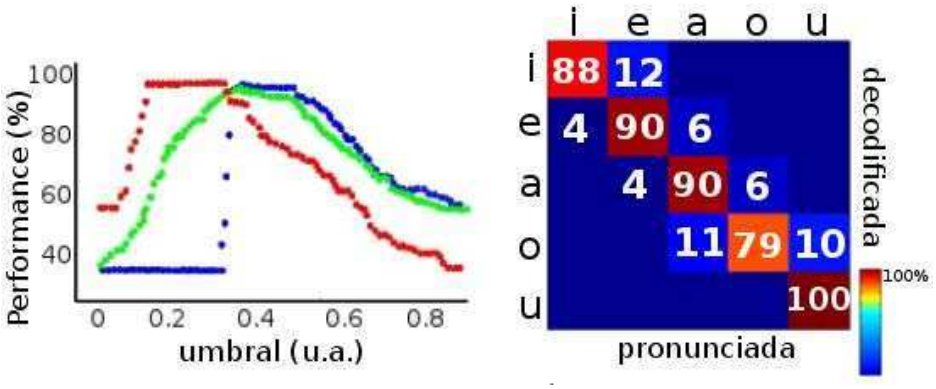

Figura 5.3: Performance de la decodificación. Izquierda: Porcentaje de aciertos para cada detector en función del valor del umbral, utilizando todas las vocalizaciones de un solo sujeto. En rojo el detector montado sobre el labio inferior, en verde el de la mandíbula y en azul el de la lengua. Derecha: Matriz de confusión calculada sobre el *conjunto de prueba*, y fijando distintos valores de umbral para cada sujeto.

Posteriormente, precisando la cuantificación, analizamos la efectividad de decodificación integrando la información de los tres detectores y teniendo en cuenta que cada sujeto requiere un set distinto de umbrales. Para esto, utilizamos la mitad de los datos disponibles para fijar un valor para cada umbral (*conjunto de entrenamiento*), y el resto para estimar el porcentaje de éxito en la decodificación (*conjunto de prueba*). Para cada participante fijamos el umbral en el centro del rango de valores de cada detector donde la decodificación, sobre el *conjunto de entrenamiento*, supera el 90 % de aciertos. La matriz de confusión obtenida sobre los *conjuntos de prueba* de todos los hablantes se muestra en el panel derecho de la figura 5.3. Encontramos que

la performance de decodificación es alta ($< 79\,\%$ *chance* $= 20\,\%$) y que los errores se encuentran limitados a los primeros vecinos de la diagonal. Esto último indica que las decodificaciones erróneas confunden solo el estado de un detector.

Mapeo del espacio neuronal al anatómico

Como ya mencionamos a lo largo de esta tesis, existe una dicotomía en el proceso de producción del habla: si bien la codificación neuronal de los gestos motores que definen los distintos fonemas parece ser discreta [12], el espacio determinado por las posibles configuraciones de tracto vocal es continuo. Sin embargo, nuestra representación logra recuperar la información discreta a partir del monitoreo de la dinámica de los distintos articuladores. A continuación mostramos que es posible mapear nuestro resultado tanto en el espacio neuronal, como en el anatómico.

En un trabajo reciente [13], se estudió la codificación neuronal para la producción de vocales. Los autores realizan una matriz de decodificación para la vocal pronunciada, a partir de mediciones de la actividad en el giro temporal superior, encontrando una estructura de vecinos idéntica a la del panel derecho de la figura 5.3. Además, reportan poblaciones de neuronas en la corteza media orbitofrontal y el cortex del cíngulo anterior que responden selectivamente a la producción de distintas vocales. La actividad de estas poblaciones puede representarse por la siguiente matriz:

$$V_N = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 \end{bmatrix}$$

donde las columnas representan las distintas vocales, de izquierda a derecha */a/, /e/, /i/, /o/* y */u/*, y las filas las distintas poblaciones neuronales activas (1) o por debajo del umbral de disparo (0). Buscamos, entonces, uni-

ficar esta representación binaria neuronal con nuestra representación motora, también binaria. Para esto, necesitamos una transformación que nos permita llevar cualquier columna de V_N, que representa la codificación neuronal de una vocal, al espacio de nuestras coordenadas motoras. Definiendo la matriz motora como:

$$V_M = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 & 0 \end{bmatrix}$$

donde las filas representan los distintos detectores, en orden decreciente: labios, mandíbula y lengua y las columnas, nuevamente, las vocales de izquierda a derecha */a/, /e/, /i/, /o/ y /u/*. Con esta representación, la conversión del espacio neuronal al motor viene dada por la operación: $V_M = T_{N\to M} V_N$ donde:

$$T_{N\to M} = \begin{bmatrix} 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ -1 & 1 & 1 & 0 \end{bmatrix}$$

De esta manera, logramos conectar la actividad neuronal correspondiente a cada vocal con el gesto motor que la caracteriza, descripto según nuestras coordenadas motoras discretas.

En el extremo opuesto de la producción vocal se encuentra el espacio anatómico continuo de las configuraciones de tracto que definen a cada vocal. Una descripción de estas configuraciones es la detallada en la sección 2.2.3, esta da cuenta de todas las configuraciones vocálicas de tracto como combinaciones lineales de dos modos ortogonales espaciales [37, 65]. Recordemos que en este modelo toda configuración de tracto vocálico viene dado por: $d(x) = \Omega(x) + q_1\varphi_1 + q_2\varphi_2$, donde x es la distancia a la glotis, d es el diámetro, Ω el tracto neutro, φ_1 y φ_2 los dos modos espaciales y son $(q_1; q_2)$ los coeficientes que describen cada vocal en este espacio. Buscamos entonces los pares $v_A^i = (q_1; q_2)^i$ que corresponden a cada vocal rioplatense i. Para esto realizamos, sobre los registros de audio de cada vocal, la transformada de fourier y extrajimos de estas los valores de las dos primeras formantes. Luego,

promediamos estos valores y buscamos los pares $(q_1; q_2)$ que generen tractos cuyas dos primeras resonancias sean compatibles con los valores experimentales de F_1 y F_2. Los resultados de este procedimiento se muestran en la tabla 5.1. Estos valores nos permiten definir una mapa para ir del espacio motor tridimensional, determinado por los articuladores, al espacio bidimensional de las configuraciones anatómicas: $T_{M\to A} v_M^i + a = v_A^i$, donde

$$T_{M\to A} = \begin{bmatrix} 2 & -4 & 5{,}5 \\ 3 & -1 & 2 \end{bmatrix}, a = \begin{bmatrix} 2 \\ -2 \end{bmatrix}$$

De esta forma logramos una descripción completa y cerrada del proceso de producción de vocales aisladas. El desafío consiste ahora en extender el resultado al resto de los fonemas e integrar el fenómeno de coarticulación de los mismos.

	F_1	$F_2(kHz)$	q_1	q_2
$/a/$	0.8	1.26	4	1
$/e/$	0.57	2.06	-1.5	3
$/i/$	0.28	2.50	-5.5	2
$/o/$	0.53	0.99	2	-2
$/u/$	0.21	0.77	-2	-3

Tabla 5.1: Ajuste de coeficientes para cada modo normal, $(q_1; q_2)$, de forma tal que las dos primeras formantes del tracto resultante se correspondan con las experimentales de las distintas vocales del español rioplatense. Para esto se realizó un mallado con $\Delta q_{1/2} = 0{,}1$ restringiendo el espacio a $q_1 \in [-6; 3]$ y $q_2 \in [-4; 4]$, para cada punto se calculó la forma del tracto y sus resonancias, inyectando una función tipo delta y calculando el espectro a la salida. Luego, se buscaron los coeficientes que generan tractos cuyas resonancias más se acercan a las dos primeras formantes experimentales.

5.2.2. Extendiendo el resultado a consonantes oclusivas

Una vez lograda una descripción motora discreta de las vocales del español, nos dispusimos a integrar las consonantes oclusivas $/p/, /t/$ y $/k/$. El reto consiste, por un lado en aumentar el número de fonemas a decodificar, y por otro en testear cuan robusta es nuestra descripción cuanto más nos acercamos al discurso natural. Abandonamos los fonemas aislados para trabajar con sílabas, incluyendo el fenómeno de coarticulación, de forma que el gesto motor de cada fonema se verá influenciado por el de sus vecinos.

Nuevamente, realizamos 3 sesiones de medición a lo largo de un mes para cada uno de los 3 sujetos. En este caso, cada sesión consta de 75 vocalizaciones distintas del tipo vocal-consonante-vocal (*vcv*), donde la vocal puede ser cualquier vocal española y la consonante cualquier oclusiva sorda del mismo idioma ($/p/, /t/$ o k).
Al analizar las señales observamos que la descripción para las vocales se mantiene. Alcanza con establecer un umbral por detector, *umbral vocálico*, para decodificar las vocales rioplatenses, aún al pronunciarse concatenadas a otros fonemas. Sin embargo, es necesario incorporar un nuevo umbral, *umbral consonántico*, por detector para dar distinguir las consonantes, como se muestra en el panel A de la figura 5.4. De esta forma, cada fonema tiene asociado un vector tridimensional, donde cada dimensión especifica en cual de los tres estados disponibles se encuentra cada articulador, panel A de la figura 5.4.
En el espacio tridimensional donde se representa el estado de cada articulador, las vocales se ubican en los vértices de un cubo, mientras que las consonantes quedan por fuera de este. Todas las consonantes estudiadas están contenidas en el plano $Mandibula = 2$, como se muestra en el panel B de la figura 5.4. La p queda determinada por estar por debajo del *umbral consonántico* fijado en la señal de los labios ($Labios = -1$), la t por estar por arriba del correspondiente a la lengua ($Lengua = 2$) y la k queda por debajo del *umbral vocálico* de la lengua ($Lengua = 0$) y en cualquier estado correspondiente a vocales en los labios ($Labios \neq 2$). Una diferencia que presentan

las consonantes con respecto a las vocales, es que quedan representadas por dos estados. Como se muestra en la figura 5.4: $p = (-1;2;1) \cup (-1;2;0)$, $t = (1;2;2) \cup (0;2;2)$ y $k = (1;2;0) \cup (0;2;0)$. En la articulación de la k y la t no intervienen los labios con lo cual esta dimensión puede tomar cualquier valor vocálico, lo mismo sucede con la p y la lengua.

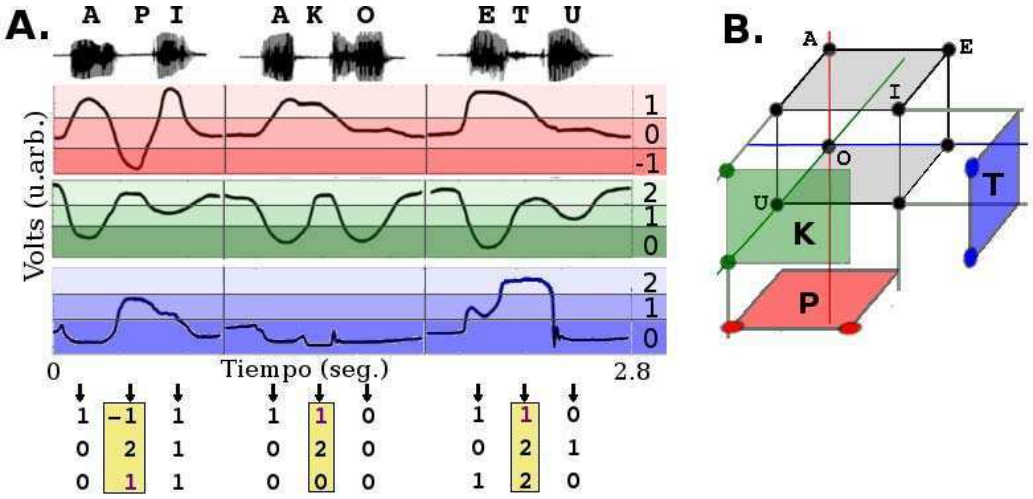

Figura 5.4: Descripción de vocales y consonantes oclusivas del español. **A.** Señales temporales para distintas combinaciones de *vcv*: panel superior presión acústica, le siguen las señales de los distintos detectores labios en rojo, mandíbula en verde y lengua en azul. La intensidad de colores remarca el estado en el que se encuentra cada articulador. En la parte inferior el vector que representa cada fonema. **B.** Cubo vocálico extendido con los planos de las distintas consonantes, cada articulador representa una dimensión con 3 estados posibles. Las aristas del cubo representa una posible vocal, mientras que las consonantes vienen dadas por dos posibles estados contenidos en el plano Mandíbula=2, en verde la K, en rojo la P y en azul la T.

Estudiamos cuan robusta es esta representación repitiendo el procedimiento de la sección anterior para construir una matriz de confusión utilizando un set de umbrales para cada participante. Nuevamente, encontramos que la performance de decodificación es alta ($< 69\,\%$ *chance* $= 12{,}5\,\%$) según se muestra en la figura 5.5.

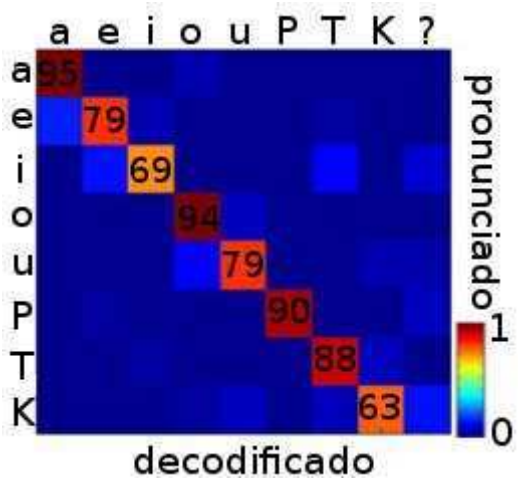

Figura 5.5: Matriz de confusión calculada sobre el *conjunto de prueba*, fijando distintos valores de umbral para cada sujeto.

Las consonantes oclusivas del español son 6, y vienen de *a pares*: una misma disposición de tracto determina dos posibles fonemas dependiendo de la actividad o no de las cuerdas vocales. Es decir, a la p y la b les corresponde el mismo perfil de tracto, lo que las diferencia es que la primera es sorda (cuerdas inactivas) y la segunda es sonora (cuerdas activas). Lo mismo ocurre para los pares $t - d$ y $k - g$. Nuestro dispositivo permite monitorear la dinámica del tracto, con lo cual no podemos distinguir sordas de voceadas. Sin embargo, esta ambigüedad podría superarse incluyendo un electroglotógrafo al sistema, instrumento de medición no invasivo que permite determinar la actividad de las cuerdas vocales [73, 74]. Esto permitiría reconocer 11 fonemas (todas las vocales y oclusivas del español), que representan el 48 % de los fonemas rioplatenses.

Coarticulación

Como mencionamos anteriormente, existen dos estados posibles para cada una de las consonantes estudiadas. Esto se debe a que no todos los articuladores participan en la articulación de estos fonemas: en la k y la t no intervienen los labios, mientras que en la p no interviene la lengua. De esta forma, una

dimensión, en cada una de estas consonantes, puede tomar cualquier valor vocálico (*0 o 1*). Creemos que este *estado libre* quedará determinado por los fonemas vecinos, permitiendo integrar el fenómeno de coarticulación en nuestra representación discreta. Recordemos que la coarticulación es el proceso por el cual los gestos motores de los fonemas se ven modificados por los de los fonemas vecinos.

En un análisis preliminar, estudiamos cómo influyen las vocales vecinas en la determinación del *estado libre* de cada oclusiva. Más precisamente, investigamos cuál es el estado correspondiente al tiempo central de la consonante, para las vocalizaciones de un sujeto, en distintos contextos vocálicos. Los resultados se muestran en la figura 5.6.

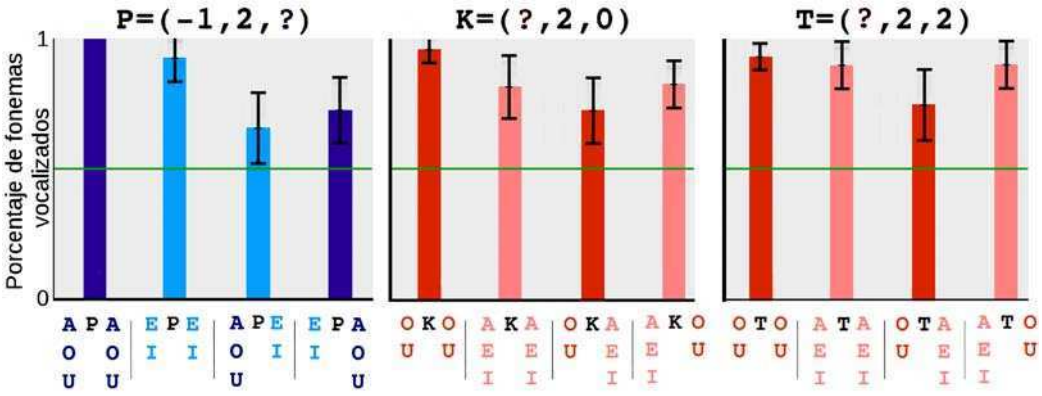

Figura 5.6: Coarticulación de las consonantes. Para cada consonante estudiamos el valor que toma el *estado libre* en función de las 4 transiciones posibles de ese estado, entre la vocal previa y la siguiente ($0 \to 0|1 \to 1|0 \to 1|1 \to 0$). Las barras representan el porcentaje en cada estado de las consonantes vocalizadas en un determinado contexto vocálico. Para la p: $Azul \to Lengua = 0$ $Celeste \to Lengua = 1$. Para la k y la t: $Rojo \to Labios = 0$ y $Rosado \to Labios = 1$. La línea verde reprecenta a chance (0,5). Resultados correspondientes a las vocalizaciones de un solo sujeto.

Lo primero que notamos es que si el *estado libre* de las vocales vecinas es el mismo, la consonante mantiene ese estado. Por ejemplo, para cualquier

vocalización del tipo $v_1 - P - v_2$: si $v_1 \wedge v_2 \in [A, O, U]$, conjunto de vocales con la señal correspondiente a la lengua en el estado 0, entonces la P viene dada por $(-1; 2; 0)$; mientras que si $v_1 \wedge v_2 \in [E, I]$, la P es $(-1; 2; 1)$. Este efecto se mantiene para las tres consonantes estudiadas, la diferencia la encontramos cuando el estado de las vocales vecinas es opuesto. En este caso, al articular una p la lengua *prefiere* el estado de la vocal siguiente mientras que en el gesto motor de la k y la t se *hereda* el estado de los labios correspondiente a la vocal previa.

Este resultado concuerda con trabajos previos que postulan que la velocidad de la lengua es mayor que la del resto de los articuladores [14,15,75,76]. Suponiendo que el gesto motor del articulador que determina el fonema, comienza al mismo tiempo que la transición entre los estados vocálicos del articulador libre, en la t y la k cuando la lengua alcanza la posición que caracteriza al fonema, los labios aún no alcanzaron el estado siguiente y continúan en el estado de la primera vocal. Mientras que en el caso de la p, como la lengua posee menos inercia, llega a su objetivo (la segunda vocal) antes de que los labios articulen la p.

Capítulo 6

Conclusiones

El estudio del proceso de producción de voz ha sido abordado desde distintas disciplinas como: biología, medicina, o lingüística. A lo largo de esta tesis se estudiaron distintos niveles de este proceso desde el punto de vista de la física, mostrando algunas contribuciones que esta disciplina puede aportar al problema.

A nivel de una descripción anatómica adecuada del sistema fonador, estudiamos la dinámica de un modelo de cuerdas vocales. Investigamos los distintos regímenes que alcanza el sistema variando parámetros anatómicos que activamente son modificados durante el habla. Reportamos la existencia de una bifurcación cercana a la zona de fonación normal que, si bien no altera la dinámica sino tan solo la topología del movimiento de las cuerdas, podría ser importante para alcanzar la riqueza espectral necesaria para pronunciar los distintos fonemas.

Combinando este modelo de cuerdas con uno de baja dimensión para el tracto vocal construimos un sintetizador articulatorio. Realizamos estudios perceptuales y de resonancia magnética funcional que muestran que la voz sintética es indistinguible de la real, resultando una herramienta adecuada y novedosa para estudiar la codificación neuronal de la voz, en términos del aparato vocal que la produce. En el marco de un programa unificado sensori-

motor del habla [12,53,77] resulta vital una herramienta que permita estudiar percepción y producción de habla en términos de parámetros anatómicos y gestos motores [16].

En los últimos años, con la aparición de técnicas no invasivas que permiten monitorear la actividad cerebral, se realizaron numerosos avances en el campo de la percepción del habla [54, 55]. Sin embargo, la codificación neuronal de la voz es estudiada, típicamente, en términos de propiedades acústicas, perdiendo de vista la hipótesis de una codificación de parámetros anatómicos. Haciendo uso del sintetizador articulatorio, mostramos que la identidad de la voz viene dada por una relación entre las dimensiones del tracto y de las cuerdas.

Combinando el modelado físico del sistema vocal con experimentos perceptuales, investigamos las fuerzas miméticas que intervienen en la generación del lenguaje. Estudiamos la existencia de componentes puramente imitativos en dos tipos de onomatopeyas. La controversia que presentan estos objetos del lenguaje es que, idealmente, se espera que la imitación de un ruido simple sea un solo sonido vocal, acústicamente similar al primero. Sin embargo, dado que las onomatopeyas pertenecen a la esfera del lenguaje están formados, como cualquier otra palabra, por un conjunto de fonemas que son sonidos con propiedades acústicas muy diversas como lo son vocales y consonantes. Si bien esta contradicción parece irreconciliable, en este trabajo, encontramos que la imitación vocal de ruidos simples descansa precisamente en el fenómeno de coarticulación. Las vocales proporcionan el contenido espectral a imitar y le otorgan sonoridad a la onomatopeya, mientras que la consonante oclusiva provee la fuente ruidosa y las características temporales adecuadas. Así, mostramos una vía plausible por la cual la imitación de un sonido simple se transforma en una estructura más compleja.

Por último, logramos una representación discreta de los gestos motores correspondientes a las vocales y consonantes oclusivas del español. Esta representación logra unificar la representación neuronal discreta con una dinámica

continua del tracto vocal. El dispositivo experimental desarrollado presenta una herramienta novedosa, simple y económica para realizar distintas investigaciones dentro del campo de la fonología experimental, y sienta las bases para posibles aplicaciones bioprostéticas.

Apéndice A

Adquisición y análisis fMRI

Los sujetos fueron escaneados en un resonador Siemens 3T Verio MRI, 32-canales TIM system. Primero se obtienen imágenes de alta resolución ponderadas en T1 para la localización anatómica (TR = 2.3 s, TE = 2.98 , matrix = 240 x 256 x 176, tamaño de voxel = 1 x 1 x 1 mm). Luego, imágenes funcionales ponderadas en T2* EPI sensitivas al contraste BOLD (TR=2.02 s, TE=25 ms, matriz = 66 x 66 x 40, tamaño de voxel = 3 x 3 x 3 mm). Un número variable de imágenes se adquiere por corrida del experimento. Los estímulos auditivos se presentan mediante auriculares RM-compatibles con un intensidad de presión acústica de 80 dB.

Para el procesamiento y análisis de las imágenes de resonancia magnética funcional utilizamos el programa SPM5 (*Wellcome Department of Cognitive Neurology, London, UK*) y código desarrollado por nosotros en MATLAB. En todos los casos los primeros 5 escaneos EPI fueron desechados. Para corregir el movimiento las imágenes fueron realineadas utilizando como referencia la primera. Se realizó una transformada estereotáxica de la imagen anatómica a las del prototipo del *Montreal Neurological Institute* en coordenadas Talairach. Luego, los escaneos funcionales son normalizados utilizando esta misma transformación. Las imágenes funcionales son suavizadas con un filtro Gaussiano espacial de 5 mm Los datos fueron modelados con SPM utilizando la función hemodinámica canónica, y los 6 parámetros de movimiento se

incluyeron como regresores sin interés. A nivel de grupo se realizó un análisis ANOVA utilizando las imágenes de los contrastes por individuo.

Apéndice B

Algoritmo genético

Un algoritmo genético es un método estocástico de optimización de parámetros inspirado en la selección natural. Básicamente, la idea es que los individuos mejor adaptados prevalecen en la reproducción, esto hace que los genes responsables de la mejor adaptación al medio predominan a lo largo de las generaciones. La información genética de las crías se construye mediante dos procesos: mezclando la de los padres (cruzamiento cromosómico) y cambios aleatorios (mutación). La aplicación de estas dos operaciones en el espacio genético de una población, presenta una forma eficiente de buscar individuos mejor adaptados al medio [42].

Esta caricatura es exportada para encontrar los parámetros de nuestro sistema de tubos que mejor reproducen un dado espectro $\hat{s}_e(f)$ (*espectro objetivo*). Para detallar el algoritmo necesitamos especificar algunos conceptos:

- Asociamos a cada set de parámetros $\{l, A\}$ un valor de éxito F (*fitness*). Para esto calculamos el sonido sintético que le corresponde ($\{l, A\} \rightarrow s(t_i)$) y su transformada de Fourier $\hat{s}(f_i)$. Luego, obtenemos la inversa de la diferencia cuadrada entre el espectro objetivo y el sintético, y calculamos F según: $F(\{l, A\}) = (\sum_i |\hat{s}_{objetivo}(f_i) - \hat{s}(f_i)|^2)^{-1}$, $f_i \leq 6{,}5$ kHz para que sea válida la aproximación de onda plana. Si incluimos restricciones anatómicas se modifica a: $F(\{l, A\}) = [\sum_i |\hat{s}_{objetivo}(f_i) - \hat{s}(f_i)|^2 + \alpha \sum_{j=1}^{10} (|a_j - a_j^{PCA}|/a_j)^2]^{-1}$ para un tracto vocal con áreas

$a_1, a_2, ..., a_{10}$. El factor α proporciona un peso relativo del 40 % a las restricciones anatómicas y del 60 % a las propiedades espectrales.

- Además cada parámetro p tiene su correlato en el espacio genético $p \in (a, b)$. Asociandole un vector $\bar{p} \equiv (\bar{p}_1, \bar{p}_2, \bar{p}_3, \bar{p}_4)$, obtenido normalizando el parámetro según: $\bar{p} = (p - a)/(b - a) \sim \bar{p}_1 10^{-1} + \bar{p}_2 10^{-2} + \bar{p}_3 10^{-3} + \bar{p}_4 10^{-4})$.

- El vector n dimensional $\{l, a_1, a_2, ..., a_{10}\}$ es reemplazado por el 4n dimensional $\{\bar{l}, \bar{a}_1, \bar{a}_2, ..., \bar{a}_{10}\}$. En este espacio el operador de intercambio genético (*crossover*) consiste en: escoger una pareja y una posición de corte al azar en el vector, e intercambiar la información genética a partir de esta posición, generando así dos nuevos ejemplares. El operador mutación es reemplazar un sitio random del vector por un nuevo valor elegido al azar.

El algoritmo arranca con una población de $n = 500$ tractos vocales, con parámetros $\{l, A\}_1, ..., \{l, A\}_n$ elegidos de manera aleatoria. Se calcula el fitness F de cada individuo y se escogen $n/2$ parejas. La probabilidad de cada individuo de ser elegido para reproducirse es proporcional a su fitness, es decir individuos con un alto F son seleccionados más veces. Para cada pareja el intercambio genético y la mutación ocurren con probabilidades del 80 % y 10 % respectivamente. Los pares resultantes conforman una nueva población de tractos sobre la que se repite el proceso, hasta que F alcanza algún umbral preestablecido o el número de generaciones es mayor a un dado valor.

Particularmente, en esta tesis se fijó como umbral que la diferencia cuadrada entre el espectro objetivo y el correspondiente al tracto sea menor al 5 % del total de la intensidad espectral del objetivo. Típicamente, después de 30 iteraciones aproximadamente el 5 % de la población se encuentra por debajo del umbral establecido.

Los resultados obtenidos con el algoritmo son los que se muestran en la tabla B.1.

	d_1	d_2	d_3	d_4	d_5	d_6	d_7	d_8	d_9	d_{10}	$10\,l$
					cm						cm
$[a]$	1.00	0.72	0.62	1.58	2.03	2.48	2.46	2.49	2.84	2.89	16.4
$[e]$	0.76	1.35	1.92	1.95	1.64	1.43	0.65	1.23	1.52	1.65	16.4
$[i]$	0.84	2.39	2.42	2.45	1.85	0.95	0.86	0.71	1.32	1.48	16.4
$[o]$	1.21	1.48	0.68	0.79	0.98	1.12	2.96	2.64	3.00	1.09	16.4
$[u]$	1.23	2.74	0.40	1.64	1.70	2.09	1.79	1.70	1.85	2.04	17.4
${}_a[x]$	1.15	0.50	1.13	1.00	1.46	0.48	1.65	2.43	2.76	2.89	16.4
${}_e[x]$	0.67	1.22	0.93	1.29	0.85	0.62	0.31	1.25	1.53	1.95	16.4
${}_i[x]$	0.59	1.77	1.72	1.65	1.80	1.40	0.45	0.76	1.21	2.45	16.4
${}_o[x]$	1.14	1.90	2.26	2.02	1.72	0.37	3.00	2.90	2.12	1.91	16.4
${}_u[x]$	0.80	1.73	1.78	1.49	1.33	0.98	0.61	2.74	1.40	1.60	17.4
click	0.86	1.43	1.86	1.73	1.70	1.72	0.22	1.72	1.54	2.45	16.4
toc	0.57	1.59	1.84	1.29	1.31	1.34	0.65	0.32	3.00	0.69	16.4
click (anat.)	0.75	1.81	1.42	1.39	1.56	1.19	0.23	1.54	1.44	2.45	16.4
toc (anat.)	0.54	2.02	1.88	1.08	2.12	0.43	2.45	1.21	2.10	1.07	16.4

Tabla B.1: Ajuste de diámetros y largos del tracto vocal. Los parámetros anatómicos hallados con el algoritmo utilizando como *objetivo* el espectro de las vocales y de las de la fricativas coarticulada $[x]_v$. La últimas filas corresponden a las mejores imitaciones de los sonidos tipo *tic* y tipo *tic* (sin y con restricciones anatómicas).

Apéndice C

Dispositivo experimental

Los participantes vocalizaron con un protector dental de plástico de 1mm de espesor sobre el que se montaron algunos de los sensores e imanes del dispositivo. Cada protector fue construido por un odontólogo de acuerdo a la anatomía de cada sujeto y consta de dos partes: una cubre dientes y muelas inferiores, otra dientes, muelas y paladar superior (ver figura 5.1).

Los transductores de efecto hall Hall utilizados son marca Allegro, más precisamente *Ratiometric Linear Hall Effect Sensor ICs for High-Temperature Operation"*. Para alterar lo menos posible la dicción, se conectaron los sensores empleando cables de electrofisiología (*Subminiature Lead Wire TDQ 44, Phoenix Wire Inc.*) recubiertos con un tubo plástico (*Silastic, Laboratory Tubing 0.76 mm x 1.65 mm*). Construímos una electrónica apropiada para recoger las señales que consta de un amplificador variable (2-30x) y un filtro pasa bajos de 20 Hz, luego pasan por una placa de adquisición de datos (*National Instruments, NI USB-6212"*) a una PC.

Empleamos imanes biocompatibles de Sm-Co, de peso entre 1 y 3 gr. de dimensiones variables. Labio superior: cilindro de 3.0 mm de diámetro y 1.5 mm de alto, mandíbula: una esfera de 5. mm de diámetro, lengua y dientes inferiores: cilindros de 3mm de diámetro y 1 mm de alto. Para adherir el imán a la lengua se usó pegamento para prótesis dentales (*Fixodent Original Denture Adhesive Cream 2.4 Oz*).

Bibliografía

[1] E. M. Arneodo, G. B. Mindlin, Source-tract coupling in birdsong production, Physical Review E 79 (6) (2009) 061921.

[2] I. R. Titze, Principles of voice production.

[3] K. N. Stevens, Acoustic phonetics, Vol. 30, MIT press, 2000.

[4] I. R. Titze, The physics of small-amplitude oscillation of the vocal folds, J. Acoust. Soc. Am. 83 (1988) 1536–1550.

[5] R. Laje, G. B. Mindlin, The Physics of Birdsong, Springer-Verlag Berlin Heidelberg, 2005.

[6] G. E. Peterson, H. L. Barney, Control methods used in a study of the vowels, The Journal of the Acoustical Society of America 24 (2) (2005) 175–184.

[7] K. N. Stevens, Vocal fold physiology, University of Tokyo Press, 1981.

[8] M. Assaneo, M. Trevisan, Computational model for vocal tract dynamics in a suboscine bird, Physical Review E 82 (3) (2010) 031906.

[9] M. F. Assaneo, M. A. Trevisan, Revisiting the two-mass model of the vocal folds, Papers in Physics 5 (2013) 050004.

[10] T. Hashimoto, N. Usui, M. Taira, I. Nose, T. Haji, S. Kojima, The neural mechanism associated with the processing of onomatopoeic sounds, Neuroimage 31 (4) (2006) 1762–1770.

[11] M. F. Assaneo, J. I. Nichols, M. A. Trevisan, The anatomy of onomatopoeia, PloS one 6 (12) (2011) e28317.

[12] K. E. Bouchard, N. Mesgarani, K. Johnson, E. F. Chang, Functional organization of human sensorimotor cortex for speech articulation, Nature 495 (7441) (2013) 327–332.

[13] A. Tankus, I. Fried, S. Shoham, Structured neuronal encoding and decoding of human speech features, Nature Communications 3 (2012) 1015.

[14] C. P. Browman, L. Goldstein, Tiers in articulatory phonology, with some implications for casual speech, Papers in laboratory phonology I: Between the grammar and physics of speech (1990) 341–376.

[15] C. P. Browman, L. Goldstein, Dynamics and articulatory phonology, Mind as motion (1995) 175–193.

[16] M. F. Assaneo, M. A. Trevisan, G. B. Mindlin, Discrete motor coordinates for vowel production, PloS one 8 (11) (2013) e80373.

[17] N. Lous, G. Hofmans, R. Veldhuis, A. Hirschberg, A symmetrical two-mass vocal-fold model coupled to vocal tract and trachea, with application to prosthesis design, Acta Acustica united with Acustica 84 (6) (1998) 1135–1150.

[18] M. E. Smith, G. S. Berke, B. R. Gerratt, J. Kreiman, Laryngeal paralyses: Theoretical considerations and effects on laryngeal vibration, Journal of speech and hearing research 35 (3) (1992) 545.

[19] X. Pelorson, C. Vescovi, E. Castelli, A. Hirschberg, A. Wijnands, H. Bailliet, Description of the flow through in-vitro models of the glottis during phonation. application to voiced sounds synthesis, Acta Acustica united with Acustica 82 (2) (1996) 358–361.

[20] K. Ishizaka, Synthesis of Voiced Sounds from a Two-Mass Model of the Vocal, Cords. Bell Syst. Tech. J.
URL http://ci.nii.ac.jp/naid/10008576149/

[21] R. Timcke, H. von LEDEN, P. Moore, Laryngeal vibrations: Measurements of the glottic wave: Part i. the normal vibratory cycle, AMA archives of otolaryngology 68 (1) (1958) 1–9.

[22] H. Hollien, R. Coleman, P. Moore, Stroboscopic laminagraphy of the larynx during phonation, Acta oto-laryngologica 65 (1-6) (1968) 209–215.

[23] M. Trevisan, M. Eguia, G. Mindlin, Nonlinear aspects of analysis and synthesis of speech time series data, Physical Review E 63 (2) (2001) 1–6.
URL http://link.aps.org/doi/10.1103/PhysRevE.63.026216

[24] X. Pelorson, A. Hirschberg, R. Van Hassel, A. Wijnands, Y. Auregan, Theoretical and experimental study of quasisteady-flow separation within the glottis during phonation. application to a modified two-mass model, The Journal of the Acoustical Society of America 96 (6) (1994) 3416–3431.

[25] J. C. Lucero, Dynamics of the two-mass model of the vocal folds: Equilibria, bifurcations, and oscillation region, The Journal of the Acoustical Society of America 94 (6) (1993) 3104–3111.

[26] T. Ikeda, Y. Matsuzaki, T. Aomatsu, A numerical analysis of phonation using a two-dimensional flexible channel model of the vocal folds, Journal of biomechanical engineering 123 (6) (2001) 571–579.

[27] A. Amador, F. Goller, G. B. Mindlin, Frequency modulation during song in a suboscine does not require vocal muscles, Journal of Neurophysiology 99 (5) (2008) 2383–2389.

[28] J. C. Lucero, L. L. Koenig, Simulations of temporal patterns of oral airflow in men and women using a two-mass model of the vocal folds under dynamic control, Journal of Acoustical Society of America 117 (3) (2005) 1362–1372.

[29] T. Baer, Investigation of phonation using excised larynxes.

[30] I. Steinecke, H. Herzel, Bifurcations in an asymmetric vocal-fold model, The Journal of the Acoustical Society of America 97 (3) (1995) 1874–1884.

[31] B. H. Story, I. R. Titze, Voice simulation with a body-cover model of the vocal folds, The Journal of the Acoustical Society of America 97 (2) (1995) 1249–1260.

[32] E. J. Doedel, Auto: A program for the automatic bifurcation analysis of autonomous systems, Congr. Numer 30 (1981) 265–284.

[33] J. Guckenheimer, P. Holmes, Nonlinear oscillations, dynamical systems, and bifurcations of vector fields, Vol. 42, New York Springer Verlag, 1983.

[34] J. C. Lucero, A theoretical study of the hysteresis phenomenon at vocal fold oscillation onset–offset, The Journal of the Acoustical Society of America 105 (1) (1999) 423–431.

[35] T. Baer, Reflex activation of laryngeal muscles by sudden induced subglottal pressure changes, The Journal of the Acoustical Society of America 65 (5) (1979) 1271–1275.

[36] L. Landau, Fluid mechanics: volume 6 (course of theoretical physics) Author: LD Landau, EM Lifshitz, Publisher: Bu, Butterworth-Heinemann, 1987.

[37] B. H. Story, A parametric model of the vocal tract area function for vowel and consonant simulation, The Journal of the Acoustical Society of America 117 (5) (2005) 3231–3254.

[38] B. H. Story, Phrase-level speech simulation with an airway modulation model of speech production, Computer speech & language 27 (4) (2013) 989–1010.

[39] N. H. Fletcher, T. Riede, R. A. Suthers, Model for vocalization by a bird with distensible vocal cavity and open beak, The Journal of the Acoustical Society of America 119 (2) (2006) 1005–1011.

[40] T. Riede, R. A. Suthers, N. H. Fletcher, W. E. Blevins, Songbirds tune their vocal tract to the fundamental frequency of their song, Proceedings of the National Academy of Sciences 103 (14) (2006) 5543–5548.

[41] C. H. Shadle, The acoustics of fricative consonants.

[42] D. E. Goldberg, et al., Genetic algorithms in search, optimization, and machine learning, Vol. 412, Addison-wesley Reading Menlo Park, 1989.

[43] T. Riede, R. A. Suthers, Vocal tract motor patterns and resonance during constant frequency song: the white-throated sparrow, Journal of Comparative Physiology A 195 (2) (2009) 183–192.

[44] T. Gardner, M. Magnasco, Instantaneous frequency decomposition: An application to spectrally sparse sounds with fast frequency modulations, The Journal of the Acoustical Society of America 117 (5) (2005) 2896–2903.

[45] T. J. Gardner, M. O. Magnasco, Sparse time-frequency representations, Proceedings of the National Academy of Sciences 103 (16) (2006) 6094–6099.

[46] B. H. Story, I. R. Titze, Parameterization of vocal tract area functions by empirical orthogonal modes, Journal of Phonetics 26 (3) (1998) 223–260.

[47] B. H. Story, I. R. Titze, E. A. Hoffman, Vocal tract area functions for an adult female speaker based on volumetric imaging, The Journal of the Acoustical Society of America 104 (1) (1998) 471–487.

[48] B. H. Story, Physiologically-based speech simulation using an enhanced wave-reflection model of the vocal tract.

[49] N. B. Pinto, D. G. Childers, A. L. Lalwani, Formant speech synthesis: Improving production quality, Acoustics, Speech and Signal Processing, IEEE Transactions on 37 (12) (1989) 1870–1887.

[50] M. J. F. Gales, Maximum likelihood linear transformations for HMM-based speech recognition, Computer speech & language 12 (2) (1998) 75–98.

[51] B. P., Voice neurocognition laboratory.
URL http://vnl.psy.gla.ac.uk/resources.php

[52] M. Latinus, P. McAleer, P. E. G. Bestelmeyer, P. Belin, Norm-based coding of voice identity in human auditory cortex., Current biology : CB 23 (12).

[53] G. B. Cogan, T. Thesen, C. Carlson, W. Doyle, O. Devinsky, B. Pesaran, Sensory-motor transformations for speech occur bilaterally, Nature.

[54] P. Belin, R. J. Zatorre, P. Lafaille, P. Ahad, B. Pike, Voice-selective areas in human auditory cortex., Nature 403 (6767) (2000) 309–12.
URL http://www.ncbi.nlm.nih.gov/pubmed/10659849

[55] P. Belin, R. J. Zatorre, P. Ahad, Human temporal-lobe response to vocal sounds., Brain Research 13 (2002) 17–26.

[56] J.-A. Bachorowski, M. J. Owren, Vocal expression of emotion: Acoustic properties of speech are associated with emotional intensity and context, Psychological Science 6 (4) (1995) 219–224.

[57] Y. Horii, Jitter and shimmer differences among sustained vowel phonations, Journal of Speech, Language, and Hearing Research 25 (1) (1982) 12–14.

[58] H. Hirose, T. Gay, Laryngeal control in vocal attack, Folia Phoniatrica et Logopaedica 25 (3) (1973) 203–213.

[59] W. T. C. for Neuroimaging, Statistical parametric mapping.
URL http://www.fil.ion.ucl.ac.uk/spm/

[60] D. A. Leopold, I. V. Bondar, M. A. Giese, Norm-based face encoding by single neurons in the monkey inferotemporal cortex, Nature 442 (7102) (2006) 572–575.

[61] C. R. Pernet, P. Belin, The role of pitch and timbre in voice gender categorization, Frontiers in psychology 3 (2012) 23.

[62] B. L. Giordano, S. McAdams, R. J. Zatorre, N. Kriegeskorte, P. Belin, Abstract encoding of auditory objects in cortical activity patterns, Cerebral Cortex 23 (9) (2013) 2025–2037.

[63] R. R. Benson, M. Richardson, D. Whalen, S. Lai, Phonetic processing areas revealed by sinewave speech and acoustically similar non-speech, Neuroimage 31 (1) (2006) 342–353.

[64] G. H. Yeni-Komshian, S. D. Soli, Recognition of vowels from information in fricatives: Perceptual evidence of fricative-vowel coarticulation, The Journal of the Acoustical Society of America 70 (4) (1981) 966–975.

[65] B. H. Story, I. R. Titze, E. A. Hoffman, Vocal tract area functions from magnetic resonance imaging, The Journal of the Acoustical Society of America 100 (1) (1996) 537–554.

[66] W. Benjamin, Selected Writings: 1913-1926, Vol. 1, Harvard University Press, 1996.

[67] B. Mesz, M. A. Trevisan, M. Sigman, The taste of music, Perception-London 40 (2) (2011) 209.

[68] V. S. Ramachandran, E. M. Hubbard, Synaesthesia–a window into perception, thought and language, Journal of consciousness studies 8 (12) (2001) 3–34.

[69] D. Maurer, T. Pathman, C. J. Mondloch, The shape of boubas: Sound–shape correspondences in toddlers and adults, Developmental science 9 (3) (2006) 316–322.

[70] A. J. Bremner, S. Caparos, J. Davidoff, J. de Fockert, K. J. Linnell, C. Spence, Bouba and kiki in namibia. a remote culture make similar shape–sound matches, but different shape–taste matches to westerners, Cognition 126 (2) (2013) 165–172.

[71] M. A. Changizi, Harnessed: How language and music mimicked nature and transformed ape to man, BenBella Books, 2011.

[72] P. Ladefoged, Preliminaries to linguistic phonetics, University of Chicago Press, 1971.

[73] M. Rothenberg, A multichannel electroglottograph, Journal of Voice 6 (1) (1992) 36–43.

[74] C. T. Herbst, W. T. S. Fitch, J. G. Švec, Electroglottographic wavegrams: A technique for visualizing vocal fold dynamics noninvasivelya), The Journal of the Acoustical Society of America 128 (5) (2010) 3070–3078.

[75] M. D. McClean, Patterns of orofacial movement velocity across variations in speech rate, Journal of Speech, Language, and Hearing Research 43 (1) (2000) 205–216.

[76] J. E. Flege, Effects of speaking rate on tongue position and velocity of movement in vowel production, The Journal of the Acoustical Society of America 84 (3) (1988) 901–916.

[77] F. Pulvermüller, M. Huss, F. Kherif, F. M. del Prado Martin, O. Hauk, Y. Shtyrov, Motor cortex maps articulatory features of speech sounds, Proceedings of the National Academy of Sciences 103 (20) (2006) 7865–7870.

Agradecimientos

A mi amigo, consejero y director, Marcos Trevisan por enseniarme a disfrutar lo que hago.
Al labo, a mi familia, y a la gente que frecuenta Vicente Lopez.

Florencia Assaneo
Buenos Aires, 2014