



**UNIVERSIDAD DE BUENOS AIRES**

**Facultad de Ciencias Exactas y Naturales**

**Carrera de Ciencias Biológicas**

**Identificación de individuos *Furnarius Rufus* utilizando redes neuronales siamesas aplicadas a propiedades acústicas del canto**

**Autor: Tomás de Udaeta  
Directora: Ana Amador**

**Lugar de trabajo: Laboratorio de Sistemas Dinámicos, Departamento de Física,  
Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires**

**Julio 2024**

**Tomás de Udaeta**

**Ana Amador**

<u>Resumen</u> .....	4
<u>Abstract</u> .....	5
<u>Agradecimientos</u> .....	6
<u>1. Introducción</u> .....	8
<u>1.1 Reseña sobre el estudio del canto aviar en suboscinos</u> .....	8
<u>1.2 Introducción a las redes neuronales artificiales</u> .....	12
<u>1.2.1 Introducción al algoritmo de backpropagation</u> .....	14
<u>1.2.2 Introducción a las redes neuronales convolucionales</u> .....	17
<u>1.2.3 Introducción a las redes siamesas</u> .....	21
<u>2. Hipótesis, predicciones y objetivos</u> .....	25
<u>3. Base de datos y trabajo de campo</u> .....	27
<u>3.1. Base de datos 2004-2005</u> .....	27
<u>3.2 Base de datos 2023</u> .....	27
<u>4. Caracterización acústica de las sílabas de hembra <i>furnarius rufus</i>: delineación de sílabas alfa y beta</u> .....	32
<u>4.1 Métodos</u> .....	32
<u>4.1.1 El espectro de cada sílaba de hembra</u> .....	32
<u>4.1.1.1 Resultados del análisis acústico de la hipótesis de trabajo 1: caracterización acústica de las sílabas de hembra</u> .....	34
<u>4.1.2. El pitch de cada sílaba de hembra</u> .....	36
<u>4.1.2.1 Extracción del pitch de la hembra en grabaciones de un solo de hembra con poco ruido</u> .....	38
<u>4.1.2.2 Extracción del pitch de las sílabas de la hembra en grabaciones con ruido y en un dueto</u> .....	38
<u>4.1.2.3 Identificación de silencios en las sílabas de hembra</u> .....	41
<u>4.2 Resultados del análisis acústico a partir del pitch de cada sílaba de hembra</u> .....	42
<u>4.3 Conclusiones sobre la caracterización acústica de sílabas alfa y beta de las hembras</u> .....	45
<u>5. Evaluación de la morfología de las sílabas y su relación con la identificación de individuos utilizando redes siamesas</u> .....	46
<u>5.1 Métodos</u> .....	46
<u>5.1.1 Implementación de la red y formato de los datos</u> .....	46
<u>5.1.2 Data augmentation</u> .....	48
<u>5.1.3 Metodología para el análisis de los resultados de una red neuronal siamesa entrenada</u> .....	50
<u>5.1.3.1 Predicción 5A</u> .....	51
<u>5.1.3.1.1 Cómo determinar la calidad de un conjunto de clusters</u> .....	52
<u>5.1.3.1.2 Cómo determinar la calidad de varios conjuntos de clusters</u> .....	52
<u>5.1.3.2 Predicción 5B</u> .....	54
<u>5.2 Resultados</u> .....	55
<u>5.2.1 Predicción 5A</u> .....	55
<u>5.2.1.1 Sílabas de hembra alfa</u> .....	55
<u>5.2.1.2 Sílabas de hembra beta</u> .....	61

<u>5.2.1.3 Sílabas de macho.....</u>	66
<u>5.2.1.4 Conclusiones sobre los resultados de la predicción 5A.....</u>	70
<u>5.2.2 Predicción 5B.....</u>	73
<u>    5.2.2.1 Visualizaciones.....</u>	73
<u>        5.2.2.1.1 Sílabas de hembra alfa.....</u>	73
<u>        5.2.2.1.2 Sílabas de hembra beta.....</u>	75
<u>        5.2.2.1.3 Sílabas de macho.....</u>	76
<u>    5.2.2.2 Cuantificación del grado de apareamiento.....</u>	78
<u>        5.2.2.2.1 Heatmaps.....</u>	78
<u>    5.4.3.2 Cuantificación de los heatmaps.....</u>	80
<u>    5.2.2.3 Conclusiones sobre los resultados de la predicción 5B.....</u>	83
<u>6. Conclusiones.....</u>	84
<u>7. Anexo.....</u>	86
<u>    7.1 Parámetros de Praat.....</u>	86
<u>    7.2 Extracción de datos a partir de los sonogramas.....</u>	88
<u>        7.2.1 Extracción del pitch de las sílabas de la hembra en grabaciones con ruido y en un dueto.....</u>	88
<u>        7.2.2 Extracción manual del pitch.....</u>	89
<u>        7.2.3 Identificación y extracción silencios dentro de una sílaba alfa.....</u>	89
<u>        7.2.3 Sonogramas en los que se observan sílabas de macho con morfología de sílaba alfa de hembra.....</u>	91
<u>        7.2.4 Subconjunto del total de los arámetros utilizados en las sílabas alfa de nidos A, B, 19, 23, 34.....</u>	92
<u>    7.3 Espectros de las sílabas sin procesar.....</u>	93
<u>    7.4 Comparaciones de medias de los scores.....</u>	93
<u>        7.4.1 Estimaciones para las comparaciones múltiples.....</u>	93
<u>        7.4.2 Supuestos distribucionales del GLM para comparación de las medias de los scores.....</u>	95

## Resumen

El hornero (*Furnarius rufus*) es un ave suboscina altamente abundante en el Sur de Sudamérica, y es el ave nacional de la República Argentina. Los estudios enfocados en aves suboscinas son pocos en comparación con los referidos a su grupo hermano: las aves oscinas, que son ampliamente utilizadas como modelos de aprendizaje vocal. La importancia del estudio del canto en suboscinos radica en que son el grupo de aves filogenéticamente más cercano a los oscinos y que, salvo a excepción de una especie, no aprende a cantar, sino que se cree que su canto es innato. Sin embargo, estudios han mostrado una riqueza en su comportamiento, en la dinámica de su canto y en la anatomía de su siringe mayor a la esperada para un suboscino, motivo por el cual hay actualmente esfuerzos de investigación enfocados en su comportamiento, fisiología, ecología y evolución. Las aves estudiadas en el suborden de los suboscinos cuentan con un control neuromuscular y complejidad silábicas apreciablemente menor que el de los oscinos y, en base a los casos estudiados hasta ahora, se cree que no es posible distinguirlas entre ellas a partir de sus sílabas. Por ende, lo que se esperaría del hornero es que sus sílabas continúen con esa tendencia. En esta tesis realizamos trabajo de campo para ampliar una base de datos existente de grabaciones de horneros. Luego de un trabajo de análisis y depuración de la base de datos, definimos trabajar con las grabaciones de cantos de 8 parejas de horneros. Estudiamos características acústicas del canto y formulamos hipótesis respecto de la posibilidad de distinguir distintos tipos de sílabas en un mismo individuo, y distintos individuos a partir de sus sílabas. Estas hipótesis fueron puestas a prueba mediante análisis cuantitativos de los espectros de frecuencias y análisis de sonogramas, a partir de los cuales extrajimos los valores de frecuencias fundamentales de las sílabas. Implementamos herramientas de aprendizaje automático para la clasificación de sílabas. En particular, entrenamos redes neuronales siamesas para la identificación de individuos a partir de sus sílabas utilizando imágenes de la evolución temporal de la frecuencia fundamental. Los resultados de esta tesis sugieren que es posible distinguir acústicamente individuos hembra utilizando características de sus cantos, mientras que las características acústicas de las sílabas de macho no son suficientes para distinguir individuos de una manera categórica. De esta manera, esta tesis motiva trabajos futuros referidos al estudio de la generación del canto en horneros, como también del control neuromuscular en suboscinos, en virtud de la capacidad de las hembras de hornero de exhibir firmas de individualidad en sus sílabas.

# Abstract

The hornero (*Furnarius rufus*) is a suboscine species widely distributed in southern South America, and is Argentina's national bird and a cultural symbol. The study of suboscine birds are few compared to those of their sister taxon: oscine birds, which are widely used as animal models for vocal learning. The importance of studying song production in suboscine birds stems from their being the closest phylogenetic group to oscines that are thought to *not* learn to sing, save for a single species; rather, their songs are thought to be innate. However, several studies have shown it displaying complexity in its behaviour, in the dynamical properties of their duetting songs and in the anatomy of their syrinx, which are greater than what should be expected. For these reasons, there are currently joint research efforts focusing on their behaviour, physiology, ecology and evolution. The suboscine species that have been studied have been shown to count on lesser neuromuscular control and complexity within their syllables than oscine species, and on the basis of the cases studied thus far, their individuals are thought to be indistinguishable from each other by means of their syllables. For these reasons, syllables sung by the rufous hornero should obey this trend. In this thesis we performed field work in order to expand an existing database of hornero recordings. After working on its analysis and depuration, we decided to work on recordings of 8 hornero couples. We studied the acoustical properties of their songs and formulated hypotheses regarding the possibility of identifying types of syllables in a given individual, and individuals by means of their syllables. These hypotheses were put to the test through quantitative analyses of their frequency spectra and sonograms, from which we extracted their fundamental frequency. We implemented machine learning methods for the task of individual detection based on their syllables. In particular, we trained siamese neural networks on images of the temporal evolution of the fundamental frequency. The results of this thesis suggest the possibility of distinguishing female horneros based on acoustical properties of their songs, while the acoustic characteristics of male song are not categorically sufficient for such a task. Thus, this thesis motivates future work on the study of the horneros' song production, as well as on the suboscines' neuromuscular control, owing to the capacity of female horneros to display signatures of individuality in their syllables.

# Agradecimientos

A mi familia, que me vio y ayudó a crecer toda mi vida. A mis padres por darme la oportunidad del tiempo para seguir mi pasión. Sin ustedes nada de esto habría sido posible. A Ale: me hiciste sentir querido como hermano, me enseñaste a sumar y multiplicar fracciones en la primaria, me enseñaste las reglas de exponentes en el CBC, y acá estoy. Gracias por ser mi hermana desde siempre. A Sesé y Sergio, por apoyarme y recibirme siempre en su casa con los brazos abiertos y por siempre preguntarme por lo que quiero hacer y estudiar, es hermoso que sean mi familia. Al Abeló, que no me verá recibirme, pero fue, es y será siempre mi abuelo. A Tatau y Connie, tío y madrina, por recibirme en su casa y motivarme en este camino.

A Lihuén. Me conociste cuando me movía en una sola dimensión, con mi foco en esta carrera, y me enseñaste a disfrutar las pausas para apreciar tu compañía. Fuiste mi roca siempre que me viste agotado. Vamos a seguir cantando a dueto, y quiero seguir escuchando tus solos. Gracias infinitas por el amor que me diste en todo este camino, gracias a Pato y Ricky, por hacerme parte de su familia, y por los innumerables días en los que mi descanso de estudiar fue la mesa y la conversación con ustedes.

En Exactas conocí a mucha gente, y a nadie lo conocí por casualidad. En este hermoso lugar intenté y logré rodearme de gente buena y que tiene ganas de crecer del mismo modo en que quise crecer yo. Que se alegra por aprender, que quiere estar en la frontera de lo que conocen, cada uno con su manera de ser en la suya. En cada uno veo un ejemplo de cómo manejarme en mi propia frontera y agradezco que sean quienes son, porque seguro que le hacen bien a la gente de la que ustedes se rodearon, incluyéndome a mí.

Agus y Gian, llegaron en la primavera de mi vida para ser ejemplos de cómo encarar la juventud y avanzar en mi frontera con ímpetu. Agus, estuviste siempre escalones arriba mío (salvo en ping pong), en muchas ocasiones en la carrera creíste en mí antes que yo, me mostraste lo bien que encarás desafíos, y lo seguís haciendo. Tu éxito es inminente. Me viste hacer mi primera secuencia de Fibonacci en Python, me hablaste maravillas del LSD, y gracias a ello estoy acá. Gian, sos el ejemplo de la osadía necesaria para buscar el crecimiento en más de una dirección, aportando evidencia a favor de la hipótesis de que buscar expandirme en más de una frontera las terminaría unificando. Te vi hacer la apuesta y ganar. Gracias a tu ejemplo aprendí más y mejor.

A Facu y Ger, que fueron mi primer cable a tierra en la carrera, y con quienes compartí hasta el final. Fue un verdadero placer caminarla con ustedes.

A Martina Radice. Gracias por aconsejarme a lo largo de mi carrera, desde mi primera materia hasta mi búsqueda por un laboratorio en el cual dar mis primeros pequeños pasos en la ciencia. Tus consejos los transmití a mucha gente, y fueron fundamentales para estar donde estoy. A Maca, Mica y Juan. Las conversaciones que iniciaron conmigo me ayudaron a disfrutar mucho más la parte de la carrera que compartimos.

A los miembros del Laboratorio de Sistemas Dinámicos. Tenía muchas ganas de poder rodearme de ustedes: gente buena y llena de ganas de entender el sistema nervioso, y la biología. Gracias a Hernán y Roberto, con la ayuda que me ofrecieron pude desatar pasos muy importantes en este trabajo. Hernán, le doy muchísimo valor a que te hayas sentado al lado mío frente a mi computadora (más de una vez) a la tarde o noche en el labo, para explicarme con tanta lucidez el funcionamiento de las herramientas que usé en esta tesis. A Javi Lassa: me recibiste en el labo y me ayudaste a sentirme cómodo desde un comienzo. Tus consejos como biólogo para mi formación fueron fundamentales para mí, y siempre te voy a estar agradecido. A Facu, a quien admiraba antes de entrar al labo por su lucidez en sus clases de Dinámica no lineal. Es un placer conocerte, que te hayas interesado en mi trabajo, y gracias por haberme recomendado descansar las veces que me viste quemado en la computadora. A Lean y Fi: las conversaciones con ustedes me enseñaron un par de cosas muy importantes, y me hicieron sentir más parte del día a día labo, algo que valoro muchísimo. Lean, fue un placer compartir charlas en el labo entrada la noche. A Tomás Bossi: no llegué a cruzarte en el labo, pero tu marca en esta tesis es muy grande. Tu trabajo fue mi primera aproximación a la herramienta que usé, y la ayuda tan clara y desinteresada que me ofreciste para la programación de la red no la doy por sentado.

Por supuesto, a Ana y Gabo. A Ana, por haberme guiado en el proceso completo de un trabajo científico y darme innumerables lecciones en el camino. Fuiste una directora excelente, y gracias a tu formación y lo que hicimos en el labo pude salir de Exactas con seguridad a intentar responder preguntas de biología. Me tuviste fe y siempre te voy a estar agradecido por lo que hiciste por mí. A Gabo, por abrirme las puertas del LSD. Antes de entrar al laboratorio, la ciencia que hacés hace mucho tiempo me mostró la forma en la que quería estudiar el cerebro, y por eso me preparé para aprovechar la oportunidad que me diste. Gracias por haberme recibido junto con Ana y con este proyecto. Haber pasado por este laboratorio es un orgullo enorme para mí.

A Hernán Améndola, quien tan amablemente nos abrió la puerta de la Estación de Cría de Animales Silvestres (ECAS) para que hagamos ciencia en la reserva ecológica.

A Grant Sanderson y Gustavo Lado. A Patricia Bottino, maestra de Biología de la secundaria. Con vos dibujé por primera vez una neurona, y desde entonces no volví atrás.

Gracias al hornero. Por mi parte, me encargué de que toda la gente que me rodea sepa quién de ustedes es la hembra y quién el macho. La literatura referida a ustedes indica que no aprenden a cantar, pero yo estoy seguro de que aprendí lo comprendido en estas 100 páginas gracias a ustedes.

A los jurados que se toman el tiempo de leer mi tesis, para culminar la carrera que tanto me dio.

A toda la comunidad de la Facultad de Ciencias Exactas y Naturales, a la ciencia y la educación pública y de calidad.

# 1. Introducción

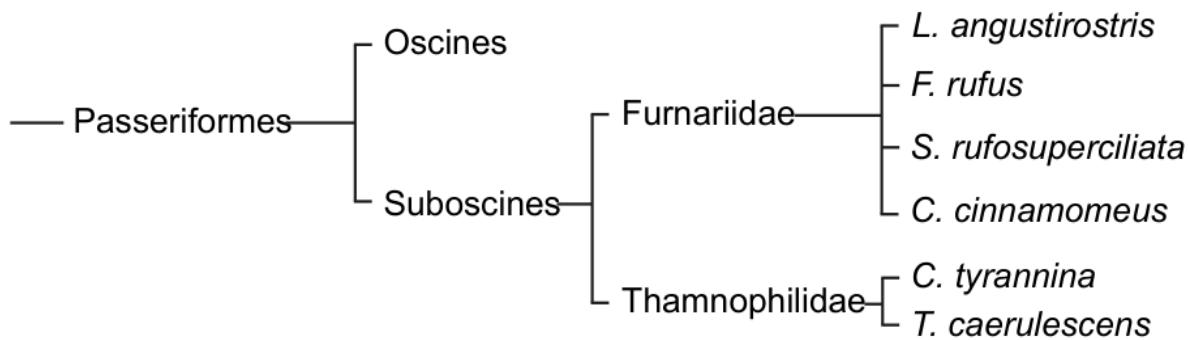
## 1.1 Reseña sobre el estudio del canto aviar en suboscinos

El estudio de un fenómeno biológico requiere de un organismo modelo que permita sistematizar la construcción del conocimiento. En el caso de la producción de vocalizaciones las aves resultan ser un excelente modelo animal, dado que el canto es crucial para su cortejo, competencia y defensa de territorio (Amador & Mindlin 2023). Las aves oscinas (del orden *Paseriforme*) son utilizadas como modelo para el estudio del control neuromuscular y el aprendizaje vocal<sup>1</sup>, que les permiten generar cantos que consisten en secuencias complejas y de una alta riqueza espectral (e.g. Zeigler & Marler 2008, Mooney 2009, Nieder & Mooney 2020, Scharff & Adam 2013). El grupo hermano de los oscinos comprende a los suboscinos (suborden *Tyranni*) (fig. 1.1), que consisten principalmente en una radiación neotropical de más de 1000 especies (Oliveros 2019). Este grupo, que representa el 26% de las aves paseriformes, presenta marcadas diferencias con los oscinos, dentro de las cuales pueden resaltarse la neuroanatomía y la producción vocal (Gahr 2000, Liu 2013). Salvo a excepción de una especie (Kroodsma 2013), no se ha reportado evidencia de aprendizaje vocal en las aves suboscinas: no se han observado dialectos en cada especie, se considera que su canto es innato, y los suboscinos estudiados no cuentan con el sustrato neuroanatómico característico del aprendizaje presente en los oscinos (Kroodsma & Konishi 1991, Touchton 2014, Amador & Mindlin 2023), aunque se han encontrado centros neuronales que podrían ser substratos rudimentarios de aprendizaje vocal (Liu 2013). El conocimiento acotado sobre suboscinos puede deberse a que la mayoría de los estudios sobre biología evolutiva y ecología se originan en el hemisferio norte, estudiando aves de esa región y, por lo tanto, limitando el conocimiento global (Nuñez 2021, Theuerkauf 2022).

En las últimas décadas grupos de Sudamérica han estudiado al hornero, aportando abundante evidencia de que el hornero puede servir como un excelente modelo para el estudio del canto en aves, y en esta tesis se espera aportar en esa dirección.

---

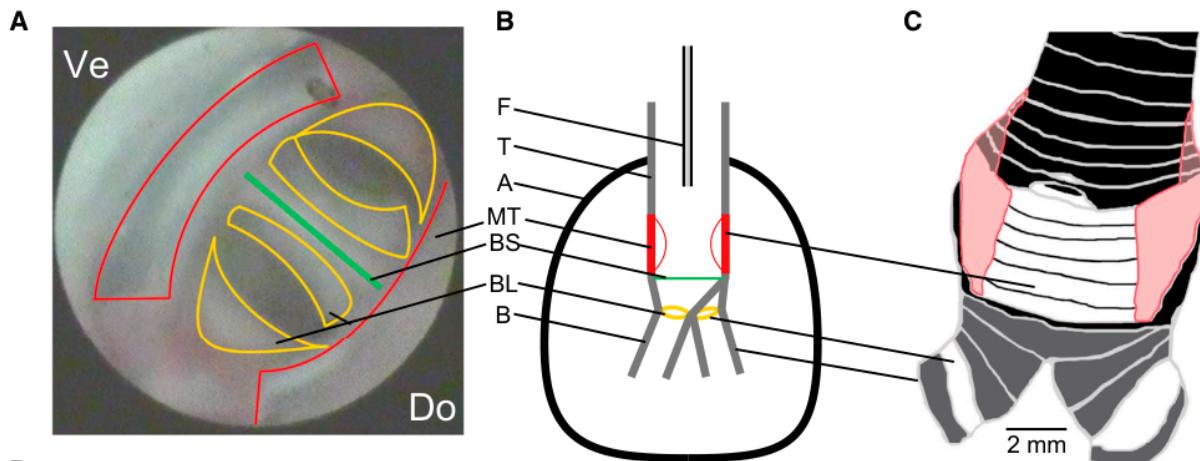
<sup>1</sup>La habilidad de modificar las vocalizaciones innatas a través del aprendizaje por imitación a un tutor.



*Fig. 1.1 adaptada de (Garcia 2017): Evolution of Vocal Diversity through Morphological Adaptation without Vocal Learning or Complex Neural Control, mostrando el lugar que el hornero (*F. rufus*) ocupa en la filogenia de los Passeriformes.*

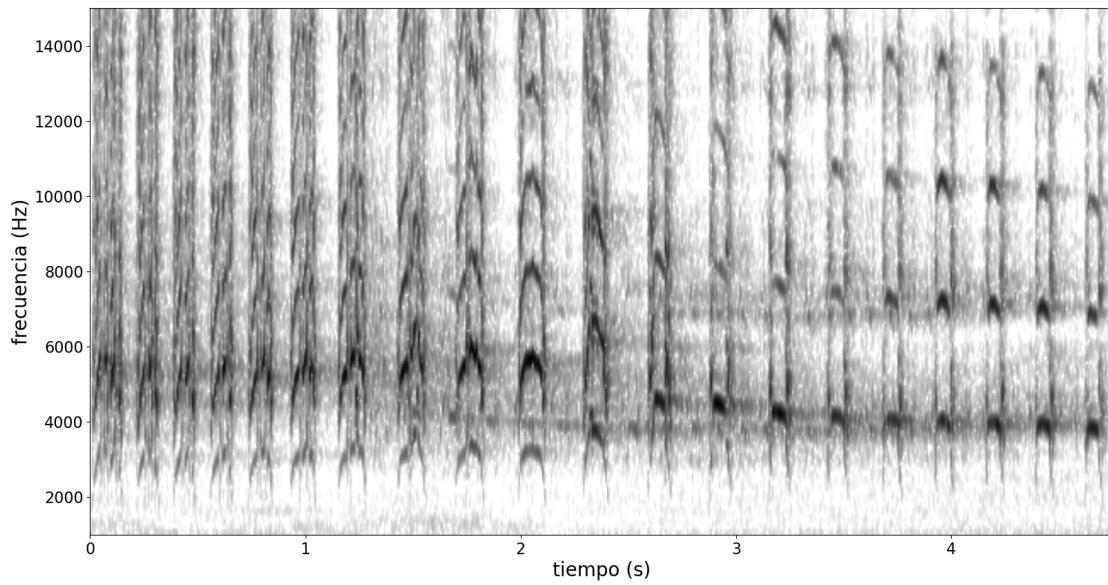
El hornero es reconocido gracias a los nidos que construye con barro, raíces y pequeñas ramas, y por su capacidad de cantar a dueto con su pareja (Adreani 2020, Fraga 1980). Desde la primera descripción de la especie (Gmelin 1798) ha habido pocos estudios sistemáticos. Sin embargo, en las últimas décadas se generó un cuerpo de trabajo sobre aves suboscinas que muestra la complejidad presente en la dinámica de sus cantos (e.g. Amador, Trevisan & Mindlin 2005, Laje & Mindlin 2003), en su comportamiento (e.g. Fraga 1980, Diniz 2018, Massoni 2012), durante el sueño (Döppeler et al. 2021) y en su aparato fonador (e.g Garcia 2017). En este último trabajo se describe la siringe de los traqueófonos, un clado dentro de los suboscinos al que pertenece el hornero, cuya sinapomorfía es un conjunto de membranas en las superficies dorsales y ventrales de la tráquea justo por encima de la bifurcación traqueal: la Membrana Tracheales (MT) (en rojo en Fig. 1.2). En dicho trabajo se expone que la siringe de los traqueófonos cuenta no solo con la MT como se creía hasta ese momento, sino que también tiene fuentes sonoras labiales en cada bronquio (amarillo en Fig 1.2), siendo el primer órgano vocal en el que se encuentran tres fuentes sonoras. En el mismo trabajo se demuestra que ésta permite producir cantos con gran riqueza acústica al interactuar con las oscilaciones de la MT. Una conclusión extraída a partir de estos resultados es que la diversidad morfológica en suboscinos permite aumentar su riqueza acústica en ausencia de aprendizaje vocal (Amador & Mindlin 2023). Por otro lado, contar con tres fuentes sonoras podría implicar requerir del sustrato neuroanatómico para controlarla. El hecho de que el descubrimiento de un componente fundamental del aparato fonador de los traqueófonos sea tan reciente,

sumado a los trabajos mencionados que evidencian una complejidad inusual en las vocalizaciones de los suboscinos, motiva continuar su estudio.



*Fig. 1.2 adaptada de (Garcia 2017): Evolution of Vocal Diversity through Morphological Adaptation without Vocal Learning or Complex Neural Control. Fig. B: esquema de una vista lateral de la siringe de los traqueófonos. F: fiberscope, T: trachea, MT: membrana trachealis, BS: bronchial septum, BL: bronchial labia, A: air sac, B: bronchus.*

En esta tesis estudiamos características acústicas de los cantos de horneros, utilizando como herramienta fundamental sus espectrogramas o sonogramas (Amador & Mindlin 2023b) (ejemplo se muestra en fig 1.3). Se realizaron grabaciones de campo originales y se llevó a cabo un estudio sistemático de las propiedades acústicas de las sílabas de horneros macho y hembra, postulándose hipótesis referidas a su potencial para funcionar como firmas de individualidad. Esta última pregunta fue respondida mediante la implementación de redes neuronales siamesas, una técnica de aprendizaje automático que en años recientes se ha convertido en una de las herramientas más utilizadas para el reconocimiento de aves, ya sea de especies o, como es en este caso, de individuos. Una propuesta original de esta tesis es dirigir esta pregunta a un ave suboscina; según la literatura referida a este suborden, los individuos no deberían ser distinguibles a partir de su canto.



*Fig. 1.3: Sonograma de un canto de hornero hembra. La intensidad de grises en cada intervalo de tiempo indica la amplitud de la frecuencia. La evolución creciente de la frecuencia se denomina upsweep, y la decreciente downsweep. Cada sílaba dura alrededor de 100ms.*

Los resultados aportan evidencia a favor de la hipótesis central propuesta en esta tesis, por la cual existen robustas firmas de individualidad en las sílabas de los horneros hembra, aportando al cuerpo de evidencia que motiva su estudio en la generación del canto y del control neuromuscular en suboscinos.

A continuación se expone una introducción a la herramienta que fue utilizada para llevar a cabo la distinción entre individuos a partir de sonogramas de sus sílabas: las redes neuronales siamesas. Para entender lo que es un red siamesa, será conveniente adquirir en primer lugar una intuición sobre la naturaleza de un tipo de red neuronal llamado perceptrón multicapa y, luego, en una red convolucional.

## 1.2 Introducción a las redes neuronales artificiales

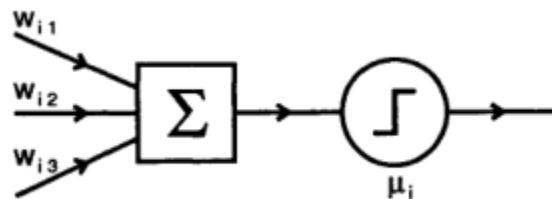
Una red neuronal artificial es un modelo de aprendizaje automático comúnmente utilizado para tareas de clasificación y regresión: tareas en las cuales se aprende a partir de datos conocidos con el fin de generalizar a datos no conocidos. “Un algoritmo de aprendizaje automático *aprende de la experiencia E* con respecto a una clase de *tareas T* y medida de *rendimiento P* si P mejora con T a mayor E” (definición adaptada de (Tom M Mitchell)). Lo que es aprendido son los parámetros de una función; una red neuronal artificial *feedforward*, también llamada perceptrón multicapa (Fig. 1.2), es un ejemplo de este tipo de funciones cuyos parámetros es posible aprender. Es el ejemplo más simple de una red neuronal *profunda*: es una función vectorial; mapea datos de entrada a una salida.

El aprendizaje de representaciones (*representation learning*) es un conjunto de métodos que permite a una máquina, a partir de datos crudos, aprender las representaciones de los datos que le permiten resolver tareas de detección o clasificación (LeCun 2015). El aprendizaje *profundo* consiste en métodos de aprendizaje con múltiples niveles de representación, obtenidos mediante la combinación de módulos simples pero no lineales (fig. 1.1) que transforman la representación en niveles sucesivos, alcanzando mayores grados de abstracción en cada uno, amplificando los aspectos de la entrada que mejoran el rendimiento en la tarea y suprimiendo los irrelevantes (LeCun 2015). El aspecto fundamental en el aprendizaje profundo es que las representaciones generadas en cada nivel no son diseñadas por humanos, sino que son descubiertas mediante un algoritmo de aprendizaje genérico (LeCun 2015).

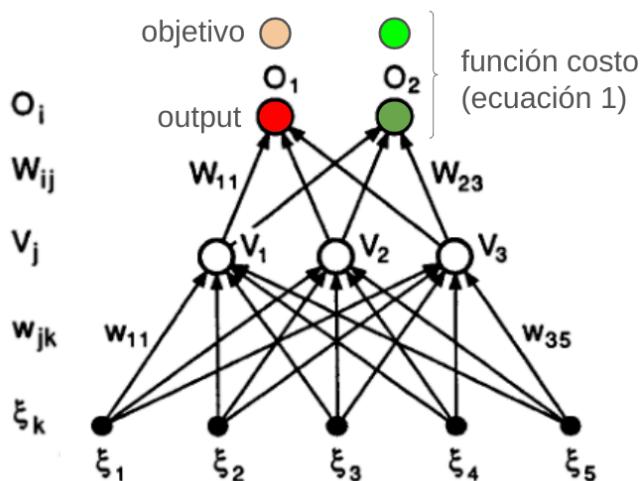
Un perceptrón multicapa consiste en una serie de vectores (al menos tres para ser un perceptrón multicapa o red neuronal feedforward profunda) conectados por matrices; las coordenadas de cada vector se llaman “neuronas” (Fig. 1.1), cada vector es llamado “capa de neuronas”, y los coeficientes de las matrices, que conectan estas neuronas, se llaman “pesos”. La red es llamada *feedforward* porque la información fluye unidireccionalmente desde el vector de entrada hasta el vector de salida (ergo la utilización del término *serie* de vectores). Las coordenadas de los vectores son llamadas neuronas porque, salvo la capa de entrada, suelen ser una función no lineal del input que reciben; ese input es la combinación lineal de las neuronas de la capa anterior con sus pesos relativos, y la función no lineal suele ser una sigmoidea, que mapea la combinación al rango 0-1, ergo el término “activación” de una neurona. Esta no linealidad en un perceptrón multicapa es fundamental para su capacidad de

aproximar cualquier función continua definida en un conjunto compacto de  $R^n$  (Funahashi 1989) dado un número suficiente de neuronas.

En la fig. 1.2 se puede apreciar que en el vector de salida hay una combinación no lineal de los datos de entrada, mediante el vector intermedio entre ellos, llamado “capa oculta”. Las combinaciones que le permiten a la red mejorar su rendimiento en la tarea son aprendidas mediante la corrección de los pesos de las matrices que mapean un vector de una capa al de la capa siguiente. Si hubiera más capas ocultas, se podrían obtener representaciones más abstractas de los datos de entrada mediante sucesivas combinaciones de las anteriores. Nótese que, con esta arquitectura, se cuenta con la posibilidad de *descubrir* combinaciones de variables relevantes (aprendizaje de representaciones), sin necesidad ni intención de conocerlas. En el aprendizaje profundo es posible no reparar en la interpretación y significancia de los parámetros que pueden mediar una tarea de, por ejemplo, clasificación, sino únicamente utilizar esta herramienta en virtud de su poder predictivo ante datos no vistos durante el entrenamiento.



*Fig. 1.1 adaptada de (Hertz): diagrama esquemático de una única neurona artificial, en la cual se aplica una función no lineal (usualmente sigmoidea, con imagen entre 0 y 1) a la combinación lineal de entrada.*



*Fig. 1.2: Figura modificada de (Hertz). Cada círculo representa una neurona del perceptrón, con una capa oculta  $V$  y capa de input  $\xi$ . La dirección de las flechas esquematiza un paso feedforward a partir del input hasta computarse el output (por ejemplo, un valor entre 0 y 1), que es comparado con el vector objetivo correspondiente al input. “W” indica “Weights”.*

En el caso del aprendizaje automático supervisado, a diferencia del no supervisado, la experiencia consiste en la comparación entre la salida de la función vectorial a partir de un dado vector de entrada (llamado *ejemplo*), y el valor que realmente debería tomar, lo cual es un dato conocido. El valor que debería tomar se conoce como *vector objetivo*, y la comparación entre la salida de la red y el vector objetivo es el argumento de una *función costo* que, como su nombre lo indica, computa el costo -o el error- de la función ante un dado ejemplo; evalúa el rendimiento de la red. Un ejemplo de una función costo es la cuadrática (1.1).

$$E[\mathbf{w}] = \frac{1}{2} \sum_{\mu i} \left[ \zeta_i^\mu - g \left( \sum_j W_{ij} g \left( \sum_k w_{jk} \xi_k^\mu \right) \right) \right]^2 \quad (1.1)$$

Sus índices se corresponden con la fig. 1.2. Es continua y diferenciable.  $\mu$  es el índice de un ejemplo,  $i$  es el índice de una neurona de salida,  $j$  el de una neurona oculta y  $k$  el de una de entrada.  $\zeta$  es el vector objetivo y la función sigmoidea es la composición de la neurona  $i$ .

De este modo, la función costo es una función compuesta por el perceptrón, que es la composición de múltiples funciones vectoriales no lineales (las neuronas) conectadas entre sí mediante pesos.

Conociendo el paso feedforward de una red neuronal, resulta pertinente retomar el objetivo de su implementación: la capacidad de generalización. Ésta puede ser evaluada con el rendimiento de la red entrenada ante datos que no hayan sido utilizados para su entrenamiento, pero para los cuales se tenga sus vectores objetivo. A continuación, se presenta una introducción al proceso de entrenamiento de la red.

### 1.2.1 Introducción al algoritmo de backpropagation.

*Un dado peso se corrige con el producto entre el valor de la neurona de entrada y el error de la neurona de salida.*

La optimización de los pesos del perceptrón (fig. 1.2) se logra minimizando la función costo; luego de un paso feedforward y del cálculo del error al comparar la salida

con el vector objetivo, cada peso de la red es corregido en base a sus contribuciones relativas al error global. El algoritmo que permite cuantificar la contribución de cada peso individual al error global, y por ende ofrecer una prescripción para la corrección de cada uno, se denomina *backpropagation*: retropropagación del error (Hertz), que no es más que una extensión del descenso por gradiente a una función vectorial con capas ocultas cuya activación es no lineal. Si el perceptrón multicapa (fig. 1.2) no tuviera capas ocultas (perceptrón simple), y fuera utilizado para una tarea de clasificación, cada unidad de salida computaría una regresión logística cuyos parámetros podrían ser optimizados mediante descenso por gradiente, y serían interpretables en términos de las variables de entrada. El término *backpropagation* sugiere que la corrección de los pesos debería contemplar el error que hayan inducido, en un paso feedforward, en las neuronas de la capa siguiente. Por ende, se puede considerar que el primer paso para la corrección de los pesos de un perceptrón implica la utilización del error **de las neuronas de la capa de salida** (1.2):

$$\delta_i^\mu = g'(h_i^\mu)[\zeta_i^\mu - O_i^\mu] \quad (1.2)$$

donde  $g(h)$  es la función de activación sigmoidea evaluada en la combinación lineal que recibe de la capa anterior. El cálculo de 1.2 luego de un paso feedforward con un dado ejemplo es trivial, ya que se cuenta con el vector objetivo  $\zeta$  para la comparación de las salidas  $O$ . Sus índices se corresponden con la fig. 1.2. En palabras, el error en la neurona  $i$  de la capa de salida es el producto entre su propia tasa de cambio respecto de su entrada (la derivada de la función de activación sigmoidea  $g(h)$ ) y la diferencia con el objetivo  $i$ .

La operación fundamental en backpropagation es la regla de la cadena, ya que lo que se busca no es más que calcular el gradiente de la función costo respecto de cada peso con el fin de encontrar los parámetros que la minimicen, y la función costo es continua y diferenciable. El punto clave que aporta este algoritmo para el aprendizaje profundo es una consecuencia directa de la regla de la cadena (1.3), y consiste en el cálculo del error cometido por una dada neurona **de una capa oculta** (la capa V en la fig. 1.2) de la red. A diferencia de lo que ocurre en las capas de salida (1.2), en las neuronas de una capa oculta no se cuenta con un valor objetivo ( $\zeta$ ) con el cual realizar la comparación. Más bien, la prescripción ofrecida para **la corrección de un peso cuya neurona siguiente es de una capa oculta** (los pesos  $W_{jk}$  en la fig. 1.2) conduce a **la definición del error en la neurona siguiente** (en violeta en 1.3) como el producto entre dos factores:

- *Su aporte al error de cada neurona de la capa de salida*, cuantificado como el gradiente de la función costo respecto de todas las contribuciones de esta

neurona oculta, en azul en (1.3). El término “contribución” de una neurona al error de una neurona siguiente se refiere a la combinación lineal del valor de la primera con el peso que la conecta a la segunda.

- *La tasa de cambio de la propia neurona respecto de la entrada que recibe*, cuantificada como su derivada respecto de la combinación lineal que recibe de la capa anterior, en rojo en (1.3). Las funciones de activación que se suelen utilizar para las neuronas de capas ocultas, como el caso de la función logística, tienen la bondad de que ya han sido en parte calculadas en el paso feedforward con cuyo error se están corrigiendo los pesos mediante la tasa de cambio de la propia neurona: la derivada de la función de activación logística que recibe un input  $z$  es  $\sigma'(z) = \sigma(z) \cdot (1-\sigma(z))$ . Esto supone una ventaja computacional muy grande.

Habiendo llegado a esta definición del error en una dada neurona oculta, **el segundo factor para el cálculo de la corrección del peso** que la precede es el valor de la activación de la neurona de entrada (verde en ecuación 1.3). Esto también es una consecuencia de la regla de la cadena, ya que el gradiente de la neurona de entrada respecto del peso que se quiere corregir es el valor de la neurona. Se define el cambio del peso que conecta a la neurona  $k$  de la entrada con la  $j$  oculta como:

$$\begin{aligned}\Delta w_{jk} &= -\eta \frac{\partial E}{\partial w_{jk}} = -\eta \sum_{\mu} \frac{\partial E}{\partial V_j^{\mu}} \frac{\partial V_j^{\mu}}{\partial w_{jk}} \\ &= \eta \sum_{\mu i} [\zeta_i^{\mu} - O_i^{\mu}] g'(h_i^{\mu}) W_{ij} g'(h_j^{\mu}) \xi_k^{\mu} \\ &= \eta \sum_{\mu i} \delta_i^{\mu} W_{ij} g'(h_j^{\mu}) \xi_k^{\mu} \\ &= \eta \sum_{\mu} \delta_j^{\mu} \xi_k^{\mu} \quad (1.3)\end{aligned}$$

$$\boxed{\delta_j^{\mu} = g'(h_j^{\mu}) \sum_i W_{ij} \delta_i^{\mu}.} \quad (1.4)$$

donde  $g'(h_j^{\mu})$  es la derivada de la función de activación (comúnmente sigmoidea) de la neurona  $j$  ante el ejemplo de entrada  $\mu$ . Los índices se corresponden con la fig. 1.2.

En resumen, **un dado peso se corrige con el producto entre el valor de la neurona de entrada y el error de la neurona de salida**, y esto es una consecuencia directa de la aplicación de la regla de la cadena. Esto provee una herramienta poderosa para el entrenamiento de modelos con relativamente bajo costo computacional.

Habiendo pasado por una introducción al perceptrón multicapa, resta adquirir una intuición de las operaciones involucradas en las redes neuronales convolucionales.

### 1.2.2 Introducción a las redes neuronales convolucionales

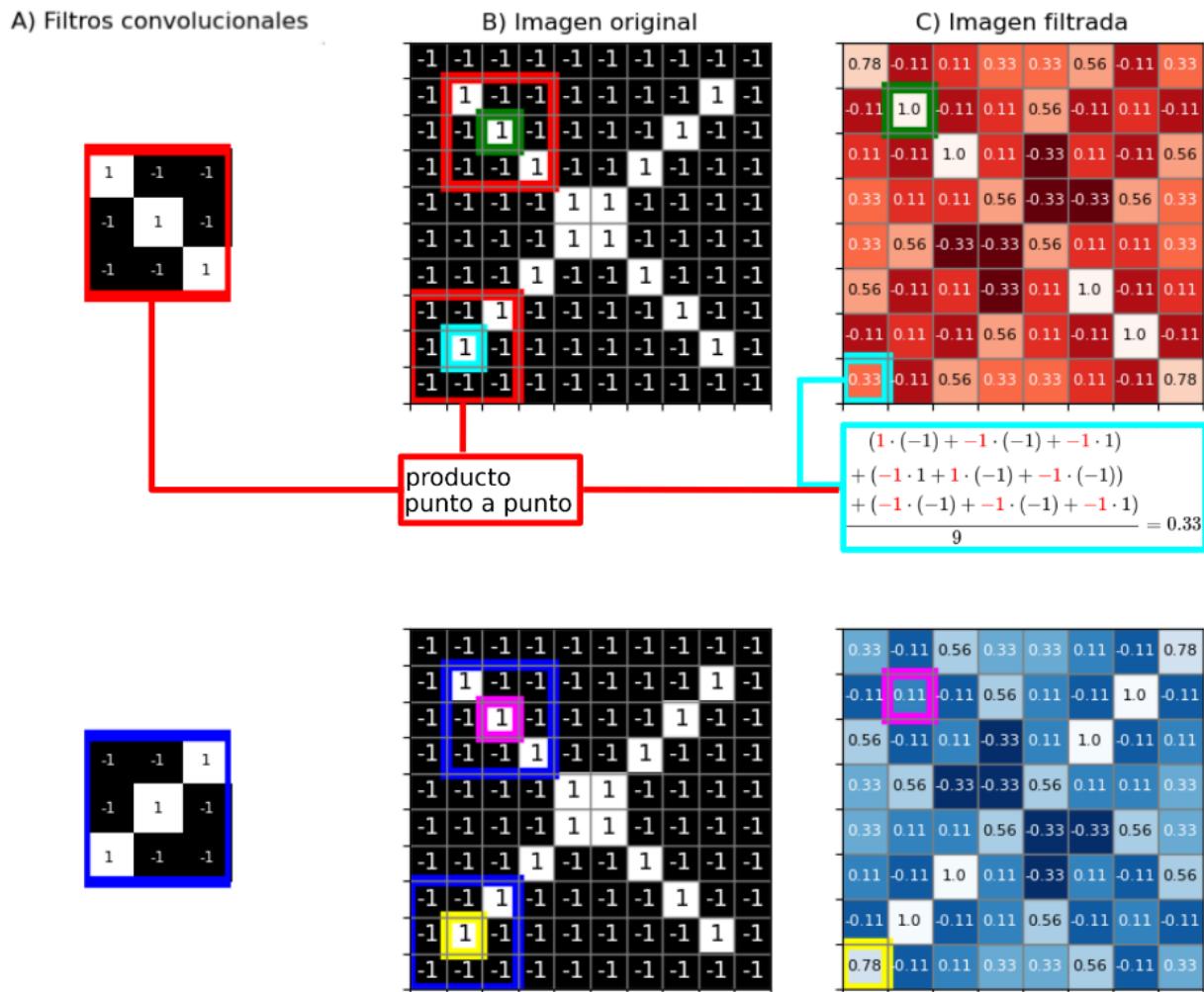
Una red neuronal convolucional también es una red profunda, lo cual permite en sucesivas capas capturar características abstractas en los datos (imágenes), de ahora en más *features*, mediante la combinación no lineal de features capturados en capas anteriores. Por ejemplo, los features aprendidos en la primera capa típicamente representan la presencia o ausencia de bordes en ubicaciones particulares de la imagen. La segunda capa suele detectar combinaciones de bordes sin importar la ubicación en la imagen, y las capas subsiguientes detectan objetos como combinación de los de la capa anterior (LeCun 2015).

Las imágenes que procesa una red convolucional son representadas como una matriz. Si la imagen es blanco y negro (fig. 1.3b), la entrada no es más que una matriz binaria. La salida de una red convolucional puede ser la misma que la de un perceptrón; a continuación se presenta una intuición de la forma en la cual se logra mapear una imagen a un vector de dimensión mucho menor conteniendo información espacial y local de la imagen.

Una neurona en una red convolucional tiene un *campo receptivo* de la entrada, a diferencia de una neurona en un perceptrón, que puede recibir mediante pesos información de todas las neuronas de la capa anterior. Es decir, es capaz de observar tan solo un parche de la imagen (fig. 1.3b): captura información local. ¿Qué información? Para responder esta pregunta, hace falta sumar el concepto de *kernel* o *filtro* (son tomados como sinónimos al tratarse de imágenes 2D), que no es más que una matriz, usualmente de 3x3 o 5x5 (fig. 1.3a) si la imagen tiene un solo canal (para una noción de escala, considérese que es común utilizar imágenes de entre 30x30 y 400x400). La particularidad de esta matriz es que se desliza solapadamente por toda la imagen; esta operación se denomina *convolución* (fig. 1.3b). Cada posición ocupada por el filtro al deslizarse por la imagen es lo que, en este contexto, es llamado una neurona. El resultado de esta operación puede pensarse como una forma de *ver la*

*imagen original* a través del filtro que se aplicó. Los parámetros que son aprendidos en esta parte de una red convolucional son, justamente, los de los filtros. Se pueden aplicar varios en la misma capa de procesamiento, y cada uno captura un feature local distinto de la imagen. El modo en que esto es codificado es mediante una operación punto a punto con la imagen (que es una matriz binaria) en cada posición que toma el filtro, se suman esos productos y se divide por el número de elementos del filtro, resultando en un único número para cada posición. En la fig. 1.3 se detalla la operación en el caso en el cual el filtro está ubicado con su centro en celeste.

Si hay un alto grado de coincidencia entre el filtro y la imagen en una dada posición, la neurona correspondiente a esa posición tendrá un valor alto (cercano a 1). La representación de la disposición de todos los puntajes en toda la imagen para un dado filtro se denomina *mapa de features* (fig. 1.3c). Esto es lo que permite ver *la imagen original a través del filtro*: permite ver en qué partes de la imagen el feature local (el filtro) está más representado.



*Fig. 1.3. Se esquematiza, para dos filtros distintos (uno en cada fila) en un paso feedforward de la red entrenada que recibe la imagen 1.2b como input, su aplicación sobre una imagen de un set de testeo; ambos en la misma capa de la red (reciben el mismo input). Dado que esta red está entrenada, los filtros se corresponden con features relevantes al espacio de los inputs. a) Dos filtros convolucionales; cada uno representa un feature aprendido. b) La imagen original indicando dos posiciones en los cuales pasa el filtro. El número del centro permite un seguimiento espacial del resultado de la operación punto a punto del filtro sobre el parche; este resultado se ubica en el mapa del filtro en c), exemplificado con la posición celeste. c) El mapa del filtro luego de haber pasado por la imagen muestra en qué regiones el feature está más representado.*

El feature identificado con rojo está presente en la imagen input (1.3b) con un match perfecto en dos ubicaciones, una de ellas con el centro verde; para facilitar la intuición de la operación de convolución, en este ejemplo se realizó un *padding* a la imagen original y se eligió un filtro de 3x3 con un stride de 1, lo cual permitió que, al superponer el filtro sobre la imagen, la posición en la misma pudiera ser identificada por el número en el centro. En el caso de la ubicación de centro verde, el filtro ya pasó por toda la fila superior y la primera columna de la imagen. En esta ubicación, al igual que en todas, se realiza la misma operación descrita anteriormente, cuyo resultado en este caso es 1: un match perfecto. Este valor es ubicado en una nueva matriz (1.3c) en la coordenada correspondiente. Como se mencionó previamente, dadas las dimensiones del filtro y de la imagen, esta nueva matriz tiene el mismo tamaño que la original, sin el padding, pero podría ser más pequeña si se eligiera un menor filtro o no se le hiciera un padding. Lo relevante es que se obtuvo en la fig. 1.3c una versión filtrada de la imagen original.

En cada etapa de convolución puede haber varios filtros, con lo cual se pueden generar muchos mapas de features para la imagen; a partir de una matriz original de  $M \times M$  se puede llegar a varias matrices de  $M \times M$  (o más pequeñas si se cambiara el padding y el stride). En este ejemplo se muestran dos filtros, resultando en dos mapas de features en la capa.

Un factor esencial en las redes neuronales convolucionales es la reducción de la dimensionalidad; en el procedimiento descrito hasta el momento, se cuenta con muchos mapas de features (se mostraron solo dos), y pueden tener el mismo tamaño que la imagen de input. Para reducir su dimensión, se procede a hacer un *pooling* sobre cada uno (fig. 1.4c). Esta operación consiste en tomar cuadrantes (en general de 2x2) sobre cada mapa de features, trasladarlos de dos en dos por cada parche de 2x2 de la imagen, y en cada uno tomar ya sea un promedio de los valores que se ven

(average pooling), o el valor máximo (maxpooling). En la fig. 1.4 se ejemplifica este procedimiento tomando un cuadrante verde del mapa de features. De este modo, luego del pooling se obtiene una imagen que puede pensarse como un resumen del mapa de features que, a su vez, es una versión filtrada de la imagen original.

El procedimiento de convolución y pooling puede ser posteriormente llevado a cabo tomando el resultado de 1.4d como input, conformando otra capa de convolución y pooling, repitiéndose los pasos B, C y D, reduciendo la dimensionalidad y obteniendo features progresivamente más abstractos. En estos esquemas se mostró una sola capa del proceso de una red convolucional, terminando en el aplanado de la fig. 1.4e.

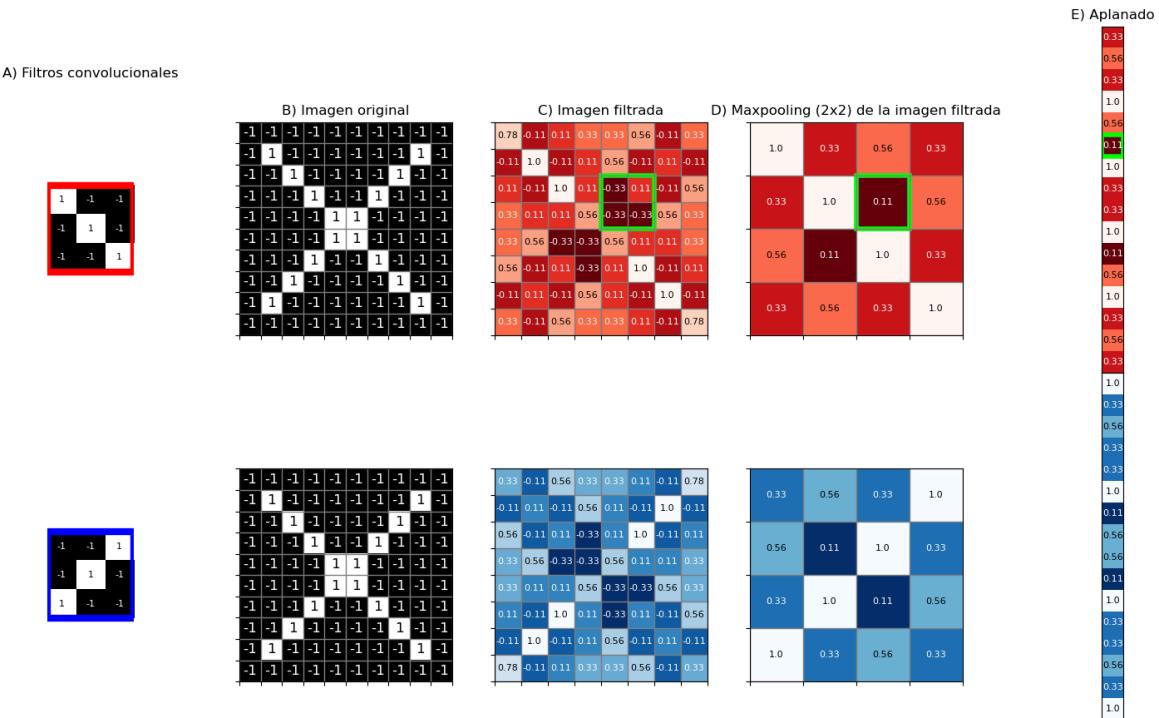


Fig 1.4: A la fig. 1.3 se le suman las operaciones de pooling (d) y aplanado (e): se esquematizan, para dos filtros distintos en un paso feedforward de la red que recibe la imagen 1.4b como input, las operaciones de convolución, pooling y aplanado. Notar que, debido a la naturaleza de los filtros que permite capturar features locales, features que estén cercanos en el espacio de entrada terminarán cerca en el vector aplanado, permitiendo invarianza ante la rotación de una imagen.

Luego de llegar a la última capa de pooling, se aplaná el resultado a un vector (fig. 1.4e), que puede ser utilizado como input para un perceptrón multicapa (fig. 1.5), que puede reducir progresivamente la dimensión del vector aplanado y permite realizar combinaciones no lineales entre todos los campos receptivos de todas las instancias de pooling de la última capa (en el ejemplo de la fig. 1.4, se pueden combinar todas las

coordenadas del pooling final de los dos filtros). De este modo, en una red entrenada cada clase de imagen redundará en una dada combinación de features en el vector de salida, y los parámetros que son entrenados son los correspondientes a los filtros y a los pesos del perceptrón multicapa (ilustrado en fig. 1.5).

Si la red fuera empleada como un clasificador, sobre el vector final del perceptrón se podría aplicar una función que lo mapearía al rango (0,1), como puede ser una sigmoidea, con el fin de que cada nodo represente la probabilidad de que el input pertenezca a la clase correspondiente al nodo. Es decir, el vector aplanado en la fig. 1.4e podría ser análogo al vector de entrada de la fig. 1.2. Este vector puede contener información espacial de la imagen de entrada.

### 1.2.3 Introducción a las redes siamesas

La idea general detrás de una red siamesa es distinta a la de un clasificador. El objetivo es encontrar una función que mapee su input a un espacio de salida en el cual la *distancia* (por ejemplo, euclídea) entre las representaciones se aproxime su distancia semántica en el espacio de entrada (Chopra 2005). En las redes siamesas el vector de salida tiene una activación lineal: sus coordenadas no representan probabilidades como lo harían en un clasificador, sino que pueden pensarse como la ubicación de *la representación que la red genera a partir del input (el embedding de la imagen)* en un espacio euclídeo (Bossi 2022). En una red entrenada, inputs con features similares deberían ser mapeados a ubicaciones cercanas en el espacio de embedding, e inputs disímiles deberían estar separados, generándose clusters. Para que la red aprenda las distancias relativas entre las clases, es necesario que aprenda una métrica de distancia. Para ello, resulta intuitivo pensar que, para corregir sus pesos, necesitaría ubicar las representaciones en el espacio de salida en un paso feedforward, y corregir sus pesos si ubicó cercanamente en el espacio representaciones de clases distintas. Esto lo logra recibiendo dos (en el caso más simple de las redes siamesas) inputs “simultáneamente” (de ahí su nombre), ubicando sus embeddings y corrigiendo sus pesos acordemente; los inputs pueden ser imágenes. Concretamente, lo que sucede es que la red hace un primer paso feedforward con la primera imagen, se retiene su embedding en memoria, luego se realiza el paso feedforward con la segunda imagen utilizando los mismos pesos que en la primera, se computa una medida de distancia entre ambos embeddings, y se corrigen los pesos en base a ella.

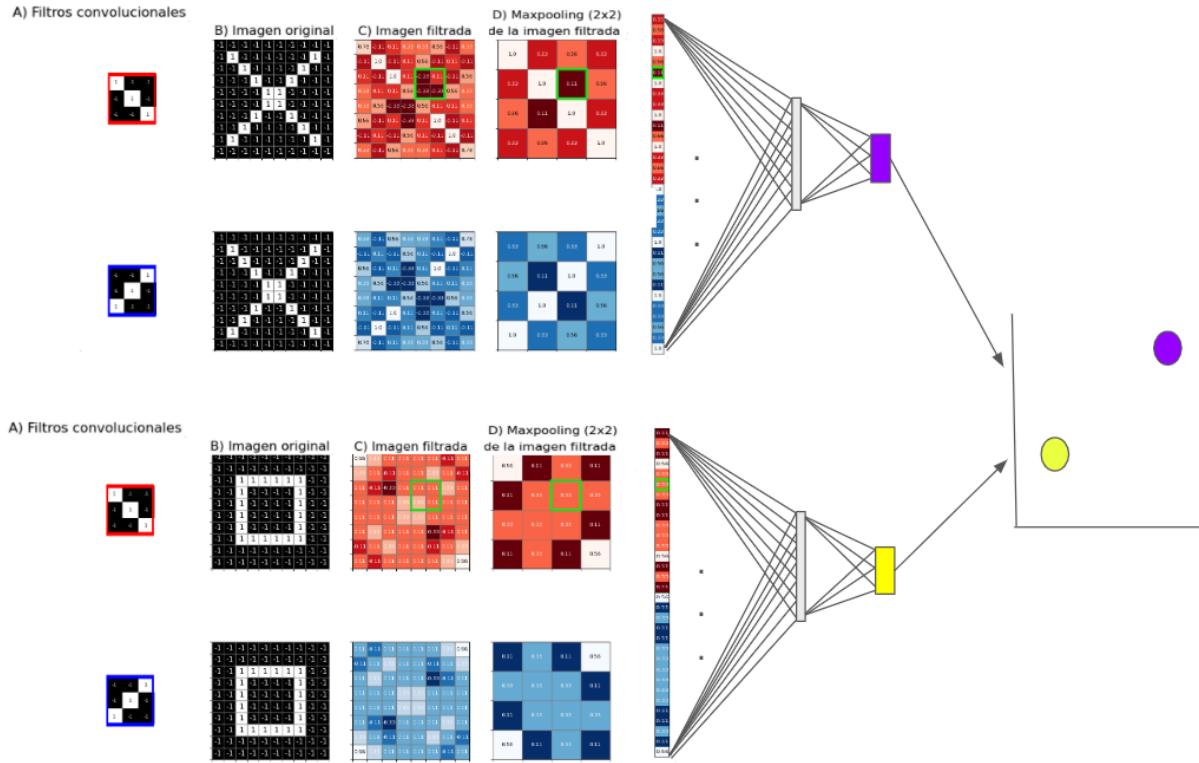


Fig 1.5: Esquema de la operación de una red convolucional con dos filtros y una sola capa de convolución + pooling, sobre dos imágenes distintas. La aplicación de cada filtro distinto sobre cada imagen lleva a un vector aplanado distinto, y por ende a un embedding distinto. En una red entrenada, cuanto más disímiles sean las imágenes, más separadas estarán en el espacio de salida, el cual es adimensional.

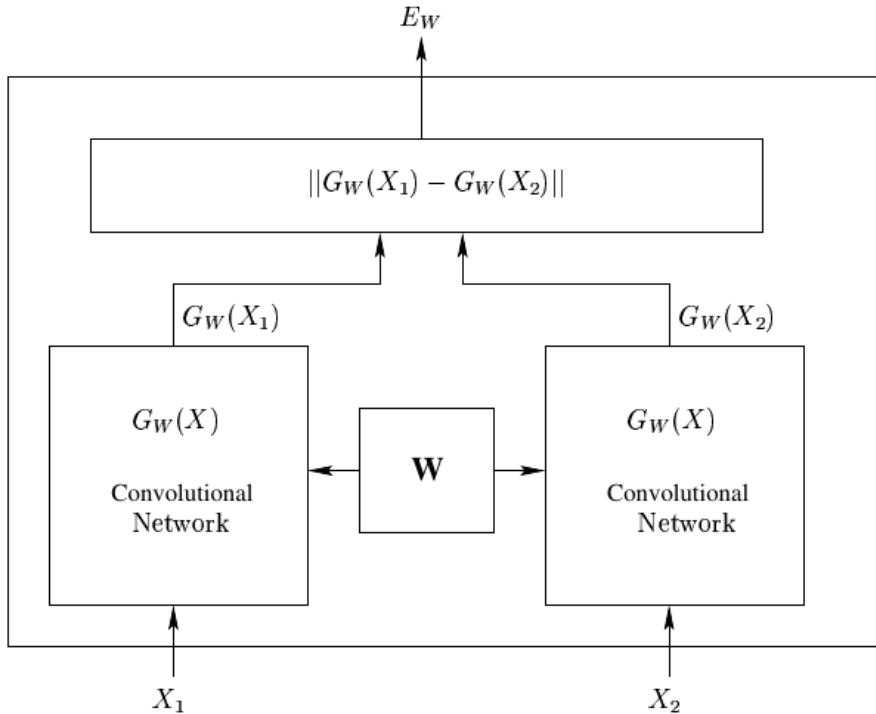
Esto es posible mediante un modo de aprendizaje llamado aprendizaje por contraste, para el cual es necesario utilizar una función costo que permita, en cada iteración del entrenamiento, contrastar dos imágenes. La función costo contrastiva fue introducida en 2005 por Yann LeCunn para abordar el problema de la verificación de rostros (Bossi 2022). La idea fue la descrita previamente: mapear inputs de alta dimensionalidad (una imagen de un rostro) a un output de mucha menor dimensión, preservando la mayor cantidad de información posible, y la manera de contrastar los inputs consiste en establecer una distancia entre ellos en el espacio de embedding. Para mejorar entre iteraciones del entrenamiento su capacidad para contrastar las clases a partir de sus features, en base a si corresponden a la misma clase o no, la red corregirá sus pesos de modo tal que, al recibir otro par de inputs que contengan features similares, la distancia entre ellos en el espacio de salida sea una representación más fiel de la distancia semántica entre ellos en el espacio de los inputs. Para tener en cuenta la distancia entre los embeddings y un *label* binario que indique si dos imágenes pertenecen o no a la misma clase, LeCunn utilizó una función costo contrastiva:

$$Cost = (1-\text{label}) * \frac{1}{2} * distancia^2 + (\text{label}) * \frac{1}{2} * \{max(0, margen - distancia)\}^2 \quad (1.5)$$

Dependiendo de si el label para el par de imágenes es 0 o 1 (si pertenecen a clases distintas o a la misma), el costo será determinado por su primer o segundo término, respectivamente. La medida de distancia entre los embeddings utilizada en esta tesis fue el cuadrado de la distancia euclídea.

Un término de la función contrastiva que fue relevante para considerar los resultados de esta tesis fue el margen: el radio desde el centro de masa de un cluster. Como puede observarse en la función, el margen se ve solo en el término de la derecha de la función (referido a las clases disímiles): un punto “será alejado por la red” de un cluster al que no pertenece si se encuentra dentro de su margen; una vez que se encuentra fuera de él, la red no se encarga de seguir alejándolo. Se puede pensar que esto es una estrategia de parsimonia que permitiría a la red concentrarse en el resto de los puntos mal clusterizados, y que la correcta clusterización de los puntos fuera de un margen al que no pertenecen dependa del término de la izquierda de la función costo (tesis tomás bossi).

Una vez explicados los componentes esenciales de una red neuronal siamesa, será de utilidad un pequeño resumen, acompañado de la fig. 1.6: una arquitectura siamesa procesa (convoluciones \* pooling \* aplanado \* multicapa) dos inputs con exactamente los mismos parámetros, cada uno resulta en un embedding, y se computa la distancia entre ellos. Esa distancia es a su vez el input de la función costo, en la que se considera si son o no del mismo individuo, y el margen preestablecido, y a partir de ella se corrigen los parámetros de la red. Por ende, el modelo entrenado puede ser usado como una medida de similaridad (Chopra 2005).



*Fig. 1.5: Arquitectura de una red siamesa, extraída de Chopra (2005).  $X_1$  y  $X_2$  son imágenes, que pueden ser de la misma o diferente clase.  $G_w$  es el codificador de la red: la misma se piensa como dos redes funcionando en paralelo, que comparten sus pesos. Cada una produce un embedding  $G_w(X)$ . Por último, se computa su diferencia, la cual no debería ser menor que un margen para pares de distintas clases.*

En esta tesis se utilizó este tipo de redes para la detección de features en sílabas de hornero que permitieran distinguir los individuos entre sí, con el fin de poner a prueba una serie de hipótesis, descritas en el capítulo 2. Este abordaje ha sido previamente utilizado para la identificación de individuos de Chingolo a partir de su canto en trabajos del Laboratorio de Sistemas Dinámicos (Bistel 2022a, Bistel 2022b, (Bossi 2022). Un aporte en esta tesis es dirigirlo a un ave suboscina.

## 2. Hipótesis, predicciones y objetivos

Lo que motivó el inicio de este trabajo fue una serie de observaciones en los sonogramas de una base de datos de cantos de cinco nidos de hornero (un hornero macho y uno hembra de cada nido) grabados entre 2004 y 2005 por Ana Amador, a partir de las cuales se sugirió que hay diferencias morfológicas entre sílabas de distintas hembras (Fig. 2.1, columna 1). A partir de estas primeras observaciones, y en base a criterios acústicos y morfológicos, se estudió detalladamente el repertorio de sílabas de los horneros en esta base de datos.

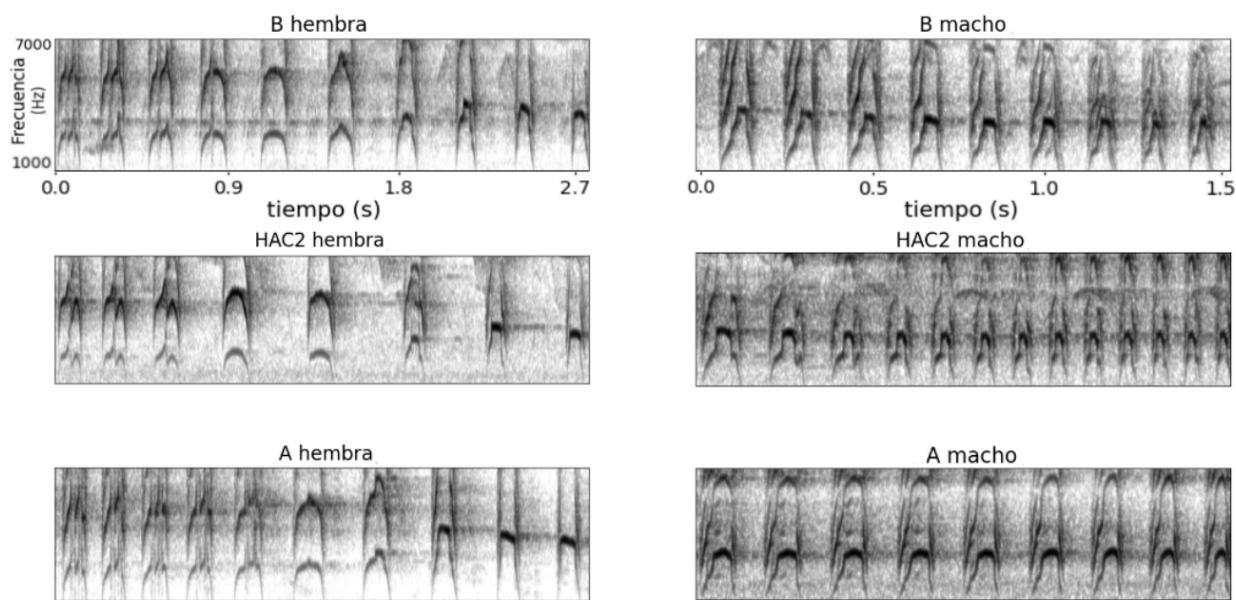


Fig 2.1: Primeras observaciones comparativas entre cantos de macho y hembra. Columna 1: hembras. Columna 2: machos. Los nidos a los que corresponden son HAC2 (2023), B (ECAS, 2004) y A (ECAS, 2004). Todos los sonogramas fueron presentados con el mismo eje de frecuencias, mientras que las grabaciones de hembras y machos tienen duraciones de 2.7s y 1.5s respectivamente.

Luego de observar los sonogramas de la base de datos con un enfoque en la morfología de las sílabas, se propuso una clasificación de los tipos de sílaba que canta un hornero típico con el fin de sistematizar su estudio. En las hembras se definieron los tipos de sílabas alfa, transición y beta (Fig. 2.2). En los machos, se observó sistemáticamente únicamente un tipo de sílaba.

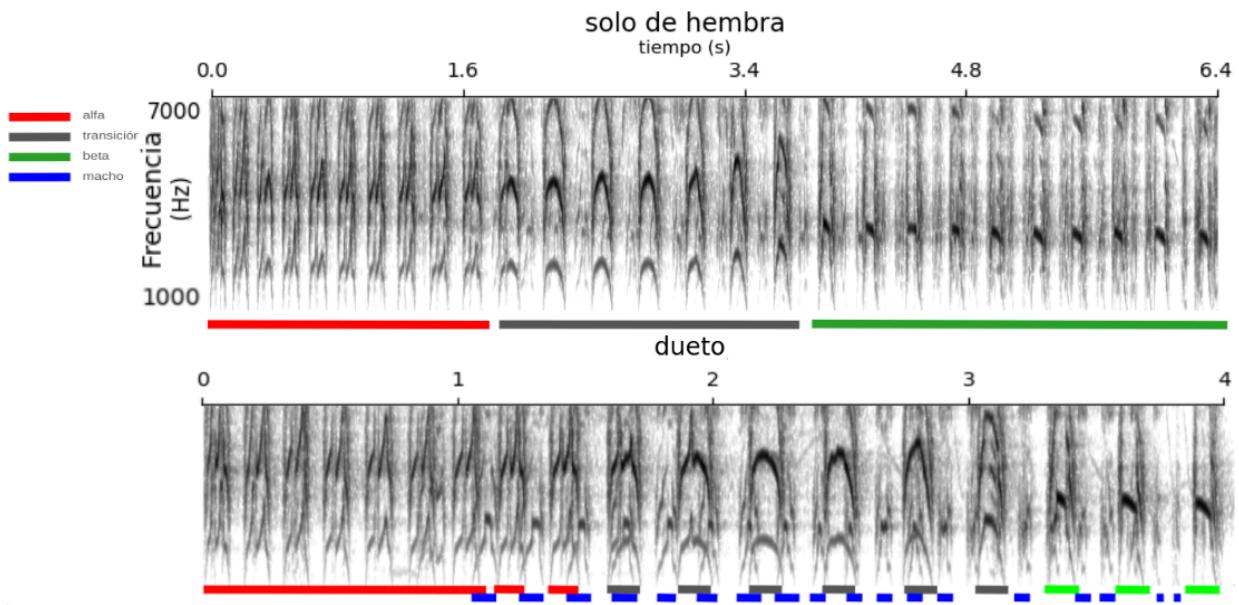


Fig 2.2: sonogramas de un solo de hembra (superior) y un dueto (inferior) del nido 19.

Estas observaciones redundaron en las siguientes hipótesis de trabajo:

**Hipótesis 1.** Cada hembra produce dos tipos de sílabas sistemáticamente en todos sus cantos (alfa y beta, fig. 2.2), y pueden ser distinguidas *en un mismo individuo* por sus propiedades acústicas<sup>2</sup>.

**Hipótesis 2.** Es posible distinguir hembras a partir de la morfología de sus sílabas alfa, pero no a partir de las sílabas beta.

**Hipótesis 3.** Los machos producen un solo tipo de sílaba que no permite distinguir a los individuos entre sí a partir de su morfología.

La primera hipótesis de trabajo fue puesta a prueba mediante los métodos detallados en el capítulo 4 de la tesis, en el cual se lleva a cabo un estudio de las propiedades acústicas de las sílabas. En el capítulo 5 se evalúan las hipótesis 2 y 3; se describe la implementación de una red neuronal convolucional siamesa con el propósito de distinguir individuos a partir de sonogramas de sus sílabas.

---

<sup>2</sup> Las sílabas de transición (ver Fig. 2.2) no se producen sistemáticamente en un mismo nido (y por ende tampoco entre nidos) en base a los criterios morfológicos adoptados en esta clasificación. En consecuencia, se descartó su estudio en este trabajo.

### **3. Base de datos y trabajo de campo**

El hornero es una especie altamente territorial, lo cual determina que, en cada árbol, hay un solo nido activo (Fraga 1980). Esto permitió, gracias al estudio del sexado de los horneros a partir de la morfología de sus sílabas (Roper 2005), identificar el sexo en el sonograma. De este modo, luego de haber llevado a cabo las grabaciones en el campo, se llegó a una base de datos con duetos, solos de macho y solos de hembra, de cada nido.

La base de datos con la que se llevó a cabo este trabajo fue generada en dos etapas. La primera fue generada por Ana Amador en 2004-2005, a partir de la cual se realizaron las observaciones preliminares que motivaron el estudio en primera instancia.

#### **3.1. Base de datos 2004-2005**

Esta base de datos inicial consistió en grabaciones correspondientes a cinco nidos, ubicados dos en la Estación de Cría de Animales Silvestres (ECAS), ubicada en el Parque Pereyra Iraola, y tres en Chascomús. Contenía 900 grabaciones, que fueron filtradas para considerar solo aquellas en las que se pudiera apreciar la morfología de las sílabas, ya que estas grabaciones podían ser demasiado ruidosas, tener duetos en los que no se pudiera apreciar las sílabas individualmente, o contener cantos de pichones, los cuales no fueron estudiados en esta tesis. De esta manera, se seleccionaron 61 grabaciones, a partir de las cuales se estudiaron las sílabas.

Una vez que obtenida la base de datos inicial, se llevó a cabo el procedimiento descrito en esta tesis para poner a prueba las hipótesis planteadas. Los resultados preliminares motivaron expandir la base de datos para otorgarles mayor robustez. Para ello, se realizaron nuevas grabaciones de campo.

Los nombres de los individuos de esta base de datos son: A, B, 19, 23 y 34.

#### **3.2 Base de datos 2023**

Tuvieron lugar en ECAS nuevamente, y en Acassuso, Pcia. de Buenos Aires, entre el Octubre y Enero de 2023. Se utilizó un micrófono direccional Audiotechnica con

un grabador Zoom H4n. Los individuos de esta base de datos son: HAC1, HAC2, y HEC1.

Antes de comenzar a realizar visitas periódicas al ECAS, fue necesario conocer el comportamiento de los horneros a lo largo de un año. Lo que resultaba de particular interés era el período en el cual los horneros tienen una cría recién nacida, ya que ésta es altamente dependiente y los parentales entran y salen del nido con muy alta frecuencia para alimentarlo y, en general, uno de ellos se queda en el nido. La gran mayoría de los cantos ocurren cuando uno de ellos llega al nido.

La principal referencia bibliográfica para conocer el comportamiento de los horneros fue el estudio de Rosendo Fraga (1980), quien hizo un estudio exhaustivo de 13 pares de horneros en Pcia de Buenos Aires entre 1970 y 1976. La información que resultó útil se resume a continuación:

1. Son altamente territoriales, y su territorio es de entre 0.25ha y 1ha.
2. El comienzo de la construcción del nido puede ocurrir entre Abril y Junio; dos o tres meses antes del comienzo de la incubación.
3. La construcción de los nidos termina unos pocos días antes del comienzo de la incubación, entre Agosto y Diciembre.
4. El período de incubación dura 16-17 días y el de nidificación 24-26 días.
5. La sequía puede retrasar la construcción del nido.
6. Todos los nidos fueron construidos a 5.5m del suelo o menos.

Tomar conocimiento de los puntos 3 y 4 fue fundamental para conocer la ventana de tiempo de la que se disponía en el trabajo de campo, habiendo comenzado en Octubre; según la referencia bibliográfica, el período en el cual los parentales alimentan a sus pichones en el nido dura 24-26 días. Esa sería la ventana de tiempo óptima para ir al campo, colocar un micrófono próximo al nido y esperar.

A su vez, las observaciones 3 y 4 de Fraga fueron corroboradas en este trabajo de campo en el nido HAC1 (el primero encontrado en Acassuso). Su seguimiento comenzó 13/11/2023; se encontró a la pareja de horneros terminando de construirlo. Por ende, se continuó su seguimiento con la expectativa de que la incubación terminase 16–17 días después. Efectivamente, se los vio llegar con comida al nido a las 06:40am el 30/11/2023, señal de que habían nacido los pichones.

Sin embargo, en este nido, al igual que en el nido HEC1 (el primero -y único-encontrado en el ECAS), hubo inconvenientes para tener suficientes grabaciones con una buena relación señal/ruido. Estos problemas se referían a la distancia del

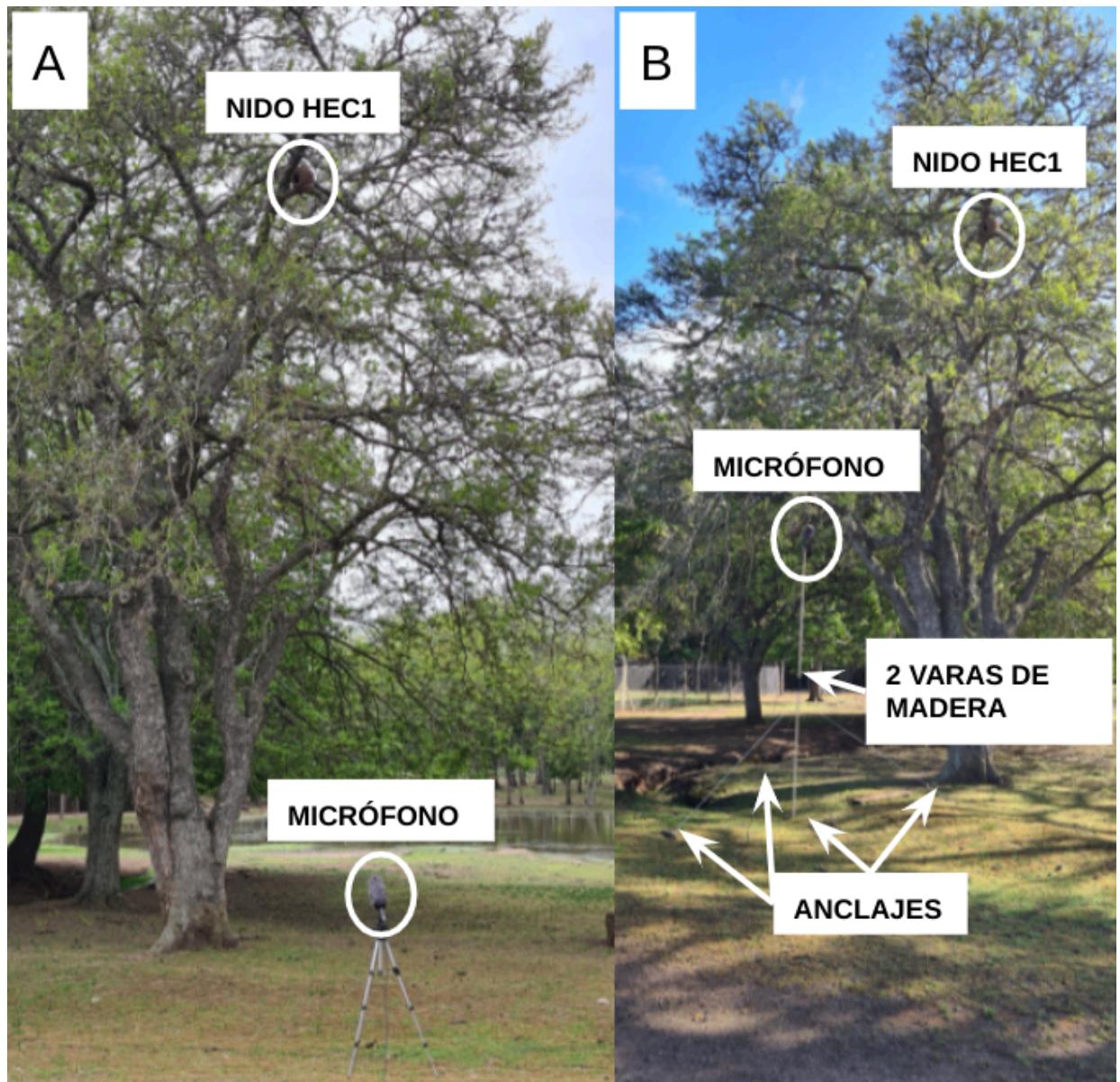
micrófono al nido y el viento. Para probar soluciones era necesario modificar las herramientas de grabación (Fig. 3.2) al final de la campaña, volver y esperar a que los horneros cantaran nuevamente. Las dificultades para determinar si el problema radicaba en la distancia entre el micrófono y el nido o en el viento se debieron a que se cumplió el punto 5 pero no el punto 6 en lo mencionado por Rosendo Fraga: en este período de sequía la abundancia de nidos en el ECAS fue extremadamente menor a la esperada, y no se encontró un solo nido activo a una distancia menor a 10m del suelo. La reserva fue recorrida extensamente. Resultó sorprendente la ausencia de cantos de hornero dentro de la reserva, la poca abundancia de nidos activos, y la altura a la cual se encontraban.



*Fig. 3.1: Mapa parcial de ECAS con las ubicaciones aproximadas de los nidos activos en el período Octubre-Enero 2023. Sin embargo, solo en los nidos marcados con 1 (HEC1) y 2 se logró obtener grabaciones, ya que en ellos los pichones seguían en el nido. Sin embargo, en el nido 2 la hembra nunca cantó y, luego de 3 visitas al ECAS, una tormenta derrumbó el nido.*

En definitiva, se corroboró que el principal problema era la altura del nido: el límite de distancia al micrófono en el cual una grabación tiene una buena relación señal/ruido es de 4-5m. Esto se corroboró con el nido HAC2, encontrado a 60m de HAC1, a una altura de 6m. Utilizando el mismo dispositivo que se muestra en la Fig. 3.2, esa distancia se redujo a 3.5m, con lo cual se llegó a la serie de grabaciones de mejor calidad en la base de datos.

Por el contrario, el nido HAC1 se encontraba a 8m de altura y el nido HEC1 a 10m. Esto redundó en trabajo adicional para lograr que los datos fueran útiles para el propósito de esta tesis, lo cual se describe en el capítulo 4.2.



*Fig. 3.2: Variaciones del setup de medición en el mismo nido: HEC1. Una vez que se había determinado, gracias a las comparaciones entre los nidos HAC1 y HAC2, que el determinante fundamental para una buena calidad de sonido era únicamente una distancia al nido menor a 5m, se hizo lo posible por llegar a cerrar esa distancia en HEC1 (a una altura de, aproximadamente, 10m, fig. a) concatenando dos varas de madera para un total de 3.2m de altura (fig. b). El viento haría imposible aumentar esa distancia con una mayor altura. La calidad del sonido resultante en HEC1 conllevó a un mayor trabajo de procesamiento para la extracción de los datos, desarrollada en el capítulo 4.2.*

En suma, se llegó a un total de 92 grabaciones nuevas, cada una correspondiente a un canto de, aproximadamente, 8 segundos.

Habiendo llegado a 8 individuos distintos en la base de datos de grabaciones, se procedió con el análisis de todas las grabaciones con el fin de poner a prueba las hipótesis planteadas. En primer lugar, fue necesario extraer los datos a partir de las grabaciones, lo cual es descrito en capítulo 4.

## 4. Caracterización acústica de las sílabas de hembra *furnarius rufus*: delineación de sílabas alfa y beta

La caracterización acústica de las sílabas de hembra fue el primer paso de este trabajo, y se refiere a la primera de las hipótesis propuestas. Para llevarla a cabo, fue necesario realizar un análisis exhaustivo de las grabaciones de solos de macho, de hembra y duetos.

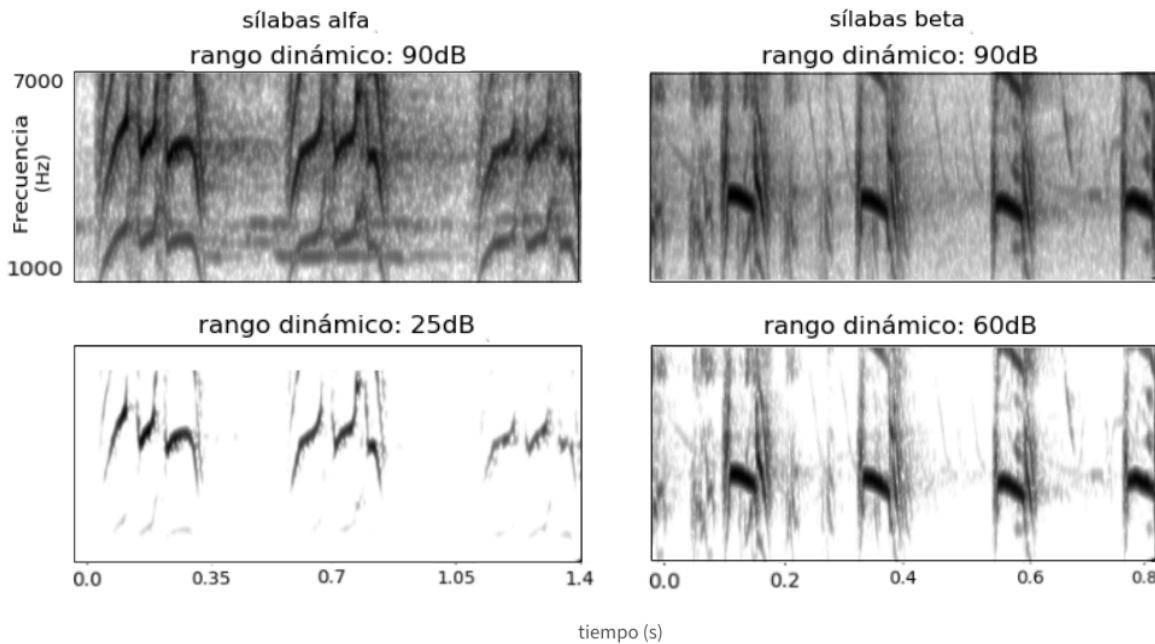
### 4.1 Métodos

Para armar la base de datos de las sílabas fue necesario obtenerlas por separado desde cada grabación. Se obtuvieron en dos representaciones distintas: el sonograma de cada sílaba por un lado, y su espectro por el otro, con lo cual se generaron dos bases de datos paralelas. Ambas representaciones de las sílabas fueron obtenidas utilizando el software de análisis de sonidos Praat (Boersma, Paul & Weenink, David 2024). Posteriormente, su procesamiento desde la limpieza hasta la presentación de resultados fue llevado a cabo en python.

#### 4.1.1 El espectro de cada sílaba de hembra

Una primera aproximación a la caracterización acústica de las sílabas que las hembras cantan sistemáticamente (alfa y beta) fue efectuada mediante un análisis comparativo de la frecuencia fundamental y el primer armónico en los espectros de cada una. No se estudió la totalidad del contenido espectral (las cantidades y pesos relativos de los componentes armónicos presentes en la señal (Mindlin & Laje: *The Physics of Birdsong*) dado que, en cada grabación, el ruido y la distancia del micrófono al nido limitan hasta qué armónico se observan las sílabas en el sonograma. Este primer abordaje a la caracterización de las sílabas mediante sus espectros puede ser considerado más rudimentario que el implementado mediante sus sonogramas (capítulo 5), ya que no hay información sobre la evolución temporal de la frecuencia (Mindlin & Laje: *The Physics of Birdsong*). Sin embargo, el foco de esta primera observación de las características acústicas estuvo en determinar si el armónico enfatizado era el de la frecuencia fundamental o el primer armónico.

El programa Praat permite observar el sonograma de una grabación identificando con distintos tonos de gris la energía de las frecuencias: blanco para la mínima, y negro para la máxima. Asimismo, cuenta con la opción de variar el rango dinámico sobre un sonograma, es decir, el valor de dB a partir del cual un punto de frecuencia es pintado en el sonograma con un tono de gris, en vez de ser considerado silencio. Por ejemplo, si el pico máximo del sonograma tiene una intensidad de 80dB y el rango dinámico es de 30dB, las frecuencias de intensidad menor a 50dB permanecerán en blanco, y las que estén dentro del rango dinámico se verán en escala de gris, siendo 80dB de color negro. De este modo, para estudiar los primeros armónicos de las sílabas alfa y beta, en cada grabación y tipo de sílaba se varió el rango dinámico, limitándolo al necesario para poder observar la morfología de cada una en su armónico más enfatizado, es decir, en el de mayor energía. Según se observa en la fig. 3, es el primer armónico en las alfa, y la frecuencia fundamental en las beta. Esta distinción acústica de las sílabas dentro de cada individuo es cuantificada en la fig. 2.4.



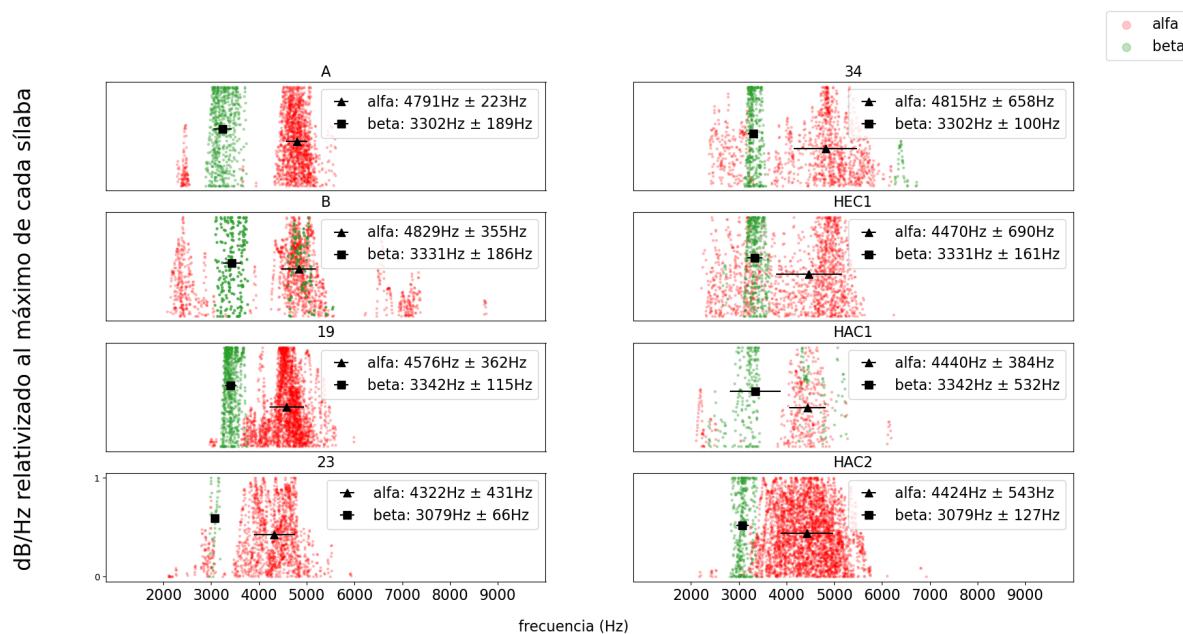
*Fig. 4.1: Se disminuyó el rango dinámico en las grabaciones con sílabas de hembra hasta que se llegara a observar una sílaba entera en un solo armónico. En esta figura se muestran fragmentos de un mismo sonograma (una misma grabación) de una hembra del nido B. En el caso de las sílabas alfa (columna 1) es el primer armónico y, en el caso de las beta (columna 2), la frecuencia fundamental.*

De este modo, habiendo definido en una dada grabación el rango dinámico para cada tipo de sílaba, se seleccionó el intervalo de tiempo sobre el sonograma en el cual se observara cada una y se extrajo su espectro (el objeto “Spectrum”) con una frecuencia de muestreo de 44100Hz, que es una tabla con dos columnas: frecuencia

(Hz) (en bins de 5Hz) e intensidad (dB). El paso siguiente para cada espectro fue filtrarlo para obtener solo la frecuencias con mayor intensidad (los picos del espectro), es decir, las de mayor energía. El espectro sin filtro se encuentra en el anexo 7.3. Posteriormente, los valores de intensidad de cada espectro fueron relativizados al rango 0-1 para facilitar su comparación dentro y entre individuos ya que, como se mencionó previamente, los valores absolutos de intensidad entre grabaciones varían considerablemente.

Los resultados de la comparación de los espectros se muestran en la Fig. 4.2.

#### 4.1.1.1 Resultados del análisis acústico de la hipótesis de trabajo 1: caracterización acústica de las sílabas de hembra



*Fig 4.2: Espectros de las sílabas alfa (rojo) y beta (verde) en cada individuo. Cada recuadro contiene los espectros de múltiples sílabas de hembra (sílabas alfa y beta); cada punto es un valor de frecuencia de uno de estos espectros. El eje de intensidad está relativizado. Fueron extraídos según el procedimiento detallado en la sección 4.1.1. Cuadrados son las medias de beta, triángulos las de alfa, y barras son desvíos estándar.*

	A	B	19	23	34	HEC1	HAC1	HAC2
macho	20	27	25	22	22	23	29	32
hembra alfa	20	12	21	9	11	28	12	30
hembra beta	25	37	27	3	29	17	22	26

*Tabla 1. Número de sílabas utilizadas en cada tipo de sílaba e individuo para los espectros de la Fig. 4.2.*

Lo que muestra la fig. 4.2 es que las frecuencias enfatizadas en las sílabas alfa y beta de la hembra son distintas. Sin embargo, lo que también resulta relevante es que el armónico de mayor energía de cada sílaba es distinto. La sílaba alfa de las hembras es enfatizada en el primer armónico; su frecuencia fundamental no se encuentra presente en estos espectros, ya que fueron filtrados para que se observe solo el armónico de mayor energía (anexo 7.3). Por el contrario, el armónico observado de las sílabas beta corresponde al de su frecuencia fundamental. Esta es la distinción acústica a la que se apuntó en la primera hipótesis planteada en esta tesis: cada hembra produce dos tipos de sílabas sistemáticamente en todos sus cantos (alfa y beta, Fig. 2.2), y pueden ser distinguidas *en un mismo individuo* por sus propiedades acústicas.

La fig. 4.2 es descriptiva, ya que los datos no siguen ninguna distribución conocida, con lo cual no fue posible estimar un modelo de comparación de medias entre frecuencias de alfa y beta, utilizando al individuo como intercepto aleatorio. A pesar de ello, se consideró suficientemente concluyente para distinguir acústicamente entre sílabas alfa y beta.

#### 4.1.2. El pitch de cada sílaba de hembra

La segunda representación de las sílabas obtenida con el programa Praat consistió en archivos que contuvieran la *frecuencia fundamental de la sílaba en función del tiempo* (de ahora en más llamada *pitch*). Esto es lo mismo que tener el sonograma de la frecuencia fundamental de la sílaba, pero sin el ruido de la grabación.

Obtener el pitch de cada sílaba permite responder la misma pregunta que la abordada mediante los espectros, pero con datos de distinta naturaleza. Una diferencia en la obtención de los datos para cada análisis es que, en el caso de los espectros, en cada bin de tiempo, se tomó más de un valor de frecuencia; se tomaron todos aquellos cuya intensidad estuviera dentro del rango dinámico establecido para la grabación. En cambio, en el caso de las extracciones de pitch, en cada bin de tiempo se tomó únicamente la frecuencia de intensidad máxima, y solo en la frecuencia fundamental.

Cada una fue extraída por separado a partir de los sonogramas en Praat utilizando la funcionalidad de extracción de pitch, detallada en el anexo 7.2.

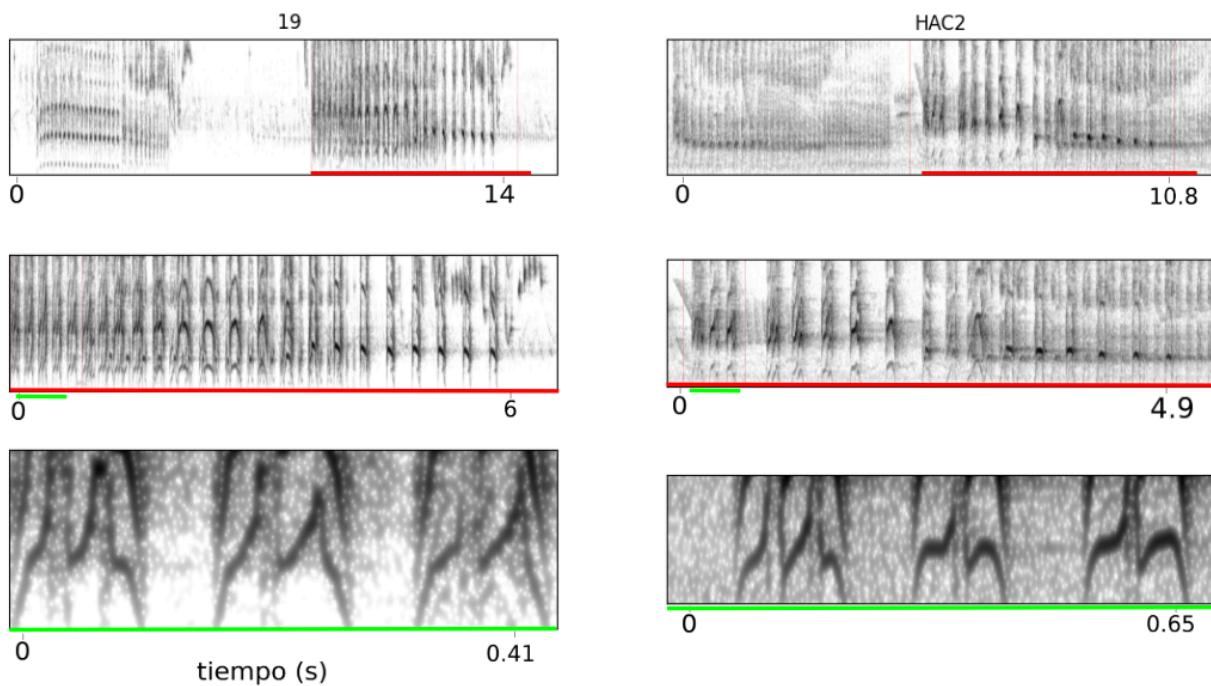
Aunque el objetivo en el capítulo 4 es estudiar las *propiedades acústicas* de las sílabas y no necesariamente su *morfología*, es necesario prestarle atención a esta última para obtener una representación fiel de los valores de frecuencia que puede tomar, y así poder estudiar también sus propiedades acústicas.

La base de datos de pitches fue utilizada para evaluar las diferencias acústicas entre las sílabas de hembra alfa y beta, al igual que los espectros. Sin embargo, en este caso también se obtuvo el pitch de las sílabas de los machos, ya que esta base de datos también será la utilizada para evaluar la morfología de cada tipo de sílaba (capítulo 5).

A continuación se expone cómo se trabajó con el algoritmo de Praat ante grabaciones de cantos con distintos niveles de ruido, siendo cantos en dueto o solos. En la descripción del procedimiento se hará énfasis en la extracción de las sílabas alfa debido a que, como se detallará a continuación, su morfología tiene una complejidad que dificultó capturar las características (o features) que se observan en los sonogramas, con lo cual exigieron el mayor trabajo en la obtención de los datos. En cada caso, el procedimiento específico de modificación de los parámetros del programa es referido al anexo.

Considerando que los horneros pueden cantar a dueto o solos, el archivo ideal en la base de datos al extraer las sílabas alfa consistía en un solo de hembra. Esto sucedió pocas veces, dado que las hembras cantan solos con mucha menos frecuencia que los machos.

La fig. 4.3 ilustra el trabajo de extracción de pitch sobre una grabación en la que no hay que modificar los parámetros de Praat para obtener cada sílaba por separado, y es un solo de hembra. En cada grabación se observó con detenimiento la forma de los upsweeps y downsweeps (modulaciones crecientes y decrecientes de frecuencia) y la presencia o no de un silencio entre ellos (características morfológicas de las sílabas), y su rango de frecuencias (características acústicas). Lo que se buscó fue modificar los parámetros de Praat con el fin de capturar estas características (los *features*). La extracción se realizó sobre la frecuencia fundamental dado que, de ese modo, se evitaba el problema del *octave jump*, ilustrado en la fig. 4.5a. Las excepciones al uso de la frecuencia fundamental se dieron en los casos descritos en la fig. 4.5b.



*Fig 4.3: Procedimiento de identificación de features para dos solos de hembra de distintos individuos (19 y HAC2) de la base de datos. En cada columna se ilustra el procedimiento con un canto distinto: al abrir el archivo en Praat se observa que a la izquierda hay un solo de macho y a la derecha uno de hembra (rojo en primera fila), con lo cual se hace foco en el segundo (segunda fila) y, por último, se inspeccionan las sílabas alfa en la frecuencia fundamental (se seleccionaron tres en ambos casos, en verde) con el fin de determinar features que hubiera en las sílabas del nido y no se encontraran en los demás (fila 3).*

#### 4.1.2.1 Extracción del pitch de la hembra en grabaciones de un solo de hembra con poco ruido

La fig. 4.4 muestra, en la imagen superior, una captura de pantalla del sonograma de las sílabas alfa de hembra en un canto del nido 19. En la imagen del centro se superpuso en Praat el pitch detectado con la combinación de parámetros de “Advanced Pitch Settings” que permitiera obtener un seguimiento del pitch de todas las sílabas a la vez. En la imagen inferior se muestra el pitch únicamente.

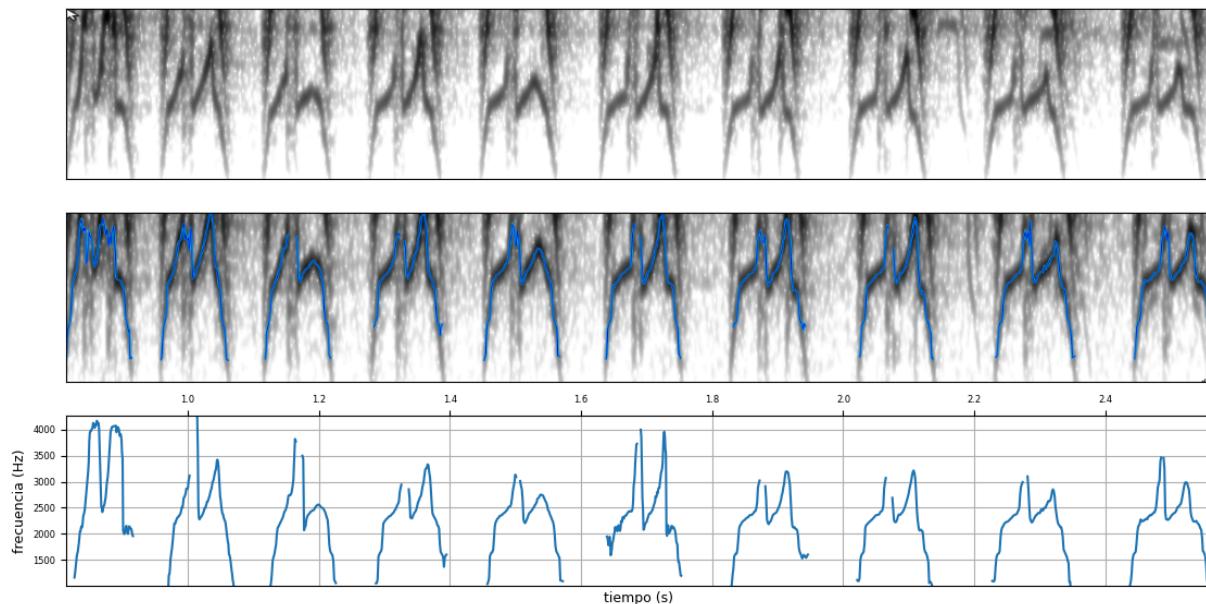
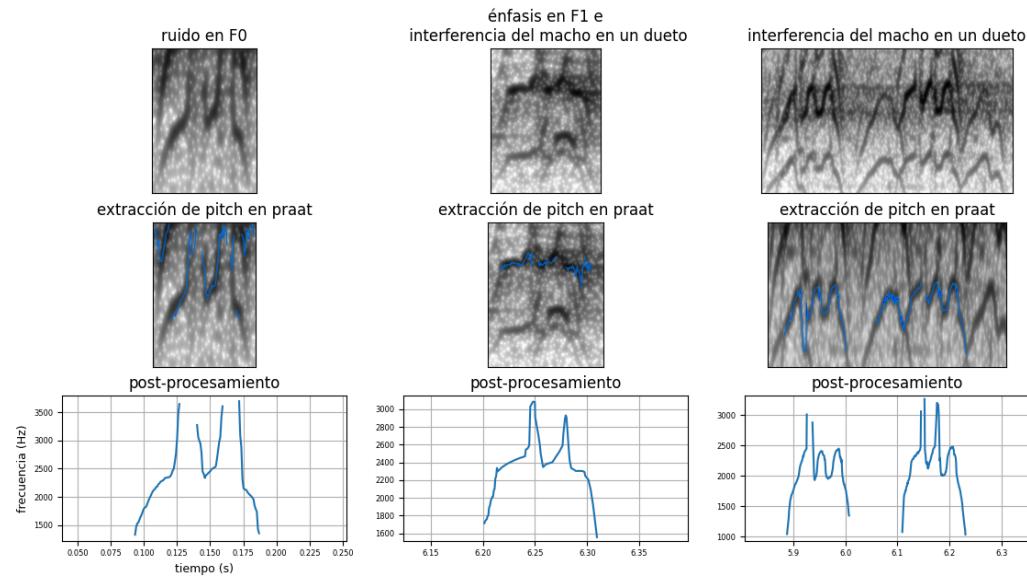


Fig. 4.4: Arriba y centro: Captura de pantalla del pitch detectado por praat superpuesto al sonograma de H19. Abajo: el pitch de cada sílaba una vez extraído. Parámetros de Praat (ver anexo 7.1): H19-031020041016.wav 0.80 2.60 600 3400 0.07 0.01 0.001 0.1 0.1 1 0

#### 4.1.2.2 Extracción del pitch de las sílabas de la hembra en grabaciones con ruido y en un dueto

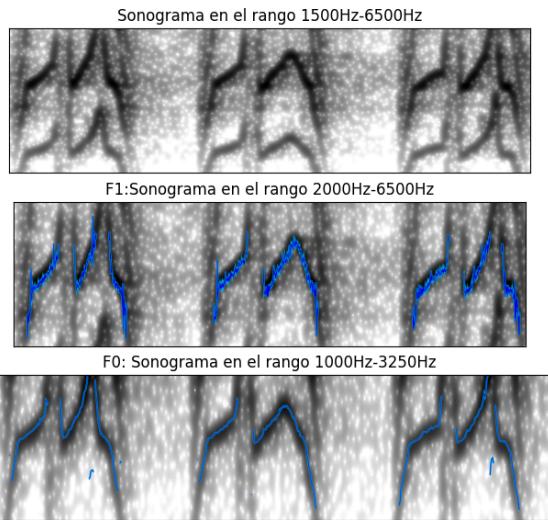
Los principales problemas en la extracción del pitch de las sílabas alfa fueron el ruido y la presencia de sílabas del macho en los duetos. Estos problemas se ven ilustrados en las figuras 4.5 y 4.6. Dependiendo del nivel de ruido que impidiera una correcta extracción del pitch (que se correspondiera con lo observado en el sonograma a pesar de que hubiera ruido) con los parámetros de Advanced pitch settings se recurrió a distintas estrategias para extraer o post-procesar los datos. En las figuras se

muestra el resultado del trabajo, mientras que el procedimiento se desarolla en el apéndice 7.2.1.



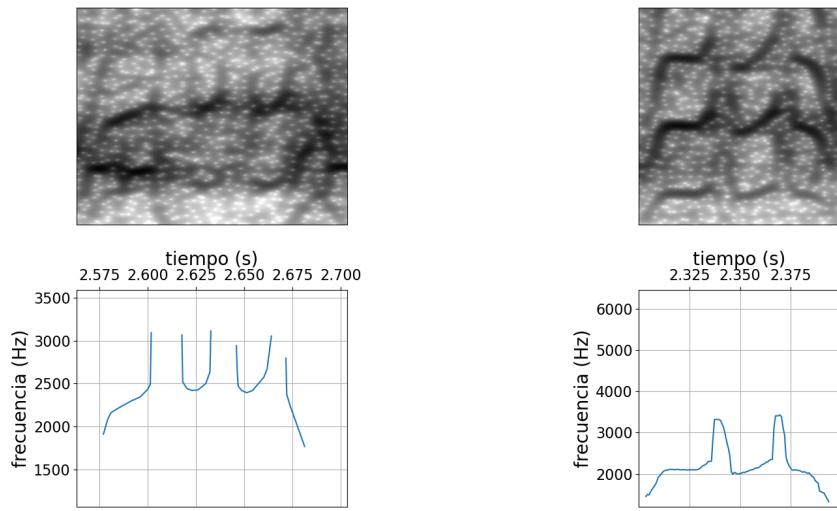
*Fig 4.5: ilustración de los problemas encontrados en la extracción del pitch de sílabas alfa a partir de grabaciones ruidosas o de duetos. A: el primer armónico de la sílaba es tomado con F0. B: La sílaba en F0 no está suficientemente enfatizada, con lo cual se extrae desde el primer armónico. C: Sílaba alfa en un dueto, con sílabas de macho interfiriendo en el pitch. Ver parámetros de Praat en apéndice 7.1.*

El pitch de F1 es extraído con una mayor frecuencia de sampleo



*Fig 4.6: Panel superior: una porción de un sonograma con la frecuencia fundamental y el primer armónico. Panel central: extracción de pitch con la información del primer armónico. Panel inferior: extracción de pitch con la información de la frecuencia fundamental. Conclusión: la frecuencia de sampleo se duplica al buscar el pitch del primer armónico.*

Hubo 18 grabaciones de los individuos HAC1 y HEC1 que fueron demasiado ruidosas y no se encontró la manera de utilizar los parámetros del programa para extraer un pitch fiel a la morfología observada en el sonograma. Sin embargo, el trazo de la sílaba era evidente. En estos casos, con el fin de no perder los datos, fue necesario recurrir a capturar manualmente la sílaba punto por punto. El modo en el que se procedió se detalla en el anexo 7.2.2; el resultado se muestra en la fig. 4.7.



*Fig. 4.7: Dos ejemplos de extracción de sílabas punto por punto, directamente sobre el sonograma. Los ejemplos son de los individuos HEC1 (izquierda) y HAC1 (derecha).*

#### 4.1.2.3 Identificación de silencios en las sílabas de hembra

Además de las particularidades en las modulaciones de frecuencia de las sílabas alfa de cada hembra, el análisis exhaustivo de esta base de datos permitió encontrar evidencia de que los horneros tienen la capacidad de generar breves (10-15ms) silencios en sus sílabas. En particular, lo que se observó en las sílabas alfa es que, una vez que se define el silencio en la sílaba o un canto entero como cierto dB, no hay un intervalo de ese dB que sea menor que 10ms. Este intervalo es visible en todos los sonogramas mostrados hasta ahora: si una sílaba alfa tiene un intervalo entre sus trazos, este es siempre mayor a 10ms.

La extracción de las sílabas incluyendo el silencio es difícil porque es una ventana de tiempo pequeña en un archivo ruidoso, y se le dedicó mucha atención porque requiere un control motor más fino que el esperado para un ave suboscina típica. En el anexo 7.2.3 se muestran los criterios y herramientas que fueron utilizados para determinar si hay un período de silencio dentro de una sílaba.

## 4.2 Resultados del análisis acústico a partir del pitch de cada sílaba de hembra

Con el pitch de cada sílaba se armó una base de datos en la que se indica, para cada individuo y tipo de sílaba, *cuántos puntos hay en cada bin de frecuencia*. En base a lo observado en los espectros (que el armónico con mayor energía en las sílabas alfa es el primero), se les sumó su propia media a las frecuencias de las sílabas alfa en los archivos extraídos con Praat, ya que fueron en primera instancia extraídas siempre utilizando la frecuencia fundamental (salvo en HEC1 y HAC1, como se indicó en el anexo 7.2). De este modo, es posible visualizar las medias y desvíos estándar de cada tipo de sílaba e individuo (fig. 4.8).

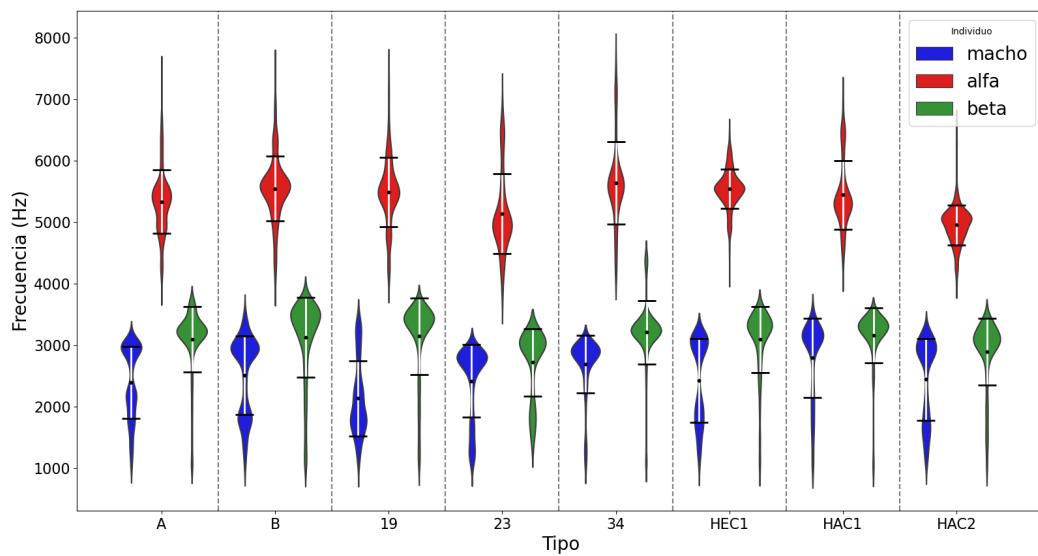


Fig 4.8: Distribución de las frecuencias (Hz) de mayor energía de cada tipo de sílaba en cada individuo. Para cada individuo y tipo de sílaba se tomaron todos puntos de frecuencia de los archivos de pitch (columnas de frecuencia y tiempo) y se graficaron a la vez en cada violín. Azul: sílabas de macho, rojo: sílabas de hembra alfa, verde: sílabas de hembra beta. Barras blancas son desvíos estándar, puntos negros son medias.

	hembra alfa (n=139)		hembra beta (n=154)		macho (n=250)	
Nido	Media (Hz)	SD (Hz)	Media (Hz)	SD (Hz)	Media (Hz)	SD (Hz)
19	4691	564	3144	621	2130	607
23	4842	651	2715	551	2415	591
34	4858	668	3205	516	2688	472
A	4476	515	3090	530	2389	588
B	4831	523	3128	649	2508	635
HAC1	4567	557	3157	448	2791	642
HAC2	4128	323	2888	545	2439	667
HEC1	4899	315	3087	535	2422	677

Tabla 2 correspondiente a la Fig. 4.8: Medias y desvíos estándar de las frecuencias (Hz) de cada tipo de sílaba e individuo utilizadas en el violinplot. El n indica los puntos de frecuencia sumando el total de las sílabas.

El motivo por el cual las distribuciones de las frecuencias de todas las sílabas tienen colas largas resulta evidente observándolas en sus sonogramas (la Fig. 2.1 tiene sílabas de hembra alfa y beta, y de macho). En los tres tipos de sílaba los mínimos de frecuencia son visitados únicamente al comienzo y final de la sílaba. En cuanto a los máximos, en las sílabas alfa son alcanzados solo al final o comienzo de un upsweep o downsweep respectivamente. Por el contrario, en las sílabas beta y de macho, los máximos son sostenidos en el tiempo durante la sílaba (figs 4.10 y 4.11).

El carácter bimodal en los violinplot de las sílabas beta y, en especial, en las sílabas de macho, se debe a las inflexiones observadas en los upsweeps y downsweeps. Se observa que son más pronunciadas y frecuentes en los machos (fig. 4.11).

La caracterización acústica sobre los espectros de los datos crudos sugiere los mismos resultados que la cuantificación obtenida mediante las sílabas extraídas a partir de su pitch: las sílabas alfa pueden ser distinguidas de las beta a partir del armónico con mayor energía, siendo la frecuencia fundamental en el caso de las sílabas beta, y el primer armónico en el caso de las alfa.

El resultado de las extracciones de pitch se muestra en las figs. 4.9-4.11.

Tipo de sílaba / nido	A	B	19	23	34	AC1	AC2	EC1
Hembra alfa	33	12	16	4	4	8	31	31
Hembra beta	18	45	20	2	14	13	26	16
Machos	39	13	15	32	29	33	50	45

Tabla 3: número de sílabas extraídas de cada tipo e individuo (Figs 4.9-4.11).

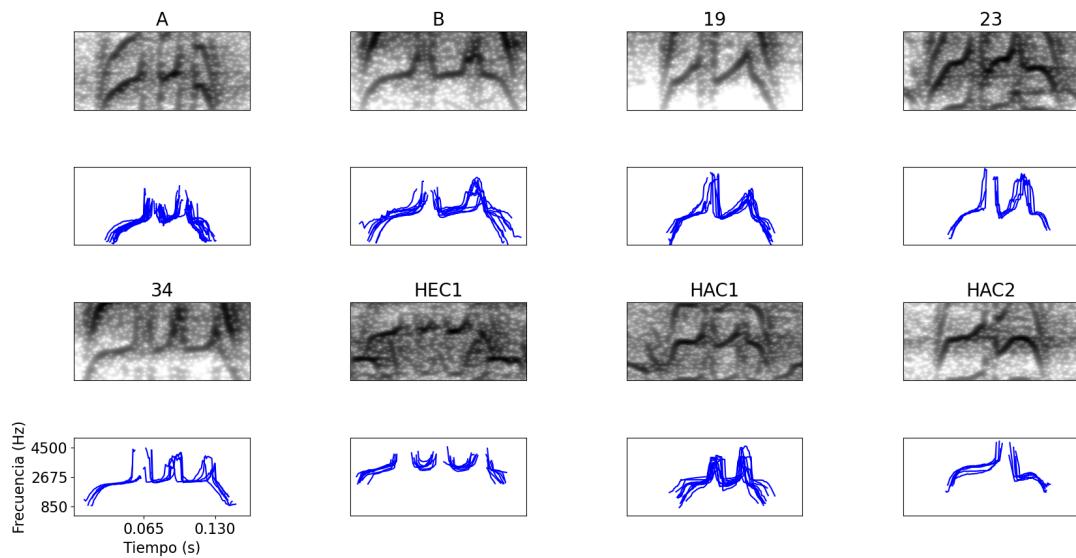
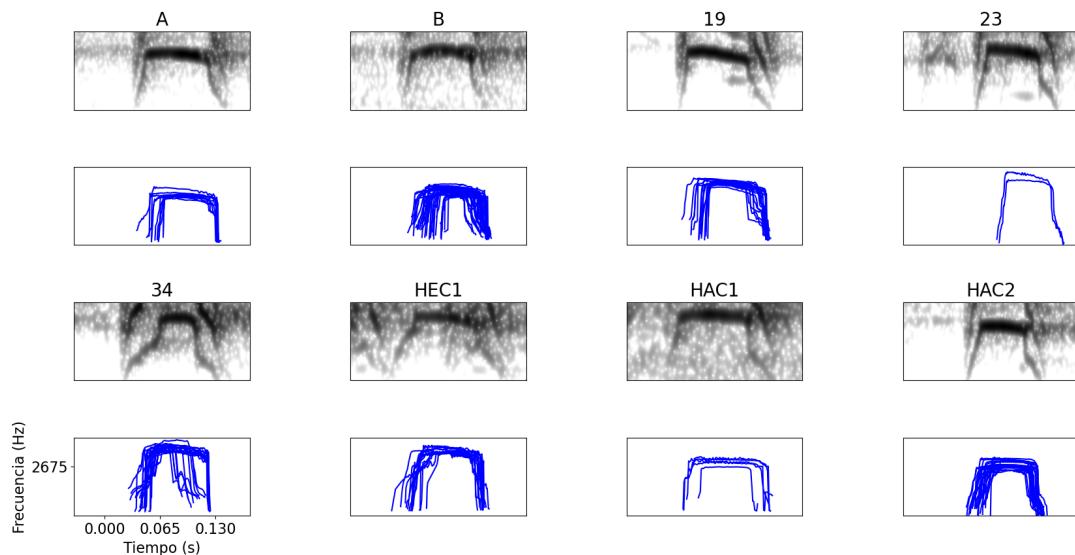
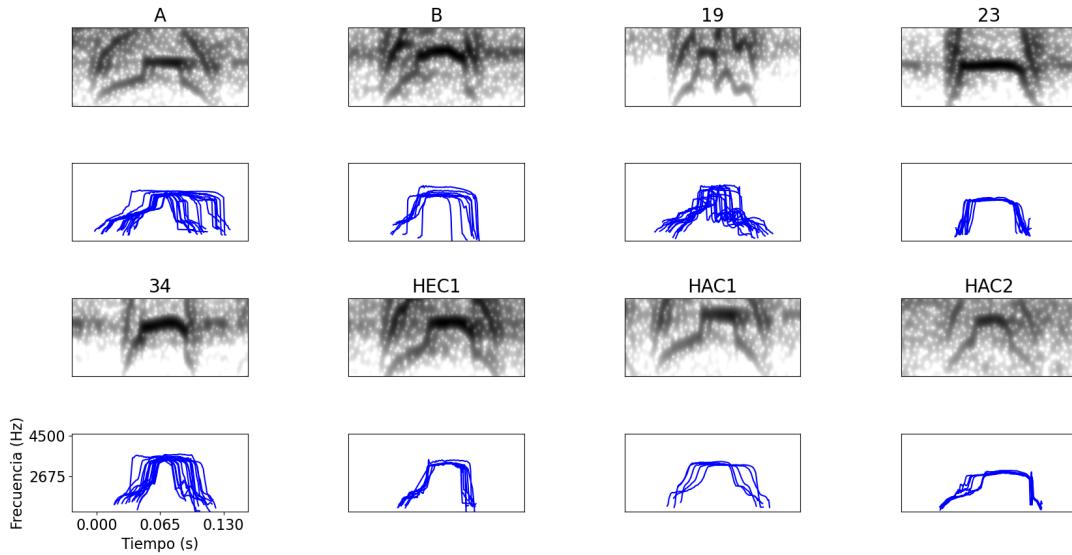


Fig. 4.9: Las sílabas alfa de cada individuo hembra extraídas con el método de 4.1: para cada individuo se muestra un sonograma de ejemplo (panel superior) y cada pitch extraído para cada tipo de sílaba.



*Fig. 4.10: las sílabas beta de cada individuo hembra extraídas con el método de 4.1: para cada individuo se muestra un sonograma de ejemplo (panel superior) y cada pitch extraído para cada tipo de sílaba.*



*Fig. 4.11: las sílabas de cada macho extraídas con el método de 4.1: para cada individuo se muestra un sonograma de ejemplo (panel superior) y cada pitch extraído para cada tipo de sílaba.*

### 4.3 Conclusiones sobre la caracterización acústica de sílabas alfa y beta de las hembras

En este capítulo se caracterizaron acústicamente las sílabas de hornero, describiendo los procedimientos específicos realizados para grabaciones de campo. Luego se describió cómo se llegó a una distinción entre sílabas alfa y beta en cada hembra individualmente. Esto permite considerarlas por separado para poner a prueba las hipótesis de trabajo 2 y 3 de la tesis, referidas a la morfología de cada tipo de sílaba y su capacidad de funcionar como firma de individualidad en estas aves suboscinas. El modo en que estas predicciones fueron puestas a prueba se describe en el capítulo 5.

# 5. Evaluación de la morfología de las sílabas y su relación con la identificación de individuos utilizando redes siamesas

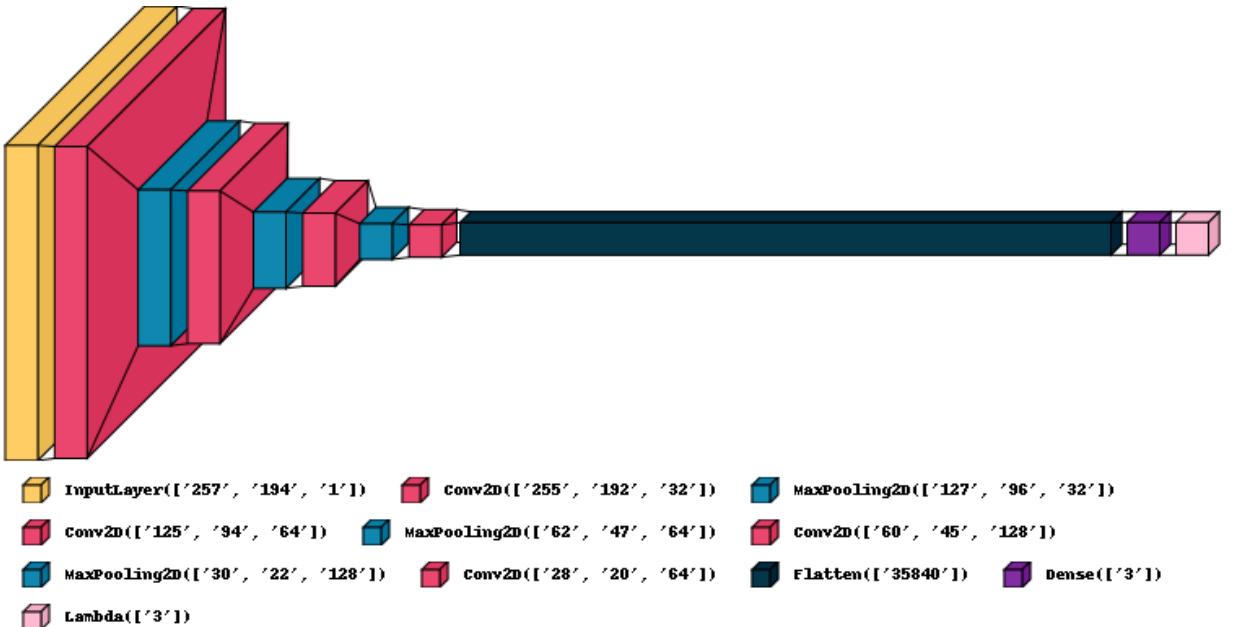
Las hipótesis de trabajo 2 y 3 de esta tesis fueron puestas a prueba utilizando redes siamesas, que fueron introducidas en el capítulo 1.2. Se entrenaron redes neuronales siamesas con sonogramas (los objetos de pitch extraídos en el capítulo 4.1) de sílabas de hembra alfa y beta, y sílabas de macho: cada red fue entrenada con un tipo de sílaba. Es decir, las preguntas apuntaron a la distinción inter-individuo a partir de un solo tipo de sílaba. No se entrenaron redes incluyendo simultáneamente sílabas alfa, beta y de macho.

## 5.1 Métodos

### 5.1.1 Implementación de la red y formato de los datos

La implementación de la red se encuentra en [github.com/ttdduu/furnarius](https://github.com/ttdduu/furnarius), en el archivo siamesa.py. Para el entrenamiento de la red se utilizó la librería tensorflow (François Chollet (2021)). En el repositorio se encuentra un archivo llamado tf.yml con el cual se puede replicar en entorno de python en el que se entrenaron las redes utilizando el environment manager conda. Requirió especial atención la compatibilidad entre las versiones de Linux, tensorflow, cuda y cudatoolkit para poder realizar los entrenamientos utilizando la GPU NVIDIA GE-FORCE RTX 3060. En la página <https://www.tensorflow.org/install/source#gpu>, salvo por la versión de Linux, cuenta con una lista de compatibilidades para los paquetes mencionados.

La parte convolucional del codificador de la red (la encargada de procesar la imagen para extraer features) fue adaptada de (François Chollet (2021), p. 216). En la fig. 5.1 se muestra una representación de la parte convolucional de la red empleada en esta tesis.



*Fig. 5.1: representación de los componentes de la red neuronal convolucional implementada en esta tesis. Amarillo: las imágenes de input, con dimensión 257x194 píxeles. Rojo: los 32 mapas de features resultantes de las operaciones de convolución. Sus dimensiones son de 255x192; se perdió una dimensión de cada lado porque no se agregó un padding a la imagen. Azul: los resultados de las operaciones de maxpooling: con matrices de 2x2 se redujo a la mitad la dimensionalidad de cada feature map. Azul oscuro: el vector aplanado de todos los mapas de features de la última capa. Violeta: el perceptrón multicapa. Rosa: la salida tridimensional en el espacio euclídeo.*

Posteriormente, el espacio de salida de la red siamesa tiene una activación lineal y es de tres dimensiones. De este modo, el resultado de una imagen como input a la red es un embedding en un espacio latente tridimensional y normalizado. Dado que la red aprende a separar clases distintas en este espacio a partir de la función costo contrastiva (ecuación 1.5), luego del entrenamiento los puntos quedan distribuidos en la superficie de una esfera unitaria (primer ejemplo en los resultados, fig. 5.3). Lo que se espera es que los puntos que corresponden a sílabas del mismo individuo estén más cerca entre sí que de otros puntos, formándose clusters con un solo individuo en cada uno. En los resultados se expone el modo en el cual se evaluó la calidad de estos clusters con el fin de poner a prueba las predicciones relativas a la capacidad de cada tipo de sílaba de funcionar como firma de individualidad.

Los datos con los cuales las redes fueron entrenadas fueron las imágenes del pitch obtenidas a partir de los métodos de la sección 4.1, ilustradas en las figs 2.11-2.13. Su dimensión fue de 257px x 194px, 50dpi, y se encuentran también en el

repositorio. Esquemáticamente, la estructura del dataset de entrenamiento es la siguiente:

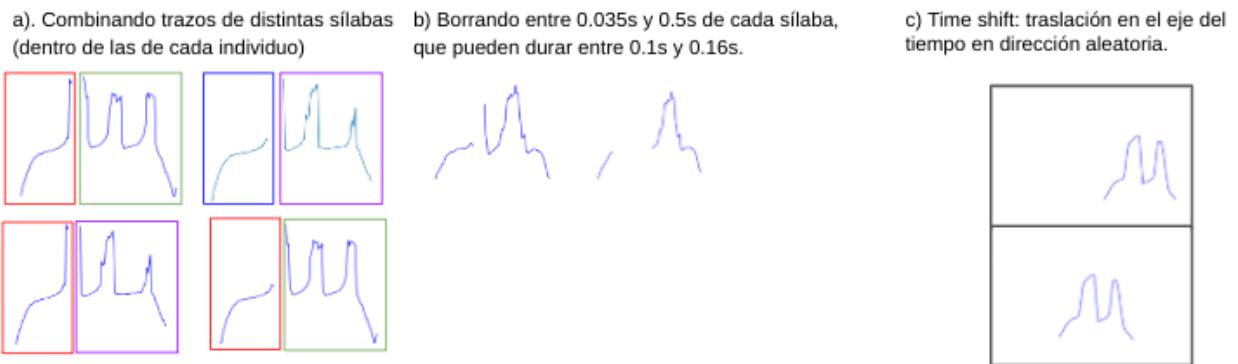
```
{  
    (sílaba de A, sílaba de B) : (0, 'individuo A', 'individuo B') ,  
    (sílaba de HAC1, sílaba sílaba de HAC2) : (1, 'individuo HAC1', 'individuo HAC2') ,  
    ... : ... ,  
}
```

Donde la lista de la izquierda contiene el par de sílabas presentadas a la red, y la lista de la derecha indica si son o no del mismo individuo, y de qué individuo son.

Con el fin de obtener resultados robustos en las cuantificaciones realizadas a partir de los resultados de la red, se realizaron múltiples entrenamientos en cada individuo. Luego de explicar las técnicas de aumentación de datos, se detalla cómo se procedió con la presentación de los datos en cada entrenamiento.

### 5.1.2 Data augmentation

Resulta evidente en la tabla 3 que la base de datos de sílabas alfa está desbalanceada. En principio, un dataset desbalanceado puede no ser un problema para entrenar una red siamesa ya que, al corregir sus pesos en base a la distancia entre los embeddings de dos sílabas, la misma sílaba es presentada múltiples veces a la red en el entrenamiento, dado que se le presentaron todas las combinaciones posibles de a pares. Sin embargo, se implementaron técnicas de data augmentation que fueran biológicamente plausibles con el fin de evitar un posible problema de falta de representación de los individuos 23, 34 y HAC1. Las maneras en las que se aumentaron los datos se presentan en la fig. 5.2. Las sílabas alfa que tienen un silencio tienen, por ende, más de un trazo en el sonograma. En un dado individuo, se mezclaron todos los trazos de las sílabas alfa entre sí (fig. 5.2a). Otra alternativa fue borrar un intervalo de tiempo de cada sílaba. La magnitud del intervalo fue aleatoria entre 0.035s y 0.05s, y su posición en la sílaba fue aleatoria (fig. 5.2b). La tercera alternativa fue trasladar toda la sílaba en el eje del tiempo en una magnitud y dirección aleatorias (fig. 5.2c).



*Fig. 5.2: Tres técnicas de data augmentation. a) Fila superior: las sílabas originales. Fila inferior: dos nuevas sílabas producto de la mezcla de los trazos de las dos de la fila superior; los trazos rojo y verde corresponden a la primera sílaba, los trazos azul y violeta a la segunda. b) Se elimina un intervalo temporal de la sílaba, cuya magnitud y posición son aleatorias dentro del margen de 0.035s y 0.05s. c) La técnica utilizada en la hembra HAC1, dado que es la única cuyas sílabas no presentan silencios; no cuentan con dos trazos que podrían ser combinados. Por ende, fueron trasladadas en el eje del tiempo.*

Luego de implementar estas técnicas, el número de sílabas alfa en cada hembra se muestra en la tabla 4.

Augmentation	A	B	19	23	34	AC1	AC2	EC1
None	33	12	16	4	4	8	31	31
Erasing	33	12	16	4	4	8	31	31
Mix	1089	144	256	16	16		625	900
Shift*Erasing						24		
Total	1555	156	272	20	20	40	656	931

*Tabla 4: se indica cómo aumenta el número de sílabas de cada individuo a medida que se implementan las técnicas de aumentado de datos.*

Dado que la base de datos extendida permite generar una inmensa cantidad de pares de sílabas para entrenar la red, hubo que restringir el número de ejemplos de cada individuo en el entrenamiento para poder llevarlo a cabo. El número máximo de sílabas de cada individuo fue 80. A continuación, se describe cómo se presentaron los datos a la red en cada entrenamiento.

- Se eligió un máximo de 80 sílabas de cada individuo para formar parte de los sets de testeo + entrenamiento. Esas 80 sílabas (en todos los individuos salvo H23, H34 y HAC1) fueron elegidas aleatoriamente.

- Se aleatorizó qué sílabas son utilizadas para entrenar y cuáles para testear, en un split 0.5. Por ende, en un entrenamiento se tiene un máximo de 40 sílabas de cada individuo.

Se entrenó con 70 épocas en las sílabas alfa y 20 épocas en las sílabas beta y de machos, ya que era el número de épocas en la cual se observaba sistemáticamente un sobreajuste. Se realizaron 20 entrenamientos para cada tipo de sílaba. Luego de cada uno, se almacenó en memoria la última capa, con lo cual sería posible darle como input las sílabas de testeo correspondientes a ese entrenamiento y visualizar sus embeddings para las sílabas alfa.

Por otro lado, una manera de controlar un posible sesgo a causa de datos desbalanceados fue entrenar modelos nulos: randomizando las etiquetas que indican si un par de sílabas son o no del mismo individuo. Lo que se hizo no fue darle a cada par una probabilidad de 0.5 de ser 0 o 1, sino que se randomizó el vector de labels existente, de modo que se mantuvo la cantidad de positivos y negativos. Es decir que, si de todos modos se observa que los individuos más representados están mejor clusterizados, se debe a que hubo más pares positivos para esa clase que para otras incluso luego de la aleatorización, con lo cual el modelo nulo sirve para compararlo con el bien entrenado para evaluar cuánto mejora el agrupamiento en este último. Las métricas en las que mejora el agrupamiento, y la manera en la que se usó el modeo nulo como un control, son detalladas en los resultados.

### 5.1.3 Metodología para el análisis de los resultados de una red neuronal siamesa entrenada

El output de una red entrenada al recibir las sílabas del set de testeo es la distribución de sus representaciones de las sílabas en el espacio de salida, es decir, la distribución de los puntos en la superficie de la esfera unitaria. Cuanto mayor sea la capacidad de un *tipo de sílaba* de funcionar como firma de *individualidad*, se esperaría que los embeddings de las sílabas de cada individuo se muestren en clusters claramente separados entre sí.

Para una dada red entrenada, la evaluación de la calidad de los clusters formados a partir de las predicciones de la red sobre la sílabas del set de testeo fue abordada complementando dos resultados, ambos obtenidos a partir de la comparación de estos clusters con los generados por un algoritmo de agrupamiento no supervisado

sobre los mismos datos. La idea detrás de este enfoque es que un algoritmo no supervisado debería encontrar los clusters "naturales" (Peter J. Rousseeuw (1987)), los cuales se esperaría que coincidieran con los reales para un tipo de sílaba que permitiera distinguir individuos. Los dos resultados complementarios se refieren al número óptimo de clusters (sección 5.1.3.1: Predicción 5A) y a su identidad respecto de los clusters formados por el algoritmo no supervisado sobre los mismos datos (sección 5.1.3.2: Predicción 5B).

El algoritmo no supervisado que se utilizó fue una variante de K-Means que utiliza la similitud coseno en vez de la distancia euclídea (Suvrit Sra et. al (2005)). Esto se debe a que, como se mencionó previamente, los embeddings resultado de una red entrenada se encuentran en la superficie de una esfera unitaria; todos tienen radio = 1. El input para este algoritmo es el conjunto de puntos que se quiere agrupar y el número de clusters que se quiere obtener. El algoritmo es inicializado con puntos al azar, que son tomados como los centros de masa iniciales de cada cluster. De este modo, cada punto es asignado al cluster con el centro de masa azaroso más cercano, lo cual define nuevos clusters y, por ende, nuevos centros de masa, esta vez en base a los puntos asignados a cada uno. Este proceso de asignación de puntos al cluster más cercano es repetido hasta que los centros de masa no cambian de lugar entre iteraciones.

A partir de este algoritmo no supervisado se pusieron a prueba las predicciones 5A y 5B enunciadas a continuación, en las secciones 5.2.1 y 5.2.2 respectivamente, para llegar a los dos resultados mencionados previamente, partiendo de la base de que se cuenta con 8 individuos: A, B, 19, 23, 34, HAC1, HAC2 y HEC1.

- Predicción 5A (sección 5.2.1): El número de clusters óptimo según el algoritmo de K-Means es el mismo número de clusters reales.
- Predicción 5B (sección 5.2.2): Que el cluster más cercano a cada cluster real sea un cluster de KMeans, y que ese cluster de K-Means sea distinto para cada cluster real.

#### 5.1.3.1 Predicción 5A

*El número de clusters óptimo según el algoritmo de K-Means es el mismo número de clusters reales*

#### 5.1.3.1.1 Cómo determinar la calidad de un conjunto de clusters

La medida de optimalidad utilizada para determinar el mejor número de clusters a partir de la distribución de los puntos correspondientes a las predicciones de una dada red entrenada sobre set de testeo se denomina coeficiente de silhouette (de ahora en más llamado score) (Peter J. Rousseeuw (1987)). El score *medio* es una medida de la calidad del agrupamiento total de un set de datos (de un conjunto de clusters) obtenida al calcular el score de cada punto X individual del set, que considera los siguientes valores para **el punto X que se encuentra en un cluster C**:

- A(X): la media de las distancias del punto X **a cada punto de C**.
- B(X): la media de las distancias de X a cada punto **del cluster más cercano a C**.
- $\max(A(X), B(X))$ : la máxima distancia de X **a un punto de C**.

El coeficiente para un punto X se calcula de la siguiente manera:

$$s(x) = \frac{B(x) - A(x)}{\max(A(x), B(x))} \quad (5.1) \text{ (Peter J. Rousseeuw (1987))}$$

Tiene un rango entre -1 y 1. Si es 0 para un punto, significa que está en la frontera entre dos clusters. Cuanto más alto sea, mejor asignado está a su cluster.

El coeficiente que se reporta es la media de los coeficientes de cada uno de los puntos, con lo cual se tiene una medida de la calidad del agrupamiento total.

#### 5.1.3.1.2 Cómo determinar la calidad de varios conjuntos de clusters

Una vez elegido el algoritmo no supervisado y el criterio de la calidad de los clusters de una dada red entrenada, el abordaje fue el siguiente: se llevaron a cabo múltiples entrenamientos (según el procedimiento descrito en la sección 5.1.2). Para cada red se obtuvo un espacio de embedding con un set de testeo único y por ende un conjunto de predicciones (de clusters) únicas. Para cada conjunto de predicciones se implementó el algoritmo de K-Means eligiendo distintos valores de K clusters: entre 2 y 14. Es decir que, sobre los puntos del testeo de una dada red entrenada, se corrió el algoritmo de KMeans 13 veces, una con cada valor de K, obteniéndose una base de datos de scores: para cada valor de K en cada entrenamiento.

La estrategia de calcular el score para distintos valores de k es comúnmente utilizada cuando no se conoce, o se quiere explorar, el número óptimo de clusters en un set de datos; la particularidad de este enfoque radica en que se tuvieron múltiples conjuntos de datos. Si no se compara el score con distintos valores de k, no es posible detectar una estructura en los datos que no sea la propuesta (la de un dado k).

Para comparar las medias, sobre todos los entrenamientos y en un tipo de sílaba, de los scores de cada K, se estimó un modelo lineal general mixto de comparación de medias (5.3). Esto implica que el número de clusters se consideró una variable categórica; a priori no se encontró un motivo para suponer una relación lineal entre el número de clusters y la media del score. Por otro lado, como se mencionó previamente, el algoritmo de K-Means se implementó 13 veces sobre cada set de predicciones (k entre 2 y 14); los scores provenientes de un mismo set de predicciones no son independientes entre sí. Por ende, se definió el identificador del entrenamiento como un bloque, dentro del cual están todos los niveles del factor K, incluyendo el primero como una variable de efectos aleatorios en el modelo, de 20 niveles. Una particularidad de este modelo es que, en un dado nivel de la variable de efectos fijos, se cuenta con una sola instancia para cada valor de la variable de efectos aleatorios. Es decir, en cada set de predicciones se implementó K-Means con un dado K una sola vez. El modelo planteado es el siguiente:

$$Y_{ij} = \mu + \alpha_i + B_l + \epsilon_{ij} \quad (5.3)$$

$$\epsilon_{ij} \sim N(0, \sigma_{\text{residual}}^2); B_l \sim N(0, \sigma_{\text{entrenamiento}}^2); i \in \{2, 14\}; j = 1; l \in \{1, 20\}$$

En el cual se estimó una media para cada k, una varianza común para todos los k, y una varianza común a los 20 niveles de la variable de efectos aleatorios.  $\alpha_i$  corresponde a la media de scores para  $k = i$ .  $\mu$  es la media general,  $B_l$  es el intercepto aleatorio del entrenamiento  $l$  y  $\epsilon_{ij}$  es el residuo entre el dato  $ij$  y la media  $\alpha_i$ .

Este modelo se estimó con el fin de inferir acerca de todos los conjuntos de entrenamiento y testeo posibles a partir del pool de datos total con el que se contó. Para este fin, se realizaron comparaciones múltiples cuyas hipótesis fueron puestas a prueba con el test de Dunnett a una cola. Este test permite realizar comparaciones de medias del tipo "todos contra el grupo control". Se tomó el grupo k=8 como control, con el fin de poner a prueba las hipótesis que proponen que la media poblacional del score de k=8 es mayor que cada una de las otras k medias.

Se realizaron las pruebas de hipótesis pertinentes para comprobar los supuestos distribucional de un modelo lineal general de comparación de medias (apéndice 7.4).

El código con el que se trabajó sobre la estadística de los datos obtenidos para la predicción 5A fue escrito en R, mientras que la implementación de K-Means, a partir de la cual se obtuvieron los datos, se realizó con la librería scikit-learn en python. Al igual que el código para entrenar las redes, el entorno en `tf_spherical.yml` contiene los paquetes compatibles necesarios para implementar estos algoritmos.

#### 5.1.3.2 Predicción 5B

*Que el cluster más cercano a cada cluster real sea un cluster de KMeans, y que ese cluster de K-Means sea distinto para cada cluster real.*

Lo que se espera de la red siamesa es que, cuantas más diferencias haya entre las sílabas de los individuos, más clusterizados estarán los datos de testeo. En consecuencia, más coincidencia habrá entre los centros de masa de los clusters reales y de K-Means. Por ende, otra manera de cuantificar la calidad del agrupamiento por parte de la red es teniendo una medida del grado de apareamiento entre los centros de masa reales y de K-Means; en un agrupamiento perfecto, a causa de una gran capacidad de las sílabas de funcionar como firma de individualidad, se esperaría una gran coincidencia entre ellos. Por el contrario, cuanto menor sea esa capacidad, los puntos estarían dispersos aleatoriamente en uno o pocos clusters y, por ende, sus centros de masa estarían cerca.

En concordancia con las hipótesis 2 y 3 de esta tesis (en las cuales se propone que solo las sílabas alfa de la hembra son distinguibles entre individuos), se esperaba que la predicción 5B se cumpliera solo en los clusters de la sílabas alfa. El primer abordaje para poner a prueba esta predicción implicó visualizar la distancia entre los clusters reales con los de K-Means, en otras palabras, el *grado de apareamiento* entre los clusters reales con los de K-Means (sección 5.2.2.1). En el segundo abordaje (capítulo 5.2.2.2) se cuantificó este grado de apareamiento.

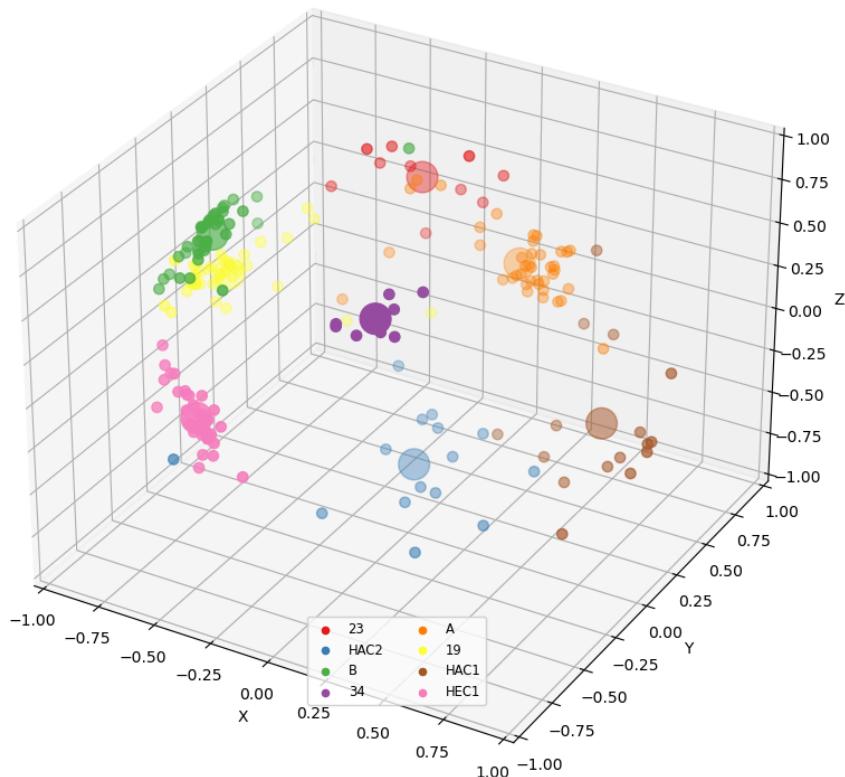
## 5.2 Resultados

Se exponen los resultados del capítulo 5 de la tesis, en la cual se formularon dos predicciones (5A y 5B) referidas a los resultados del entrenamiento de redes siamesas sobre los sonogramas de las sílabas de hembras y machos hornero.

### 5.2.1 Predicción 5A

*El número de clusters óptimo según el algoritmo de K-Means es el mismo número de clusters reales*

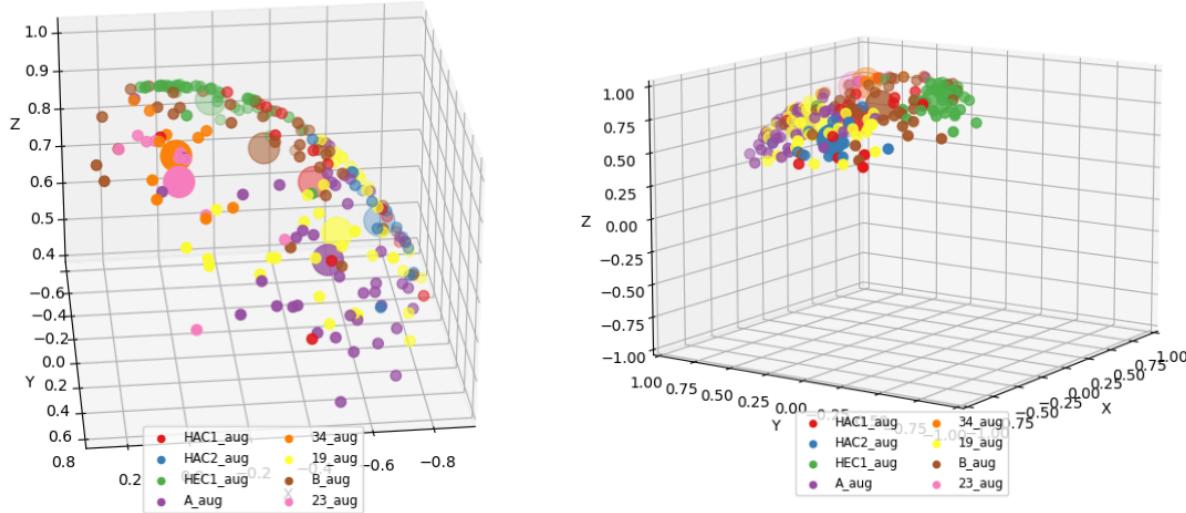
#### 5.2.1.1 Sílabas de hembra alfa



*Fig. 5.3: Espacio de embedding de una red entrenada con sílabas alfa, identificando cada sílaba del set de testeo con el individuo al que pertenece, formándose los clusters bien diferenciados. Los ejes de este espacio son adimensionales, y están normalizados entre -1 y 1: en el espacio de salida de la red entrenada, cada punto se encuentra en la superficie de una esfera de radio 1.*

Se observa una gran calidad en el agrupamiento a partir del set de testeo de las sílabas alfa, la cual es cuantificada a continuación.

A continuación se muestra un ejemplo de los clusters reales formados a partir de modelos nulos entrenados con sílabas alfa (fig. 5.4).



*Fig. 5.4: Embeddings de una red nula con sílabas alfa. a) Se muestra el detalle de la distribución de los puntos y su distribución esférica. Notar la menor separación de los clusters relativo a las sílabas alfa, b) muestra el mismo resultado pero en la escala de (-1,1).*

Como se mencionó previamente, se obtuvieron 20 sets de predicciones como las de la fig. 5.3. Sus scores para cada K se muestran en la fig. 5.5 junto con los modelos nulos.

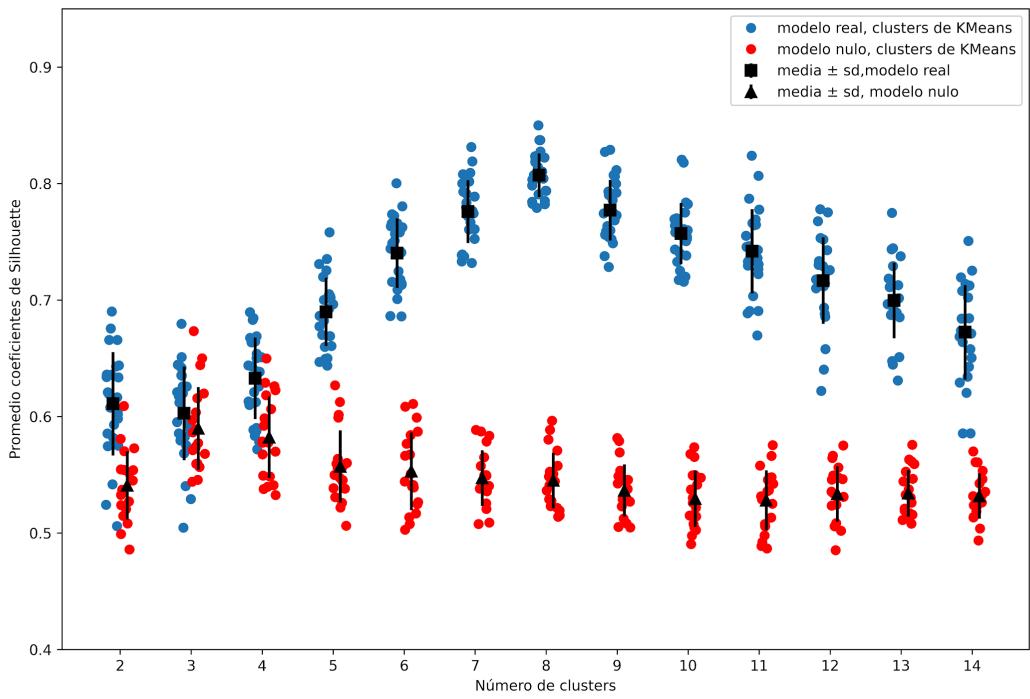


Fig. 5.5: En el eje horizontal se muestra el número de clusters  $K$  con el cual se implementó el algoritmo de KMeans en cada set de predicciones. Cada punto corresponde al score para el valor de  $K$  en ese entrenamiento. En modelos bien entrenados se realizaron 20 entrenamientos (azul) y en los modelos nulos 15 (rojo). Las barras son desvíos estándar, los cuadrados y triángulos son las medias de los scores de cada  $K$  para modelos bien entrenados y nulos, respectivamente.

En primer lugar, se estimó diferencia de medias de los scores de **un modelo bien entrenado y uno nulo**, a partir de los clusters formados por K-Means, **con  $k=8$** . Es decir, tomando  $k=8$  en la fig. 5.5 se compararon las medias de los grupos rojo y azul. Los resultados se muestran en la tabla 5. Se encontró que el score de K-Means con  $k=8$  clusters en un modelo bien entrenado es entre 0.25 y 0.298 unidades de score mayor que el del modelo nulo, con un 95% de confianza.

contrast	estimate	SE	df	<a href="#">lower.CL</a>	<a href="#">upper.CL</a>	t.ratio	p.value
bien entrenado-nulo	0.276	0.0109	24	0.253	0.298	25.3	8.25E-19

Tabla 5 referida a la fig. 5.5: comparación de medias marginales entre los score de  $K=8$  nulos y de los modelos bien entrenados. Se concluye que, a partir del pool de sílabas alfa con el que se cuenta, una red bien entrenada puede generar una distribución de

*puntos en la cual K-Means podrá encontrar 8 clusters cuyo score aumente entre 0.25 y 0.298 respecto de la distribución de puntos generada por un modelo nulo, con un 95% de confianza.*

Según se observa en la fig. 5.4, la estrategia de la red al encontrarse con datos nulos es juntarlos todos en un parche pequeño de la superficie de la esfera, disminuyendo la dispersión de cada uno. El compromiso entre juntar todos los puntos para que los de una misma clase estén cerca, pero que de este modo también todas las clases estén cerca, es el que minimiza la función costo. Intuitivamente, si los puntos estuvieran separados aleatoriamente ocupando una mayor superficie, no solo sus centros de masa seguirían juntos, sino que también cada punto estaría lejos del suyo. Otra forma de considerar este resultado es que la red no fue capaz de aprender nada, ya que recibió distintos tipos de sílaba que aleatoriamente tenían etiqueta de ser del mismo o de distinto individuo. Dado que no fue capaz de corregir sus pesos en las direcciones que le permitieran separar tipos de sílaba distintos, en el set de testeo agrupó todas en un mismo cluster. La red no formará clusters bien diferenciados en el entrenamiento (relativo al margen establecido en la función costo) si no son los correctos, es decir, si no están compuestos por sílabas del mismo individuo. Si en un paso feedforward ubicara puntos de individuos distintos en un mismo cluster, la función costo sería alta, con lo cual la red corregiría los pesos de tal modo que se maximice la distancia entre puntos de sílabas que tengan los features que la red detectó en una época posterior. Por ende, se puede pensar que el cambio neto en los pesos de la red sería cero. Esto redundaría en que, sobre un set de testeo, la red ubique todos los puntos en un mismo cluster.

El score de las predicciones del modelo nulo con los clusters formados por K-means (en rojo) por sí solo indica el máximo grado de estructura que se puede encontrar en los datos nulos, ya que lo que hace K-means es particionar equitativamente el conjunto de puntos (fig. 5.4), y K-Means no tiene en cuenta el margen de la función costo utilizada durante el entrenamiento.

Por otro lado, comparar el score de los clusters de K-means formados a partir de las predicciones del modelo nulo (rojo) con las formadas por el modelo bien entrenado (azul) es una forma de medir cuánta información o estructura hay en el dataset de sílabas, dado que indica cuánta estructura detecta K-Means.

A continuación se presentan las comparaciones múltiples de los scores **para todos los valores de K en los modelos bien entrenados con las sílabas de hembra alfa**, lo cual constituye el resultado más relevante en esta sección. En la fig. 5.6 se muestran los resultados del test de Dunnett, bajo las hipótesis nulas de ausencia de

diferencias significativas entre las medias poblacionales de  $k=8$  y las demás. Se concluye que la calidad de los clusters formados por K-Means a partir de un dado set de predicciones de la red sobre sílabas alfa se maximiza tomando 8 clusters.

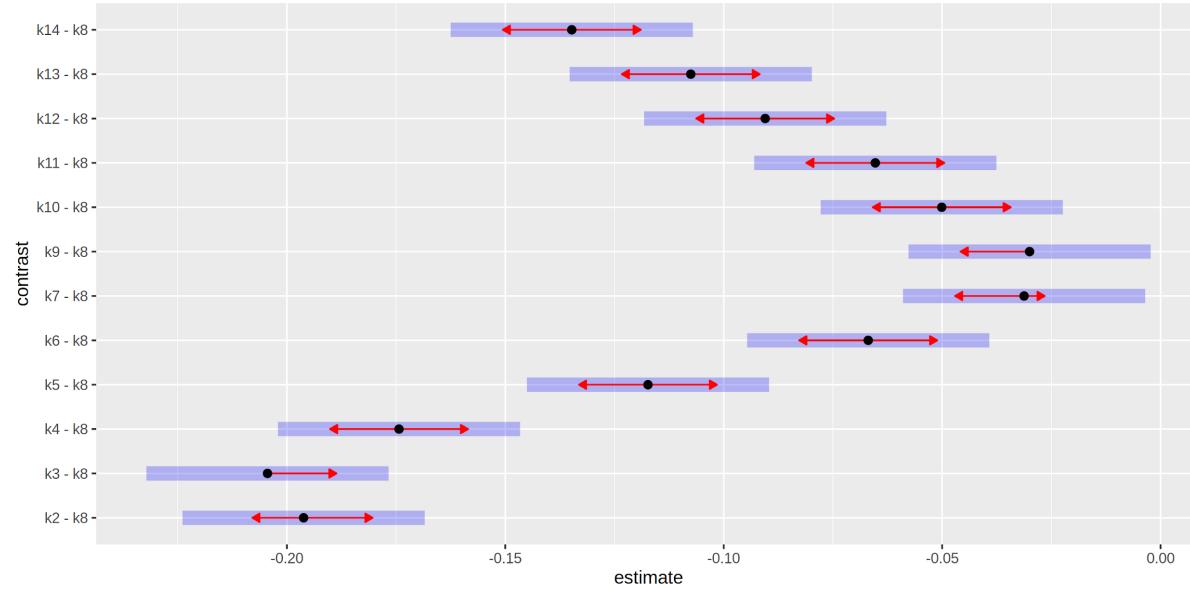


Fig. 5.6: Test de Dunnett a una cola tomando los score de  $k=8$  como grupo de referencia. En negro, media estimada de cada  $k$  expresada como la diferencia con la media estimada de  $k=8$ . En violeta, el intervalo de confianza para la media estimada de cada  $k$ . Flechas: intervalo de confianza para la diferencia de medias. Se evidencian diferencias significativas en la métrica de agrupamiento a partir del algoritmo de K-means entre todos los valores de  $k$  y  $k=8$  ( $pval < 0.05$ ); ningún intervalo de confianza para una diferencia de medias incluye al cero. El 83% de la varianza aleatoria es explicada por el número de clusters, el 17% por el entrenamiento del cual proviene un dado score.

La fig. 5.7 muestra las medias y errores estándar estimados para el score de los agrupamientos de K-Means con cada número de clusters a partir de las predicciones de la red sobre sílabas alfa.

Coeficientes de silhouette para cada K  
Media ± error estándar de cada K a partir del modelo de comparación de medias

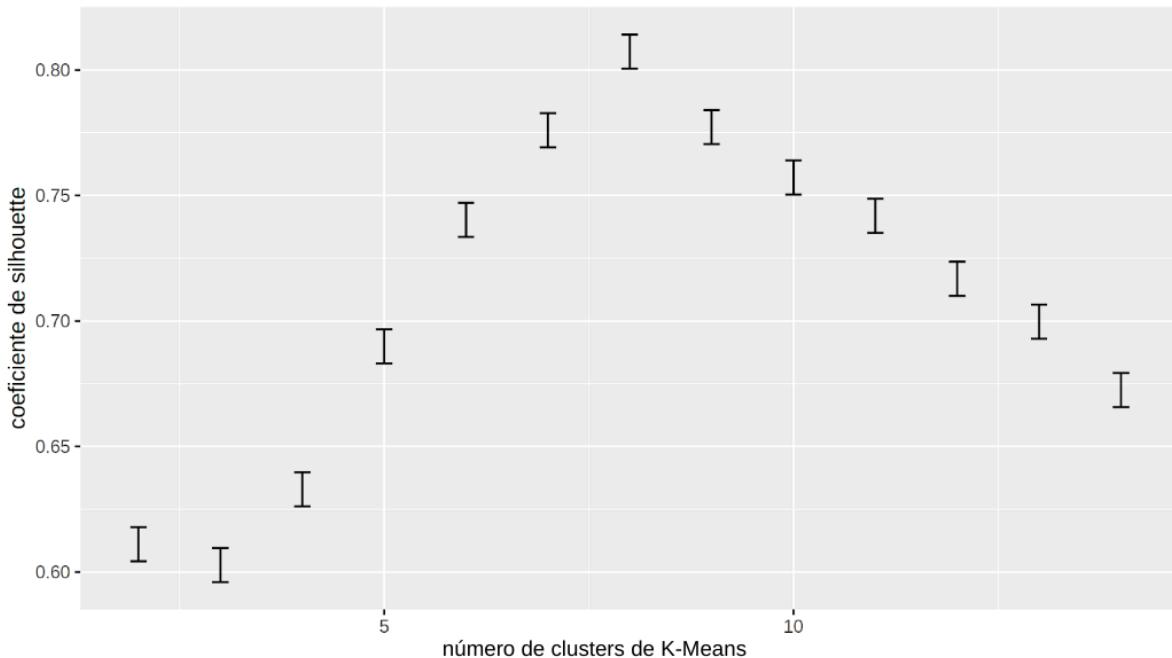
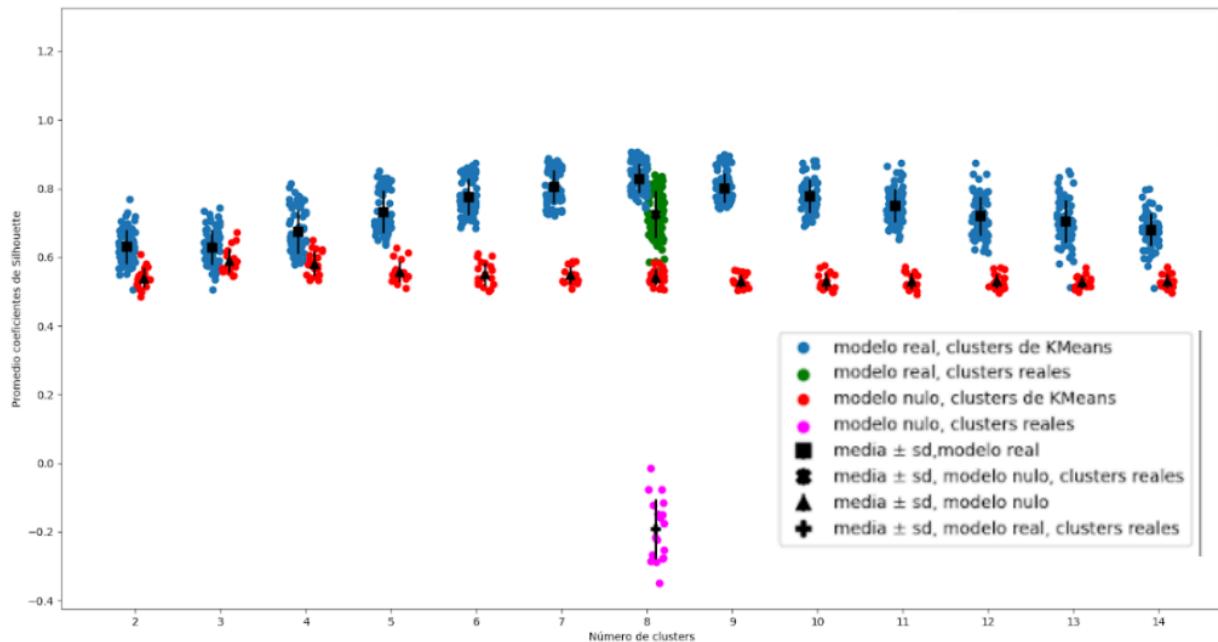


Fig. 5.7: Medias estimadas con errores estándar del coeficiente de silhouette para los clusters formados con K-Means con cada número de clusters (entre 2 y 14) en predicciones de la red sobre las sílabas alfa.

El resultado del test de Dunnett permite concluir que, **para las sílabas alfa, el número óptimo de clusters** de K-Means es 8. Estos clusters son formados en el espacio latente tridimensional de un dado entrenamiento de una red siamesa, y con un dado subset de datos de entrenamiento y validación del pool de subsets posibles. De este modo, el número óptimo de clusters es igual número real de clusters (8 horneros hembra). En cuanto a la menor diferencia de medias, la de k=8 y k=7 se encuentra entre -0.0598 y -0.0082 con un 95% de confianza.

A continuación (fig. 5.8) se muestran los mismos datos que los de la fig. 5.5 pero incluyendo ahora los scores de los 8 clusters reales, en vez de los clusters formados por K-Means.



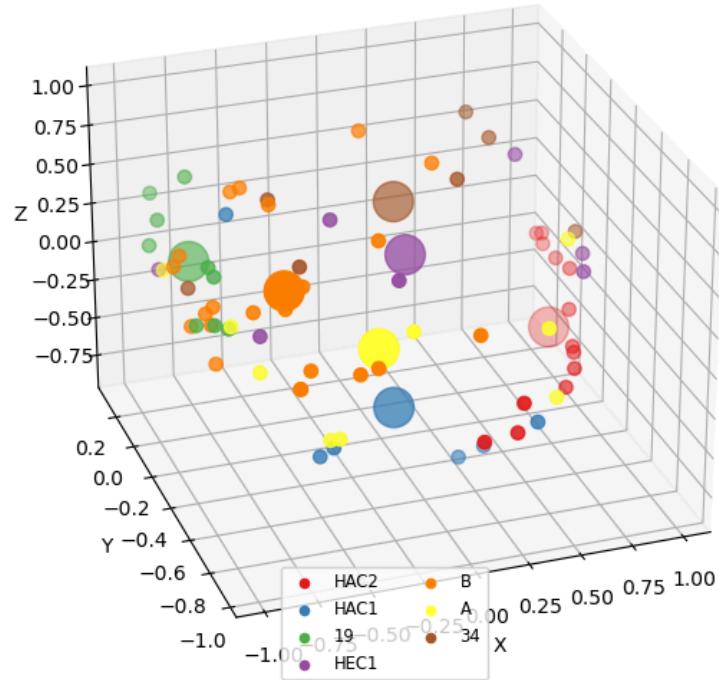
*Fig. 5.8: Extensión de la fig. 5.5. En rojo y azul se muestran los scores de los clusters formados por KMeans en los modelos nulos (en rojo) y los modelos bien entrenados (azul). En verde y rosa se muestran los scores de los clusters reales en los modelos bien entrenados (verde) y los clusters reales en los modelos nulos (rosa).*

La diferencia entre el score de los clusters reales en el modelo nulo (rosa) con los clusters de K-means en el mismo modelo (rojo) y, sobre todo, el hecho de que el primero sea cero, es una manera de determinar que no hay información sobre los clusters reales en el modelo nulo. Esto permite descartar un sesgo en el éxito de un modelo bien entrenado causado por un desbalance en los datos de entrenamiento.

La diferencia entre el score de los clusters reales (verde) y los de K-means (azul) (el score de los clusters ideales) en el modelo bien entrenado es abordada en las conclusiones sobre los resultados de la predicción 5A (sección 5.2.1.2.4)

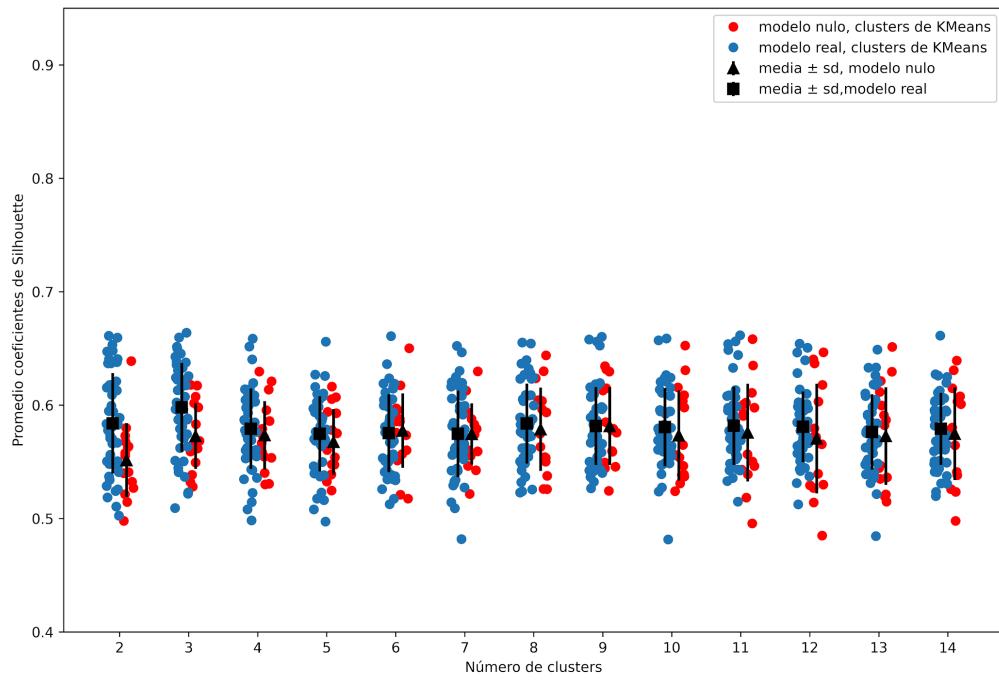
#### 5.2.1.2 Sílabas de hembra beta

A continuación se muestran los resultados con las sílabas beta, del mismo modo que se presentaron para las sílabas alfa. Se observará que se cumple la **hipótesis de trabajo 2**: No se observa firma de individualidad con las sílabas beta.



*Fig. 5.9: Espacio de embedding de una red entrenada con sílabas beta identificando cada sílaba con el individuo al que pertenece, formándose los clusters reales. Como se esperaba, la calidad de los clusters aparenta ser mucho menor que en las sílabas alfa.*

Del mismo modo que la calidad de los clusters obtenidos mediante las sílabas beta aparenta ser menor que la de las sílabas alfa (fig. 5.3), en los scores (fig. 5.10) no se observan diferencias ni entre distintos valores de K de modelos bien entrenados (azul) ni entre los bien entrenados y los nulos (rojo).



*Fig. 5.10: En el eje horizontal se muestra el número de clusters con el cual se implementó K-Means en cada set de predicciones de la red entrenada a partir de sílabas beta en un modelo bien entrenado (azul) y nulo (rojo).*

El test de Dunnett para las comparaciones múltiples de los scores de K=8 con todos los demás en un modelo bien entrenado (fig. 5.11) permite concluir que no hay un solo número de clusters que maximice el score a partir de las sílabas beta. En particular, entre K=4 y K=14 la media de los scores es igual a la de K=8.

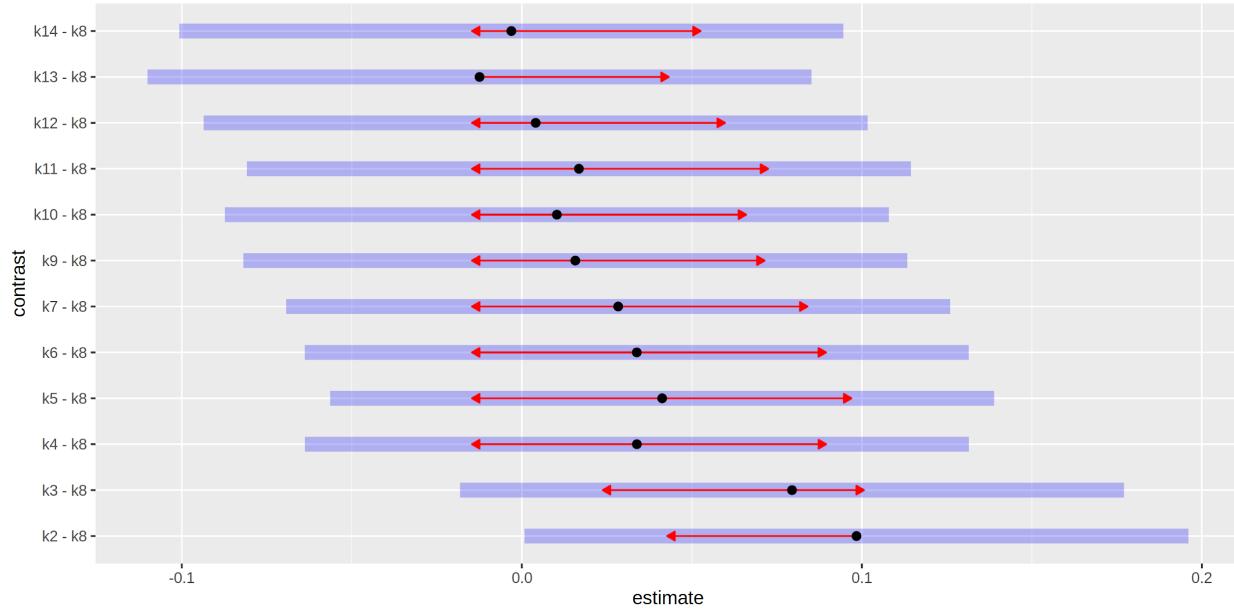


Fig. 5.11: Resultado del test de dunnett. No se observan diferencias significativas entre las medias de  $k_8$  y las demás, al igual que sucede en un modelo nulo. El 60% de la varianza aleatoria es explicada por el número de clusters, el 40% por el entrenamiento del cual proviene un dado score.

La fig. 5.12 muestra las medias y errores estándar estimados para el score de los agrupamientos de K-Means con cada número de clusters a partir de las predicciones de la red sobre sílabas beta.

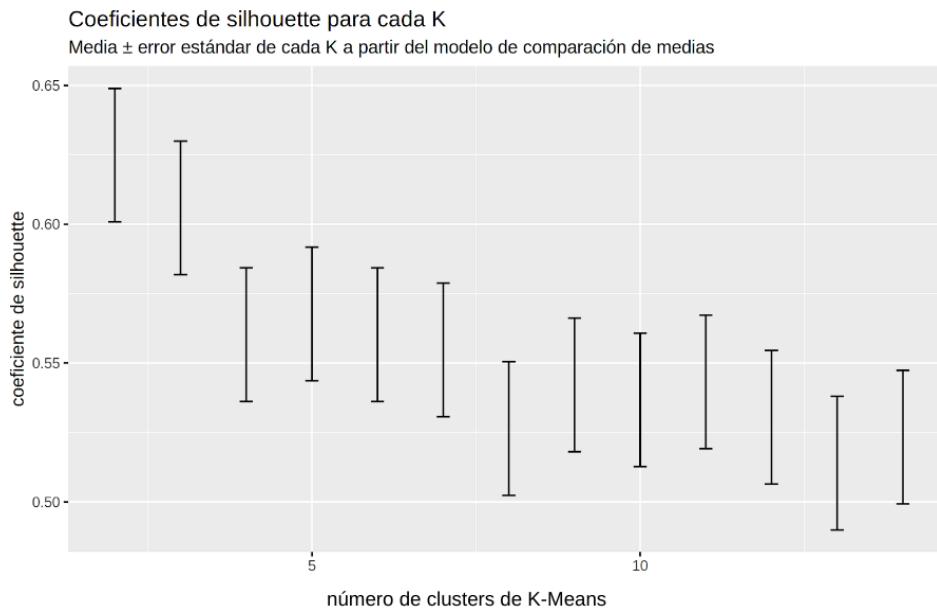


Fig. 5.12: Medias estimadas con errores estándar del score de los agrupamientos formados con K-Means, y distintos valores de K-clusters, a partir de predicciones de la red sobre sílabas beta.

A continuación (fig. 5.13) se muestran los mismos datos que los de la fig. 5.10 pero incluyendo ahora los scores de los 8 clusters reales, en vez de los clusters formados por K-Means. Como se mencionó previamente, la diferencia entre los scores para 8 clusters reales (verdes) y de K=8 (azules) se aborda en las conclusiones sobre los resultados de la predicción 5A (sección 5.2.1.2.4)

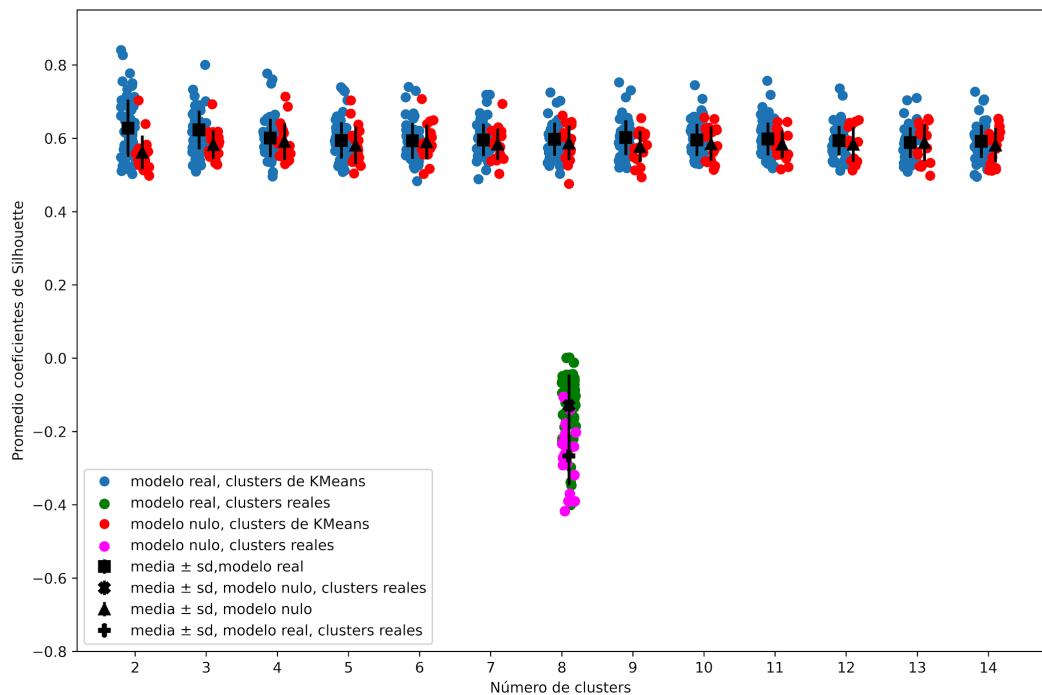


Fig. 5.13: Extensión de la fig. 5.10 para incluir los score de clusters reales: del modelo bien entrenado (verde) y modelo nulo (rosa).

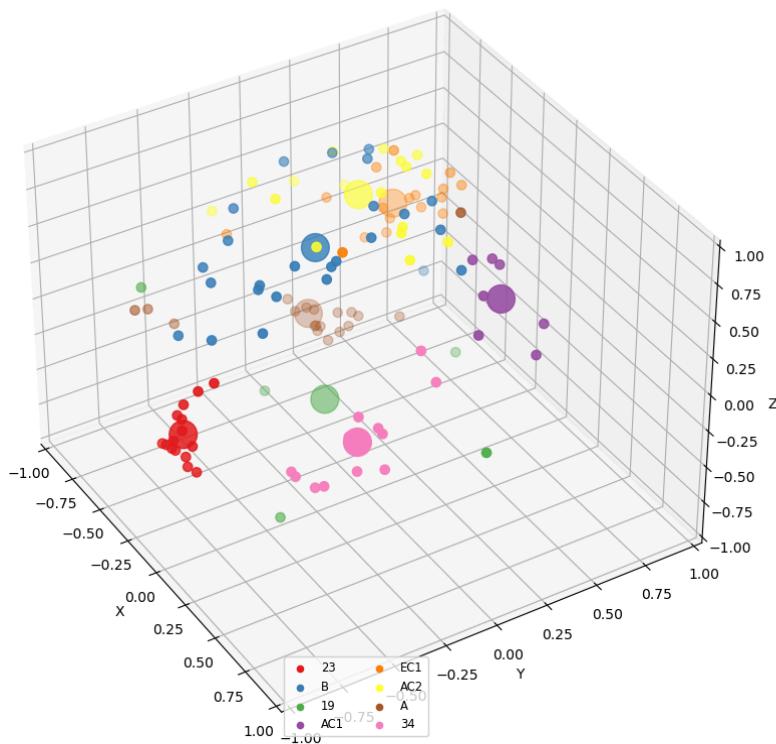
contrast	estimate	SE	df	<u>lower.CL</u>	<u>upper.CL</u>	t.ratio	p.value
bien entrenado - nulo	0.00266	0.013	39	-0.0235	0.029	0.205	0.839

Tabla 9: comparación de medias marginales entre los score de K=8 nulos y de los modelos bien entrenados con sílabas beta. Se concluye que, a partir del pool de sílabas beta con el que se cuenta, una red bien entrenada no puede encontrar un

*espacio de embedding en el cual el score de sus clusters cambie respecto del modelo nulo.*

Los resultados del test de Dunnett para el score de las sílabas beta se corresponden con lo esperado bajo la predicción relativa a la incapacidad de las sílabas beta de permitir la distinción de individuos a partir de su morfología.

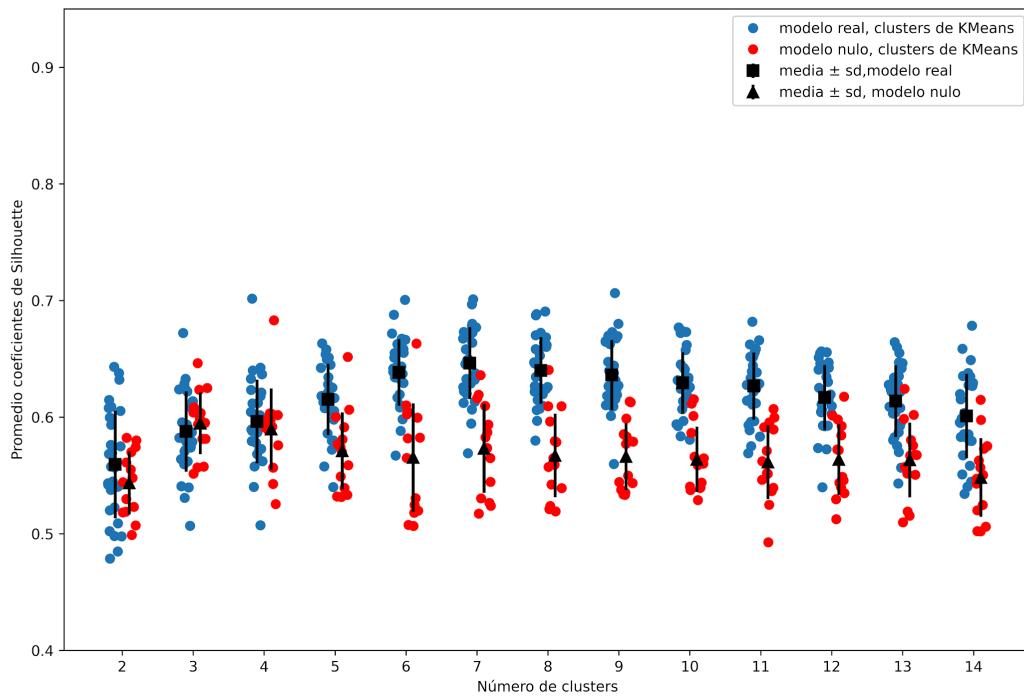
#### 5.2.1.3 Sílabas de macho



*Fig. 5.14: Espacio de embedding de una red entrenada con sílabas de macho identificando cada sílaba con el individuo al que pertenece, formándose los clusters reales.*

La visualización del espacio de embedding en la red entrenada con sílabas de macho (fig. 5.14) sugiere que hay un grado de agrupamiento que, aunque aparenta presentar un alto grado de mezcla entre clusters, es superior al azar, es decir, superior al esperado según la hipótesis de trabajo 3 de esta tesis, y a partir de las observaciones iniciales de las sílabas de macho (fig. 2.13). Se observa lo mismo en la fig. 5.22.

Los scores de la red sobre sílabas de macho (5.15) sugieren que puede haber un leve grado de estructura en los datos, ya que los modelos bien entrenados sugieren ser, para ciertos valores de K, superiores a los correspondientes modelos nulos.



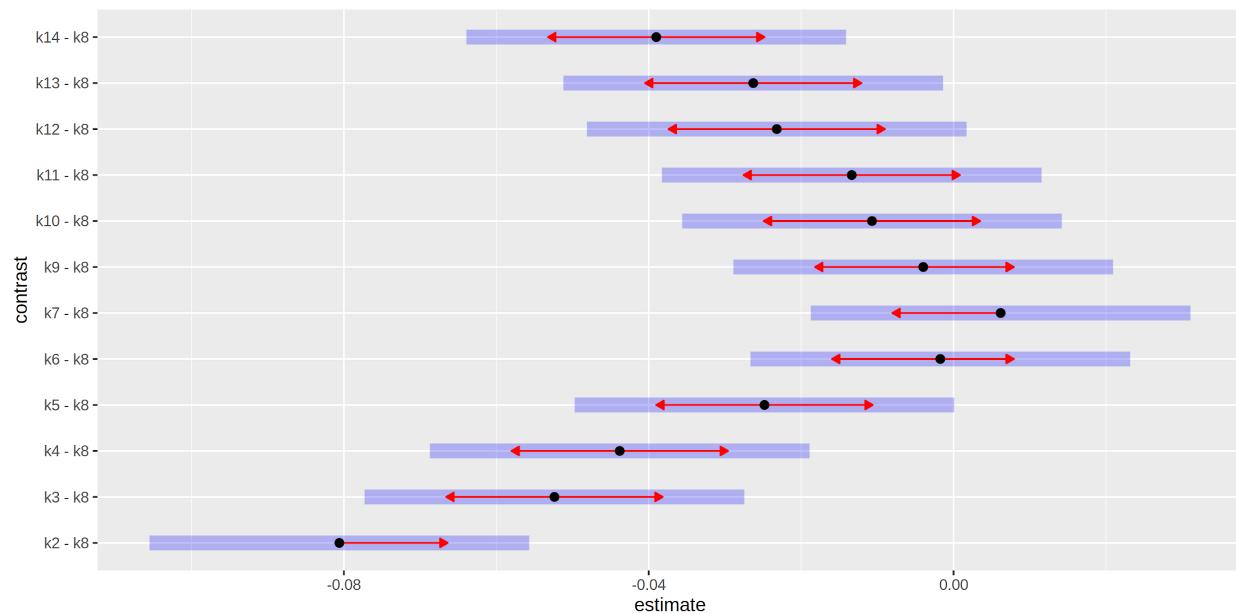
*Fig. 5.15: Score para cada valor de K en 20 entrenamientos de sílabas de macho (modelo bien entrenado, azul) y 15 entrenamientos (modelo nulo, rojo).*

A continuación se estima la diferencia de medias entre los scores de K=8 para los modelos bien entrenados y los nulos. Se encuentran diferencias significativas en las medias de estos scores.

contrast	estimate	SE	df	lower.CL	upper.CL	t.ratio	p.value
modelo bien entrenado - nulo	0.0784	0.01	41	0.059	0.097	8.54	1.22E-10

*Tabla 7 para la fig. 5.15: comparación de medias marginales entre los score de K=8 nulos y de los modelos bien entrenados con sílabas de macho. Se concluye que, a partir del pool de sílabas de macho con el que se cuenta, una red bien entrenada puede encontrar un espacio de embedding en el cual el score de sus clusters aumente, respecto de un modelo nulo, entre 0.06 y 0.097, con un 95% de confianza.*

Por otro lado, las comparaciones múltiples con K=8 para clusters formados a partir de predicciones sobre sílabas de macho permiten concluir que, entre K=6 y K=12, la media de los scores es igual a la de K=8.



*Fig. 5.15: Test de Dunnett a una cola tomando los score de k=8 como grupo de referencia. No se observan diferencias significativas entre las medias de k=8 y las de k=6, 7, 9, 10, 11 y 12: a partir de esta métrica, en el espacio de las sílabas puede haber entre 6 y 12 individuos. El 54% de la varianza aleatoria es explicada por el número de clusters, el 46% por el entrenamiento del cual proviene un dado score.*

La fig. 5.16 muestra las medias y errores estándar estimados para el score de los agrupamientos de K-Means con cada número de clusters a partir de las predicciones de la red sobre sílabas de macho.

### Coeficientes de silhouette para cada K

Media ± error estándar de cada K a partir del modelo de comparación de medias

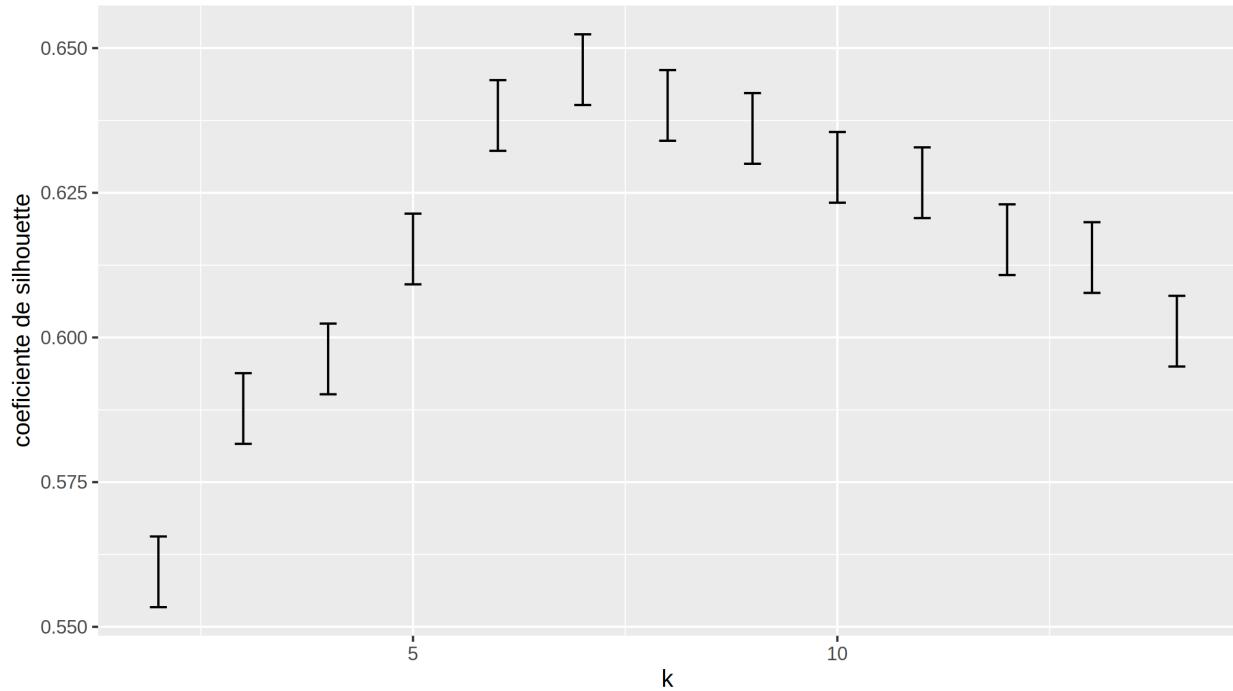


Fig. 5.16: Medias estimadas del score para los clusters formados con K-Means con errores estándar, sílabas de macho.

El resultado del test de Dunnett permite concluir que, para las sílabas de macho, el número óptimo de clusters de K-Means formados en el espacio latente tridimensional de un dado entrenamiento de una red siamesa, y con un dado subset de datos de entrenamiento y validación del pool de subsets posibles, es cualquier cantidad entre 6 y 12, con un 95% de confianza.

A continuación se muestran los mismos datos que los de la fig. 5.15 pero incluyendo ahora los score de los 8 clusters reales, en vez de los clusters formados por K-Means.

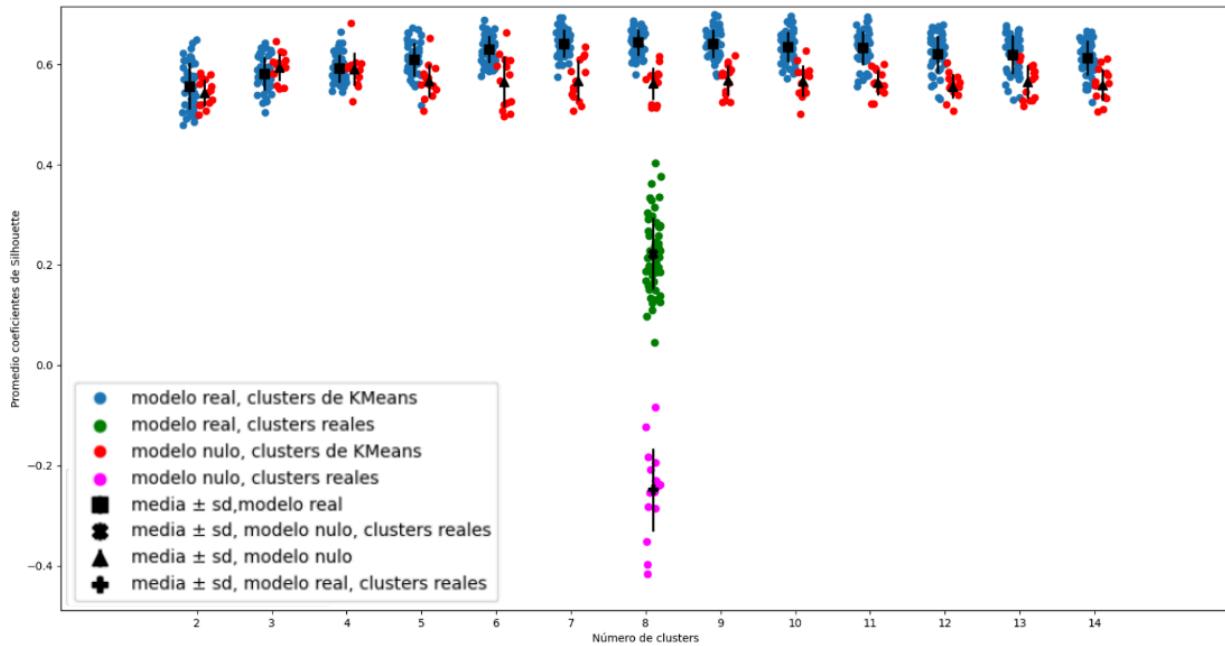


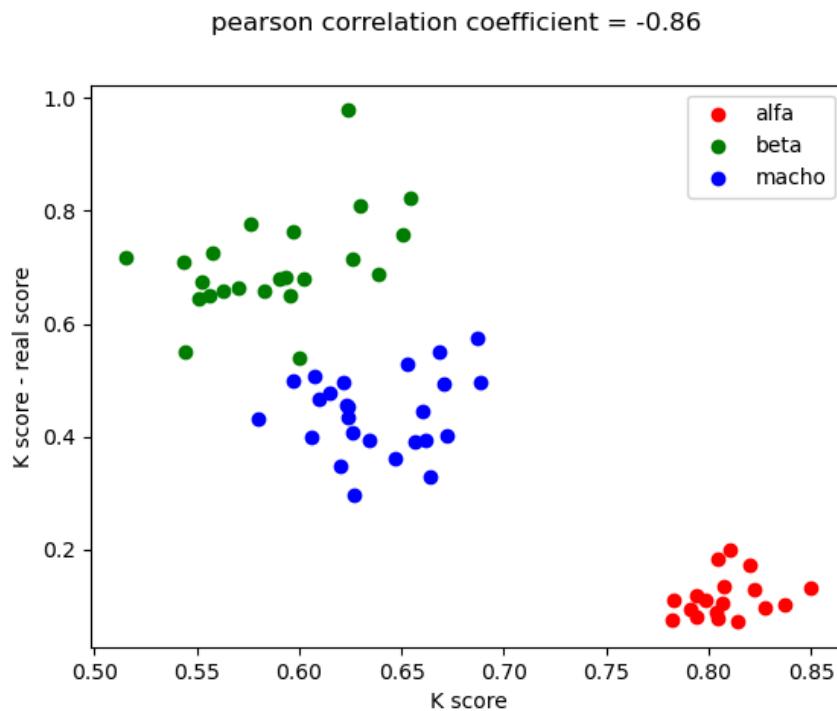
Fig. 5.17: Extensión de la fig. 5.15 para incluir los score de clusters reales: del modelo bien entrenado (verde) y modelo nulo (rosa).

#### 5.2.1.4 Conclusiones sobre los resultados de la predicción 5A

Se realizaron comparaciones múltiples de las medias de los scores para evaluar la calidad de los agrupamientos en base a sílabas alfa, beta y de macho para distintos números de clusters. Se encontró que el número óptimo de sílabas a partir del agrupamiento generado por la red en el set de testeo en las sílabas alfa es 8 (siendo este el número real de individuos). Esto no sucedió para las sílabas de macho y las sílabas beta. Por ende, no se encontró evidencia para rechazar la hipótesis nula asociada a la predicción 5A: la primera de las dos predicciones enunciadas para poner a prueba las hipótesis 2 y 3 de la tesis, en las cuales se propuso que las sílabas de hembra alfa presentarían firmas de individualidad, y no así las sílabas de hembra beta y las de los machos.

A continuación se interpreta la diferencia entre los scores de los clusters formados por K-Means con  $k=8$  y los reales (verde) (azul y verde respectivamente, en figs. 5.8, 5.13 y 5.17). La fig. 5.18 demuestra que la diferencia entre los scores de clusters reales y los de K-Means fue menor en las sílabas alfa que en las beta y las de macho. Esta diferencia se puede pensar como una medida de la calidad del clustering real respecto del ideal, dada la distribución de los datos en el espacio de salida. Es decir, es una medida del grado de aprendizaje a partir de las sílabas. Sin embargo, la distribución de las sílabas del set de testeo es, en sí misma, función del grado de

aprendizaje de la red: cuanto más haya aprendido, se esperaría que los puntos estén mejor agrupados. Esta correlación se muestra en la fig. 5.18.



*Fig. 5.18: correlación entre la diferencia del score de K-means con el score real, y el score real. Para cada entrenamiento de cada tipo de sílaba se tomó la media de los scores de  $k=8$  reales y de K-means.*

Aunque es esperable que esta correlación pueda ser considerada trivial, es presentada porque es una forma de demostrar:

- que la calidad del agrupamiento real se acerca al ideal a medida que la calidad del ideal aumenta (en vez de que, por ejemplo, la diferencia sea constante),
- que el score de K-Means por sí solo puede servir como una medida de calidad de los clusters reales, pero también lo puede ser la diferencia entre los clusters reales y los de KMeans..

Además, el motivo por el cual se planteó únicamente una correlación es que no se considera que una variable sea función de la otra. Más bien, ambas son función del grado de agrupamiento de los datos generado por la red lo cual, a su vez, depende del margen de la función costo y de la capacidad de las sílabas de funcionar como firma de individualidad. Lo que se puede concluir de esta correlación es que la resta entre scores reales y de K-Means, y el score de K-Means son intercambiables para ser utilizadas como medida de calidad de un agrupamiento.

En otras palabras, cuanto mayor sea el agrupamiento de los puntos en clusters, **mayor será la coincidencia entre los clusters formados por la red y los formados por K-Means**, ya que estos segundos son los clusters ideales. Este es un motivo por el cual puede ser útil **cuantificar la cercanía de los clusters reales a los de K-Means como otra medida de éxito en el clustering**; en vez de utilizar el score, se podría utilizar una medida que considere la distribución espacial de los clusters. En la evaluación de la predicción 5B se propone una manera de medirla, y en la sección 5.2.2.2 se observa la correlación entre ellas: entre el score y la cercanía de los clusters reales a los de K-Means.

A partir de los resultados referidos al score de las sílabas de macho (Predicción 5A), se puede considerar que no hay evidencia para rechazar la hipótesis 3 de la tesis. Sin embargo, también se considera que hay un contraste entre los resultados del score de las sílabas de macho (fig. 5.15) y la observación cualitativa de sus clusters (fig. 5.14).

Esto es un motivo por el cual puede resultar útil buscar otra medida de la calidad del agrupamiento que tenga en cuenta la estructura espacial de los clusters y no solo su score. Por un lado, esto permitiría llegar a una manera distinta de definir la calidad del agrupamiento generado por la red, independientemente del score en cada caso. Por otro lado, puede resultar útil para calibrar qué es lo que efectivamente se puede concluir a partir de la observación de los clusters, en referencia a la medida de calidad utilizada, que es el score.

Esta discrepancia puede deberse a que el score es sensible a la varianza de los clusters y, a su vez, la varianza de los clusters formados *por la red* está sujeta al margen que se elija en el entrenamiento. Como se mencionó previamente, una vez que un punto se encuentra más allá del margen de un cluster al que no pertenece, deja de ser alejado del mismo. Por ende, se procedió a buscar una forma de independizarse del margen del entrenamiento, y por ende del score; se buscó una medida de la distribución espacial de los clusters reales (verdes) relativa a los ideales (azules) para una dada distribución de puntos, que es presentada en la sección 5.2.2.

En suma, a partir de la fig. 5.15 y la tabla 7 se puede concluir que, nuevamente, el modelo bien entrenado con sílabas de macho distribuye los puntos en el espacio de un modo ligeramente mejor que un modelo nulo, ya que ambos modelos nulos son inferiores a su respectivo modelo bien entrenado.

A continuación, se buscará cuantificar la calidad de los clusters independientemente del score, o bien, de la diferencia los scores reales y de K-means.

De este modo, se podrá conciliar el resultado de esta medida con lo observado en los resultados de las sílabas de macho. Para eso, se considerarán las distancias de los centros de masa reales y de K-means.

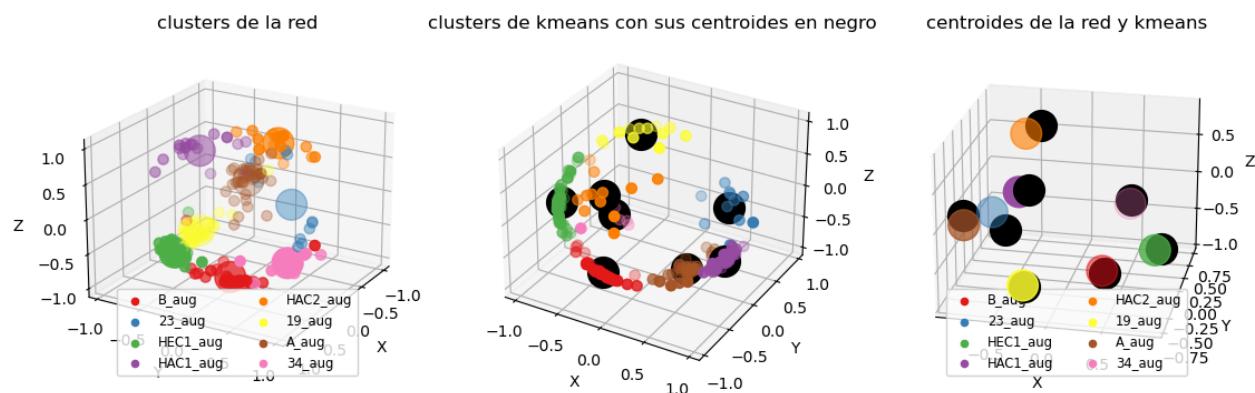
## 5.2.2 Predicción 5B

*Que el cluster más cercano a cada cluster real sea un cluster de KMeans, y que ese cluster de K-Means sea distinto para cada cluster real.*

En acuerdo con las hipótesis 2 y 3 de esta tesis (en las cuales se propone que solo las sílabas alfa de la hembra son distinguibles entre individuos), se esperaba que la predicción 5B se cumpliera solo en los clusters de la sílabas alfa. A continuación se expone una serie de visualizaciones en las cuales se observan las distancias entre los clusters reales y los de KMeans para cada tipo de sílaba. La cuantificación de estas distancias (del grado de apareamiento entre los clusters reales y los de KMeans) se expone en la sección 5.2.2.2, con el fin de facilitar la comparación tanto de las visualizaciones como de las cuantificaciones entre tipos de sílaba.

### 5.2.2.1 Visualizaciones

#### 5.2.2.1.1 Sílabas de hembra alfa



*Fig. 5.19: Visualización de la correspondencia entre los clusters reales y de K-Means en el espacio latente de un modelo bien entrenado con las sílabas de hembra alfa. Los centros de color indican los centros de masa de los clusters reales, y los centros negros indican los de los clusters formados por K-Means con K=8.*

Al considerar un modelo nulo (con las etiquetas aleatorizadas) entrenado con sílabas alfa (fig. 5.20), la comparación entre las figuras 5.20a (con los 8 clusters reales) y 5.20b (con los 8 clusters de K-Means) muestra cómo K-Means partitiona el espacio de los datos para minimizar la varianza. Esto ilustra el motivo por el cual, con las

predicciones de un modelo nulo, el score de K-Means (rojo en figs 5.8, 5.13, 5.17) es superior al score real (rosa en figs 5.8, 5.13, 5.17)

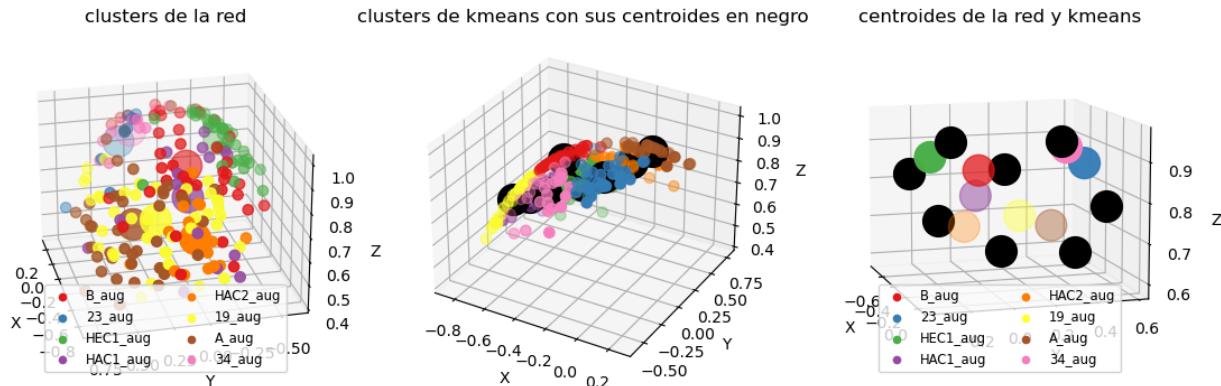


Fig. 5.20: Comparación entre los clusters reales y de K-Means de un **modelo nulo con sílabas de hembra alfa**. Los centros de color indican los centros de masa de los clusters reales, y los centros negros indican los de los clusters formados por K-Means con  $K=8$ .

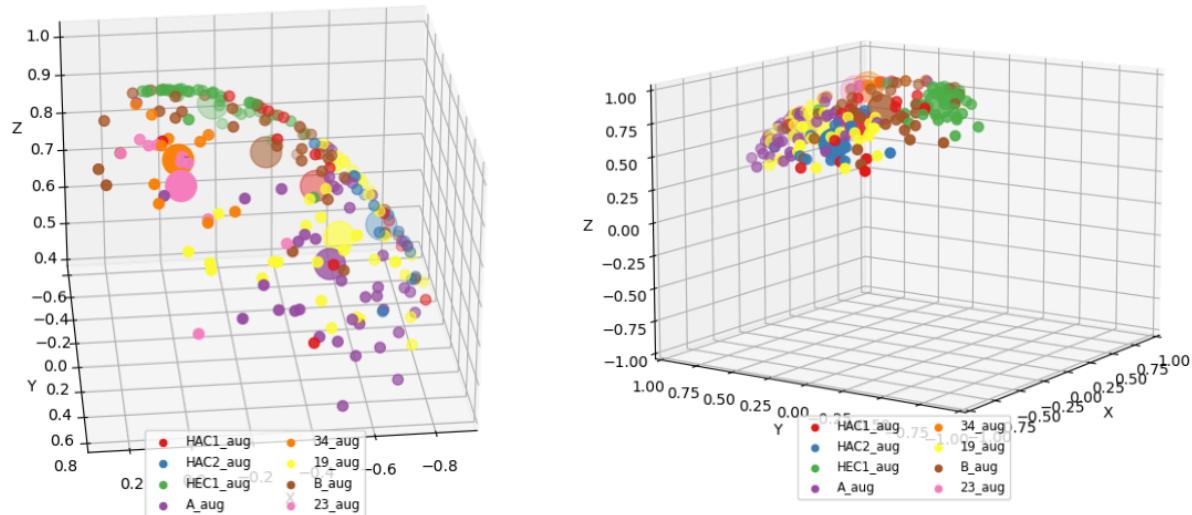
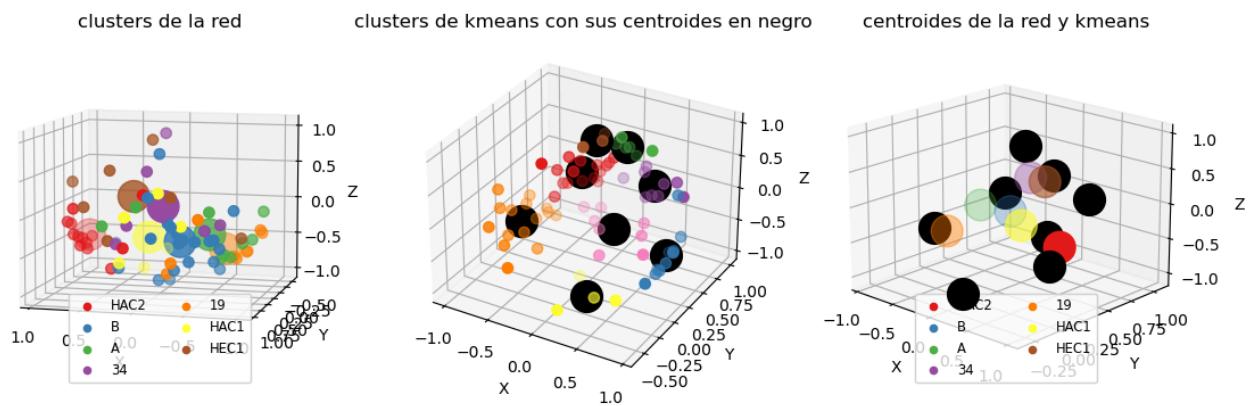
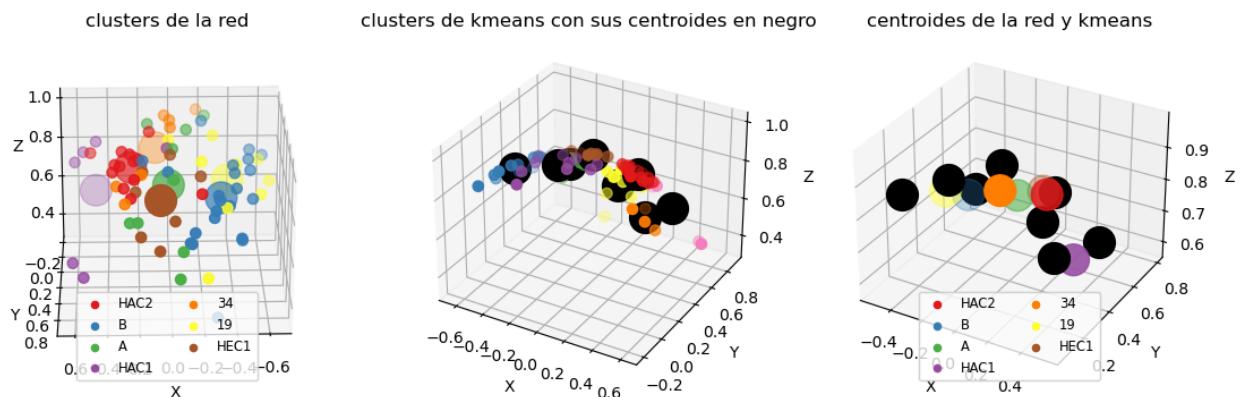


Fig. 5.21: Acercamiento de la figura 5.20a para mostrar el tamaño del parche ocupado por los puntos en el modelo nulo.

### 5.2.2.1.2 Sílabas de hembra beta

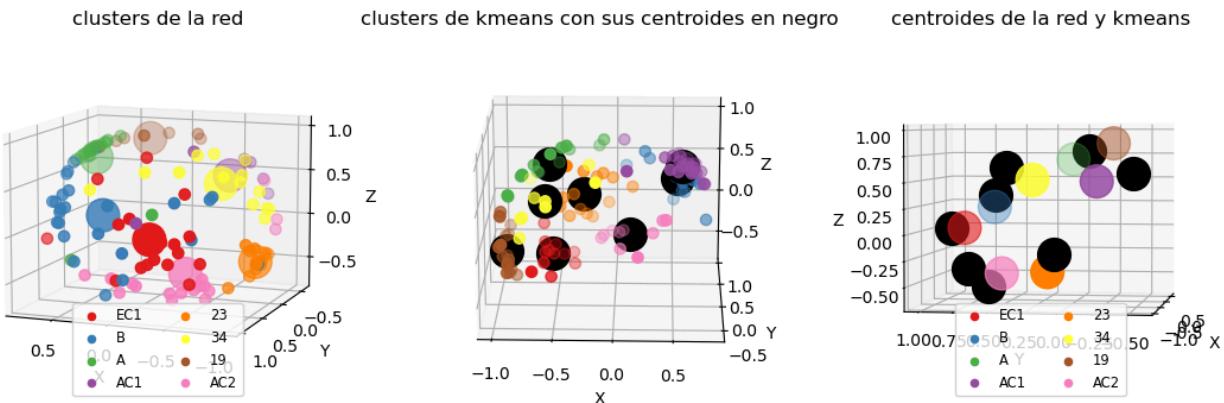


*Fig. 5.22: Visualización de la correspondencia entre los clusters reales y de K-Means en el espacio latente de un modelo bien entrenado con las sílabas beta. Los centros de color indican los centros de masa de los clusters reales, y los centros negros indican los de los clusters formados por K-Means con K=8.*



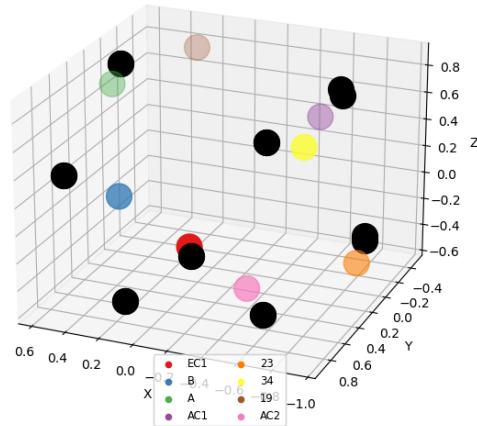
*Fig. 5.23: Comparación entre los clusters reales y de K-Means de un modelo nulo de las sílabas beta. Se observa nuevamente la falta de correspondencia entre los clusters reales y de K-Means en un parche pequeño de la esfera, además de un score=0. Los centros de color indican los centros de masa de los clusters reales, y los centros negros indican los de los clusters formados por K-Means con K=8.*

### 5.2.2.1.3 Sílabas de macho



*Fig. 5.24: Visualización de la correspondencia entre los clusters reales y de K-Means en el espacio latente de un modelo bien entrenado con las sílabas de macho. Los centros de color indican los centros de masa de los clusters reales, y los centros negros indican los de los clusters formados por K-Means con  $K=8$ .*

Fue a partir de las figuras 5.25 y 5.24 que se propuso una métrica que permitiera cuantificar el grado de apareamiento entre los centros de masa de los clusters con los ideales para una dada distribución de puntos.



*Fig. 5.25: Acercamiento de la fig. 5.24c para mostrar en detalle el grado de apareamiento en las sílabas de macho. Los centros de color indican los centros de masa de los clusters reales, y los centros negros indican los de los clusters formados por K-Means con  $K=8$ .*

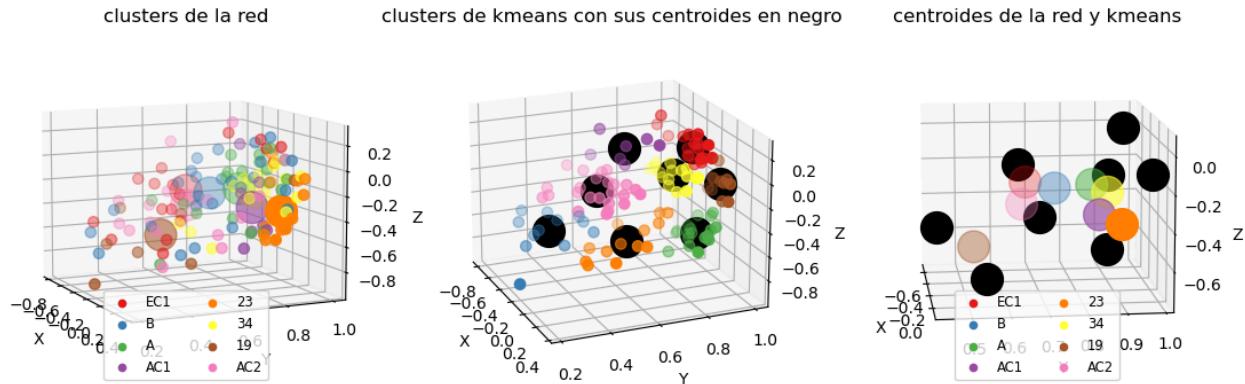


Fig. 5.26: Comparación entre los clusters reales y de K-Means de un **modelo nulo** en sílabas de macho. Se observa nuevamente la falta de correspondencia entre los clusters  $r$  y  $k$  en un parche pequeño de la esfera, además de un score bajo. Los centros de color indican los centros de masa de los clusters reales, y los centros negros indican los de los clusters formados por K-Means con  $K=8$ .

El resultado que resalta en las figs. 5.19-5.26 es el mencionado acerca de las sílabas de macho: se observa una correspondencia mayor que la esperada según su score con los centros de masa de K-means. Es a partir de esta observación que se profundizó para llegar a otra medida de éxito en el clustering que aportara información distinta y más afín al resultado que puede observarse: una medida del grado de apareamiento entre los clusters reales y los ideales para una dada distribución de puntos, ya que la misma está condicionada al margen establecido durante el entrenamiento.

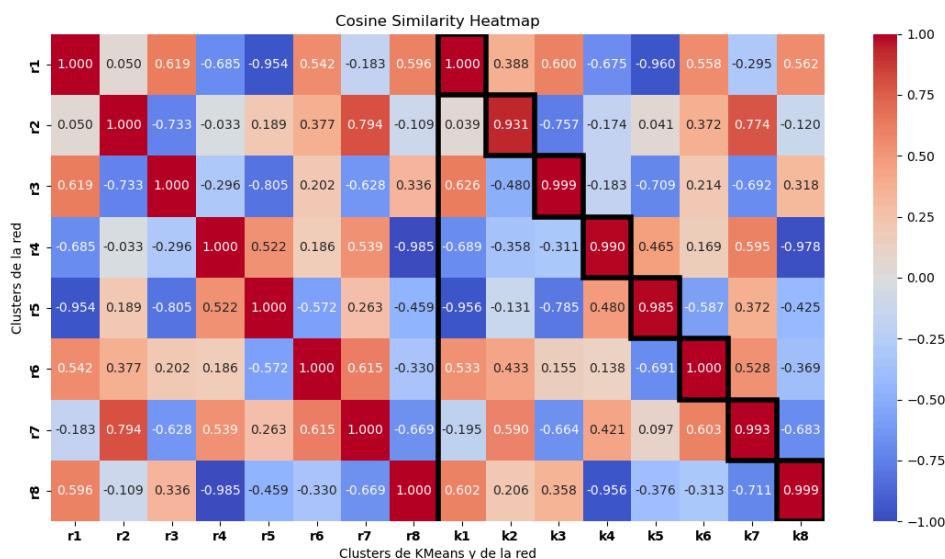
### 5.2.2.2 Cuantificación del grado de apareamiento

#### 5.2.2.2.1 Heatmaps

Una aproximación para poner a prueba la predicción 5B, asociada a las figs. 5.18-5.25, es observar una matriz de similitud coseno para todo los clusters presentes en el espacio latente: los r (reales) y los k (de K-means). La similitud coseno entre dos puntos es el cálculo del coseno del ángulo entre ellos. Es 1 cuando los puntos se encuentran en el mismo lugar, y -1 si están en ubicaciones opuestas. Dado que los puntos están en la superficie de una esfera, el ángulo comprendido entre ellos y el origen contiene toda la información sobre su ubicación.

Para facilitar la lectura de los heatmaps, se define un cluster r como cluster real, y un cluster k como cluster de KMeans. Lo que se espera observar en el heatmap correspondiente a un buen agrupamiento es que se forme una diagonal en el segundo panel: que la máxima similitud para un dado cluster r sea un cluster k, y que ese k sea distinto para cada r. De este modo, se puede observar una estructura de a pares en la matriz: para un dado par de clusters R-K, los dos clusters más cercanos a ellos son otro par r-k, y la distancia a cada par es muy similar, de modo que se observa una estructura de a pares.

Del mismo modo que en las visualizaciones anteriores, primero se muestran las figuras correspondientes a cada tipo de sílaba, y luego se discuten los resultados.



*Fig. 5.27: Heatmap de similitudes coseno entre los clusters de k-means y reales en las sílabas alfa. Se observa un muy alto grado de correspondencia al notar la diagonal en*

el segundo panel. En cada fila se tiene un cluster  $r$ . En el panel izquierdo se tienen los mismos clusters. Por supuesto, la diagonal en el panel de la izquierda indica que el cluster  $r$  más cercano al  $r$  de la fila es sí mismo.

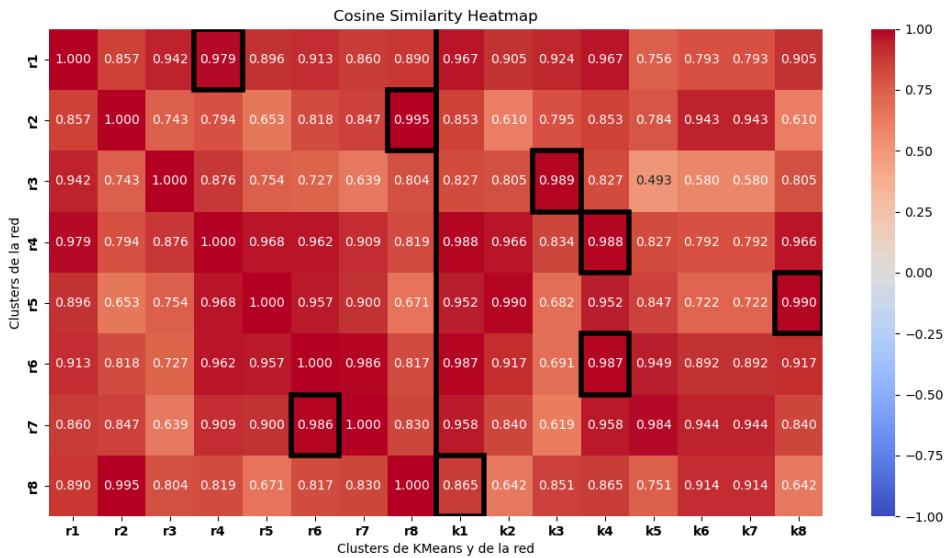


Fig. 5.28: Heatmap del **modelo nulo, sílabas alfa**. No se detecta una estructura de a pares, como se esperaba.

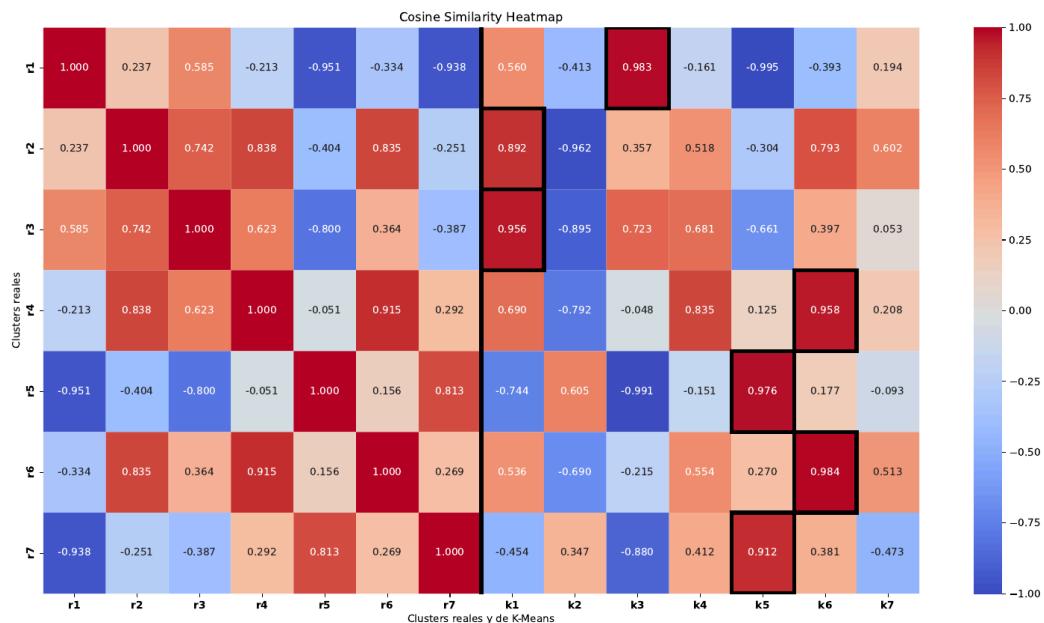
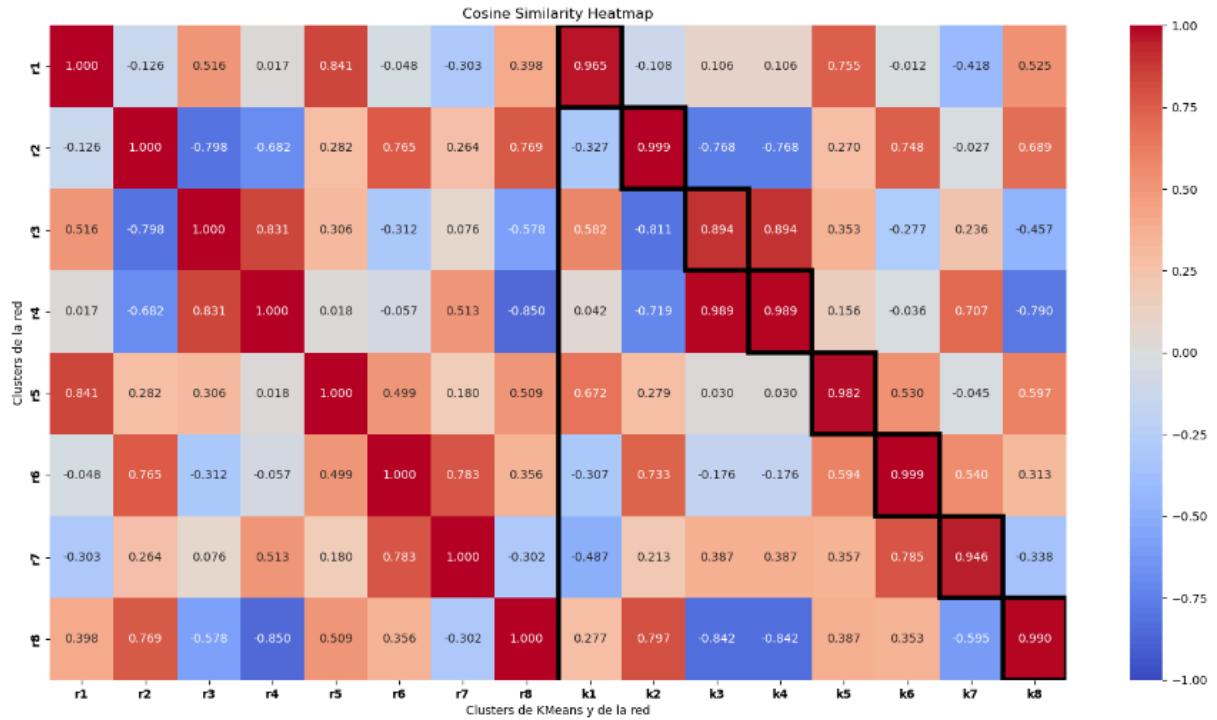


Fig. 5.29: Heatmap de similitudes coseno entre los clusters de las **sílabas beta**. No se cumple la **Predicción 5B**: el cluster más cercano a cada cluster  $r$  es el mismo para varios  $r$ , por ende, no se observa una diagonal en el segundo panel del heatmap.



*Fig. 5.30: Heatmap de similitudes coseno entre los clusters de k-means y reales en las sílabas de macho. Se observa un cierto grado de correspondencia al notar la diagonal en el segundo panel. Sin embargo, parece ser menor que en las sílabas alfa, y esta observación no considera el grado de dispersión de los puntos, sino sus centros de masa.*

La matriz de similitudes de las sílabas de macho (fig. 5.30) denota un menor grado de apareamiento que las sílabas alfa, pero mayor que el esperado según el score de los clusters reales del macho (fig. 5.17). Este resultado es el esperado a partir de las observaciones sobre sus clusters: que el grado de apareamiento sea menor que las alfa, pero que sea mucho mayor que el de un modelo nulo, o que el de las sílabas beta.

#### 5.4.3.2 Cuantificación de los heatmaps

A partir de esta primera aproximación a una medida de similitud entre clusters r y k, se procedió a definir una cantidad que permitiera cuantificar el grado de apareamiento observado en los heatmaps. Puede resultar trivial para determinar si hubo correspondencia con los clusters de K-Means en base a lo observado en los clusters de las sílabas alfa, pero puede ser útil para comparar la calidad de los clusters en base al grado de apareamiento con los de k-means entre tipos de sílaba en los que el grado sea menor que en este caso, pero mayor que en un modelo nulo.

Para esta cuantificación, fijando un dado cluster r (llamado R) en un dado set de predicciones, se compararán dos ángulos (fig. 5.31). Lo que se considerará es la similitud entre los puntos involucrados. La primera similitud es la del ángulo formado entre el origen, R, y el k más cercano: ( $\gamma_k$ ). El segundo ángulo se forma entre el origen, R, y el r más cercano; la similitud R-r se denominará  $\gamma_r$ . Lo que se espera en un buen apareamiento es que, para todos los clusters R,  $\gamma_k > \gamma_r$ . Tomando el cociente  $\gamma = \gamma_k / \gamma_r$ , se espera que sea  $> 1$  para buenos agrupamientos.

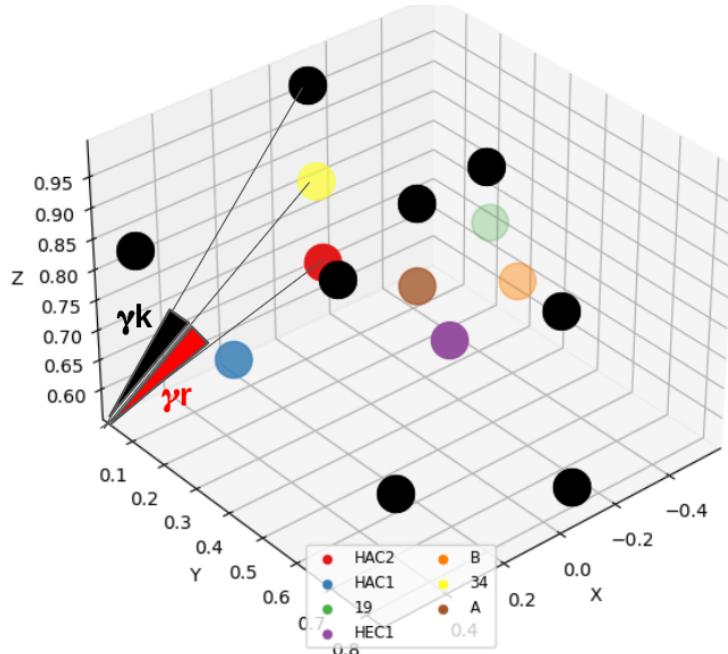
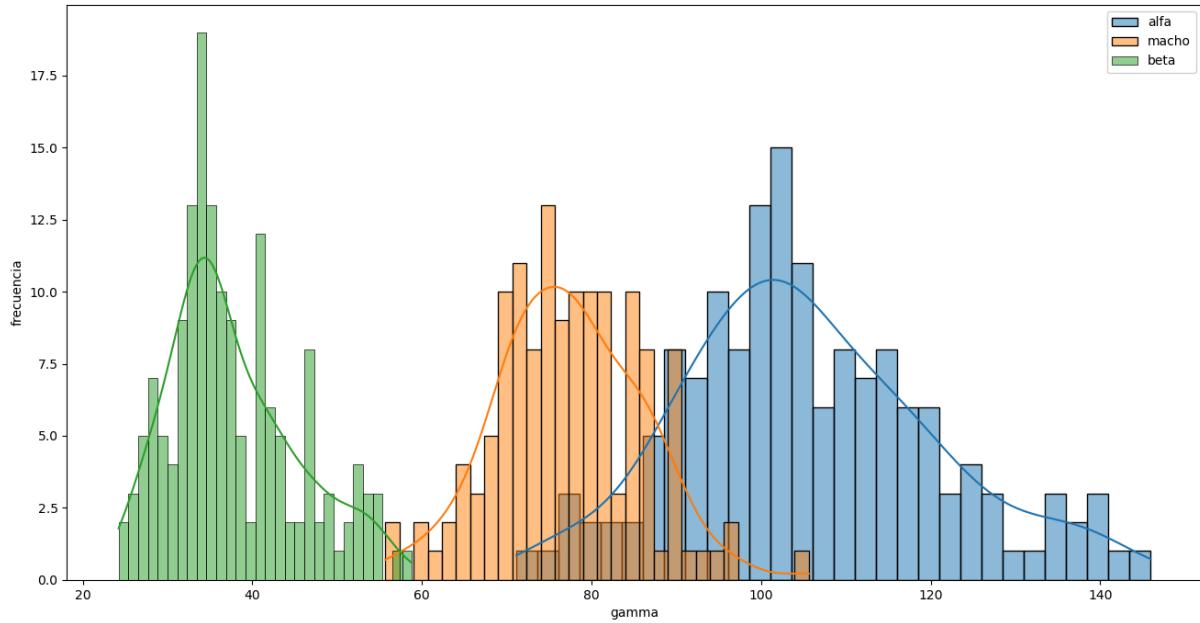


Fig. 5.31: Se ilustra el procedimiento de obtención de  $\phi$  con un conjunto de centros de masa (K-Means en negro, r en colores). Se utilizó un modelo nulo con el fin de que hubiera una distancia cómoda entre los clusters para mostrar los ángulos; justamente, en un modelo nulo, se espera que los ángulos de un par  $\gamma_k$ - $\gamma_r$  sean iguales. Cuanto menos pareados estén los clusters r-k, menor será  $\gamma$ . Si llama la atención la escala de los ejes es porque se trata de un modelo nulo.

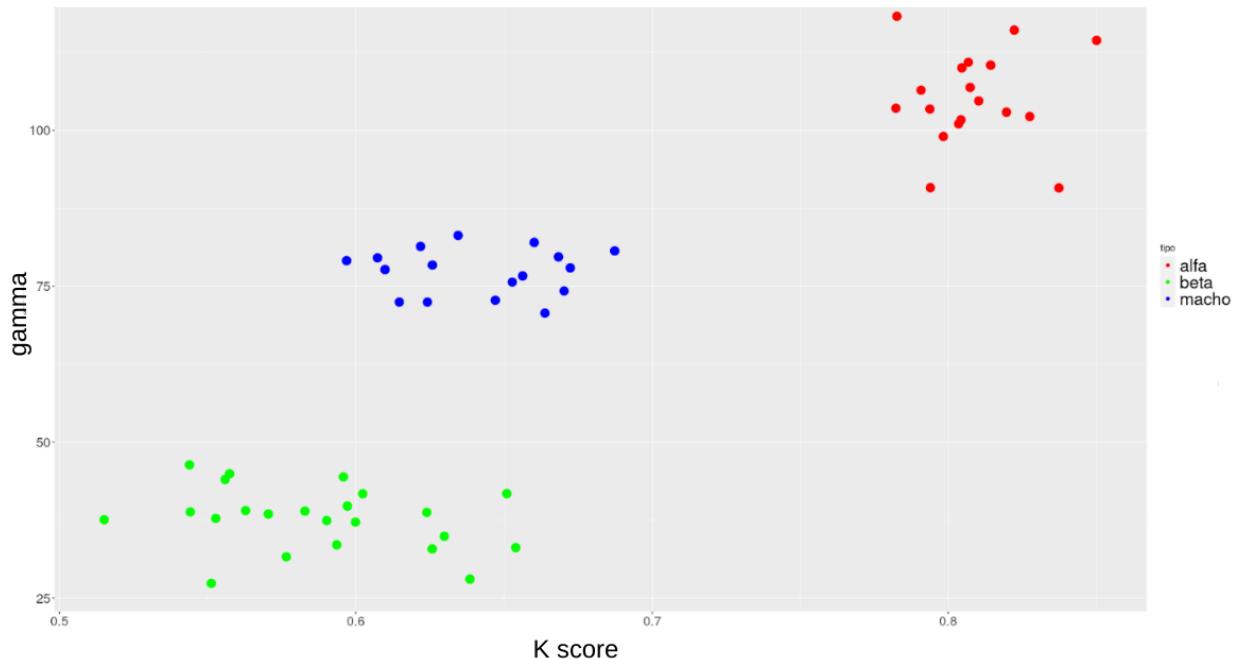
Con este procedimiento se pueden obtener 8 valores de  $\gamma$  en un set de predicciones. Al tener 20 sets para cada tipo de sílaba, se tuvo la intención de comparar las medias poblacionales de los  $\gamma$  de cada tipo mediante un test de tukey con un modelo de comparación de medias, utilizando el identificador del entrenamiento como una variable de efectos aleatorios, dado que los  $\gamma$  de un dado entrenamiento no serían independientes entre sí. Sin embargo, no se cumplieron los supuestos de normalidad de los residuos ni de igualdad de varianzas. Por ende, la fig. 5.32 es descriptiva.



*Fig. 5.32: distribución de gamma en 20 entrenamientos para los tipos de sílaba alfa, beta (verde) y macho (naranja); las curvas son estimaciones de la función de densidad de probabilidad de la variable aleatoria para cada tipo de sílaba utilizando un kernel gaussiano. La magnitud gamma es adimensional; es el cociente de dos ángulos.*

Esta manera de medir la calidad del agrupamiento puede sugerir que la de los machos es intermedia entre la de las sílabas alfa y beta (e incluso se superpone más con los  $\gamma$  de las alfa), relativa a la distribución de los puntos generada por la red con cada tipo de sílaba la cual, a su vez, depende del grado de estructura en los datos. Este cercanía de las sílabas de macho con las alfa difiere de la observada en las diferencias entre los scores del cluster real y cluster k en cada tipo de sílaba.

Dado que las métricas presentadas conducen a resultados algo diferentes, se procedió a explorar la correlación entre ellos.



*Fig. 5.33: Gráfico exploratorio de la correlación entre el grado de apareamiento entre clusters reales en el eje y, el score de K-Means en k=8 en eje x. Se tomó el gamma promedio de cada entrenamiento. Coeficiente de correlación de Pearson: 0.87.*

La correlación entre ellos sugiere que estos indicadores de calidad del agrupamiento deberían ser intercambiables. Sin embargo se reitera que, mientras que los resultados de los scores de los machos (fig. 5.15) indicarían que hay un grado de individualidad mucho más cercano al de las sílabas beta que las alfa, la visualización de las predicciones de la red (fig. 5.14) y la cantidad  $\gamma$  indicarían que el grado de individualidad en estas sílabas es intermedio entre alfa y beta.

#### 5.2.2.3 Conclusiones sobre los resultados de la predicción 5B

A partir de estas observaciones, se considera que los resultados de la **Predictión 5B** referida a las sílabas de macho no son enteramente concluyentes. En una conclusión conservadora se puede afirmar que se observó en las sílabas de macho una muy leve, pero no nula, capacidad de funcionar como firma de individualidad a partir de su morfología. Esta capacidad no permitió distinguir 8 individuos, pero sí fue superior al azar.

Por el contrario, todos los resultados asociados a la **Predictión 5B** referidos a las sílabas alfa y beta se cumplieron. Se puede concluir que las sílabas alfa permiten distinguir individuos a partir de su morfología, y no así las sílabas beta.

## 6. Conclusiones

En esta tesis se llevó a cabo una caracterización de las sílabas que componen el canto de los horneros y se postularon hipótesis referidas a su capacidad de funcionar como firmas de individualidad. La base de datos utilizada estuvo compuesta por grabaciones de 5 individuos llevadas a cabo por Ana Amador en Chascomús y ECAS entre 2004 y 2005, y por grabaciones de 3 individuos llevadas a cabo en el trabajo de campo de esta tesis, en 2023. Las hipótesis formuladas fueron las siguientes:

**Hipótesis 1.** Cada hembra produce dos tipos de sílabas sistemáticamente en todos sus cantos (alfa y beta, Fig. 2.2), y pueden ser distinguidas *en un mismo individuo* por sus propiedades acústicas<sup>3</sup>.

**Hipótesis 2.** Es posible distinguir hembras a partir de la morfología de sus sílabas alfa, pero no a partir de las sílabas beta.

**Hipótesis 3.** Los machos producen un solo tipo de sílaba que no permite distinguir a los individuos entre sí a partir de su morfología.

A partir de las grabaciones de 8 individuos se realizó el análisis de los datos en el marco de la Hipótesis 1, definiendo los tipos de sílaba que canta cada hembra: las sílabas alfa y beta, en el capítulo 4 de la tesis. Las herramientas utilizadas para diferenciarlas se basaron en los pesos relativos de sus dos primeros componentes armónicos. Se encontró que, mientras que las sílabas alfa enfatizan el primer armónico, la frecuencia de mayor energía en las sílabas beta es la fundamental. Por otro lado, el procedimiento de extracción de las sílabas a partir de los sonogramas requirió dedicarle especial atención a las sílabas alfa, debido a su complejidad y su carácter sistemático dentro de cada hembra individual. Al prestar atención a la evolución de la frecuencia fundamental de las sílabas alfa en cada hembra con el fin de lograr una fiel extracción de los datos se encontró que, siempre que se encuentran pausas dentro de la sílaba, duran entre 10ms y 15ms. El estudio del control motor en estas aves puede motivar retomar los trabajos referidos al modelado de la interacción entre los elementos neurales y anatómicos que subyacen a los duetos de los horneros para la producción de cantos sintéticos (Amador, A., Trevisan, M. A., Mindlin, G. B (2005)), considerando también el efecto del control motor sobre la morfología de las sílabas sintéticas.

---

<sup>3</sup> Las sílabas de transición (ver Fig. 2.2) no se producen sistemáticamente en un mismo nido (y por ende tampoco entre nidos) en base a los criterios morfológicos adoptados en esta clasificación. En consecuencia, se descartó su estudio en este trabajo.

Una vez cuantificadas las diferencias entre las sílabas alfa y beta, fue posible poner a prueba las hipótesis 2 y 3 en el capítulo 5 de la tesis, que motivaron este trabajo en primera instancia. Se utilizó una red neuronal siamesa como herramienta para identificar horneros a partir de los sonogramas de sus sílabas. Por un lado, se observó que no fueron capaces de distinguir las sílabas beta de distintas hembras entre sí, como se esperaba según la hipótesis 2. En cuanto a las sílabas alfa, la capacidad de la red de separarlas en el espacio de salida se considera una evidencia a favor de la hipótesis 3, por la cual sirven como firma de individualidad en los horneros hembra. Todos los tests estadísticos y exploraciones visuales de los datos referidos a las sílabas alfa permiten no rechazar la hipótesis 2.

En cuanto a las sílabas de macho, la comparación de las medias de score de KMeans para distintos K (azul en Fig. 5.15), y la diferencia entre el score real y el de K=8 (azul y verde en Fig. 5.17), mostraron que la calidad de su agrupamiento según el score es muy leve en comparación el modelo nulo. Por sí solo, este resultado habría sido considerado suficiente como para no rechazar la hipótesis 2. Por otro lado, dado que se utilizó K-means como medida de calidad de clustering, en la sección 5.2.2.1 (Visualizaciones) se mostraron los gráficos de los centros de masa de un solo set de predicciones para cada tipo de sílaba. Se pudo observar un grado de apareamiento entre clusters reales y de K-means en las sílabas de macho mayor que el esperado al observar las sílabas de la fig. 2.13 y, principalmente, al observar el score de los machos.

Esto motivó una cuantificación del grado de apareamiento, con el fin de tener una manera distinta de definir la calidad del agrupamiento generado por la red, que considerase la identidad de los clusters reales relativa a los de K-means, independientemente del score en cada caso. Por otro lado, esta búsqueda también se consideró útil con el fin de calibrar lo que uno podría concluir a partir de la observación de los clusters con la medida de calidad utilizada, que es el score. La observación de los heatmaps fue una primera forma de cuantificar la cercanía de los clusters de los machos a los ideales, para la distribución de los puntos generada por la red. Notando que había un grado de correspondencia alto entre los centros de masa r y k de los machos, esto redundó en el cálculo de las similaridades entre ellos, generando la cantidad  $\gamma$ . El resultado (fig. 5.32) fue parcialmente contradictorio con los referidos al score de los machos. Una conclusión conservadora entre ambos es considerar que se encontró un leve grado de individualidad en estas sílabas de machos (Fig. 5.15: no se encontraron diferencias entre las diferencias de medias del score de K=8 con los scores de K=6 a K=11), pero puede motivar futuros trabajos acerca del rol ecológico que puede tener en el cortejo, identificación de parejas o territorialidad.

Las evidencias de firmas de individualidad en las sílabas alfa de horneros hembra presentadas en esta tesis se posan sobre los estudios previos referidos a su aparato neuromuscular, la dinámica de su canto y su comportamiento, para continuar motivando el estudio de la producción vocal y el control neuromuscular en el hornero.

## 7. Anexo

### 7.1 Parámetros de Praat

El algoritmo de extracción de pitch que se implementa en Praat se basa en un método (Boersma 1993) que considera, en ventanas de tiempo sucesivas, valores de frecuencia candidatos mediante los máximos locales de una función de autocorrelación. El problema en las grabaciones de los cantos de hornero es que el pitch cambia demasiado rápido en una dada ventana, pero aumentar la frecuencia de sampleo no es una solución, ya que también se toma más ruido. La mejor forma de remediar esta situación es mediante los parámetros del menú “Advanced pitch settings”, descritos a continuación en base al manual de Praat.

El menor valor de la frecuencia fundamental de las sílabas de los horneros es de alrededor de 1000Hz. Además, al observar el sonograma de cada individuo es posible determinar cuál es el rango de frecuencias de las sílabas que están siendo definidas como alfa y beta. Por ende, si se quiere estudiar la frecuencia fundamental de una sílaba, se puede configurar la visión del sonograma para observar solo ese rango. En el caso de las sílabas alfa, se encuentra entre 1000Hz y 4500Hz, pero el máximo depende del individuo.

Dado que se quiere extraer un seguimiento de la frecuencia en cada sílaba (su pitch), no solo es necesario determinar el rango de frecuencias en la visión del sonograma (por supuesto, se querrá elegir 1000Hz - 4500Hz), sino también en la opción del *pitch range*. El mínimo valor es llamado *pitch floor*. Es muy importante, ya que determina el tamaño de la ventana de tiempo que Praat considera para determinar los valores candidatos de frecuencia. Parafraseando el manual de Praat:

“Si el *pitch floor* es de 75 Hz, el método de análisis de *pitch* requiere una ventana de análisis de 40 milisegundos, es decir, para medir el *F0* en un momento dado, por ejemplo, 0.850 segundos, Praat necesita considerar una parte del sonido entre 0.830s y 0.870s. Estos 40 milisegundos corresponden a 3 períodos de tono máximo ( $3/75 = 0.040$ ). Si se disminuye el *pitch floor* a 25 Hz, la ventana de análisis se ampliará a 120 milisegundos (que nuevamente son 3 períodos de tono máximo), es decir, se considerarán todos los tiempos entre 0.790 y 0.910 segundos. Esto hace que sea más difícil ver cambios rápidos en el *F0*. “

Esto implica que, si se quiere extraer el pitch en el primer armónico en vez de en F0, se tendrá el doble de puntos de pitch en cada tiempo. En grabaciones con ruido esto supone un problema, ilustrado en la fig. 2.8.

Fijar el *pitch range* en la frecuencia fundamental (1000Hz-4500Hz para las sílabas alfa) soluciona algunos problemas. La causa del primero es que el rango de las sílabas es suficientemente grande como para que haya una gran superposición entre distintas partes de la sílaba con el primer armónico. Este problema es el abordado en la fig. 6a. Los *Advanced pitch settings* permiten controlar el seguimiento del pitch en estos casos. Otro problema abordado modificando el *pitch range* es que la frecuencia fundamental de una sílaba se vea muy poco enfatizada en la grabación (fig. 2.7b), en cuyo caso sería necesario extraer su pitch en el primer armónico.

*Time step (s)*: Es el tamaño de la ventana de tiempo. Se utilizó el predeterminado: 0.75/*pitch floor*.

*Silence threshold*: ventanas con amplitudes debajo de este valor (relativo a la amplitud máxima) no son considerados.

El *voicing threshold* es un parámetro que es necesario disminuir en grabaciones ruidosas.

*Octave cost*: este parámetro resultó fundamental. Es el grado en el que se favorece, en una dada ventana de tiempo, al candidato de frecuencia que sea más alta.

*Octave jump cost*: También resultó fundamental. Es el grado en el que se desfavorece un cambio de pitch: desfavorece grandes saltos de frecuencia. Este valor es corregido por el *time step* que, a su vez, depende del *pitch floor*. Dado que los sweeps de las sílabas de los horneros son extremadamente rápidos relativo a la capacidad de Praat de seguir el pitch, fue necesario configurarlo exhaustivamente para capturar los features observados en el sonograma.

*Window length*: permite jugar con el tradeoff entre la resolución temporal y la resolución de la frecuencia en el sonograma. Resultó de suma utilidad en la sección

Para manejar los archivos de audio con facilidad se escribió un script de Praat (`praat_vim.praat`, encontrado en <https://github.com/ttdduu/furnarius>) que se puede llamar desde una terminal con los siguientes argumentos como ejemplo:

```
praat –send /path_to/script.praat /path_to/grabacion.wav 0.54 0.68 900 3350 0.06 0.17  
0.2 0.14 0.04 1 0.54
```

Los dos primeros números son el comienzo y fin del intervalo de tiempo donde están las sílabas, los dos últimos son:

- utilizar o no un filtro pasabandas (1 o 0) que deja pasar el mínimo y máximo de frecuencia seteado: 900Hz-3350Hz en este caso.
- hacer un denoising (desde t=0 hasta t=0.54 en este caso)
- y el resto son los parámetros de “Advanced Pitch Settings” en orden.

De este modo, una vez que se encuentran los parámetros de Advanced pitch settings que permiten una fiel representación del seguimiento de la frecuencia en el tiempo de una dada sílaba o conjunto de sílabas, es posible abrir Praat nuevamente con esos mismos parámetros en el intervalo en el cual son útiles.

## 7.2 Extracción de datos a partir de los sonogramas

Se desarrolla el trabajo realizado con el software Praat para extraer el pitch de la frecuencia fundamental de las sílabas en grabaciones que presentaron distintos obstáculos para la obtención rápida de los datos.

### 7.2.1 Extracción del pitch de las sílabas de la hembra en grabaciones con ruido y en un dueto

En los casos en los que hubiera ruido en F0 (fig. 4.5a) a causa de que el pitch tomado por Praat correspondiera al primer armónico de la misma sílaba, fue necesario limpiar los archivos posteriormente a haberlos extraído con Praat. Es decir, se utilizaron los parámetros que permitieran la mejor extracción de pitch posible, y se post-procesó la sílaba manualmente. Parámetros utilizados en el ejemplo 4.5a: HA280920041353-solo.wav 1.57 2.52 900 3800 0.03 0.17 0.001 0.2 0.1 1 1

En el caso del dueto en la fig. 4.5c, es necesario mirar el canto entero e identificar dónde se está superponiendo la sílaba del macho para determinar qué partes del trazo corresponden a la sílaba de la hembra, y cuáles no.

En el caso de sílabas cuyo trazo solo se observa claramente en F1 (fig. 4.5b), no es plausible únicamente determinar en Praat un mínimo de frecuencia (pitch floor) acorde al primer armónico, dado que esto achica a la mitad el tamaño de cada ventana, con lo cual se obtiene un pitch como el observado en la fig. 4.6, con el doble de datos por unidad de tiempo. En estos casos, es necesario suavizar el pitch obtenido y dividirlo por dos. Cuando fue necesario, se obtuvieron buenos resultados con una media móvil.

## 7.2.2 Extracción manual del pitch

Hubo 18 grabaciones de los individuos HAC1 y HEC1 que fueron demasiado ruidosas y no se encontró la manera de utilizar los parámetros del programa para extraer un pitch fiel a la morfología observada en el sonograma. Sin embargo, el trazo de la sílaba era evidente. En estos casos, con el fin de no perder los datos, fue necesario recurrir a capturar manualmente la sílaba punto por punto, registrando en un archivo de texto los valores de frecuencia observados en el sonograma. Para ello, *no* se utilizó la misma frecuencia de sampleo que la utilizada en las demás grabaciones: 75ms. Dada la duración de una sílaba (entre 0.1s y 0.16s), esto implicaba completar con precisión, aproximadamente, 200 puntos para cada una. Para acelerar el trabajo, se marcaron manualmente únicamente los puntos necesarios (aproximadamente 40), y los demás fueron interpolados y suavizados con una media móvil para obtener cada sílaba según se observaba en el sonograma. En un dado instante se buscó marcar el punto en la mitad del trazo sobre sonograma, ya que podía llegar a ser de 400Hz en modulaciones rápidas de frecuencia (los upsweeps y downsweeps), fijando una ventana temporal de 0.01s.

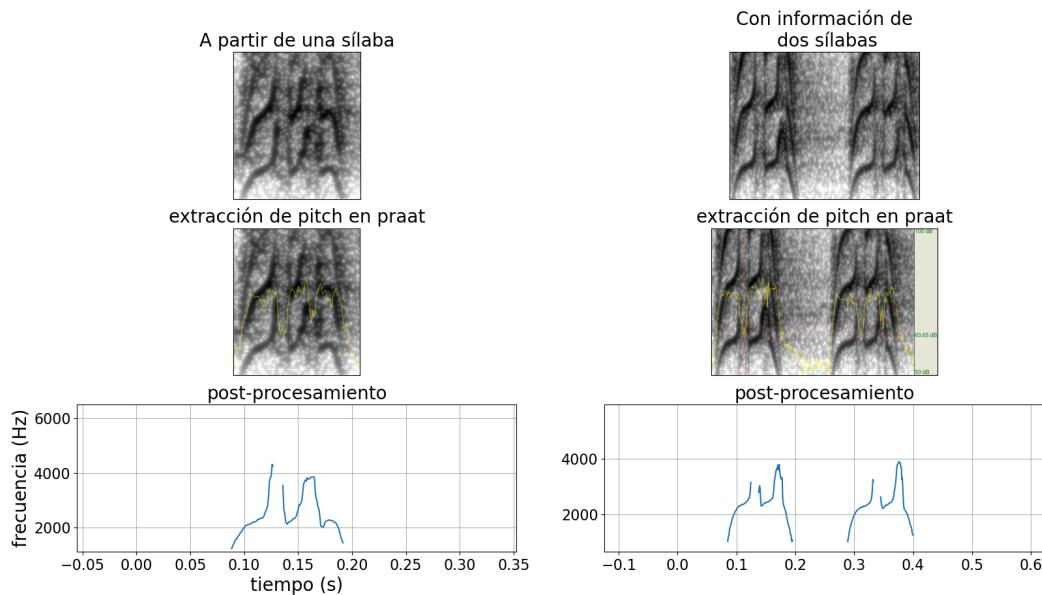
Este método supuso una pérdida del carácter sistemático de la obtención de los datos, ya que no queda registro de los parámetros usados en Praat. De todos modos, incluso en los casos donde se pudo hacer una extracción automática, el criterio en cada sílaba para limpiar los datos provenientes del programa fue propio, observando cada sílaba en el sonograma.

## 7.2.3 Identificación y extracción silencios dentro de una sílaba alfa

Para poder determinar si, dentro de una sílaba, hay o no un intervalo de silencio, fue necesario definir el silencio como un umbral de decibeles específico para cada grabación. Además, un parámetro muy útil en esta situación fue el *window length* del sonograma. Una ventana mayor aumenta la resolución de la frecuencia, es decir, disminuye el ancho de la banda de frecuencia en el sonograma. Por el contrario, disminuir el tamaño de la ventana otorga una mejor resolución temporal, lo cual permite determinar si, en un período de tiempo dentro de una sílaba, hay o no silencio.

Otra herramienta utilizada para definir el silencio en una sílaba fue la posibilidad de superponer sobre el sonograma la curva de intensidad en función del tiempo.

Con esas herramientas, si se está observando una sílaba tratando de determinar si hay o no un intervalo de silencio, se puede proceder según la fig. 7.1.



*Fig. 7.1: Cómo determinar la presencia o no de silencio utilizando información del primer armónico y de la curva de intensidad, estableciendo un umbral de 63.5dB. Parámetros de Praat para la primera columna: H19-031020040851.wav 9.093 9.207 900 4500 0.05 0.025 0.1 0.1 0.05 0 0. Parámetros para la segunda columna: HB121120041436-solo-hembra.wav 1.36 1.52 600 4500 0.06 0.025 0.013 0.06 0.1 0 0*

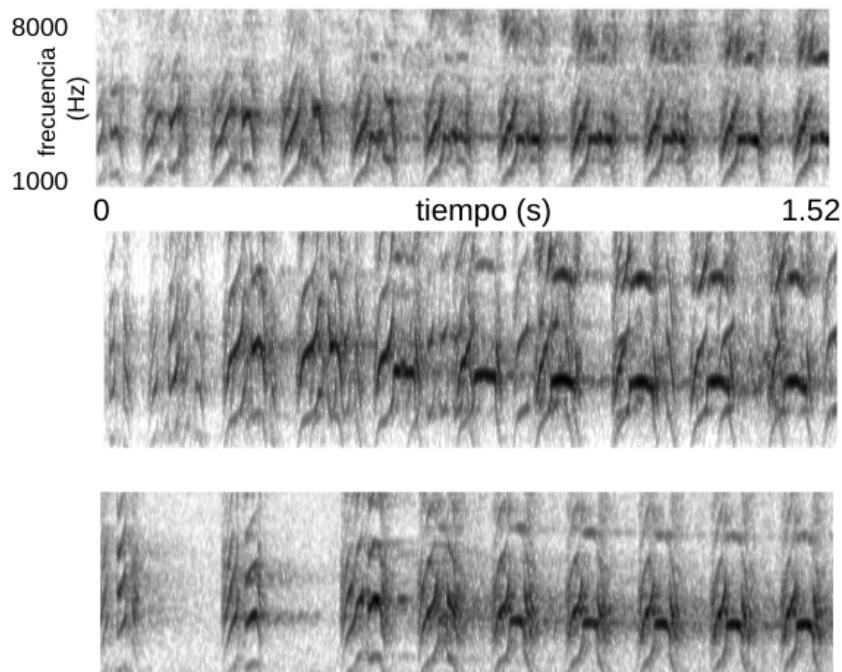
En el caso de H34, si se quiere determinar si hay o no silencio después del segundo upsweep, una opción es valerse del hecho de que en esa misma sílaba hay uno evidente en el primero, a partir del cual se define un umbral de intensidad que permite determinar si también lo hay en el segundo, o no.

En el caso de HB no solo se tiene información de la misma sílaba, sino de la previa, que tiene el mismo nivel de ruido por estar en la misma grabación inmediatamente antes. En ambas sílabas hay un silencio después del primer upsweep, pero no está claro si también sucede en el segundo upsweep de la segunda sílaba. Utilizando el valor de referencia de 63.5 dB establecido para esta grabación, se determina que no hay silencio en ese intervalo.

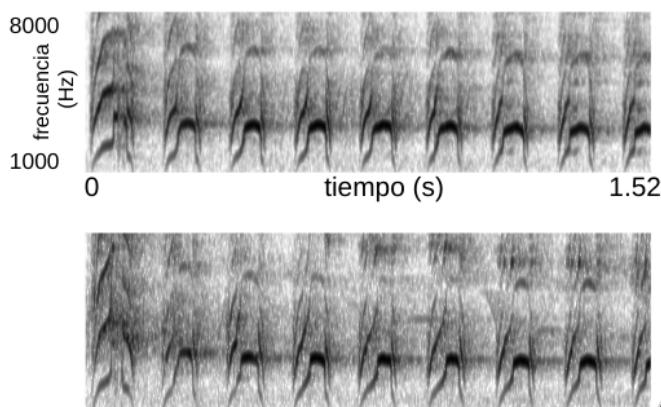
A partir de esa conclusión, se extrae el pitch de la sílaba procurando capturar los features encontrados. Para eso, en varias ocasiones fue necesario iterar exhaustivamente sobre para llegar a combinaciones específicas de los parámetros de *Advanced Pitch Settings* en cada sílaba, dependiendo del nivel de ruido de la grabación.

### 7.2.3 Sonogramas en los que se observan sílabas de macho con morfología de sílaba alfa de hembra

Por último, dado que en este estudio se profundizó por primera vez en la morfología de las sílabas de una especie de aves cuyo canto está poco estudiado, un detalle que es necesario aclarar es que los machos también tienen la capacidad de generar un silencio en sus sílabas. Lo hacen con muy poca frecuencia y no sistemáticamente. Y, si lo hacen, sucede en sus primeras sílabas. Sin embargo, no se profundizó en esta observación.



*Fig 7.2: Sonogramas del macho HEC1 en el cual se observa un silencio mayor a 100ms en sus primeras sílabas.*

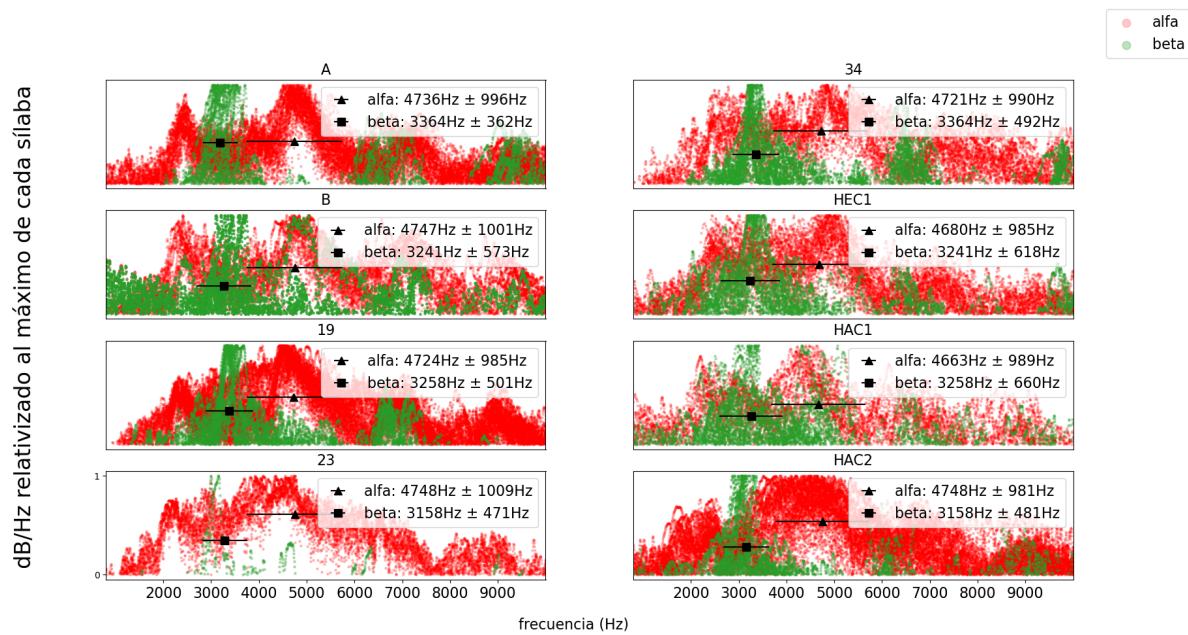


*Fig 7.3: Sonogramas del macho HAC2 en el cual se observa un silencio mayor a 0.01s en sus primeras sílabas.*

## 7.2 4 Subconjunto del total de los arámetros utilizados en las sílabas alfa de nidos A, B, 19, 23, 34

HA061020041409-solo.wav 2.59 3.38 900 3800 0.07 0.1 0.05 0.3 0.2 1 1.61 0 0  
HA090920041347-solo.wav 0.77 0.98 900 3300 0.06 0.17 0.1 0.05 0.06 1 0.8  
HA090920041347-solo.wav 1.64 1.83 900 3300 0.07 0.17 0.1 0.1 0.06 0 0.8  
HA090920041347-solo.wav 1.64 1.83 600 3300 0.07 0.17 0.1 0.1 0.06 1 0  
HA280920041338-solo.wav 1.70 1.81 600 3800 0.1 0.1 0.01 0.1 0.06 1 0  
HA280920041338-solo.wav 2.38 2.975 900 3800 0.08 0.1 0.05 0.27 0.1 1 0  
HA090920041649.wav 1.486 1.84 900 4000 0.09 0.1 0.1 0.2 0.1 1 0  
HA140920041149.wav 1.35 2.66 900 4200 0.05 0.17 0.15 0.1 0.3 1 0  
HA140920041337.wav 2.15 2.83 900 4500 0.06 0.15 0.045 0.07 0.03 0 0  
HA140920041337.wav 1.06 1.20 900 4500 0.05 0.15 0.03 0.07 0.03 0 0  
HA140920041337.wav 1.78 1.88 900 4200 0.05 0.15 0.1 0.05 0.1 1 1  
HA140920041337.wav 1.277 1.42 900 3900 0.06 0.15 0.65 0.05 0.004 1 0  
HA140920041337.wav 1.41 1.53 900 3900 0.08 0.1 0.13 0.06 0.1 1 1  
HA230920041349.wav 1.59 2.71 900 3700 0.05 0.17 0.05 0.2 0.06 1 0  
HA230920041349.wav 1.75 2.71 900 3700 0.1 0.17 0.01 0.07 0.06 1 0  
HA230920041435b.wav 1.14 2.36 900 3500 0.07 0.17 0.01 0.07 0.09 1 0  
HA280920041153.wav 0.54 1.35 900 3600 0.04 0.17 0.01 0.2 0.1 1 0.54  
HA280920041153.wav 1.59 1.87 900 3600 0.04 0.17 0.0001 0.2 0.1 1 0.54  
HA280920041232-solo.wav 0.82 1.85 600 3700 0.03 0.17 0.02 0.2 0.1 1 0  
HA280920041353-solo.wav 1.57 2.52 900 3800 0.03 0.17 0.001 0.2 0.1 1 1  
HA280920041419-solo.wav 2.07 2.20 900 4500 0.05 0.02 0.1 0.09 0.03 0 0  
HA280920041419-solo.wav 1.88 2.0 900 4500 0.05 0.02 0.1 0.08 0.03 0 0  
HA061220041755.wav 1.07 1.46 900 4500 0.05 0.02 0.17 0.04 0.02 0 0  
HA061220041812.wav 1.75 2.10 600 4200 0.03 0.025 0.017 0.05 0.05 1 0  
HB121120041436-solo-hembra.wav 1.36 1.52 600 4500 0.06 0.025 0.013 0.06 0.1 0 0  
HB151120041104.wav 2.36 2.50 900 4500 0.06 0.02 0.22 0.02 0.05 0 0  
HB271220041018-solo-pichon.wav 10.64 10.84 600 4100 0.07 0.025 0.001 0.14 0.2 1 0  
HB271220041018-solo-pichon.wav 10.92 11.10 600 4100 0.06 0.025 0.1 0.4 0.2 1 0  
HB100120051644.wav 45.77 45.89 900 4500 0.06 0.025 0.14 0.11 0.2 0 0  
H19-031020040851.wav 9.093 9.207 900 4500 0.05 0.025 0.1 0.1 0.05 0 0  
H19-031020040851.wav 9.25 9.37 900 4500 0.05 0.025 0.1 0.06 0.05 0 0  
H19-031020040851.wav 9.40 9.52 900 4500 0.034 0.025 0.2 0.21 0.4 0 0  
H19-031020040851.wav 9.57 9.67 900 4500 0.01 0.025 0.1 0.2 0.3 0 0  
H19-031020041016.wav 0.82 0.93 600 4300 0.06 0.01 0.01 0.3 0.1 1 0  
H19-031020041016.wav 1.63 1.76 900 4300 0.06 0.01 0.01 0.23 0.1 1 0  
H23-031020041549.wav 0.54 1.23 900 4500 0.09 0.025 0.1 0.05 0.03 0 0  
H34-031020041147.wav 1.64 1.78 900 4500 0.07 0.03 0.01 0.03 0.1 0 0

## 7.3 Espectros de las sílabas sin procesar



*Fig 7.4: Los espectros de todas las sílabas de hembra utilizadas, sin procesar. En cada gráfico, cada punto es un valor de frecuencia, a su amplitud relativizada, de un espectro del individuo.*

## 7.4 Comparaciones de medias de los scores

### 7.4.1 Estimaciones para las comparaciones múltiples

contrast	estimate	SE	df	<a href="#">lower.CL</a>	<a href="#">upper.CL</a>	t.ratio	p.value
k2 - k8	-1.96E-01	8.95E-03	3.51E+02	-2.22E-01	-1.71E-01	-2.20E+01	9.29E-67
k3 - k8	-2.05E-01	8.95E-03	3.51E+02	-2.31E-01	-1.80E-01	-2.30E+01	8.24E-71
k4 - k8	-1.74E-01	8.95E-03	3.51E+02	-2.00E-01	-1.48E-01	-1.94E+01	1.42E-56
k5 - k8	-1.19E-01	8.95E-03	3.51E+02	-1.45E-01	-9.33E-02	-1.33E+01	6.31E-32
k6 - k8	-6.97E-02	8.95E-03	3.51E+02	-9.55E-02	-4.39E-02	-7.79E+00	9.29E-13
k7 - k8	-3.40E-02	8.95E-03	3.51E+02	-5.98E-02	-8.21E-03	-3.80E+00	2.03E-03
k9 - k8	-3.33E-02	8.95E-03	3.51E+02	-5.91E-02	-7.50E-03	-3.72E+00	2.76E-03
k10 - k8	-5.38E-02	8.95E-03	3.51E+02	-7.97E-02	-2.80E-02	-6.02E+00	5.35E-08
k11 - k8	-6.98E-02	8.95E-03	3.51E+02	-9.56E-02	-4.40E-02	-7.80E+00	8.79E-13

k12 - k8	-9.38E-02	8.95E-03	3.51E+02	-1.20E-01	-6.80E-02	-1.05E+01	1.81E-21
k13 - k8	-1.08E-01	8.95E-03	3.51E+02	-1.34E-01	-8.23E-02	-1.21E+01	2.80E-27
k14 - k8	-1.34E-01	8.95E-03	3.51E+02	-1.59E-01	-1.08E-01	-1.49E+01	2.86E-38

Tabla 6: estimaciones para las diferencias de medias en las sílabas alfa con  $k=8$  con p-valores, correspondiente a la fig. 5.6 (**sílabas de hembra alfa**).

contrast	estimate	SE	df	t.ratio	p.value
k2-k8	0.0984	0.0339	988	2.895	0.04
k3-k8	0.0794	0.0339	988	2.337	0.23
k4-k8	0.0338	0.0339	988	0.995	1
k5-k8	0.0412	0.0339	988	1.214	1
k6-k8	0.0338	0.0339	988	0.994	1
k7-k8	0.0283	0.0339	988	0.833	1
k9-k8	0.0157	0.0339	988	0.462	1
k10-k8	0.0103	0.0339	988	0.303	1
k11-k8	0.0167	0.0339	988	0.493	1
k12-k8	0.004	0.0339	988	0.119	1
k13-k8	-0.012	0.0339	988	-0.36	1
k14-k8	-0.003	0.0339	988	-0.09	1

Tabla 8: IC95% del test de Dunnett para sílabas beta, correspondientes a la fig. 5.11 (**sílabas beta**).

contrast	estimate	SE	df	t.ratio	p.value
k2-k8	-0.08	0.0086	364	-9.33	1.269
k3-k8	-0.052	0.0086	364	-6.063	4.008
k4-k8	-0.043	0.0086	364	-5.071	7.57
k5-k8	-0.024	0.0086	364	-2.872	0.051
k6-k8	-0.001	0.0086	364	-0.201	1
k7-k8	0.0061	0.0086	364	0.7152	1
k9-k8	-0.003	0.0086	364	-0.46	1
k10-k8	-0.01	0.0086	364	-1.239	1
k11-k8	-0.013	0.0086	364	-1.546	1

k12-k8	-0.023	0.0086	364	-2.685	0.09
k13-k8	-0.026	0.0086	364	-3.043	0.03
k14-k8	-0.039	0.0086	364	-4.516	0

Tabla 8 para la fig. 5.15: estimaciones para las diferencias de medias con k=8 con p-valores (sílabas de macho).

#### 7.4.2 Supuestos distribucionales del GLM para comparación de las medias de los scores

El modelo planteado es el siguiente:

$$Y_{ij} = \mu + \alpha_i + B_l + \epsilon_{ij} \quad (5.3)$$

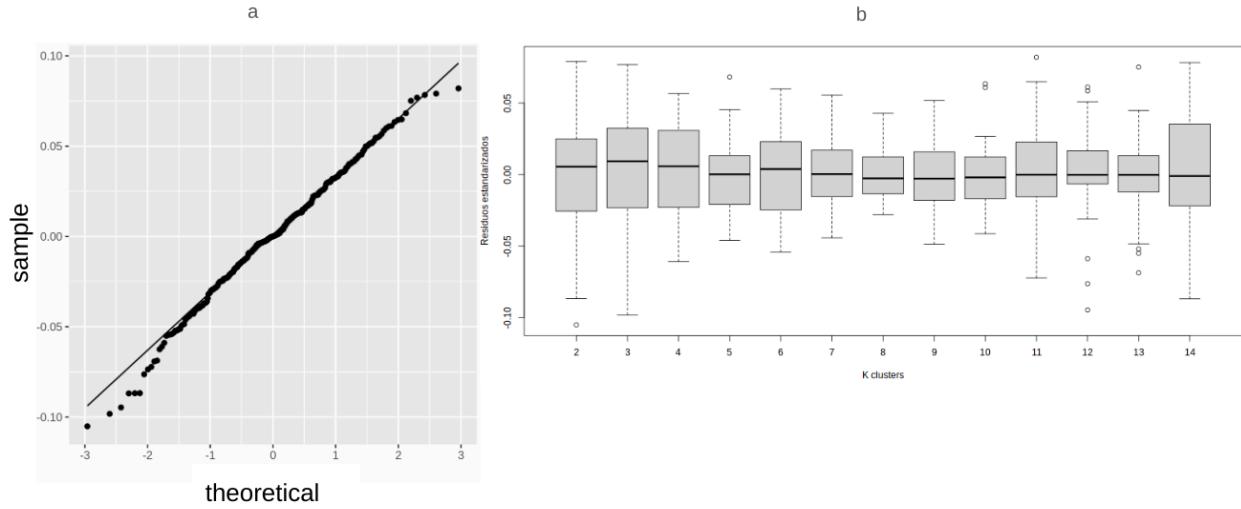
$$\epsilon_{ij} \sim N(0, \sigma_{\text{residual}}^2); B_l \sim N(0, \sigma_{\text{entrenamiento}}^2); i \in \{2, 14\}; j = 1; l \in \{1, 20\}$$

Las muestras son aleatorizadas en el split de entrenamiento-testeo.

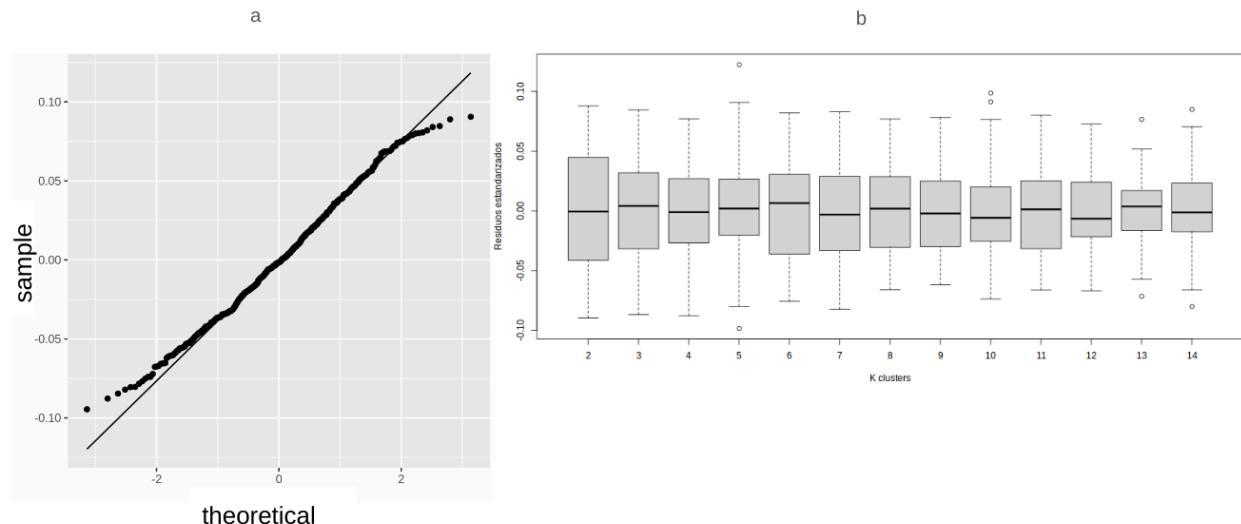
Tipo de sílaba	Normalidad de residuos	Normalidad de Bj (Shapiro-Wilks)	Homocedasticidad (Levene)
hembra alfa	W = 0.9928, p-value = 0.1245	W = 0.9737, p-value = 0.6823	F = 1.7614, p-value = 0.053
hembra beta	W = 0.9943, p-value = 0.028	W = 0.9694, p-value = 0.2761	F = 1.740, p-value = 0.0551
machos	W = 0.9946, p-value = 0.2150	W = 0.9668, p-value = 0.4767	F = 1.4986, p-value = 0.1221

Tabla 7.4.2.1: Se realizaron pruebas de hipótesis para poner a prueba los supuestos de un modelo lineal general mixto. La normalidad de los residuos y de la variable de efectos aleatorios fue puesta a prueba mediante el test de Shapiro-Wilks, y el supuesto de igualdad de varianzas mediante la prueba de Levene.

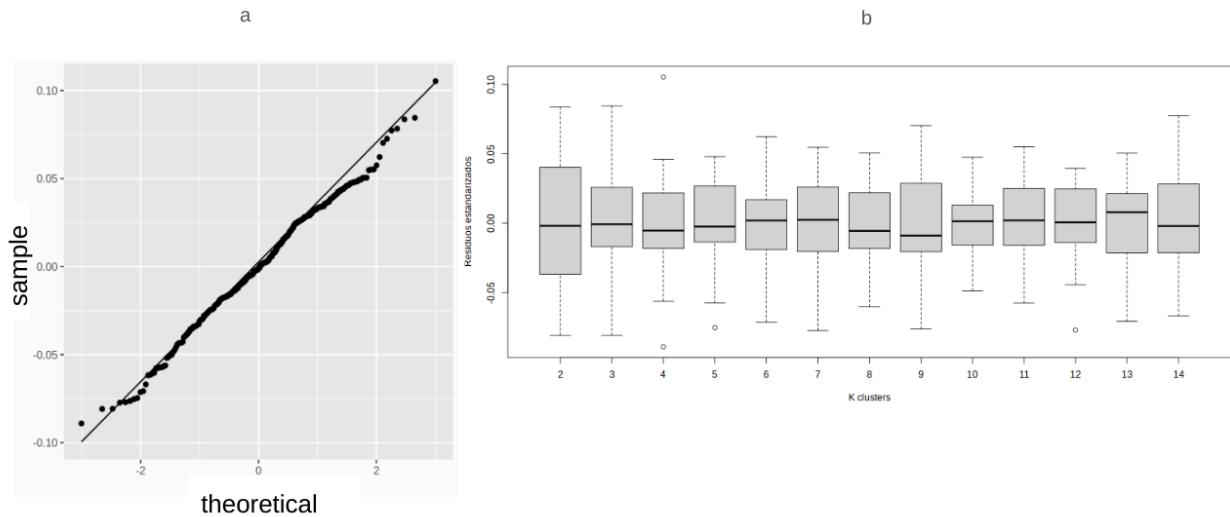
A continuación se presenta el estudio del cumplimiento de los supuestos del modelo mixto mediante herramientas gráficas (figs. 7.3, 7.4, 7.5).



*Fig. 7.3: verificación de supuestos de normalidad (7.3a) y homocedasticidad (7.3b) en los scores de los clusters de KMeans con distintos valores de K en sílabas de hembra alfa. Se observó una menor varianza en los valores de K que llevan a un mayor score.*



*Fig. 7.4: verificación de supuestos de normalidad (7.4a) y homocedasticidad (7.4b) en los scores de los clusters de KMeans con distintos valores de K en sílabas de hembra beta.*



*Fig. 7.5: verificación de supuestos de normalidad (7.5a) y homocedasticidad (7.5b) en los scores de los clusters de KMeans con distintos valores de K en sílabas de macho.*

## 8. Referencias bibliográficas

- Adreani, N. M., Valcu, M., Scientists, C., & Mentesana, L. (2022). *Asymmetric architecture is non-random and repeatable in a bird's nests*. Current Biology, 32(9), R412-R413.
- Amador, A., Trevisan, M. A., Mindlin, G. B (2005): *Simple neural substrate predicts complex rhythmic structure in duetting birds*. Phys. Review E 72, 031905.
- Amador, Mindlin 2023b. *The science of birdsong and the spectrogram, the technique that changed it all*. Molecular Psychology: Brain, Behavior, and Society, 2, 9.
- Ana Amador, Gabriel B. Mindlin (2023): *The dynamics behind diversity in suboscine songs*. Journal of Experimental Biology.
- Bistel Esquivel, Roberto Andrés; Martinez, Alejandro; Mindlin, Bernardo Gabriel (2022a): *An analysis of the persistence of Zonotrichia capensis themes using dynamical systems and machine learning tools*. Chaos, Solitons and Fractals, Vol. 165, Part 1.
- Bistel Esquivel, Roberto Andrés; Martinez, Alejandro; Mindlin, Bernardo Gabriel (2022b);: *Neural networks that locate and identify birds through their songs*. Springer; European Physical Journal: Special Topics; 231; 3; 4-2022; 185-194.
- Boersma, Paul & Weenink, David (2024). *Praat: doing phonetics by computer (Computer program)*. Version 6.4.15, retrieved 26 July 2024 from <http://www.praat.org/>.
- Döppler, J. F., Peltier, M., Amador, A., Goller, F., & Mindlin, G. B. (2021). *Replay of innate vocal patterns during night sleep in suboscines*. Proceedings of the Royal Society B, 288(1953), 20210610.
- François Chollet (2021): *Deep Learning with Python*, 2ed. Manning.
- Fraga, R (1980). *The breeding of rufous horneros (furnarius rufus)*. Condor, 82:58-68.
- Fukuzawa Yukio, Marsland Stephen, Pawley Matthew, Gilman Andrew: *Segmentation of Harmonic Syllables in Noisy Recordings of Bird Vocalisations*.
- G. B. Mindlin, R. Laje: *The Physics of Birdsong*. Springer Biological and medical physics, biomedical engineering.
- Gahr, M. (2000). *Neural song control system of hummingbirds: comparison to swifts, vocal learning (songbirds) and nonlearning (suboscines) passerines, and vocal learning (budgerigars) and nonlearning (dove, owl, gull, quail, chicken) nonpasserines*. Journal of Comparative Neurology, 426(2), 182-196.

Gabriel B. Mindlin (2023): *The dynamics behind diversity in suboscine songs*. Journal of Experimental Biology.

Hertz, John A. *Introduction to the theory of neural computation*.

James. J. Roper (2005): *Sexually distinct songs in the duet of the sexually monomorphic Rufous Hornero*. J. Field. Orniithol. 76(3):234-236, 2005.

K.L Funahashi. *On the Approximate Realization of Continuous Mappings by Neural Networks*, Neural Networks 2 (1989), 183-192.

Kroodsma, D. E., & Konishi, M. (1991). *A suboscine bird (eastern phoebe, Sayornis phoebe) develops normal song without auditory feedback*. Animal Behaviour, 42(3), 477-487.

Kroodsma, D., Hamilton, D., Sánchez, J. E., Byers, B. E., Fandiño-Mariño, H., Stemple, D. W., ... & Powell, G. V. (2013). *Behavioral evidence for song learning in the suboscine bellbirds (Procnias spp.; Cotingidae)*. The Wilson Journal of Ornithology, 125(1), 1-14.

Laje, R., Mindlin, G. B (2003): *Highly Structured Duets in the Song of the South American Hornero*. Physical Review Letters, Vol 91, Number 25.

Liu, W. C., Wada, K., Jarvis, E. D., & Nottebohm, F. (2013). *Rudimentary substrates for vocal learning in a suboscine*. Nature communications, 4(1), 2082.

LeCun, Yoshua Bengio, Geoffrey Hinton (2015): *Deep Learning*. Nature Review,

Massoni, V., Reboreda, J. C., López, G. C., & Aldatz, M. F. (2012). *High coordination and equitable parental effort in the Rufous Hornero*. The Condor, 114(3), 564-570.

Mooney, R. (2009). *Neural mechanisms for learned birdsong*. Learning & memory, 16(11), 655-669.

Nieder, A., & Mooney, R. (2020). *The neurobiology of innate, volitional and learned vocalizations in mammals and birds*. Philosophical Transactions of the Royal Society B, 375(1789), 20190054.

Nuñez, M. A., Chiuffo, M. C., Pauchard, A., & Zenni, R. D. (2021). *Making ecology really global*. Trends in Ecology & Evolution, 36(9), 766-769.

Oliveros, C. H., Field, D. J., Ksepka, D. T., Barker, F. K., Aleixo, A., Andersen, M. J., ... & Faircloth, B. C. (2019). *Earth history and the passerine superradiation*. Proceedings of the National Academy of Sciences, 116(16), 7916-7925.

Pedro Diniz, Edvaldo F. da Silva Júnior, Michael S. Webster and Regina H. Macedo (2018): *Duetting behavior in a Neotropical ovenbird: sexual and seasonal variation and adaptive signaling functions*. Journal of Avian Biology e01637.

Peter J. Rousseeuw (1987): *Silhouettes: A graphical aid to the interpretation and validation of cluster analysis*. Journal of Computational and Applied Mathematics, Vol. 20.

Sarah M. Garcia, Cecilia Kopuchian, Gabriel B. Mindlin, Matthew J. Fuxjager, Pablo L. Tubaro, Franz Goller (2017): *Evolution of Vocal Diversity through Morphological Adaptation without Vocal Learning or Complex Neural Control*. Current Biology 27, 2677–2683.

Saranathan, V., Hamilton, D., Powell, G.V.N., Kroodsma, D.E., and Prum, R.O. (2007). *Genetic evidence supports song learning in the three-wattled bellbird Procnias tricarunculata (Cotingidae)*. Mol. Ecol. 16, 3689–3702.

Scharff, C., & Adam, I. (2013). *Neurogenetics of birdsong*. Current opinion in neurobiology, 23(1), 29-36.

Sumit Chopra Raia Hadsell Yann LeCun (2005): *Learning a Similarity Metric Discriminatively, with Application to Face Verification*.

Suvrit Sra et. al (2005): *Clustering on the Unit Hypersphere using von Mises-Fisher Distributions*. Journal of Machine Learning Research 6 (2005) 1345–1382.

Theuerkauf, J., Villavicencio, C. P., Adreani, N. M., Attisano, A., Craig, A., D'Amelio, P. B., ... & Masello, J. F. (2022). *Austral birds offer insightful complementary models in ecology and evolution*. Trends in Ecology & Evolution, 37(9), 759-767.

Tom M Mitchell et al. *Machine learning*. 1997. In: Burr Ridge, IL: McGraw Hill 45.37 (1997), pp. 870–877.

Tomás Bossi (2022): *Tesis de Licenciatura en Ciencias Biológicas: “Identificación de aves individuales por sus cantos”*. Laboratorio de Sistemas Dinámicos, Departamento de Física, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires.

Touchton, J. M., Seddon, N., & Tobias, J. A. (2014). *Captive rearing experiments confirm song development without learning in a tracheophone suboscine bird*. PLoS One, 9(4), e95746.

Zeigler, H., and Peter Marler. *Neuroscience of birdsong*. Cambridge University Press, 2008.