UNIVERSIDAD AUTÓNOMA DE ZACATECAS
"FRANCISCO GARCÍA SALINAS"
Posgrado del Área de Ingeniería

CORRELACIÓN SOLO DE FASE LIMITADA EN BANDA Y USO DE COEFICIENTES CEPSTRALES INVERSOS: APLICACIÓN EN RECONOCIMIENTO DE VOZ Y BIOACÚSTICA

Tesis de Doctorado

presentada como requisito parcial para obtener el grado de

DOCTOR EN CIENCIAS DE LA INGENIERÍA

ORIENTACIÓN: PROCESAMIENTO DE SEÑALES Y MECATRÓNICA

Presentada por:

Ángel David Pedroza Ramírez

Directores de tesis:

Dr. José Ismael de la Rosa Vargas , Dr. José de Jesús Villa Hernández. y Dr. en C. Rogelio Rosas Valdez

UNIDAD ACADÉMICA DE INGENIERÍA ELÉCTRICA

Zacatecas, Zac., 1 de Septiembre

RESUMEN

El procesamiento digital de señales consiste en la aplicación de distintas operaciones matemáticas a una cierta información bajo análisis. Debido a su estrecha relación con otras ciencias, el procesamiento de señales conforma la base de otras áreas de investigación como el reconocimiento automático de voz y la bioacústica. El procesamiento de señales en este contexto conforma una herramienta de solución para diversas problemáticas. Una de las que más relevancia tiene es el denominado: Cambio climático. En este contexto, las aves juegan un rol fundamental y por tanto la conservación e identificación de las especies de aves es de suma importancia. A pesar de que el reconocimiento de voz provee de soluciones para sistemas concretos en bioacústica, algunas de las técnicas utilizadas fallan en la capacidad de reconocimiento en ambientes naturales.Tomando en cuenta lo anterior, dado que el procesamiento de señales es una herramienta de solución en diversos contextos y tomando en cuenta que algunas especies de aves poseen patrones acústicos, entonces es posible el desarrollo de una nueva metodología en reconocimiento de voz que luego puede ser extrapolada como parte del diseño de un nuevo sistema de reconocimiento automático para la identificación de aves (para algunas especies específicas) e individuos, con una eficiencia de reconocimiento por encima del 70%. En la primera fase de investigación fue propuesto una adaptación de la función BLPOC (correlación solo de fase limitada en banda) para la verificación automática de hablantes con datos limitados. Luego de las pruebas, la función BLPOC mostró ser también un método efectivo para un sistema de verificación de hablantes bajo la condición de datos limitados. Tomando como base estos resultados, en la segunda fase se propuso una nueva técnica para la identificación individual de aves mediante la función BLPOC. De las pruebas de desempeño se puede concluir que este es un método eficiente para la identificación de individuos de especies. En esta segunda fase se propuso un método adicional de clasificación automática de especies de aves basado en la extracción de las características IMFCC (coeficientes cepstrales inversos en la frecuencia mel) de las vocalizaciones. De los resultados obtenidos se concluye que la información acústica de las vocalizaciones de aves en las altas frecuencias (capturadas por los IMFCC's)

es tan significativa como la información acústica en las bajas frecuencias (capturadas por los MFCC´s) para la clasificación de aves a través de vocalizaciones.

ABSTRACT

Digital signal processing is the application of mathematical operations to a piece of certain information. Because of its close relationship with other sciences, signal processing is the base of other science methodologies: automatic speech recognition and bioacoustics. Digital signal processing in this context is a solution tool. One of the most urgent problems is climate change. In this context, birds play a significant role where their identification and conservation are essential tasks. Even though automatic speech recognition provides specific solutions in bioacoustics, some of the traditional techniques fail in the capability of real field recognition. Since digital signal processing is a solution tool in a few contexts and bird species have acoustic patterns, it is possible to develop a new methodology in automatic speech recognition applied for recognition of species and individuals of birds (proposed efficiency over 70%). On the first phase of this research was to propose a new method for the speaker verification under limited data using the BLPOC function. After experiments, the BLPOC function confirmed to be an effective method. Taking these results into account, in the second phase a new technique for the individual identification of birds using the BLPOC function was proposed. The experiments confirmed that the BLPOC function is also an effective method for the individual identification of birds. Finally, in the same phase, another method for the automatic classification of species based on the IMFCC features was offered. Experiments conclude that the acoustic information of vocalizations in the high frequencies (captured by the IMFCC features) is as significant as the information in the low frequencies (captured by the traditional MFCC features).

A Dios principalmente, por acompañarme y guiarme en cada momento de mi vida.

A mi papá, a mi mamá, a mis hermanas, a mis abuelos, a toda la familia, a mis compañeros y maestros.

A todas aquellas personas que siguen ofreciendo su vida al servicio de los demás.

Agradecimientos

Muchas son las personas e instituciones que han participado en este trabajo y a quienes quiero expresar mi gratitud por el apoyo y la confianza que me han prestado de forma desinteresada.

En primer lugar quiero agradecer al Programa de Posgrado en Ciencias de la Ingeniería de la Universidad Autónoma de Zacatecas por el apoyo recibido durante este periodo en el que he desarrollado mi labor investigadora.

Agradezco a mis directores de tesis: El Dr. José Ismael de la Rosa Vargas y el Dr. en C. Rogelio Rosas Valdez, por el apoyo recibido. También el Dr. José de Jesús Villa y al Dr. Daniel Alaniz Lumbreras por su valiosa ayuda y apoyo durante toda la investigación

Un especial agradecimiento a **CONACYT** por su apoyo al concederme la beca de estudios, la cual me facilitó el sustento y movilidad necesario para desarrollar mi labor investigadora.

Gracias.

Contenido General

Pag.

Lista de figuras

Figura Pag.

Figura Pag.

Figura Pag.

Lista de tablas

Capítulo 1

Introducción

Con el fin de analizar las señales provenientes de distintas fuentes, el procesamiento digital de señales consiste en la aplicación de distintas operaciones matemáticas y lógicas a dicha información bajo análisis. Debido a la estrecha relación del procesamiento de señales con otras ciencias, este puede ser aplicado en áreas de investigación tan diversos como el investigador tenga la capacidad de observar su uso como herramienta de solución.

El reconocimiento de voz, como rama de aplicación del procesamiento digital de señales, se basa en el conocimiento sobre el proceso del habla y tiene como fin el de hacer que las computadoras logren un nivel suficiente para expresar y comprender la comunicación humana. Las técnicas utilizadas en reconocimiento de voz utilizan, desde las técnicas de procesamiento de señales de voz clásicas (modelado acústico fonético lineal), hasta las técnicas modernas en donde se proponen combinaciones de métodos estocásticos con métodos conexionistas (modelado acústico fonético no lineal) [7].

Aunque algunos sistemas de reconocimiento de voz se han enfocado en utilizar información auxiliar a la voz con el fin de mejorar el desempeño de los sistemas de reconocimiento (especialmente en condiciones adversas de ruido [8–11]), el aprendizaje automático (enfoque de frontera en el reconocimiento de voz y una rama cada vez más importante de la inteligencia artificial (IA)) ha permitido nuevos desarrollos tecnológicos. Muchos investigadores consideran al aprendizaje automático como la mejor forma de avanzar la IA hacia una equiparable inteligencia humana.

Es por su alta flexibilidad de aplicación y su consistencia matemática que las técnicas de procesamiento de señales han dado origen a otras ramas de aplicación donde el análisis y tratamiento de señales es una necesidad. En este sentido, la metodología y técnicas aplicadas en procesamiento de voz forma parte del fundamento de una rama de investigación de reciente creación: la Bioacústica. La Bioacústica, como aplicación del procesamiento digital de señales, es la ciencia en donde confluyen la biología y la acústica para investigar de forma no invasiva la producción y recepción del sonido biológico, así como los métodos para la transmisión de información por medio de la acústica y su respectiva propagación en el medio. En esta área de estudio se relacionan otras como la biología, taxonomía, sistemática, ecología del paisaje sonoro, computación y electrónica [12–16]. Ella provee de técnicas de bajo costo para el monitoreo de la biodiversidad de áreas remotas [13]. Algunos de los estudios más comunes se enfocan en la identificación de especies tales como murciélagos, insectos, aves y anfibios [17]; así como en la determinación de la actividad biológica del fondo marino [18]. El continuo desarrollo en esta área ha permitido crear estudios más avanzados para así determinar la forma de prevenir la reducción de poblaciones del hábitat silvestre por actividad antropogénica (todo ello sin perturbar el ecosistema). El avance implica por tanto, el monitoreo de población, estimación de abundancia, riqueza y demografía; así como la identificación de especies vulnerables o en riesgo de extinción, estudios taxonómicos, selección sexual y ecología del comportamiento [19].

Como se mencionó en un inicio, el procesamiento de señales conforma una herramienta de solución para diversas problemáticas. Una de las que más relevancia tiene y que de mayor urgencia necesita de una solución es el denominado: Cambio climático. Este problema ha llevado en poco tiempo a la pérdida de habitats y la extinción de distintas especies [20]. En este contexto, las aves juegan un rol fundamental para la estructura de un ecosistema y por tanto de la biodiversidad [1, 21]. Su importancia radica en que su comportamiento (condición de sus habitats) está sensiblemente relacionado con patrones de perturbación, es decir, estas pueden ser indicadores sensibles a cambios mínimos en la naturaleza (razón por la cual la conservación e identificación de las especies de aves es de suma importancia).

Específicamente enfocado en el estudio de la conservación de las aves, México ocupa el décimo lugar a nivel mundial en riqueza de aves. Esto coloca a nuestro pais en el onceavo

lugar entre los países mega diversos de acuerdo a su riqueza avifaunística (entre 1,123 y 1,150 especies) y en el cuarto lugar en cuanto a la proporción de especies endémicas, es decir que, entre 194 y 212 especies de aves solo se encuentran en nuestro país [1] (y por tanto son las que poseen mayor importancia en la conservación). La riqueza de especies endémicas se concentra en las zonas montañosas del eje Neo volcánico, la Sierra Madre Occidental y del Sur, y la Planicie Costera del Pacífico (ver Figura 1.1).

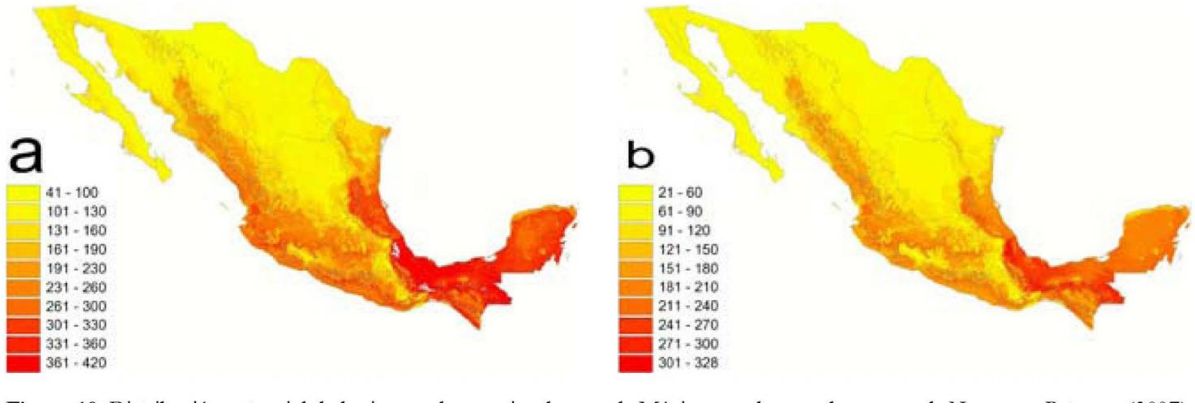


Figura 1.1 Distribución de riqueza potencial de aves en México a) total y b) de especies residentes [1].

En un inicio, y con el fin de determinar la riqueza y abundancia de algunas especies de aves, la identificación se realizaba por medio de inspección anatómica mediante la observación y captura de individuos (contando siempre con el camuflaje adecuado a una distancia considerable con el fin de no perturbar al entorno). Otro método tradicional es la identificación de especies de aves mediante la identificación de cantos. En este contexto, la mayoría de los escuchas entrenados solamente fallan en la identificación de aves poco comunes o raras [22]. Sin embargo, esto conlleva ciertas desventajas como la dispersión del conocimiento, el tiempo de adquisición para cubrir diferentes áreas de los ecosistemas y, el posible sesgo introducido por el escucha al detectar, identificar y registrar las especies de un área.

Tras el nacimiento de la bioacústica como método para la identificación de aves, se identificaron grandes áreas de aplicación entre las cuales se pueden mencionar las siguientes (algunas de ellas aún sin resolver hoy en día de forma eficiente) [13]:

Identificación de especies en campo: En la mayoría de las aves se ha identificado que son vocalmente más activas durante los periodos de reproducción. Esto permite la detección de aves aún a grandes distancias (lo cual representa un método económico para estimar biodiversidad). Ejemplo de la aplicación de la bioacústica en este contexto, fue la identificación del 85% de las especies de aves en el amazonas boliviano mediante un registro de vocalizaciones de apenas 5 días (en contraste con el método de identificación mediante captura con redes en el mismo lugar que requirió 54 días de campaña llevada acabo por 7 ornitólogos).

Identificación de individuos, sexo y mapeo de territorios: Para el caso de la identificación individual es necesario, dentro de la generalidad uniforme de sonidos, la diferenciación entre matices de canto y por tanto, tomando como base esas diferencias, lograr una identificación individual (lo cual posibilita el conteo de animales de una especie por zonas). Para poder diferenciar entre sexos en una población de una misma especie es necesario conocer diferencias específicas en los parámetros de canto (las cuales pueden ser la frecuencia fundamental, presencia de armónicos, entre otros). En relación al mapeo de territorios, una vez que se lleva a cabo la identificación individual, esto consiste en realizar registros de identificación mediante "perchas" o por análisis en los cambios de respuesta (parecidos a los de identificación de especies en campo) en diferentes posiciones geográficas.

Estimación de densidad absoluta y relativa: En algunas condiciones será necesario el conocer la densidad absoluta de la población la cual se calcula mediante la detección de cada uno de los individuos de una especie que estén próximos al observador; así como la distancia a la cual se encuentra de cada uno de ellos. Por otro lado, para calcular el índice relativo de la densidad solo es necesario que una misma proporción de individuos de la población sea registrada. Sin embargo, dado que la detectabilidad de los sujetos (individuos de una especie) no es siempre la misma y varía según sea el sexo, época del año, condición climática, hora del día, tipo de ambiente, entre otros, estas estimaciones resultan tareas no triviales.

Estimación de la tasa de mortalidad: La existencia de los machos a través del tiempo en un área geográfica representa un índice para la estimación de la tasa de mortalidad (cuya identificación tradicionalmente se hacia utilizando elementos artificiales como la captura por medio de redes y la marcación de individuos). En contraste, la bioacústica utiliza una marca

de canto del individuo macho invariante a través del tiempo (es decir, un rasgo vocal invariante a lo largo de toda la vida del sujeto estudiado).

Atracción y ahuyentamiento por playback: La atracción por playback se aplica para la marcación de individuos y en captura selectiva (dado que la grabación es especie-específica). De esta forma, las aves son atraídas por los parlantes hacia sitios seguros. Por otro lado, el ahuyentamiento es utilizado cuando se requiere evitar en cierta área geográfica la presencia de aves (siendo esto el sustituto de técnicas que implicaban barreras o métodos pirotécnicos). Consiste en producir sonidos naturales o de gran intensidad capaces de espantar a las aves y evitar así pérdidas económicas o posibles accidentes en áreas donde su presencia sería perjudicial. Para este último caso, es necesario tomar en cuenta la posible extinción de respuesta asociada al estímulo; por lo que será necesario cambiar de playback con frecuencia. Se recomienda la combinación de varios métodos para asegurar la efectividad del ahuyentamiento.

En este contexto, algunos avances significativos se han logrado gracias a la aplicación de algunas de las técnicas en bioacústica. En los años 90's, por ejemplo, se estudió la composición de aves que migraban en viajes nocturnos (lo cual hubiera resultado casi imposible mediante las técnicas clásicas de observación). Otra investigación fue desarrollada para la identificación del Turdus Migratorius en la cual, a partir de una base de datos acústica de 59 minutos y utilizando Coeficientes Cepstrales para el entrenamiento de una máquina de soporte vectorial, se creó un dispositivo robótico para el monitoreo y estimación poblacional del ave antes mencionada [6]. Esto conlleva, al igual que en reconocimiento de voz, a que exista una gran variabilidad de estructuras y enfoques para llevar acabo el reconocimiento en bioacústica las cuales varian desde los métodos de adquisición hasta los esquemas completos de reconocimiento [23–28] .

Vale la pena destacar que, independientemente de la herramienta que se utilice para llevar a cabo la identificación, los dispositivos básicos en el estudio de bioacústica constan de micrófonos, grabadoras, fuentes de energía, mecanismos de inicio y fin de grabación, y las cubiertas propias para el equipo [29]. Para la obtención de las vocalizaciones se hace uso de micrófonos; los cuales capturan el sonido de las aves en cualquier dirección y posición [28]. Al respecto, algunas de las investigaciones recientes han sido enfocadas en la distribución de los micrófonos, formatos de grabación, entre otros [22]. Hoy en día, los métodos de grabación

más avanzados utilizan unidades de grabación automáticas (ARUs) para grabar información de acuerdo a la programación establecida por el investigador (algunas se pueden activar por medio de dispositivos móviles). Estas unidades requieren de grandes volúmenes de almacenamiento debido a la cantidad de información acústica que en ocasiones se desea capturar. Algunas de las ARUs comerciales poseen cronogramas, conexiones a diferentes micrófonos equipados con sistemas GPS, transmisión de datos por internet, entre otros.

Por su parte, algunos centros de investigación a nivel mundial están desarrollando proyectos de monitoreo bioacústico automatizado (todo ello partiendo de la identificación de especies). Ejemplo de ellos son [29]:

- En Estados Unidos, el " programa de investigación en bioacústica BRP (Bioacoustic Research Program) del Laboratorio de Ornitología de la Universidad Cornell, Nueva York ".
- En Puerto Rico, el " proyecto ARBIMON (Automated Remote Biodiversity Monitoring Network)" .
- En Europa, el " proyecto AMIBIO (Automatic acoustic Monitoring and Inventorying of Biodiversity)".
- El " Laboratorio de Evaluación Ambiental Remota REAL (Remote Environmental Assessment Laboratory) de la Universidad del Estado de Michigan".

Entre los objetivos que finalmente se pretenden alcanzar es el crear aplicaciones portables para teléfonos inteligentes capaces de realizar la identificación de especies en tiempo real mediante la grabación, almacenamiento, identificación y geo-localización de vocalizaciones (esto con el fin de, una vez ampliadas las bases de información disponibles en los centros de investigación, lograr imágenes satelitales de monitoreo continuo para la recolección de información ambiental [25]).

1.1 Planteamiento del problema y Justificación

En nuestro pais, entre 298 y 388 especies (26-33% del total) de la avifauna se encuentra en alguna categoría de amenaza (de acuerdo a autoridades nacionales e internacionales [1,6]). En datos globales, una de cada 10 aves está en categoría de amenaza por lo cual se cree que para los próximos 100 años el 12% de la fauna mundial enfrentará riesgos (por lo cual la extinción de alguna de las especies representa una pérdida global significativa). Por tanto, es necesario el mejoramiento de los sistemas de monitoreo de los ecosistemas alrededor del mundo [17]; los cuales han de ser capaces de responder a los cambios de la naturaleza [22]. Todo esto con la finalidad de lograr una implementación más eficiente de estrategias para la conservación de aves a niveles locales, nacionales e internacionales.

A pesar de que existen avances tecnológicos para la aplicación en la conservación de especies, algunas de ellas resultan muy costosas en espacio y tiempo (como el caso de la telemetría la cual es utilizada para el seguimiento de aves, seguimiento de rutas migratorias, resguardo de anidación, entre otros); contando además con la dependencia del investigador el cual introduce errores por la variabilidad en sus habilidades. En este sentido, para poder identificar aves mediante el método tradicional es necesario educar al oído; lo cual es eficiente para cantos característicos fáciles de aprender pero limitado para aquellos de aves que imitan a otras (sin tomar en cuenta la introducción de dialectos en las vocalizaciones de aves entre distintas zonas geográficas [20]). Aunado a esto, para el caso de nuestro país en donde se creía que la mayoría de las aves ya habían sido identificadas, estudios recientes han demostrado que, lo que antes se consideraba como una especie, era en realidad una unidad evolutiva [1]; lo cual puede llegar a indicar que la diversidad de avifauna es mucho mayor de lo que se pensaba.

Una herramienta que ayude al observador en la identificación de especies mediante análisis de espectrogramas lograría reducir el tiempo de procesamiento de la información (como se sugiere en [22]). Sin embargo, la mayoría de los usuarios a ser beneficiados con los avances en bioacústica no poseen habilidades de programación o de manejo de la matemática para el desarrollo de algoritmos de clasificación automática especie-objetivo [17]; es por ello que los estudios encaminados al monitoreo y conservación deben ser herramientas accesibles, de fácil

manejo y de bajo costo y, de ser posible, adaptable para la identificación de múltiples especies de anfibios, insectos y aves.

Por otro lado, el desarrollo continuo en el área de la bioacústica sigue siendo promisorio "en teoría" [22]. Para el caso específico del reconocimiento de especies de aves, cuya finalidad es entre otras el estudio de la biodiversidad aviar, se enfrenta el problema de que las vocalizaciones (en ocasiones único medio para la identificación) son variadas de un individuo a otro. Por otro lado, para la identificación de individuos los problemas que limitan la eficiencia de las técnicas acústicas se deben a la naturaleza de la fuente de identificación (como el que algunas aves pueden ser escuchadas más fiablemente a distancias mayores de las que se pueden observar, la continua movilidad de los individuos y la propia extracción de características adecuadas que permitan su identificación). Por ello, los algoritmos aplicados en la identificación de especies de aves (independientemente de su aplicación) deben ser capaces de procesar vocalizaciones similares y ser adaptables al reconocimiento (siendo además robustas en las diferentes condiciones de identificación). Es necesario aclarar sobre este punto que es poco frecuente que la identificación se logre sobre todas las especies de fondo en las grabaciones (por el desconocimiento de vocalizaciones o por que no son relevantes para la investigación) [29].

No se debe perder de vista que las técnicas que se utilizan en el reconocimiento del habla son la base para el monitoreo bioacústico [29]. A pesar de que estas proveen de soluciones para sistemas concretos, algunas de las técnicas utilizadas en sistemas de reconocimiento automático (cuyo enfoque es la conversión de voz a texto) fallan en la capacidad de reconocimiento en ambientes naturales (por lo que es necesario hacer uso de estrategias y técnicas avanzadas de reconocimiento adaptadas al reconocimiento en bioacústica). Por otro lado, algunas de las técnicas más avanzadas de reconocimiento de voz (como las redes neuronales) tienen como sustento el uso de grandes volúmenes de datos para el entrenamiento y prueba de los sistemas. Desafortunadamente, la mayoría de las bibliotecas de sonido de aves (o bases de datos), se encuentran limitadas en la identificación de especies e individuos por grabación en la base de datos; dificultando así el uso de los métodos que presuponen información ilimitada y etiquetada para el entrenamiento del sistema.

Algo importante a aclarar es que, aunque existe más de una década de investigaciones en el enfoque de la bioacústica, existen diversos enfoques para llevar a cabo la identificación debido a que la literatura concerniente se encuentra dispersa entre biología e ingeniería. Aunque la eficiencia de reconocimiento es alta para algunas investigaciones, dicha eficiencia no posee valores estándares para investigación en ambientes reales ni para la aplicación industrial; dejando este valor de eficiencia para algunas en porcentajes mayores al 70%.

Por último, es útil remarcar que independientemente de la finalidad para la que se desee orientar la aplicación en bioacústica, es necesario de un conocimiento detallado y específico de la o las especies sobre las cuales se desea emplear y sobre las técnicas de reconocimiento de voz (y procesamiento de señales) existentes para llevar a cabo la identificación. Aunque se desconoce la cantidad de investigadores dedicados al monitoreo y conservación de la biodiversidad [6], en México no existen aún registros abundantes sobre sistemas que identifiquen especies de aves de manera automatizada o que identifiquen individuos de los mismos; creando un nicho de oportunidad de nuevas aplicaciones útiles para esta rama de la investigación.

Tomando en cuenta lo anterior, en este documento se propone el uso del procesamiento de señales como una herramienta de solución para atacar algunas problemáticas. Específicamente se abordan los siguientes temas:

Algoritmo de verificación del hablante basado en la función de correlación solo de fase limitada en banda(BLPOC) La identificación automática del locutor consiste en la identificación de hablantes mediante la clasificación de señales por medio de maquinas [47–50]. Por otro lado, si la tarea consiste en confirmar si un/a hablante es quien dice ser se trata de un sistema de verificación. Estos sistemas requieren de la extracción de información relevante de las señales de habla y de un esquema de entrenamiento el cual puede estar basado en modelos estadísticos [14, 52]. Este modelado requiere de grandes volúmenes de información acústica (grandes bases de datos) con el fin de calcular algunos parámetros de distribución. Sin embargo, para datos limitados estos sistemas proporcionan bajos desempeños [53]. En este sentido, en este documento se propone usar una técnica biométrica de empate de imágenes (tradicionalmente aplicada en la identificación humana) llamada **función de correlación solo de fase limitada en banda(BLPOC)** con la finalidad de validar su uso como una nueva técnica

de verificación de hablantes para datos limitados.

Algoritmo de clasificación automática de especies de aves basado en Coeficientes Cepstrales Inversos en la frecuencia Mel El reconocimiento de voz es la base de las técnicas aplicadas en bioacústica sin embargo, para el caso específico de las aves, aún no es bien entendido cuales características en las vocalizaciones deben ser extraídas para obtener tasas de clasificación semejantes a las obtenidas en reconocimiento de voz [73]. Una de las características tradicionalmente extraídas tanto en reconocimiento de voz como en bioacústica son los Coeficientes Cepstrales en la Frecuencia Mel (MFCC) [45, 74]. A pesar de que esta técnica ha mostrado buenos resultados en ambas disciplinas, algunos estudios recientes en reconocimiento de voz utilizan una variante de los MFCC conocida como **Coeficientes Cepstrales Inversos en la Frecuencia Mel (IMFCC)** con el fin de extraer los componentes de alta frecuencia que pierden los MFCC's [14, 15, 51, 52]. En este contexto, y dada la cercana relación entre la bioacústica y el reconocimiento automático de voz, se propone un sistema de clasificación basada en la extracción de componentes IMFCC de las vocalizaciones de aves con el fin de validar esta técnica para su uso en bioacústica.

Algoritmo de identificación acústica individual en aves basado en la función de correlación solo de fase limitada en banda Para medir la densidad de aves en un area natural, la cual está basada en la identificación de individuos, tradicionalmente pueden ser utilizadas diferentes técnicas artificiales [76–78]. A pesar de que estas técnicas han sido ampliamente utilizadas, la no estandarización de los métodos, la propia naturaleza de las observaciones y la limitada disponibilidad de observadores expertos reducen su efectividad [78]. La identificación acústica individual en las aves consiste en la extracción de características de las vocalizaciones (cantos y llamados) de un individuo que permitan su identificación entre otras vocalizaciones de individuos de la misma especies (o incluso de otras especies) [82]. En este sentido, algunas de las técnicas estadísticas para el reconocimiento automático de hablantes comúnmente son aplicadas para la identificación individual en bioacústica [83, 84]. Aunque estas técnicas poseen un buen desempeño, algunas de ellas pueden proveer un desempeño inadecuado en condiciones de datos limitados (lo cual es una condición común en la identificación de individuos). En este

contexto, se propone la aplicación de la función BLPOC (aplicada en este mismo documento como una técnica para verificación de hablantes) como una nueva técnica para la identificación individual de aves. Dicho algoritmo es ofrecido como un método de identificación acústica individual en aves capaz de lograr la identificación usando datos limitados.
Para llevar a cabo la investigación reportada en este documento se proponen las siguientes fases:

1. La aplicación de una nueva metodología en procesamiento de voz como eje fundamental de la investigación.

2. La extrapolación de dicha técnica a la bioacústica teniendo como fundamento su desempeño en procesamiento de voz.

3. La aplicación de una técnica tradicional en reconocimiento de voz que logre aderezar el sistema integral aplicado en bioacústica.

Cabe destacar que estas fases siguen el orden lógico que dio nacimiento a la bioacústica como ciencia, es decir, una técnica eficiente desarrollada para aplicaciones en reconocimiento de voz es luego aplicada de manera eficiente en bioacústica. De esta manera se tiene el fundamento para su inclusión total como una nueva técnica en ambas áreas de aplicación del procesamiento de señales.

1.2 Hipótesis

Dado que el procesamiento de señales es una herramienta de solución en diversos contextos y tomando en cuenta que algunas especies de aves poseen patrones acústicos, entonces es posible el desarrollo de una nueva metodología en reconocimiento de voz que luego puede ser extrapolada como parte del diseño de un nuevo sistema de reconocimiento automático para la identificación de aves (para algunas especies específicas) e individuos, con una eficiencia de reconocimiento por encima del 70%.

1.3 Objetivos

Objetivo General: Diseño de un nuevo sistema de reconocimiento automático para voz y bioacústica basado en la correlación de fase limitada en banda y el uso de coeficientes cepstrales inversos.

Objetivos Específicos:

- Recopilación y análisis de esquemas recientes para la creación de sistemas eficientes en reconocimiento de voz y bioacústica.
- Diseño de una nueva metodología para la aplicación de la correlación de fase limitada en banda en verificación del hablante.
- Adecuación de la nueva metodología desarrollada para su aplicación en un sistema eficiente en bioacústica enfocada en el reconocimiento de especímenes de aves.
- Aplicación de los coeficientes cepstrales inversos para el desarrollo de un nuevo sistema para reconocimiento de especies de aves.

1.4 Metas

Meta general: Desarrollar un algoritmo para la verificación de hablantes y otro basado en este para el reconocimiento de especies y de individuos de aves mediante la implementación de la función de correlación solo de fase limitada en banda(BLPOC) y los Coeficientes Cepstrales Inversos en la frecuencia Mel (IMFCC)
Metas Específicas:

- Recopilación de esquemas eficientes en la literatura para reconocimiento de voz y bioacústica.
- Recopilación de esquemas para implementación de la función BLPOC en procesamiento de imágenes.

- Adecuación y aplicación de la función BLPOC para verificación de individuos de hablantes.
- Recopilación de corpus de voces.
- Entrenamiento y prueba del sistema.
- Cálculo de desempeño del sistema.
- Diseño de un nuevo esquema para reconocimiento de individuos de especies de aves basado en las pruebas anteriores de la función BLPOC.
- Recopilación de Corpus de especies-específicas de aves.
- Entrenamiento y prueba del sistema.
- Cálculo de desempeño del sistema.
- Diseño de un nuevo esquema para la identificación de especies de aves basados en coeficientes cepstrales inversos.
- Entrenamiento y prueba del sistema.
- Cálculo de desempeño del sistema.

1.5 Estructura de la tesis

En el capítulo 2 se introducen los antecedentes sobre las metodologías clásicas de reconocimiento de voz para el diseño de sistemas de reconocimiento; también se dan las bases para la identificación de especies de aves desde el contexto biológico estableciendo también algunos de los fundamentos para la identificación acústica. Luego, en el capítulo 3 se presenta la teoría concerniente al primer método propuesto sobre la aplicación de la función de correlación solo de fase limitada en banda para verificación de hablantes. El capítulo 4 presentan los resultados y conclusiones de la aplicación de esta metodología. Después, en el capítulo 5 se presenta la metodología de la función BLPOC para la identificación de individuos de especies de aves y la respectiva para la identificación de especies de aves utilizando los coeficientes cepstrales

inversos. El capítulo 6 presenta las conclusiones de estas implementaciones. Finalmente, en la sección de apéndices se presentan los artículos productos de estas investigaciones.

Capítulo 2

Antecedentes

Una de las finalidades del análisis de señales es la aplicación del conocimiento en diferentes contextos donde el procesamiento de señales sea una necesidad. Con el fin de determinar el estado de un ecosistema se ha buscado diagnosticar a las especies en el hábitat así como sus relaciones macro sistemáticas y densidades de población relativa. Esto conlleva al desarrollo de herramientas tecnológicas que permitan la identificación de las especies con el fin de crear sistemas complejos de monitoreo lo más automatizados posibles [20, 21, 30]. Sin embargo, no es posible la creación de estos sistemas de monitoreo sin tener como base la ingeniería (es decir, las técnicas de análisis de señales y reconocimiento de voz) y, de igual manera, no es posible crear aplicaciones enfocadas en la conservación de las especies sin tener como base las mismas relativas a su estudio (es decir, el contexto de la biología). A continuación se describen estos contextos con el fin de introducir al lector en el área específica bajo estudio.

2.1 Técnicas de análisis de señales y reconocimiento de voz

El reconocimiento de voz se basa en el conocimiento sobre el proceso del habla y la estructura que conlleva el lenguaje en el ser humano [31]. El fin último del reconocimiento de voz es hacer que las computadoras logren un nivel suficiente para expresar y comprender la comunicación como lo hace en su naturaleza el ser humano (reconocimiento del habla espontanea [32]).

Los sistemas de reconocimiento de voz buscan poseer algunas características importantes tales como robustez (buen desempeño en diversos entornos), flexibilidad (extendible a diferentes vocabularios), de gran vocabulario (un léxico superior a 1,000 palabras), entre otras. Los sistemas también se pueden clasificar como dependientes e independientes del vocabulario. Para el caso de los sistemas dependientes (hablante cooperativo), el mensaje que se identifica (y por tanto articula el locutor), son palabras o mensajes fijos (pre-establecidos); mientras que para el caso de los sistemas independientes (hablante no cooperativo), se posee la libertad para articular cualquier mensaje para ser reconocido. Debe aclararse que, para este último caso, la biblioteca es mucho más extensa en relación al vocabulario que el locutor posee por lo que se puede decir que tiene la libertad para pronunciar cualquier mensaje. La eficiencia de los sistemas dependientes es mayor que la de los sistemas independientes.

El estudio de reconocimiento de voz se basa en tres principios:

1. La información de la señal de voz se puede representar por el espectro en amplitud a corto plazo de la forma de onda de la voz. Esta es la base para la construcción de patrones.

2. El contenido de la voz se puede expresar en forma escrita (una secuencia de símbolos fonéticos o caracteres del alfabeto).

3. El habla es un proceso cognoscitivo, es decir, la comprensión está ligada a la gramática, semántica y estructura del lenguaje.

La señal de voz puede considerarse como cuasi-estacionaria (que varía lentamente) si se analiza en periodos de tiempo relativamente cortos (entre 5 y 100 ms), mientras que para periodos más largos (mayores a 0.2 s) se dice que la señal es no estacionaria. En este contexto, el analizar la señal de voz en periodos cortos (mediante ventanas de análisis) se utiliza además para preservar la resolución en frecuencia y reducir los efectos de las variaciones temporales.

Los sistemas de reconocimiento de voz utilizados en la actualidad pueden variar en estructuras tan diversas como el diseñador lo desee [33], sin embargo, de entre la variabilidad existente, algunos son los procesos fundamentales que deben de poseer para poder ser considerados como sistemas de reconocimiento (ver Figura 2.1).

2.1.1 Adquisición y preprocesamiento

Este proceso consiste, entre otros, en la recopilación de información por medio de la grabación de sonido y la aplicación de filtros capaces de suprimir el ruido externo posiblemente captado en las grabaciones.

2.1.2 Extracción de características

Una vez que la señal de voz ha sido adecuada, la etapa siguiente corresponde a la extracción de la información contenida en la señal de voz. Esto se realiza con el fin de obtener una secuencia de vectores o parámetros que describan la información de interés contenida en la señal [34]. Es necesario aclarar que entre más parámetros posean estos vectores, más difícil se vuelve su procesamiento en los sistemas.

En este sentido, el análisis de la señal acústica se puede realizar por medio de dos métodos clásicos:

1. Una representación espectral: Esta representación se enfoca en conocer el contenido de frecuencias y la distribución de la energía en el espectro.

2. Representación puramente temporal: Se fundamenta en el análisis de la forma de onda de la señal.

Una vez analizada la señal acústica, el proceso de parametrización o caracterización consiste en obtener las características propias o distintivas de la señal de voz para llevar a cabo el reconocimiento. Sin embargo, el método utilizado para llevar a cabo este proceso depende de la aplicación en que se busca aplicar el reconocimiento.

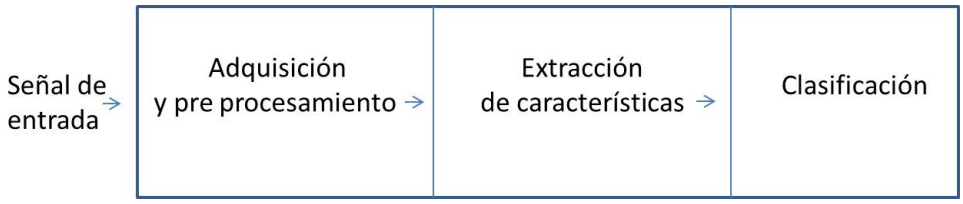


Figura 2.1 Esquema general de reconocimiento.

A continuación se describen brevemente algunos de los métodos principales de extracción de características utilizados en el reconocimiento de voz.

Bancos de Filtros: Es uno de los modelos clásicos perteneciente a los métodos de análisis espectral. En el, la señal de voz es filtrada a través de bancos de filtros pasa-banda agrupados en rangos de frecuencia con la finalidad de medir y analizar la energía de la señal de voz en esas frecuencias de interés. Para cumplir el propósito del banco de filtros, cada señal se pasa a través de una no-linealidad (como si fuera un rectificador de media u onda completa). Dicha no-linealidad introduce un corrimiento en la señal pasa banda a pasa bajas lo cual crea imágenes de la señal en altas frecuencias. Finalmente se aplica un filtro pasa bajas para eliminar dichas imágenes y poder obtener finalmente una estimación de la energía en cada una de las bandas.

Entre las ventajas que posee el utilizar bancos de filtros se encuentra que con pocos parámetros se puede representar la envolvente espectral; aunado al poder tener resoluciones de frecuencias diferentes para cada filtro (esta ultima característica es utilizada en simulación de procesos auditivos). Otra alternativa es usar bancos de filtros no uniformes los cuales se diseñan de acuerdo a algún criterio de espaciamiento.

Codificación Lineal Predictiva (LPC): El análisis por predicción lineal o modelo LPC ha brindado buenos resultados en reconocimiento y procesamiento de voz desde hace mucho tiempo. Su principal característica es que con esta técnica se modela el sistema traqueal que produce la voz humana. El análisis LPC considera que la señal que se produce no ocurre de forma aleatoria por lo que existe una relación entre cada una de las muestras de voz.

Las principales razones por las que se utiliza el análisis LPC sobre otros métodos son:

1. Permite distinguir una separación entre la fuente (señal de excitación) y el tracto vocal (función de transferencia del tracto vocal).

2. Es analíticamente manipulable para su implementación sencilla en software y hardware.

3. Puede ser ligada al análisis cepstral.

Análisis Cepstral: La señal de voz se puede analizar por medio de características espectrales por dos razones fundamentales [35]:

- La señal es cuasi estacionaria (en periodos de tiempo corto) y puede ser aproximada por medio de la sumatoria de ondas senoidales.

- La información en la percepción de la voz por medio del oído humano incluye información espectral.

La envolvente espectral cambia lentamente en función de la frecuencia y refleja no solo las características resonantes (o antiresonantes), sino también algunas otras características como la radiación del sonido en labios y nariz. Dadas estas características, una técnica de análisis comúnmente utilizada es el análisis cepstral la cual es una buena técnica de deconvolución (separación entre la función de excitación y la respuesta del tracto vocal). El cepstrum esta definido como la transformada inversa de la amplitud del espectro logarítmico en corto tiempo.

La ventaja principal del uso de los coeficientes cepstrales es que aprovechan los beneficios del análisis LPC pero eliminando ruido y perturbación propiamente contenida en las tramas de información.

Coeficientes Cepstrales en la frecuencia Mel (MFCC): Uno de los métodos para la extracción de características más utilizado comercialmente son los Coeficientes Cepstrales en la Frecuencia o escala de Mel (MFCC, por sus siglas en inglés) [35]. El objetivo de esta técnica es obtener características de la señal basadas en criterios perceptuales (es decir, en la percepción del sistema auditivo en el ser humano).

Para la obtención de los coeficientes se utiliza la transformada de Fourier (para la representación espectral de la señal) y el diseño de banco de filtros (para seleccionar las bandas de frecuencia de la señal en análisis). Luego, la idea básica es calcular la energía que aporta a cada banda de frecuencia la señal que se analiza y calcular sus correspondientes en términos de una secuencia de coeficientes (coeficientes cepstrales).

2.1.3 Clasificación

En esta etapa se lleva a cabo propiamente la identificación de las palabras (o secuencias). Este proceso requiere de dos fases: entrenamiento y prueba. En el entrenamiento, se crean lo parámetros de referencia mediante la extracción de las características inherentes a los sonidos

a identificar. En la prueba, una vez creados dichos parámetros, se procede a la identificación de otros sonidos de la misma clase a la que pertenece la referencia mediante la comparación de características extraídas en estos otros sonidos con los de referencia.

Existen cuatro enfoques principales que se pueden utilizar para realizar el reconocimiento del habla [36]:

1. Contraste de patrones:
    - Supone al habla como una secuencia de palabras, cada una con un patrón o conjunto de patrones.
    - En este, se compara la unidad de habla entrante con los patrones de referencia almacenados.
    - Es comúnmente utilizado en reconocimiento de palabras aisladas o conectadas.
    - Al no poderse generalizar los patrones, es necesario crear una biblioteca para cada hablante (lo cual reduce su nivel de aplicación en tareas de reconocimiento avanzadas).
2. Sistemas basados en el conocimiento:
    - Emula y aplica los conocimientos sobre el habla en tareas de reconocimiento que utiliza el ser humano.
    - Usa técnicas con reglas y sistemas expertos, desde el nivel acústico-fonético hasta niveles más complejos.
3. Modelos estocásticos:
    - Hace uso de modelos estocásticos en vez de modelos deterministas.
    - Utilizado comúnmente para el reconocimiento de palabras continuas.
4. Modelos neuronales o conexionistas:
    - No posee tantas restricciones como los modelos estocásticos.

- Sus tiempos de entrenamiento son razonables.
- Se han utilizado también como parte de fusiones entre varios enfoques.

A continuación se describe brevemente, según los enfoques anteriores, algunas de las técnicas más utilizadas en reconocimiento de voz. Es necesario aclarar que el uso de cada una o combinación entre ellas dependerá de cual es el objetivo que el reconocimiento pretende alcanzar.

Entrenamiento y prueba

Cuantificación Vectorial (VQ): El utilizar cualquiera de las técnicas de extracción de características da como resultado vectores característicos que representan las características espectrales variantes en el tiempo de la señal de voz. Denotaremos a los vectores espectrales como X_i, para $i = 1, 2, 3, \ldots, L$, en donde típicamente cada vector es de una dimensión *p* (o al menos *p*):

$$\begin{aligned} X_1 &= A_0^1, A_1^1, A_2^1, \ldots, A_p^1, \\ X_2 &= A_0^2, A_1^2, A_2^2, \ldots, A_p^2, \\ \vdots \quad & \quad \vdots \quad \quad \vdots, \\ X_L &= A_0^L, A_1^L, A_2^L, \ldots, A_p^L. \end{aligned} \tag{2.1}$$

La cuantificación vectorial es la generalización de la cuantificación escalar y en ella la cantidad total de información es tratada y almacenada en vectores, los cuales representarán al total de información [37]. Dicha técnica fue ampliamente aplicada y estudiada en los años 80´s especialmente en la rama de la comunicación. Andrés Buzo es uno de los mexicanos que ha logrado mayor popularidad en esta área de estudio. Esta técnica brinda una representación parcialmente eficiente de la información espectral de la señal de voz; gracias a esto se puede llevar a cabo una codificación de la señal comúnmente aplicada en codificación fuente para tareas de telecomunicaciones y reconocimiento. La codificación fuente tiene como fin el lograr la mínima distorsión para una tasa de bits de transmisión predeterminada. En cuanto a las tareas

de reconocimiento, ha sido aplicada en sistemas independientes y dependientes del vocabulario con tasas de error bajas.

Entre las ventajas de la representación por cuantificación vectorial se encuentran la reducción de almacenamiento para el análisis de la información, el tiempo de cálculo para determinar los vectores de análisis espectral y, el que los cálculos de similitud son reducidos a solo una tabla de estimación. Por otro lado, una de las desventajas es que se crea una distorsión espectral inherente en la representación del análisis de cada vector. En este sentido, para poder saber la eficiencia de un cuantificador, es necesario conocer la distorsión promedio producida por dicho cuantificador (ver ecuación (2.2)). Se habla de un cuantificador óptimo cuando la distorsión promedio es minima al cuantificar las secuencias de vectores.

$$D = \frac{1}{L} \sum_{l=1}^{L} d(X_l, q(x - l)), \tag{2.2}$$

en donde *D* es la distorsión promedio y $d(\cdot)$ es una medida de distorsión o distancia espectral.

Modelos Ocultos de Markov (HMM): La aproximación de patrones no modela estadísticamente a las señales de voz por lo que posee ciertas restricciones. Por otro lado, las técnicas estadísticas han sido aplicadas en el agrupamiento para crear patrones de referencia. Mediante ello se mejora la clasificación de patrones y se realiza una simplificación dado que se utilizan multiples señales de referencia y se caracterizan mejor las variaciones de diferentes pronunciaciones.

Una de las herramientas que permiten caracterizar estadísticamente las propiedades de un patrón o una señal de voz son los Modelos Ocultos de Markov (MOM ó HMM, por sus siglas en inglés) [38] [39]. Los HMM son considerados una extensión de las Cadenas de Markov. Los HMM fueron desarrollados por primera vez por un grupo de IBM en 1970 y para 1976 se incorporaron y desarrollaron investigaciones con redes neuronales, DTW y HMM para incorporar información semántica y de sintaxis con el fin de lograr, más allá del reconocimiento del habla, la comprensión del lenguaje. Los HMM tuvieron gran impacto hasta mediados de 1980 donde casi cada investigación de reconocimiento se hacia mediante esta técnica [40].

El fundamento de esta técnica consiste en considerar que la señal de voz es un proceso estocástico paramétrico y que dichos parámetros pueden ser estimados y definidos de manera

precisa (ver Figura 2.2). De manera específica, una cadena de Markov queda definida como un conjunto de variables aleatorias que convergen en probabilidad (si conocida su distribución en el instante t, ésta es la misma para el instante t+1, y hasta el infinito).

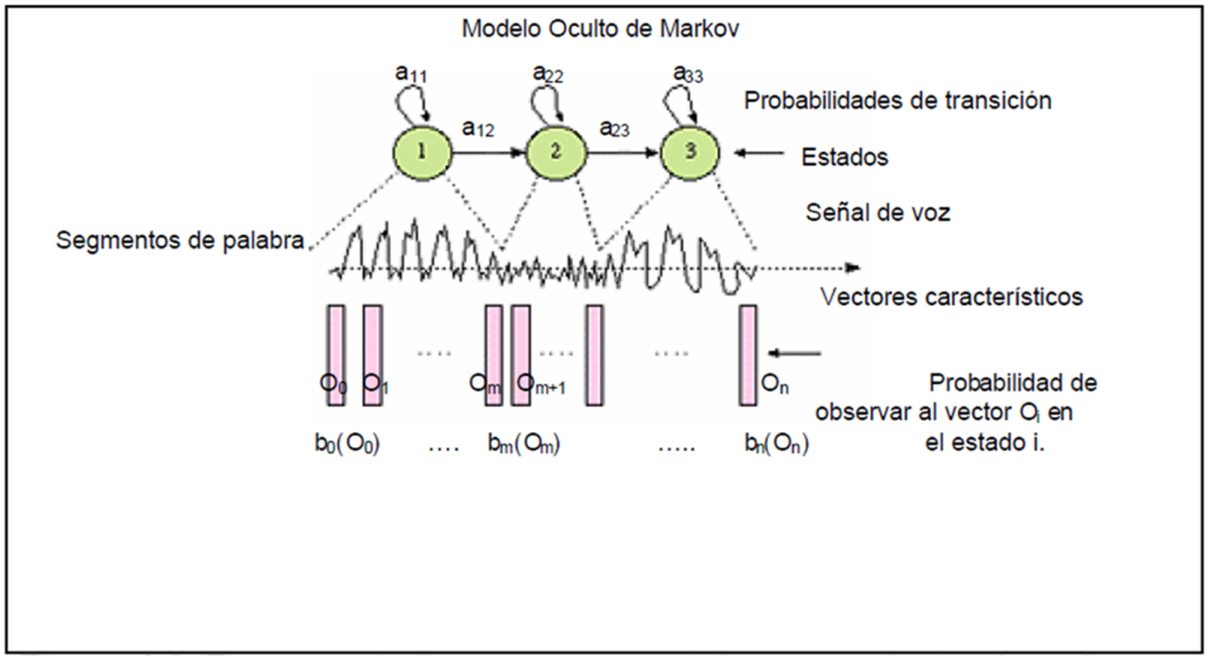


Figura 2.2 Extracción de características para conformar un HMM [2].

Un modelo oculto de Markov, por otro lado, está caracterizado por [41]:

- El **número de estados en un modelo** (N). Los estados individuales se denotan con la letra q en el tiempo t.
- El **número de símbolos distintos de observación en cada estado** (M). Al conjunto de observaciones se le denota como $O = o_1, o_2, o_3, \ldots, o_M$, mientras que a los símbolos individuales se les denota como $V = v_1, v_2, v_3, \ldots, v_M$.
- La **matriz de probabilidad de transición de estados** ($A = [a_{ij}]$). La suma de los elementos de fila de la matriz es la unidad, es decir, misma probabilidad de pasar de un

estado a otro. Expresada como:

$$a_{ij} = P[q_{t+1} = j | q_t = i], \quad 1 \leq i \leq N, \quad 1 \leq j \leq N. \tag{2.3}$$

- La **distribución de probabilidad de los símbolos en el estado** ($B = [b_j(k)]$), es decir:

$$b_j(k) = P[O_t = V_k | q_t = j], \quad 1 \leq k \leq M. \tag{2.4}$$

- La **distribución de probabilidad inicial de estados** ($\pi = [\pi_i]$). Probabilidad de iniciar en un estado cualquiera definida como:

$$\pi_i = P[q_i = i], \quad 1 \leq i \leq N. \tag{2.5}$$

Con estos valores, el HMM puede generar una secuencia de observaciones:

$$O = O_1, O_2, \ldots, O_T. \tag{2.6}$$

De manera compacta, un HMM se puede expresar completamente como:

$$\lambda = (A, B, \pi), \tag{2.7}$$

donde λ es el nombre de un HMM particular ó secuencia (palabra) específica.

Para aplicar un HMM (y en general, como se describió anteriormente para las tareas de reconocimiento), es necesario tomar en cuenta dos etapas: entrenamiento y prueba. En la etapa denominada *entrenamiento* se crean los modelos de referencia mediante la extracción de parámetros de las secuencias especificas para cada modelo. Por otro lado, la etapa de reconocimiento propiamente dicha se lleva a cabo en la fase de *prueba*, en la cual se realiza la

comparación entre el modelo entrenado y la muestra a identificar. Esto consiste en obtener el valor del logaritmo de la probabilidad de observar una secuencia de símbolos contenida en el vector de símbolos en el modelo Markoviano entrenado previamente; esto es, calcular la probabilidad de que la muestra (secuencia) a identificar haya sido generada por un modelo en específico.

Modelos de Mezcla de Gaussianas (GMM): Para el caso de HMM continuos, las observaciones se modelan con funciones de densidad de probabilidad que son una mezcla o combinación de funciones de densidad de probabilidad gaussianas [2]. El utilizar funciones de este tipo mejora el modelado estadístico del conjunto de vectores de observación que se pueden producir desde un estado cuando la dispersión de esos vectores es grande. En reconocimiento de voz es el caso cuando se intenta reconocer a un número grande de locutores utilizando un solo modelo de Markov por palabra de vocabulario (configuración de los sistemas de reconocimiento de voz independientes del locutor). En el caso especifico de los HMM continuos se calcula un vector de medias y de covarianza para cada una de las Gaussianas de la mezcla de cada estado. Esto se logra conociendo el vector de observaciones correspondiente a la Gaussiana de un estado específico. Experimentalmente esta técnica es adecuada para el reconocimiento del habla pero es necesario una gran cantidad de datos de entrenamiento (dada la cantidad de parámetros a entrenar) y utilizar un tiempo de ejecución significativamente mayor que con otros métodos.

Redes Neuronales: Las redes neuronales (NN, del inglés Neural Networks) consisten en un gran número de unidades de cálculo simple, las cuales se encuentran altamente interconectadas entre si [42]. Ellas pretenden interconectar un conjunto de unidades de proceso (o neuronas) en paralelo de forma similar a como ocurre en un cerebro humano; obteniendo también prestaciones similares en reconocimiento tanto en tiempo de respuesta como en tasa de error (ver Figura 2.3). Este método es útil cuando se desea evaluar varias hipótesis en paralelo y como herramienta de clasificación de vectores característicos; así como para el filtrado de señales de voz.

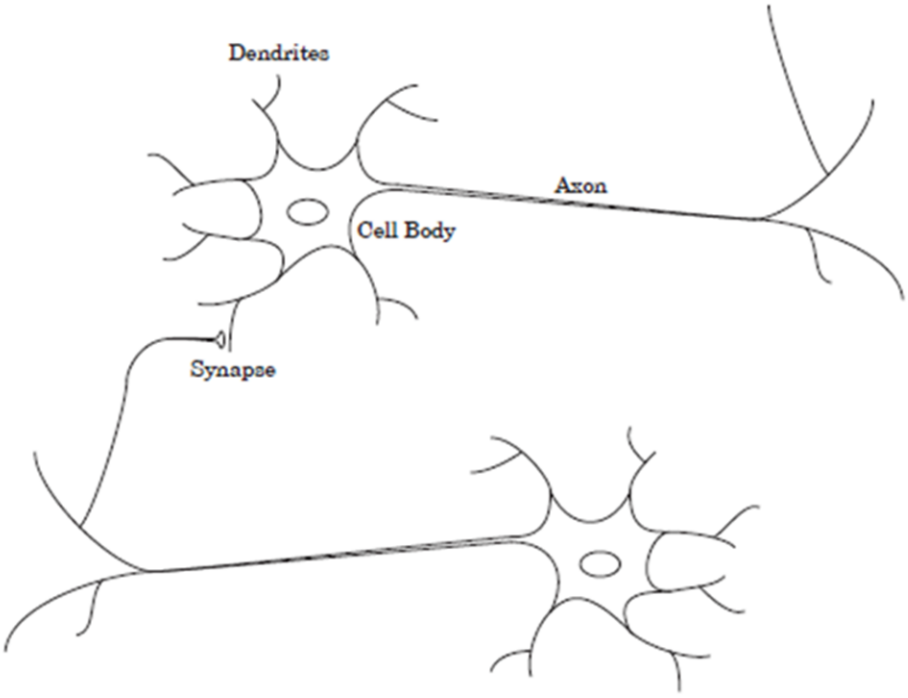


Figura 2.3 Modelo de conexión entre neuronas [3].

La idea básica detrás de una red neuronal es que, dados una serie de parámetros, estos pueden ser combinados con la finalidad de predecir un cierto resultado. El modelo de la neurona queda descrito mediante:

$$a = f(pw + b). \tag{2.8}$$

En la ecuación 2.8, $f()$ representa la función de transferencia o de activación la cual puede ser lineal o no lineal, w representa el peso (o sinapsis de entrada), p es la señal de entrada, b es la señal umbral y a es la salida o axón. En esta, las funciones de transferencia pueden ser lineal, simétricas, log-sigmoidal, positivo lineal, entre otros. Por otro lado, el tipo de red se define según la forma en que se conectan las neuronas, el tipo de neurona que lo conforma y la forma de entrenamiento de la red.

Para el caso de reconocimiento de patrones, se tienen tres tipos de redes principales:

1. Red Perceptrón (feed forward),

2. Red Hamming,

3. Red de memoria asociativa (Hopfield).

De estas, el tipo de red que mejor resultado ha brindado en reconocimiento automático del habla es el "perceptrón multicapa" (MLP) [43]. En ella las neuronas poseen capas: una capa de entrada (operando con los vectores de observación directamente), una capa de salida (apuntando a la palabra reconocida), y las respectivas capas intermedias. El enlace entre capas se da bajo pesos específicos entre neuronas. En el caso en que se requiere de un número de capas intermedias con varias neuronas, cada una se clasifica como "red neuronal profunda" [44]. Para este tipo de redes, la idea básica es que con más capas y neuronas, cada una produce una mejor predicción del conjunto de datos.

El entrenamiento de una red neuronal consiste en, dada una entrada conocida, calcular la salida de la red y compararla con la salida esperada para después calcular el error obtenido. Luego, dicho error se propaga y se ajustan los pesos de las conexiones entre neuronas. Si la propagación se realiza de atrás hacia adelante se le denomina "feed-forward", es decir, en un solo sentido desde la capa de entrada a la capa de salida; si la propagación se realiza de modo descendente se le denomina "Back-propagation". El proceso se repite varias veces hasta que la red aprenda qué respuesta debe dar para cada entrada en la fase de reconocimiento. Esto permite que los elementos neuronales puedan ser utilizados en más de una ocasión con el fin de obtener un resultado más complejo y eficaz. En otras palabras, el proceso de entrenamiento consiste en adaptar la red a la información oculta que contienen los datos de entrenamiento. Durante este proceso cada capa aprende y detecta las características que mejor ayudan a clasificar los datos.

El siguiente paso en la escala de complejidad son las denominadas "redes convolucionales" las cuales han mostrado buenos resultados en reconocimiento de voz y de imágenes. Por ejemplo, para el caso de procesamiento de imágenes en el que se quiere encontrar un número en una foto, se utilizaría una neurona por cada pixel de la imagen y después se utilizarían varias capas con varias neuronas (todas conectadas entre si). La idea es buscar características locales en pequeños grupos de entrada (para el ejemplo, consistiría en buscar características en pequeñas

regiones y no en toda la imagen). Con ello se busca detectar la misma característica en todos los grupos con el fin de repetir esa estructura y reducir el ajuste que se debe hacer; lo cual reduce el tiempo de entrenamiento. Las capas de convolución suelen utilizarse junto con una red neuronal sencilla.

2.2 Contexto Biológico (Ornitología)

Como se mencionó anteriormente, se ha buscado diagnosticar a las especies en el hábitat mediante el desarrollo de herramientas tecnológicas lo más automatizadas posibles. En este sentido, la mayoría de las metodologías que se utilizan en el reconocimiento del habla son la base para el monitoreo bioacústico [17,29,45,46]. Para el desarrollo de aplicaciones enfocadas en el análisis de señales provenientes de las aves es necesario tomar en cuenta una de las principales áreas de estudio que la conforman: el contexto biológico.

Las aves son el grupo biológico con más diversidad en el planeta tierra. Estas ponen huevos, su temperatura corporal es superior a los 40 °C, presentan un pico (por su carencia de dientes), poseen agudeza visual, tienen el cuerpo provisto por plumas (algunas asociadas al vuelo en las extremidades) [5]. Además, la mayoría de ellas son diurnas [20].

Para lograr el vuelo, algunas especies de aves (en mayor o menor medida) poseen algunas adaptaciones. Ejemplo de ello es su ligereza en peso, patas adaptadas al hábitat, y su sistema optimizado de respiración [4, 5] (Ver Figura 2.4).

(a) (b) (c)

Figura 2.4 Modificaciones evolutivas en las aves: a) Los huesos más largos son huecos, el cráneo está fusionado (para tener ligereza) y al esternón se adhieren los músculos con los que aletea. b) Al no tener en la boca huesos ni dientes, se muele con la molleja; por no tener vejiga urinaria, al defecar se elimina el exceso de agua. c) Poseen sacos aéreos interconectados con los pulmones con lo cual se optimiza la asimilación de oxígeno; lo cual es muy útil a grandes alturas [4].

Las aves pertenecen a un taxón de reptiles diápsidos llamados Arcosauros ([1]) y, hasta 1942, se conocían aproximadamente 8,200 especies diferentes. Sin embargo, una de las clasificaciones más utilizadas desarrollada en 2003, clasificó un total de 9672 especies. Hoy en día se conocen entre 9720 y 9845 especies diferentes [1].

En cuanto a la identificación de la avifauna en México (como se indica en [1]), el estudio más completo se publicó entre los años 1950 y 1957. En el se describió detalladamente la distribución geográfica de la avifauna y se actualizaron los datos taxonómicos. En este sentido, México es considerado un pais megadiverso gracias a su alta diversidad biológica (resultado de cambios ecológicos, evolutivos, entre otros). Además de ocupar el décimo lugar en el mundo en riqueza avifaunística, México alberga aves migratorias durante la mayoría de las estaciones del año (por lo cual ocupa el primer lugar en este tipo de riqueza) [4].

La clasificación de la avifauna, de modo general en México, es la siguiente [1]:

1. Residentes Permanentes: Presentes en nuestro país todas las estaciones del año.

2. Residentes Temporales: Presentes solamente alguna época del año. Estas a su vez se pueden subdividir en:

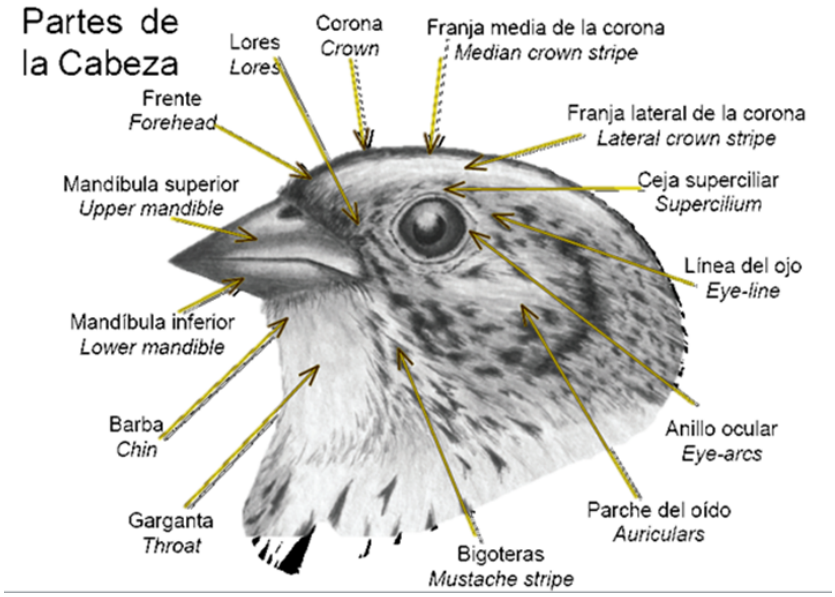
    - Residentes de verano: Llegan en primavera para anidar y regresan a sus orígenes en otoño.
    - Visitantes de invierno: Vienen de sus sitios de anidación en invierno y parten en la primavera.
    - Transitorios: Se detienen temporalmente durante su migración (en otoño o primavera).
    - Vagabundas: Utilizan nuestro territorio para alguna necesidad (como la reproductiva) y continúan su camino.
    - Accidentales: Se registran en ocasiones cuando pierden su ruta.

Del total de la avifauna clasificada, las especies exclusivas de un país (especies endémicas) son las de especial importancia para el estudio de la conservación de las aves. Nuestro país ocupa el cuarto lugar en riqueza de especies endémicas a nivel mundial con lo que coloca a estas especies en alguna categoría de "Protección especial".

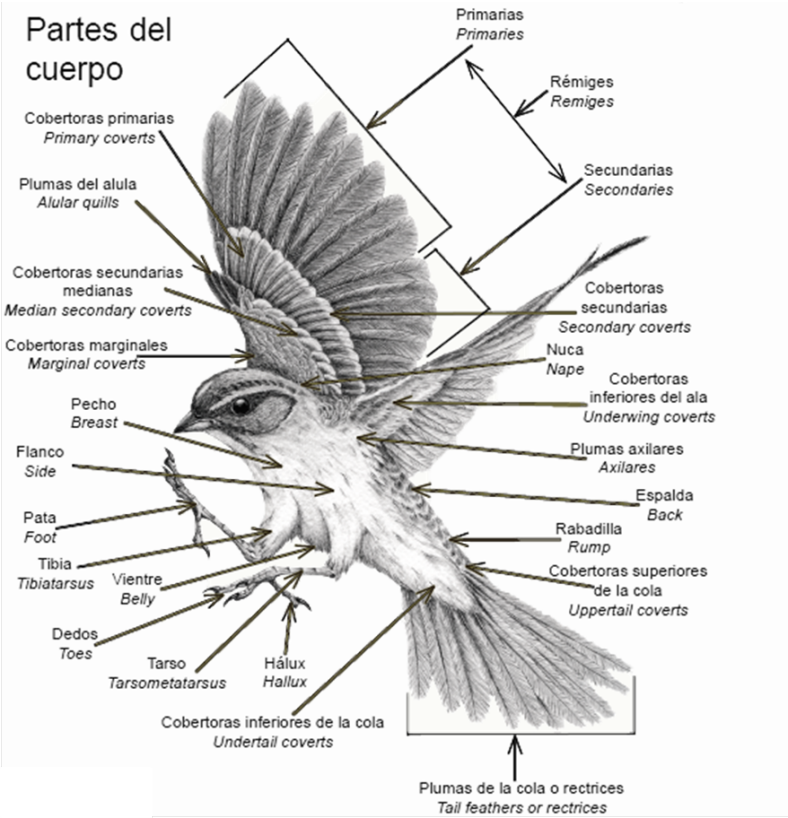
2.2.1 Patrones de clasificación

La importancia de las aves (como se indica en [5]), es su participación en la regeneración natural y en la dispersión de las semillas. Por ejemplo, la presencia o ausencia ("¿Qué debería estar y no está? o, lo que está, ¿qué puede indicar?") de algunas aves indican la calidad o conservación de algún lugar. En este sentido, las aves están presentes en casi todos los habitats y varían por tipo de especie en cada uno. Así por ejemplo, las especies encontradas en el norte de nuestro pais (en la región montañosa) son diferentes a las encontradas al sur (en la región selvática). Es aquí donde radica la importancia de la identificación de especies pues, al tener diferentes ecosistemas, la identificación de especies proporciona importantes datos sobre los mismos.

Existen algunas características propias de una familia o especie que funcionan como identificador entre la gran variedad de especies de aves. Para lograr esta identificación la observación detallada resulta una tarea fundamental (ver Figuras 2.5). Para una explicación detallada sobre el reconocimiento de especies en campo se recomienda consultar los manuales de identificación en campo [4, 5] o similares.



(a)



(b)

Figura 2.5 Anatomía de las aves: a) partes de la cabeza y b) partes del cuerpo [4].

De forma general para esta identificación se deben tomar en cuenta las siguientes características morfológicas (ver Figura 2.6):

- El pico: Esta característica relaciona al ave observada con sus hábitos alimenticios. Su color también ayuda en la identificación.

- Cuerpo, alas, cola y tarsos: La identificación del cuerpo se basa en la observación de la forma del ave observada (silueta). Por otro lado, la identificación del tipo de alas funciona como identificador de los hábitos de vuelo en al especie. Siendo así, las aves de alas más grandes se relacionan a aquellas aves que realizan largos y veloces vuelos.

  El tamaño de la cola se relaciona con el tipo de vuelo que realiza la especie (recto, quebrado, etc.).

  Por último, el tamaño del tarso se relaciona a la pertenencia a una especie de ave vadeadora (garzas, flamencos y cigueñas).

- Rasgos particulares: Estos razgos permiten, además de la identificación de especies, conocer la edad y sexo en las aves. Las características a observar como rasgos particulares contemplan desde el color del iris hasta la presencia de manchas en el pecho de la especie observada.

- Vuelo: La forma de vuelo se relaciona al tipo de familias. Los tipos de vuelos puede ser desde los muy especializados (como en el caso del colibrí) hasta los rectos (como el caso de las palomas).

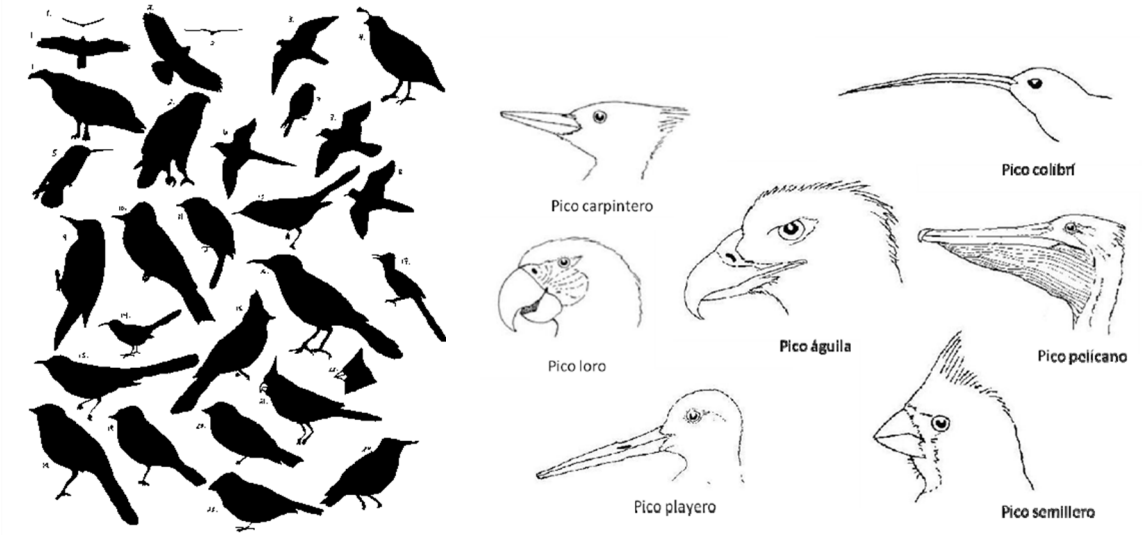
- Siluetas en vuelo: Este parámetro de clasificación es especialmente utilizado para la identificación de zopilotes y rapaces.

- Siluetas de aves acuáticas: Se utilizan para conocer la pertenencia de un ave acuática a una familia específica (útil especialmente en la observación a distancia).

- Plumaje: Las características del plumaje sirven para la identificación del sexo, época reproductiva y edad en ejemplares de una misma especie.

- Conducta: Esta característica abarca desde los hábitos alimenticios hasta el lugar en donde se posa el ave.
- Habitat: Según el hábitat donde se ha observado el ave, es la especie a la que pertenece el ave.

Cada una de las características anteriores son útiles en la clasificación de especies por observación, sin embargo, en algunos casos la identificación por medio del análisis acústico de las vocalizaciones de algunas aves puede llegar a representar una valiosa herramienta (sobre todo cuando el ave no es visible y es la única fuente de información para la identificación).



(a) (b)

Figura 2.6 Ejemplos de características morfológicas: a) formas, tamaños, siluetas y, b) picos [5].

2.2.2 Clasificación acustica

Los animales utilizan la comunicación (con otros animales) para la reproducción (atracción), alarma (advertencia de posibles depredadores), agrupación (declaración de territorio e identificación de miembros de la familia) y localización (especialmente para identificar fuentes de alimento) [28]. Cada clase animal produce frecuencias de sonidos específicos (ver Figura 2.7). Según un estudio, la frecuencia óptima de comunicación es inversamente proporcional a la potencia de 0.4 la masa corporal del animal terrestre (una frecuencia baja produce un sonido

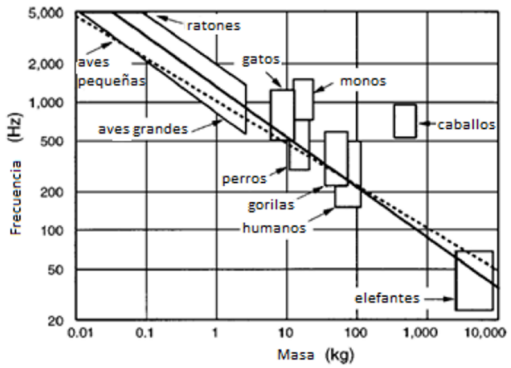


Figura 2.7 Relaciones masa-frecuencia en la comunicación [6].

grave y corresponde a un animal grande; una frecuencia alta produce un sonido agudo y corresponde a un animal pequeño). Para el caso de las vocalizaciones de aves el intervalo de sonidos se clasifica entre 700 y 2,200 Hz o 500 y 10,000 Hz [6].

La señal acústica para la comunicación en aves, de forma general, puede ser dividida como cantos y llamados. Sin embargo, como se indica en [4], no es lo mismo ave que pájaro. Las aves se diferencian de los pájaros en que estos últimos tienen desarrollado el canto. En este sentido, el canto en los pájaros, se caracteriza por un modelo morfo-neuro-etológico. El sonido es producido a partir de la vibración de las membranas timpaniformes (en la siringe) y es así que a mayor cantidad de músculos en la siringe, mayor compresión y descompresión para modulación del sonido y finalmente, una vocalización más armoniosa [6]. Mediante la sistemática (disciplina que se encarga de estudiar las relaciones de parentesco entre especies [13]) se ha llegado a la conclusión de que, al experimentar con aves en aislamiento acústico, para algunas especies el canto es una cualidad innata.

El canto de los pájaros se caracteriza por ser de larga estructura y complejidad variante en frecuencia. El canto es utilizado principalmente en las épocas de cría y es utilizado para el aprendizaje, atracción sexual y defensa de territorio [4, 5]. Las unidades básicas de un canto son las notas y las sílabas, lo cual le sigue a la frase y a la canción. La especificidad de las canciones en cada especie (como se indica en [19]) se basan en el número de sílabas, duración, ritmo, frecuencia y calidad tonal; las sílabas individuales y los intervalos de silencios crean el

ritmo de las canciones (siendo posiblemente los cantos más elaborados indicadores de edad, destreza o adecuación del individuo).

Por otro lado, los llamados en las aves están conformados por pequeñas notas simples y sílabas de frecuencia invariante. Estos pueden ser utilizados, entre otros, para alarma o agrupación, y son de carácter conductual [6]. Los llamados (los cuales son emitidos por las especies durante todo el año [5], no varían significativamente entre poblaciones y por ello, junto con las notas del canto que tampoco varían, pueden servir para el reconocimiento especie-especifico [29]. Sin embargo, una nueva clasificación surge respecto de la forma de aprendizaje de los cantos (detectado especialmente en la literatura en loros, colibríes y aves canoras [19]) y la estructura cerebral subyacente (los cantos también han evolucionado a lo largo del tiempo [20]). Por ejemplo, el llamado del gavilán de socorro se difiere en duración y estructura armónica de los gavilanes del continente americano [19]; esto representa una variabilidad individual y geográfica (algo similar a los dialectos en los seres humanos) lo cual dificulta la identificación de las especies [13]. Por tanto, además de lo ya mencionado, un sistema de reconocimiento basado en una clasificación acustica debe tomar en cuenta que la gran variedad de sonidos entre especies es debida a las necesidades de cada una (y por ello son específicos para cada especie) [17].

Por último, los sonidos no vocales (picoteo en árboles, sonido de frotamiento, entre otros) conforman otro tipo de análisis fuera de esta investigación por su naturaleza distinta a los sonidos acústicos (cantos y llamados) utilizados en el análisis para la identificación de especies de aves.

A continuación, en el capítulo 3 se presenta la metodología propuesta para la aplicación de la correlación de fase limitada en banda en procesamiento de voz. En este capítulo se plantean la adaptación de esta técnica (utilizada comúnmente en procesamiento de imágenes) como método aplicado en el área de procesamiento de voz para verificación de hablante. Luego, en el capítulo 4 se presentan los resultados y conclusiones de esta metodología. Posteriormente, y gracias a la aplicación en procesamiento de voz, en el capítulo 5 se plantea la adaptación de esta nueva técnica (en procesamiento de voz) para la aplicación en bioacústica para el reconocimiento de especies de aves. En este mismo capítulo también se presenta la aplicación

de la técnica IMFCC para la identificación de especies (logrando así una metodología de reconocimiento integral de especies de aves). Después, en el capítulo 6 se presentan los resultados y conclusiones de esta metodología

Capítulo 3

Algoritmo de verificación del hablante basado en la función de correlación solo de fase limitada en banda(BLPOC)

3.1 Introducción

La comunicación es un proceso elemental que involucra la coordinación de algunos órganos, la articulación de palabras y procesos cerebrales (por medio de los cuales la humanidad evolucionó como sociedad). El reconocimiento automático de voz consiste en el reconocimiento de palabras habladas por medio de maquinas. La identificación automática del locutor se centra en la identificación de hablantes mediante la clasificación de señales de habla como dentro de una base de datos. Dado que el sistema conoce la identidad de los hablantes impostores, este es una identificación de conjunto cerrado [47–50]. Por otro lado, la tarea de verificación del hablante consiste en confirmar si un/a hablante es quien dice ser (una decision de aceptación o rechazo) mediante la extracción y el análisis de algunos parámetros de los hablantes. Dado que para este caso el sistema no conoce a los impostores por que no pueden ser definidos, este es un problema de identificación de conjunto abierto [47].

Un sistema de verificación de hablante requiere entre otros pasos de extraer la información relevante de las señales de habla y de un esquema apropiado de entrenamiento y prueba. En cuanto a la extracción de información, una de las características de las señales de voz usualmente extraídas son los Coeficientes Cepstrales en la Frecuencia Mel (MFCC). Las características MFCC se basan en el sistema de audición humano y son regularmente extraídos en los

sistemas de identificación de hablantes [51]. Por otro lado, para comparar dicha información extraída de entre varias señales del habla, se usan comúnmente modelos estadísticos [14, 52]. Este modelado requiere de la extracción de algunos parámetros de distribución de las señales de habla en la base de datos y la utilización de un esquema de entrenamiento (para utilizar la información como un patron estadístico). Aunque los modelos estadísticos han sido ampliamente utilizados (debido a su alto desempeño en tareas de reconocimiento de habla), para estos casos generalmente se asume disponible una base de datos de gran escala. Sin embargo, para datos limitados (o bases de datos pequeñas) algunos de estos métodos de alto desempeño, tradicionalmente aplicados en verificación de hablantes, pueden proporcionar un desempeño inferior [53].

Recientemente, algunos nuevos métodos se han desarrollado con la finalidad de extraer la información relevante de las señales del habla [54, 55]. La función de correlación solo de fase limitada en banda(BLPOC) es una técnica biométrica de empate de imágenes tradicionalmente aplicada en la identificación humana mediante huellas dactilares [56]. En este sentido, con la finalidad de validar su uso como una nueva técnica de verificación de hablantes para datos limitados, se propone usar la función BLPOC como un algoritmo de emparejamiento.

3.2 Función de correlación solo de fase limitada en banda(BLPOC)

La función de correlación solo de fase (POC) es una técnica eficiente comúnmente aplicada en tareas de empate de imágenes como una medida de similitud entre dos secuencias. Dado que la información principal en el espectro de frecuencia de una señal dada usualmente está acotada a una banda de frecuencia, algunos de los componentes de fase menos significativos de la señal reducen la eficiencia del método. En este sentido, la función de correlación solo de fase limitada en banda (BLPOC) elimina dicha información menos significativa mediante la extracción de una región efectiva dependiente del análisis de las secuencia comparadas [57–61]. La teoría básica de la función BLPOC se describe a continuación.

Dadas dos secuencias 1D de N-datos ($f(n)$ y $g(n)$) cuyos indices de rango son $n = 0, \ldots, N-1$, se calcula la Transformada Discreta de Fourier en 1D (1D DFT) de cada secuencia ($F(k)$ y $G(k)$ respectivamente) como [56]:

$$F(k) = \sum_{n=0}^{N-1} f(n) W_N^{kn} = A_F(k) \exp(j\theta_F(k)), \tag{3.1}$$

$$G(k) = \sum_{n=0}^{N-1} g(n) W_N^{kn} = A_G(k) \exp(j\theta_G(k)), \tag{3.2}$$

donde $k = 0, \ldots, N-1$, y $W_N = \exp(-j2\pi/N)$; los componentes de amplitud se denotan como $A_F(k)$ y $A_G(k)$ y, $\theta_F(k)$ y $\theta_G(k)$ denotan los componentes de fase. Después, el espectro de fase cruzada ($Rn_{FG}(k)$) entre $F(k)$ y $G(k)$ se define como:

$$Rn_{FG}(k) = \frac{F(k)\overline{G(k)}}{\|F(k)\overline{G(k)}\|}, \tag{3.3}$$

donde $\overline{G(k)}$ denota el complejo conjugado de $G(k)$. Para eliminar la información menos significativa en el espectro de fase cruzada se calcula una región efectiva (dependiendo de la secuencia a comparar) [62, 63]. Así, asumiendo el componente de frecuencia cero de la secuencia a comparar desplazada al centro del espectro, el rango de la región efectiva (definida por los limites F_1 y F_2) se calcula como:

$$k = F_1, \ldots, F_2, \tag{3.4}$$

donde $0 \leq F_1 \leq F_2$ y $F_1 \leq F_2 \leq N-1$. Finalmente, la función BLPOC ($\hat{r}_{fg}(n)$) es la Transformada Inversa de Fourier en 1D (1D IDFT) de $Rn_{FG}(k)$, la cual se calcula como:

$$\hat{r}_{fg}(n) = \frac{1}{L} \sum_{k=F_1}^{F_2} Rn_{FG}(k) W_L^{-kn}, \tag{3.5}$$

donde L es el tamaño de la región efectiva la cual se define como $L = (F_2 - F_1) + 1$. La función BLPOC es una medida de similitud entre secuencias y, en este sentido, si $f(n)$ y $g(n)$ son la misma secuencia, la función BLPOC resultante es la función delta de Kronecker (δ(n)); por otro lado, si $f(n)$ y $g(n)$ no son la misma secuencia, la similitud resultante se calcula a través de un parámetro de coincidencia.

3.3 Algoritmo de verificación de hablante usando la función BLPOC

El objetivo de la verificación de hablantes es, como se mencionó anteriormente, identificar si un/a hablante es quien dice ser. Este sistema puede ser dependiente del texto para el caso de hablantes cooperativos que repiten frases pre-establecidas o, independientes del texto, para el caso de hablantes no cooperativos los cuales pronuncian frases independientes (no pre-establecidas) [64]. El sistema propuesto es un algoritmo de verificación de hablantes dependientes del texto el cual compara una señal de voz pre-establecida (de un hablante autorizado) contra otra señal de entrada (de la misma palabra pronunciada) e identifica si la señal de entrada es del mismo hablante (para el caso de un emparejamiento genuino) o no (para el caso del emparejamiento impostor). Tradicionalmente, la magnitud del pico máximo en la función BLPOC es usada como una medida de similitud en la identificación humana mediante emparejamiento de imágenes; para el emparejamiento de hablantes, algunas otras características fueron utilizadas en el método propuesto. Los pasos principales en el algoritmo son: a) detección de la región del habla, b) ajuste de tiempo, c) empate de hablantes y d) decisión de verificación individual.

3.3.1 Detección de la región del habla

Basado en un detector de actividad vocal (VAD por sus siglas en inglés) [65], se propone un nuevo método para la identificación de la región del habla mediante un umbral de energía. Después de la conversión analógica a digital de la señal del habla, el primer paso consiste en analizar e identificar cuales segmentos de la señal del habla corresponden a información (región del habla) y cuales corresponden a silencios (ver Figura 3.1). Considere $f(n)$ como una señal de habla. La región del habla (SF) se extrae como sigue: i) calcular el promedio de energía de toda la señal como $E = \frac{1}{N}\sum_{n=0}^{N-1}|f(n)|^2$, ii) dividir la señal $f(n)$ en M tramas ($fr_j(i)$) de m muestras de acuerdo a la frecuencia de muestreo (para asegurar estacionariedad en las características de la señal), iii) calcular el promedio de energía de cada $j-th$ trama como $e(j) = \frac{1}{m}\sum_{i=1}^{m}|fr_j(i)|^2$, iv) determinar un valor adecuado de umbral de energía (thr_E) e v) identificar las tramas que contienen el habla de acuerdo con las siguientes ecuaciones:

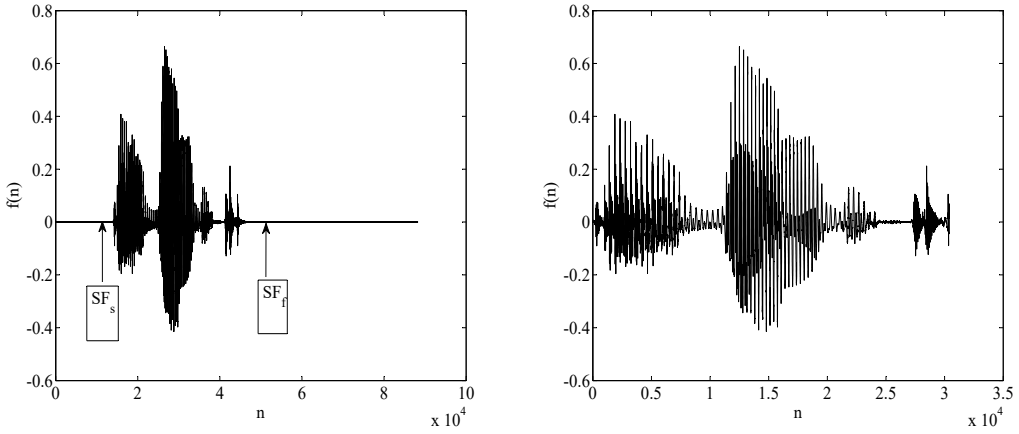
$$SF_s = first(\{fr_j(i)|e(j) \geq thr_E, 1 \leq j \leq M, 1 \leq i \leq m\}), \tag{3.6}$$

$$SF_f = last(\{fr_j(i))|e(j) \geq thr_E, 1 \leq j \leq M, 1 \leq i \leq m\}), \tag{3.7}$$

donde SF_s y SF_f son la primera y ultima tramas que contienen información del habla. Finalmente, la región del habla se calcula como:

$$SF = [SF_s, \ldots, SF_f]. \tag{3.8}$$

En otras palabras, una vez que la señal del habla es enventanada, usando las ecuaciones 3.6, 3.7 y 3.8 para extraer la región del habla (SF) y remover el silencio o ruido de la señal, es posible automáticamente identificar cuales segmentos ($fr_j(i)$) en la señal poseen la cantidad de energía ($e(j)$) por encima del umbral (thr_E). El valor óptimo de umbral en los experimentos fue $thr_E = 0.001 * E$. Sin embargo, en algunos casos para distinguir entre el final de la señal del habla y una pausa media (debido a la pronunciación fonética natural), se necesita un análisis de tramas consecutivas de energía y el calculo de un criterio de aceptación de tramas (basado en la tasa de muestreo).



(a) (b)

Figura 3.1 Ejemplo de identificación automática de la región del habla: a) señal de voz antes y b) después de la extracción de la región del habla.

3.3.2 Ajuste de tiempo

Debido a la variabilidad del habla en el dominio del tiempo introducida por el hablante (variabilidad en la velocidad del habla), la señal requiere un ajuste en este dominio. Este proceso se describe como sigue: i) comparar el número de muestras entre la señal pre-almacenada y la señal a comparar y ii) ajustar la variabilidad en el dominio del tiempo mediante un relleno de ceros al final de la señal de voz más corta de acuerdo al numero de muestras que contiene la más extensa (ver Figura 3.2).

3.3.3 Empate de hablantes

El siguiente paso consiste en calcular la banda de frecuencia efectiva en el espectro de fase cruzada como sigue (ver Figura 3.3): i) usando de la ecuación (3.1) a (3.3), calcular el espectro de fase cruzada ($Rn_{FG}(k)$) entre la señal pre-establecida y la señal entrante (a comparar) y ii) calcular la densidad espectral de potencia normal ($NPSD$ por sus siglas en inglés) del vector $Rn_{FG}(k)$ usando las siguientes ecuaciones:

$$NPSD(k) = \frac{Rn_{FG}(k)\overline{Rn_{FG}(k)}}{\max(Rn_{FG}(k)\overline{Rn_{FG}(k)})}, \tag{3.9}$$

donde $\overline{Rn_{FG}(k)}$ es el complejo conjugado de la secuencia $Rn_{FG}(k)$. Después, iii) calcular y almacenar el índice de la banda de frecuencia usando las ecuaciones:

$$F_1 = \min(\{i|NPSD(i) \geq thr_p, 0 \leq i \leq N-1\}), \tag{3.10}$$

$$F_2 = \max(\{i|NPSD(i) \geq thr_p, 0 \leq i \leq N-1\}), \tag{3.11}$$

donde thr_p es el umbral de potencia (ver Tabla 3.1). El valor del umbral thr_p permite al algoritmo eliminar las componentes de frecuencia con menos energía que el mínimo aceptable. Para calcular el thr_p, una comparación de TAR y TRR se llevó a cabo variando el valor del umbral. De la Tabla 3.1 se puede observar que thr_p=0.001 es el valor con el desempeño más alto en TAR y TRR.

Finalmente, iv) calcular la función BLPOC (usando la ecuación (3.5)) y asumir la componente de frecuencia cero de la secuencia $\hat{r}_{fg}$ desplazada al centro del espectro.

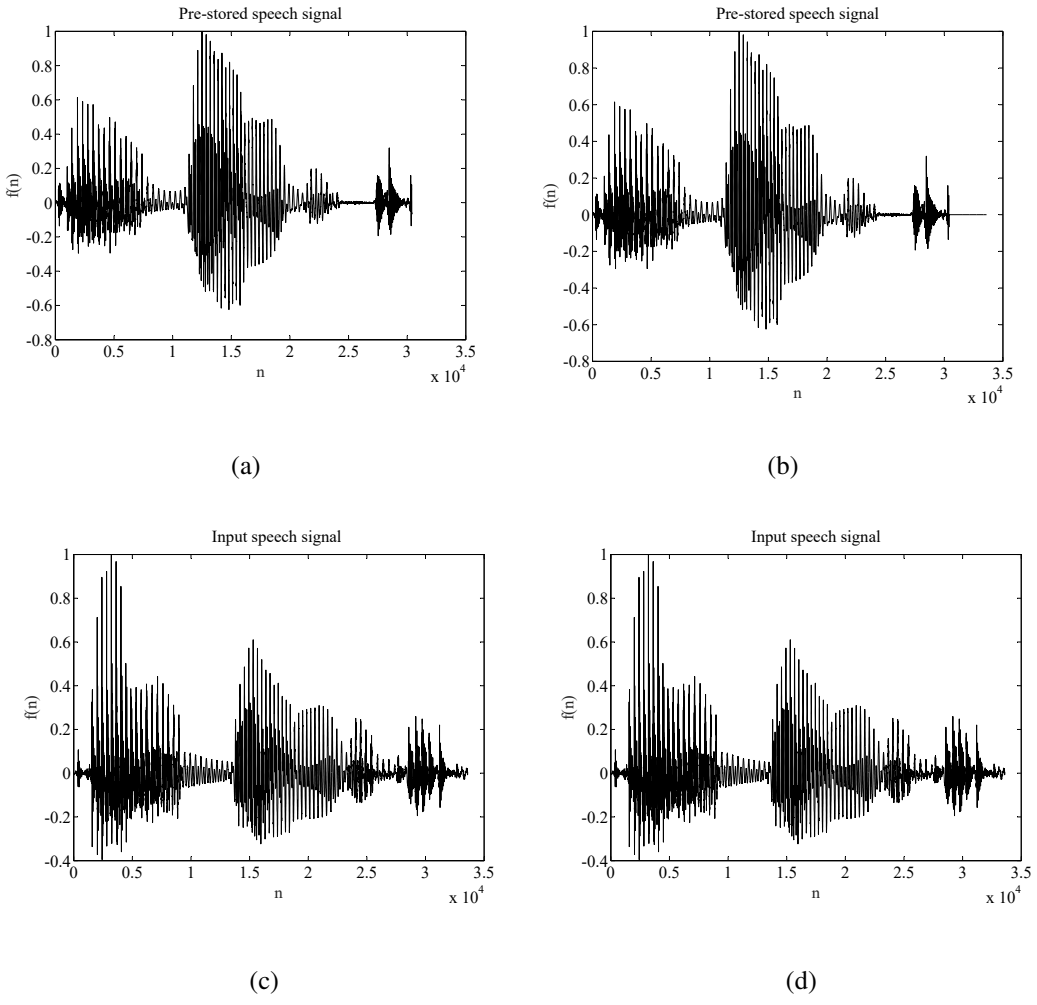


Figura 3.2 Ejemplo de ajuste del tiempo: Señal del habla pre-almacenada (pre-stored speech signal) a) antes y b) después del ajuste en el dominio del tiempo y, señal a comparar (input speech signal) c) antes y d) después del ajuste en el dominio del tiempo.

Tabla 3.1 Ejemplo del desempeño del algoritmo para diferentes valores del umbral thr_p.

Valor de thr_p	TAR(%)	TRR(%)	Valor de thr_p	TAR(%)	TRR(%)
0.0005	100	90	0.005	100	60
0.001	**100**	**100**	0.055	100	60
0.0015	100	90	0.006	100	10
0.0020	100	90	0.0065	100	10
0.0025	66	90	0.007	100	10
0.003	66	100	0.0075	100	20
0.0035	66	90	0.008	100	10
0.004	100	60	0.0085	100	20
0.0045	100	60	0.009	100	10

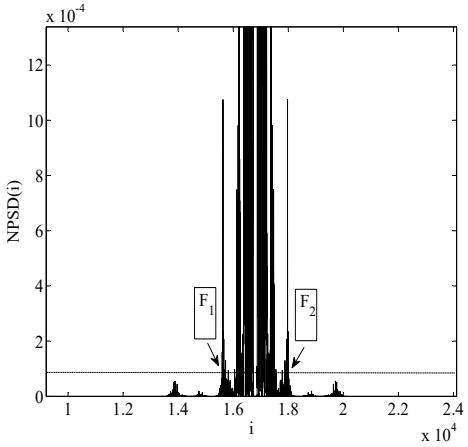


Figura 3.3 Ejemplo de identificación de la banda de frecuencia efectiva del espectro de amplitud de $NPSD(k)$. Las líneas punteadas representan el umbral de potencia (thr_p).

3.3.4 Decisión de verificación individual.

Dado que la función BLPOC en la verificación del hablante posee picos múltiples, algunas características de discriminación necesitan ser calculadas para determinar un empate genuino e impostor. Como un criterio de evaluación, este método propone el uso de una serie de umbrales de decisión: 1) relación del pico máximo, 2) varianza y 3) energía promedio.

Para cada caso se realiza un proceso de delimitación de pronunciación del hablante autorizado y una prueba de pronunciación del hablante. El primer caso consiste en la caracterización del umbral del hablante autorizado mediante la extracción de dicho umbral en varias pronunciaciones de una misma palabra del hablante. Por otro lado, la prueba de pronunciación del hablante consiste en, una vez caracterizado el umbral, extraer el mismo umbral de otras pronunciaciones de la misma palabra junto con la de otros hablantes.

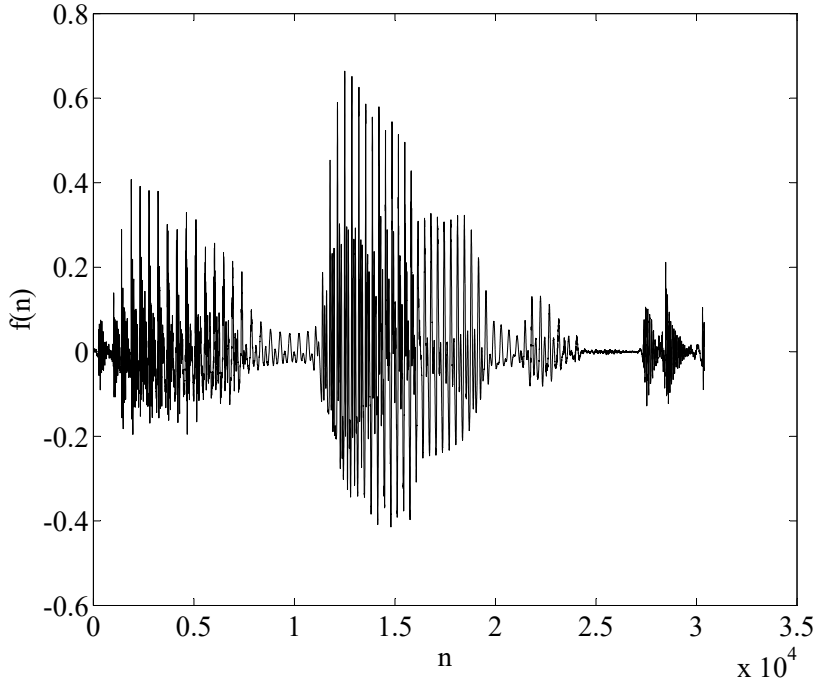
Relación del pico máximo: La relación del pico máximo (highest peak relationship, Hp por sus siglas en inglés) es calculado para evaluar la relación entre el pico máximo en la función BLPOC (P_M) y los otros picos menos significativos (mediante el calculo de su energía media $\overline{S}$) usando las siguientes ecuaciones:

$$P_M = \max |\hat{r}_{fg}|, \tag{3.12}$$

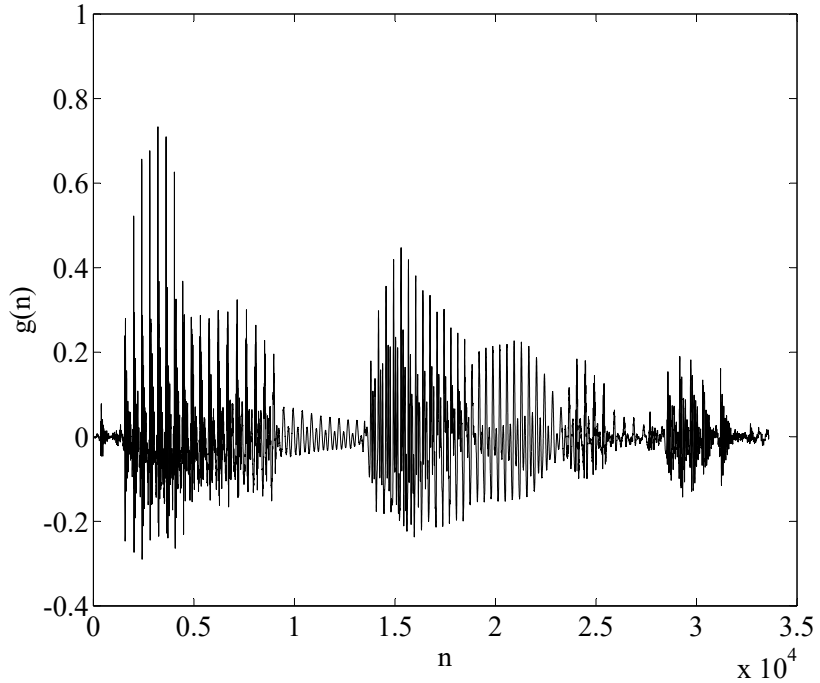
$$\overline{S} = \frac{1}{N} \sum_{i=1}^{N} |\hat{r}_{fg}(i)|, \tag{3.13}$$

$$Hp = \frac{\overline{S} * 100}{P_M}. \tag{3.14}$$

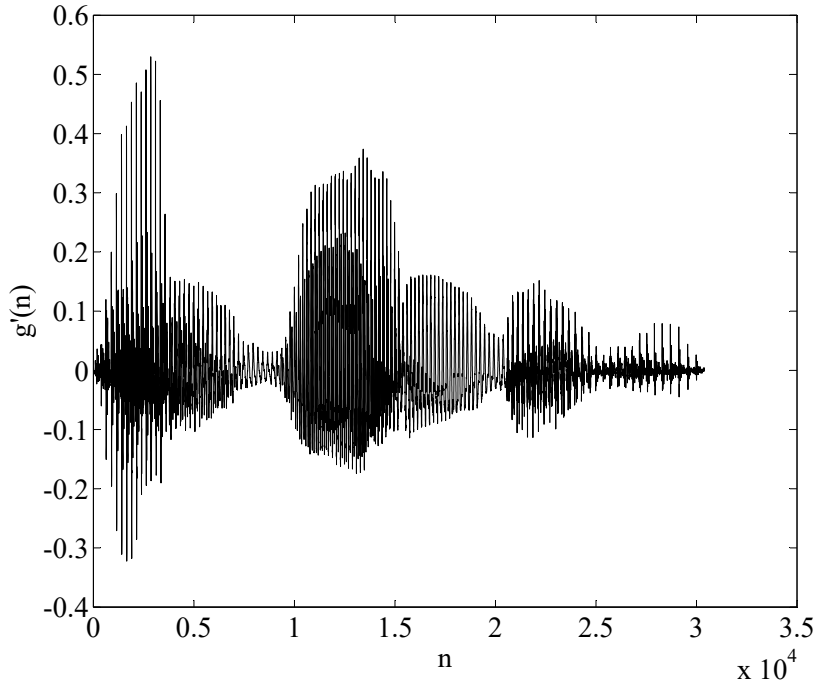
Además, se calcula una sección de aceptación (s_{Hp}) de Hp. Un valor de aceptación de Hp es un porcentaje de aceptación entre la magnitud de P_M y $\overline{S}$ de la función BLPOC. El valor de s_{Hp} en los experimentos se calculó automáticamente seleccionando el valor mínimo y máximo de Hp. Este valor cambia dependiendo de la palabra pronunciada, la cantidad de datos disponibles para el entrenamiento, el hablante y la calidad de las grabaciones. La Figura 3.5 muestra un ejemplo de la delimitación de s_{Hp} (líneas horizontales punteadas) y una prueba de pronunciación donde solo las magnitudes de Hp en s_{Hp} son aceptadas.


0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
f(n)
0
0.5
1
1.5
2
2.5
3
3.5
n
x 10^4

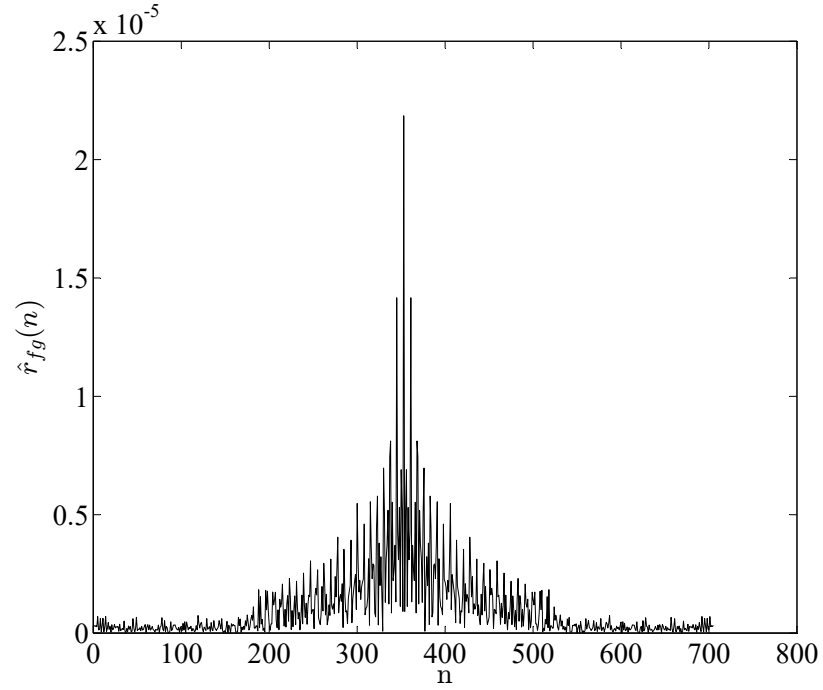
(a)


1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
g(n)
0
0.5
1
1.5
2
2.5
3
3.5
n
x 10^4

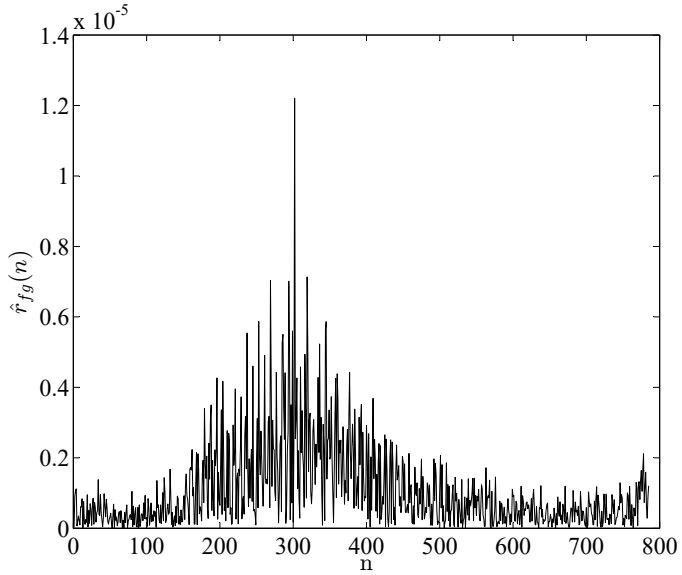
(b)


0.6
0.5
0.4
0.3
0.2
0.1
0
-0.1
-0.2
-0.3
-0.4
g'(n)
0
0.5
1
1.5
2
2.5
3
3.5
n
x 10^4

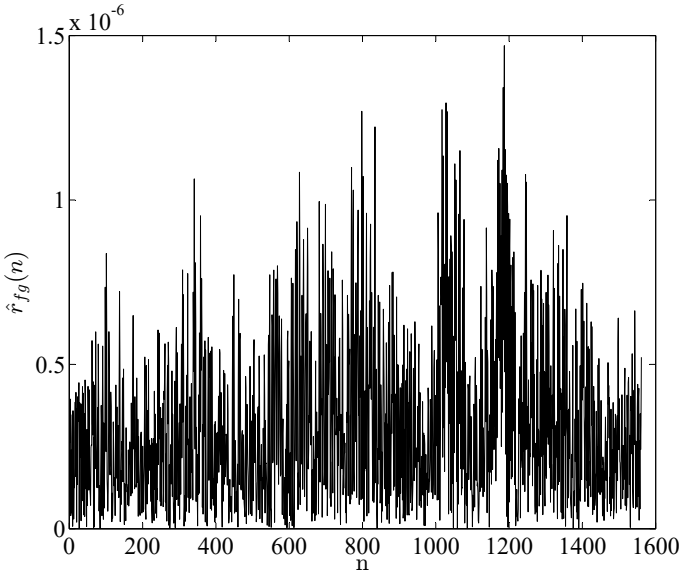
(c)


x 10^-5
2.5
2
1.5
1
0.5
0
$\hat{r}_{fg}(n)$
0
100
200
300
400
500
600
700
800
n

(d)

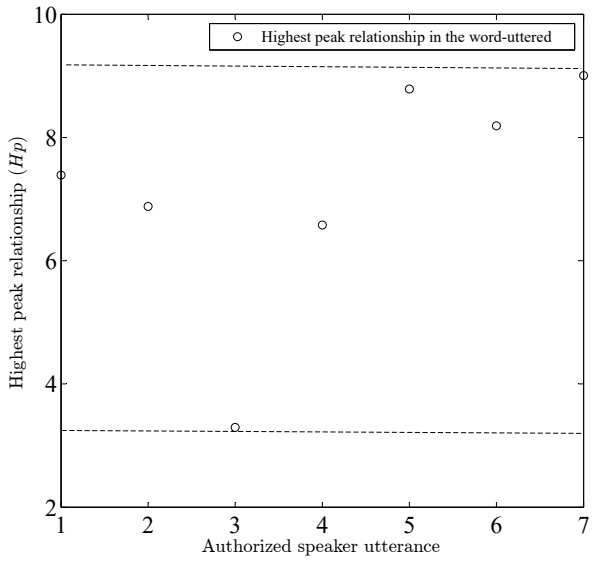


(e)

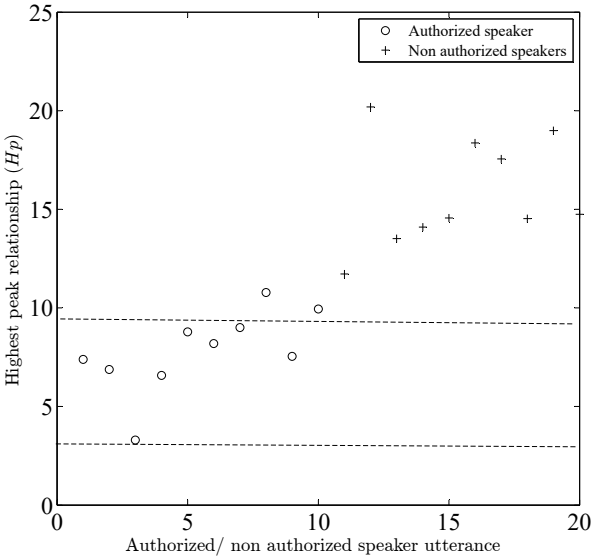


(f)

Figura 3.4 Ejemplo de un empate genuino e impostor de la misma palabra pronunciada: a) señal del habla de un hablante autorizado ($f(n)$), b) otra señal de habla del hablante autorizado ($g(n)$), c) señal de habla de un hablante impostor ($g'(n)$), d) función BLPOC entre dos señales idénticas de voz (la señal $f(n)$ del hablante autorizado), e) función BLPOC entre las dos señales del hablante autorizado ($f(n)$ y $g(n)$) y, f) función BLPOC entre la señal del hablante autorizado ($f(n)$) y el impostor ($g'(n)$).



(a)



(b)

Figura 3.5 BLPOC Hp: a) delimitación de pronunciación de un hablante autorizado y, b) prueba de pronunciación del hablante.

Varianza: La característica de varianza (σ^2) se propone para determinar que tan dispersas están la muestras en la función BLPOC. Esta característica se calcula mediante:

$$\sigma^2 = \frac{\sum_{i=1}^{N}(\hat{r}_{fg}(i) - \overline{S})^2}{N-1}. \tag{3.15}$$

La sección de aceptación (s_{σ^2}) de σ^2 establece el máximo y mínimo valor de (σ^2) en la función BLPOC considerada como un valor aceptable. El valor de s_{σ^2} utilizado en los experimentos fue calculado automáticamente determinando el máximo y mínimo valor de σ^2 (estos valores también cambian dependiendo de la palabra pronunciada por cada prueba de hablante). La Figura 3.6 muestra un ejemplo de la delimitación de s_{σ^2} (líneas punteadas horizontales) y una prueba de pronunciación donde solo las magnitudes de σ^2 en s_{σ^2} son aceptadas.

Energía promedio: La función BLPOC tiene (de acuerdo a los experimentos) más energía promedio (E) en la prueba de empate genuino que en el empate impostor. Esta característica puede ser calculada mediante:

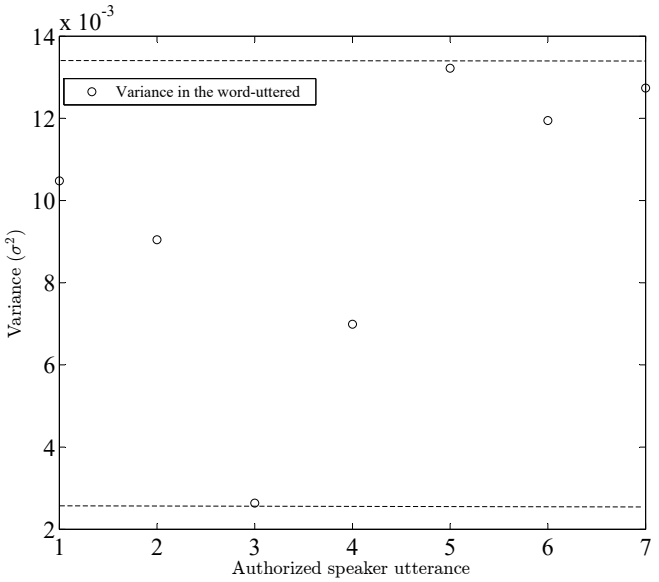
$$E = \frac{1}{N}\sum_{i=1}^{N}|\hat{r}_{fg}(i)|^2. \tag{3.16}$$

La sección de aceptación (s_E) de E establece el máximo y mínimo de E en la función BLPOC considerada como un valor aceptable. El valor de s_E utilizado en los experimentos fue calculado automáticamente determinando los valores máximos y mínimos de E (estos valores también cambian dependiendo de la palabra pronunciada por cada prueba de hablante). La Figura 3.7 muestra un ejemplo de la delimitación de s_E (líneas punteadas horizontales) y una prueba de pronunciación donde solo las magnitudes de E en s_E son aceptadas.

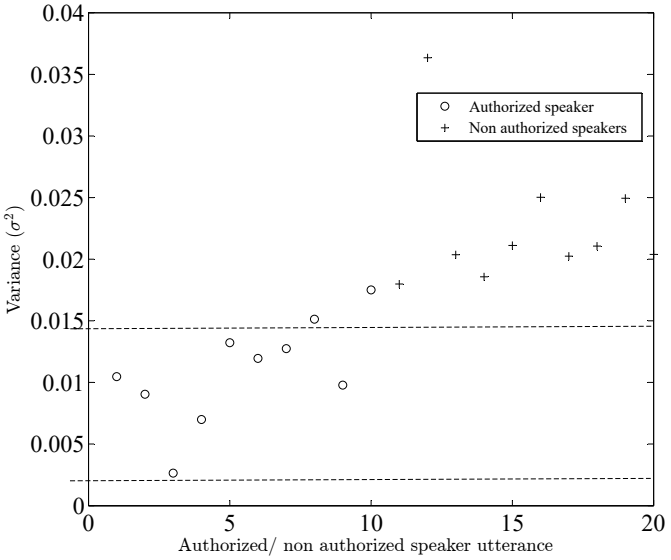
3.4 Técnica de verificación mediante MFCC-HMM

Con el fin de comparar el algoritmo propuesto, a continuación se describe una de las técnicas tradicionales de verificación: Coeficientes Cepstrales en la Frecuencia Mel y Modelos Ocultos de Markov (MFCC-HMM, por sus siglas en inglés).

Entre otras, la principal ventaja de los MFCC's son [51]:

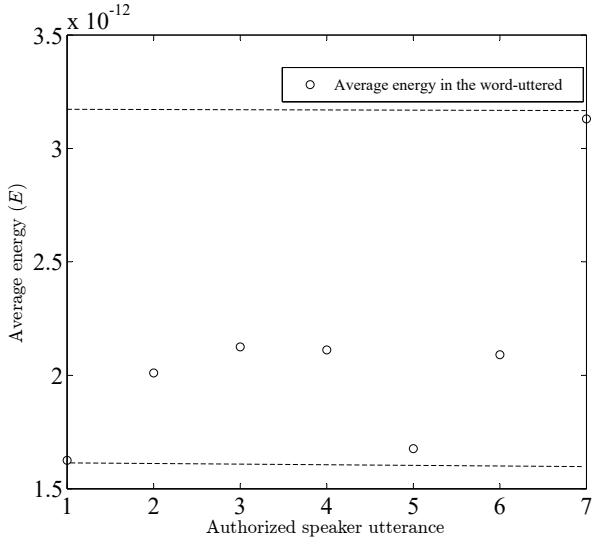


(a)

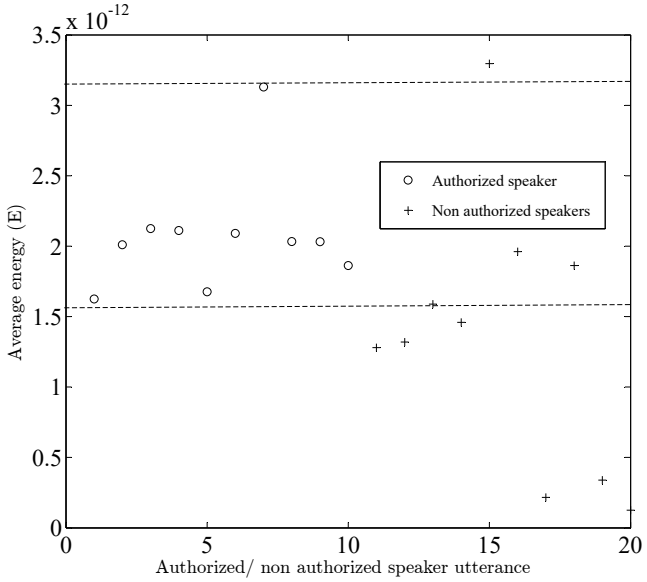


(b)

Figura 3.6 BLPOC σ^2: a) delimitación de pronunciación de un hablante autorizado y, b) prueba de pronunciación del hablante.



(a)



(b)

Figura 3.7 BLPOC E: a) delimitación de pronunciación de un hablante autorizado y, b) prueba de pronunciación del hablante.

- Esta es una de las características comúnmente extraídas en los sistemas de reconocimiento de voz.
- Comparada con otras técnicas, esta conlleva matemática relativamente sencilla.
- Los coeficientes extraídos representan una reducción de datos.

Los MFCC's están basados en la percepción humana del sonido (llamada la escala Mel) y son, el mapeo de las frecuencias en Hertz (f_{Hz}) en la escala Mel (f_{mel}) de acuerdo con (ver Figura 3.8):

$$f_{mel} = 2595 \log_{10}(1 + (f_{Hz}/700)). \tag{3.17}$$

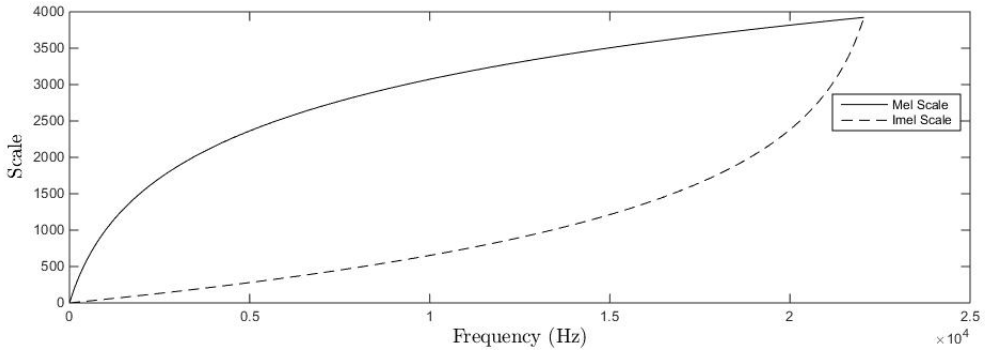


Figura 3.8 Curvas Mel e Imel.

Antes de la extracción de los MFCC's de una señal de voz, un conjunto de p filtros triangulares Mel necesitan ser calculados (ver Figura 3.9). Para calcular este banco de filtros se procede de la siguiente manera: i) determinar los limites frecuenciales (en Hz) del banco de filtros (f_{max} y f_{min}) y convertirlos a la escala Mel ($f_{mel}(f_{max})$ y $f_{mel}(f_{min})$ respectivamente) usando la ecuación (3.17). Luego, ii) determinar $p + 2$ puntos linealmente separados entre los puntos Mel calculados, y calcular la transformada inversa de dichos puntos (f_{mel}) a frecuencias (en Hz) mediante:

$$K_{sp}(j) = \frac{Ns}{Fs}(700 * (10^{(f_{mel}(j)/2595)} - 1)), \tag{3.18}$$

donde Fs es la frecuencia de muestreo y, $\{K_{sp}(j)\}_{j=0}^{p+1}$ son los puntos Mel linealmente espaciados en la escala de frecuencia en Hz. Finalmente, iii) la respuesta del $M_j(k)$ filtro se calcula mediante:

$$M_j(k) = \begin{cases} 0, & k \leq K_{sp}(j-1) \\ \frac{k-K_{sp}(j-1)}{K_{sp}(j)-K_{sp}(j-1)}, & K_{sp}(j-1) \leq k \leq K_{sp}(j) \\ \frac{K_{sp}(j-1)-k}{K_{sp}(j+1)-K_{sp}(j)}, & K_{sp}(j) \leq k \leq K_{sp}(j+1) \\ 0, & (k \geq K_{sp}(j+1)) \end{cases} \tag{3.19}$$

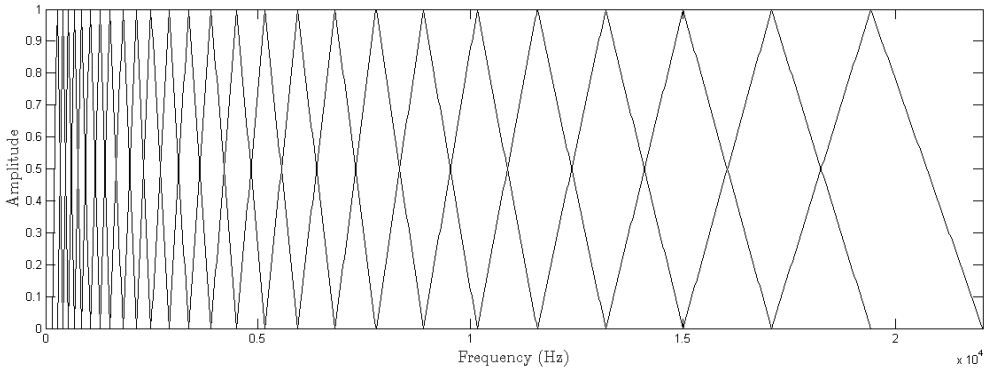


Figura 3.9 Estructura del banco de filtros MFCC.

Después, para extraer los MFCC's de la señal de voz: i) calcular la salida del banco de filtros ($E(j)$) mediante la multiplicación del espectro de energía de cada segmento ($y(k)$) y la ganancia de los filtros ($M_j(k)$) a través de:

$$E(j) = \sum_{k=1}^{Ns/2-1} y(k)M_j(k), \tag{3.20}$$

donde j es el índice de cada filtro en el banco de filtros. Después, ii) calcular la Transformada Discreta de Coseno (DCT-II) de acuerdo a:

$$C_l(t) = \sum_{j=0}^{p-1} \log[E(j)] \cos[t(j - \frac{1}{2})\frac{\pi}{p}], \tag{3.21}$$

donde $C_l(t)$ es el coeficiente MFCC de t^{esimo} orden del l^{esimo} segmento y, t=1,...,13. Finalmente, iii) calcular el valor medio de los coeficientes por cada segmento para formar un vector de magnitudes. Este proceso se repite para cada conjunto de archivos de audio del mismo tipo de palabra (los cuales conforman a su vez una matriz de magnitudes).

La información extraída (MFCC's) de cada conjunto de archivos de audio del mismo tipo de palabra (y por locutor), es usada para crear un Modelo Oculto de Markov (HMM). En la etapa de entrenamiento, se calculan los parámetros del modelo de acuerdo con [41, 66, 67]:

$$\lambda = (A, B, \pi), \tag{3.22}$$

donde λ es un modelo específico (por ejemplo el modelo de una palabra específica de un locutor específico), A es una matriz de probabilidad de la transición entre estados, B es una matriz de probabilidad de las emisiones dado los estados y, π es un vector de probabilidades de cada estado en la secuencia. Este conjunto de parámetros es entrenado mediante un proceso iterativo usando el algoritmo Baum-Welch. El resultado de este algoritmo son los parámetros A, B y π optimizados. En otras palabras, para cada palabra y por orador, se entrena un HMM que modela la información para la secuencia dada. Finalmente en la etapa de prueba, dada una secuencia (una palabra de un hablante), se calcula la probabilidad de que dicha secuencia haya sido generada por un hablante especifico (λ).

La verificación de hablante propuesta se desarrolló usando un radio de probabilidad (likelihood radio sets en inglés) [64]. Así, dados N modelos de hablantes de fondo ($\lambda_1, \lambda_2, ..., \lambda_N$), el modelo de hipótesis se calcula como:

$$p(X|\lambda_{hyp}) = f(p(X|\lambda_1), p(X|\lambda_2), ..., p(X|\lambda_N)), \tag{3.23}$$

donde λ_{hyp} es el hablante hipótesis y $f(.)$ es el máximo de los valores de probabilidad de hablantes de fondo. En otras palabras, esto consiste en establecer una lista de "hablantes no autorizados" en las pruebas de hablante (para cubrir una lista de modelos alternativos de hablantes) y clasificar si una cierta señal a clasificar (señal entrante) es de quien dice ser (una decision del tipo **si** o **no**). El algoritmo HMM fué implementado usando software libre online [68].

Capítulo 4

Resultados y conclusiones del algoritmo de verificación del hablante

4.1 Resultados y discusión

Dos diferentes bases de datos fueron utilizadas en los experimentos:

Base de datos personalizada: Esta es una base con un corpus de voz de datos limitados dependiente del hablante (en Español de México) utilizada con anterioridad para una aplicación de reconocimiento de palabras aisladas [69]. La base de datos contiene 400 repeticiones (10 repeticiones por palabra por cada uno de los 4 hablantes mexicanos en la base de datos). Las muestras de habla fueron grabadas utilizando un porta estudio digital marca TASCAM DP-008 a una frecuencia de muestreo de 44.1 KHz en formato .wav (sin consideraciones especiales sobre las características del micrófono). Las muestras fueron grabadas en una cámara de video conferencias para garantizar una alta relación señal a ruido (SNR por sus siglas en inglés).

ELSDSR: El corpus ELSDSR de discurso leído fue un esfuerzo conjunto de la facultad y estudiantes de maestría del Departamento de Informática y Modelado Matemático (IMM) y la Universidad Tecnológica de Dinamarca (DTU) [70, 71]. El corpus de esta base está en idioma Inglés y fue leída por 20 Daneses, 1 Islandes y 1 Canadiense (10 mujeres, 12 hombres). La base de datos contiene 198 repeticiones (sugeridas 154 y 44 repeticiones para el conjunto de entrenamiento y prueba, respectivamente). Aunque tradicionalmente esta base de datos fue implementada en tareas de reconocimiento de hablantes, recientemente fue aplicada para un

sistema de verificación de hablantes [72]. En este sentido, para los experimentos mostrados fueron colectados de la base de datos, solamente las repeticiones de palabras utilizadas con mayor frecuencia por los hablantes.

Los siguientes parámetros fueron utilizados como métricas de desempeño para evaluar los algoritmos a comparar:

- Tasa de Verdadera Aceptación (TAR, por sus siglas en inglés): La probabilidad de verificar como aceptado (autorizado) un hablante (persona) autorizado.

- Tasa de Verdadero Rechazo (TRR, por sus siglas en inglés): La probabilidad de verificar como rechazado (no autorizado) un hablante (persona) impostor.

Para cada base de datos se consideró para la prueba TAR (por cada prueba de hablante dependiente de la palabra pronunciada) el 70% de los archivos de audio como el conjunto de entrenamiento (7 repeticiones); y el resto de audios como el conjunto de prueba. Por otro lado para la prueba TRR, se utilizó el total de archivos de audio de los otros hablantes (repeticiones por palabra de los hablantes no autorizados en cada prueba TAR). El número total de pruebas de empate genuino e impostor por cada base de datos (y por cada método propuesto) fue de 38 y 73 pruebas para la base de datos personalizada y ELSDSR, respectivamente.

El algoritmo propuesto fue comparado usando cuatro diferentes configuraciones de umbrales: relación del pico máximo, relación del pico máximo-varianza, relación del pico máximo-energía promedio y, una característica de discriminación total (relación de pico máximo-varianza-energía promedio).

La Figura 4.1 muestra la comparación de los algoritmos usando la base de datos personalizada. El desempeño general en TAR fue cercanamente uniforme para los tres primeras configuraciones de umbrales de BLPOC. De la prueba de relación del pico máximo, se muestra una mejora (especialmente en el desempeño de TRR) agregando la varianza a la relación del pico máximo. Así, un hablante no autorizado tiene una probabilidad del 85% de ser rechazado. Además, en comparación a esta misma prueba, se muestra una mejora del 9% en el desempeño de TRR utilizando el umbral relación del pico máximo-energía promedio.

Por otro lado, tomando en cuenta la característica de discriminación total, se puede observar un 98.24% en el desempeño de TAR y un 87.17% en el desempeño de TRR. Finalmente, el algoritmo MFCC-HMM tiene un 1.75% y 100% en el desempeño TAR y TRR, respectivamente (siendo así, de las comparaciones mostradas en la Tabla 4.1, la técnica con la más baja probabilidad de verificar correctamente a un hablante autorizado bajo la condición de datos limitados).

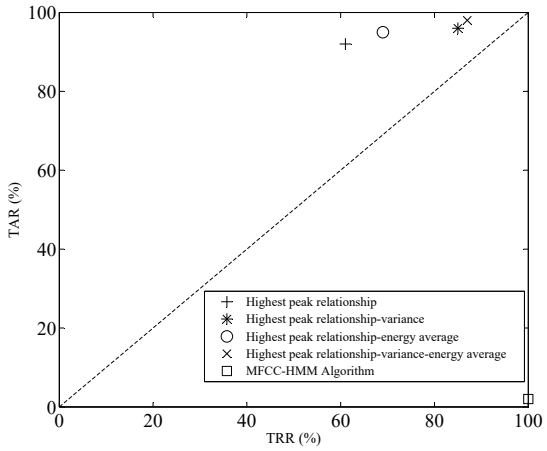


Figura 4.1 Desempeño de los algoritmos en la base datos personalizada.

La Figura 4.2 muestra la comparación de los algoritmos utilizando la base de datos ELS-DSR. Aquí se muestra una mejora de desempeño agregando umbrales al umbral de relación de pico máximo. En este sentido, tomando en cuenta todos los umbrales, se puede observar un 93.75% en TAR y un 67.05% en TRR. Finalmente, el algoritmo MFCC-HMM tiene un 8.97% en TAR y un 99.83% en TRR (ver la Tabla de comparaciones 4.2).

Como se menciona arriba, de la Figura 4.1 y la Figura 4.2, se puede observar una mejora de desempeño aplicando diferentes configuraciones de umbrales. Así que, como se esperaba, la aplicación de umbrales de decisión apropiados a la función BLPOC determinan en el algoritmo, en gran medida, cuando o no autorizar a un hablante.

La banda de frecuencia inherente es tradicionalmente extraída analizando la DFT de la señal pre-almacenada [62]. Sin embargo, para evitar inconsistencias en la función BLPOC

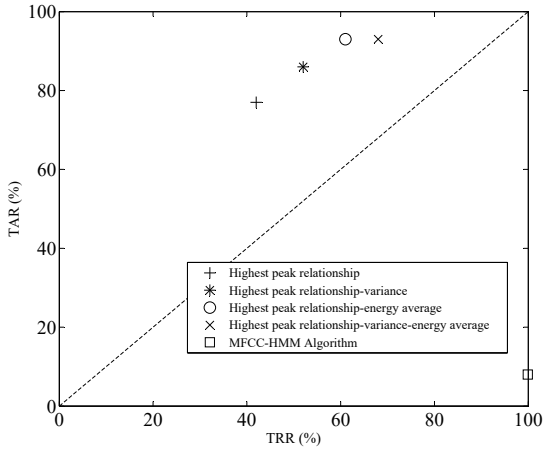


Figura 4.2 Desempeño de los algoritmos en la base datos ELSDSR.

Tabla 4.1 Desempeño de los algoritmos en la base de datos personalizada.

Usando el umbral relación de pico máximo-variaza-energía promedio	
TAR	TRR
98.24%	87.17%
Desempeño del algoritmo MFCC-HMM	
TAR	TRR
1.75%	100%

Tabla 4.2 Desempeño de los algoritmos en la base de datos ELSDSR.

Usando el umbral relación de pico máximo-variaza-energía promedio	
TAR	TRR
93.75%	67.05%
Desempeño del algoritmo MFCC-HMM	
TAR	TRR
8.97%	99.83%

dependiendo de cual es tomada como la señal pre-almacenada y la señal entrante, el algoritmo propuesto detecta la banda de frecuencia inherente en el espectro de fase cruzada (ver sección 3.3.3).

De la Tabla de comparaciones 4.1 y la Tabla 4.2 se puede observar que la configuración de discriminación total tiene un mejor desempeño en la base de datos personalizada que en la base ELSDSR. Sin embargo, el algoritmo MFCC-HMM posee el desempeño más bajo en ambas bases de datos debido a que la cantidad de datos no logra entrenar de forma adecuada los modelos creados. Así, la cantidad de datos de entrenamiento juega un rol importante en la eficiencia de este método tradicional estadísticos. En este sentido, dado que no se tiene un "entrenamiento formal" en el algoritmo propuesto, la selección de la sección de aceptación para cada umbral representa el paso de entrenamiento para datos limitados. Tomando en cuenta los resultados, la función BLPOC, mediante la configuración de algunos umbrales, puede ser ofrecida como una técnica automática de verificación de hablantes con datos limitados.

4.2 Conclusiones

En esta fase de investigación referente a la aplicación en reconocimiento de voz, fue propuesto un nuevo método para la verificación automática de hablantes con datos limitados usando la función BLPOC. Tomando en cuenta el desempeño de las pruebas, puede ser logrado cerca de un 98% en TAR y 87% en TRR en la base de datos personalizada (y un 93% y 67% respectivamente en la base ELSDSR). El algoritmo MFCC-HMM rechazó a más del 90% de los hablantes. Este resultado específico da soporte a que bajo la condición de datos limitados (especialmente para el entrenamiento), la técnica de verificación HMM puede proveer un bajo desempeño (lo cual explica los resultados del algoritmo mostrados en las Figuras 4.2 y 4.1). Por lo tanto, una comparación justa del algoritmo propuesto sería en contra de otro método de verificación de hablante de datos limitados.

Así, la función BLPOC es un método eficiente tradicionalmente aplicado en la identificación humana mediante empate de imágenes que, dado el análisis de los resultados obtenidos, es también un método efectivo para un sistema de verificación de hablantes bajo la condición de datos limitados para reconocimiento de voz. Por ende, este método también podría ser aplicado en otras tareas de identificación biométrica por sonido donde los datos limitados sean una condición.

Capítulo 5

Reconocimiento de especies y especímenes

Como se mencionó en la introducción, la gran mayoría de las técnicas actualmente utilizadas en bioacústica tienen como base las técnicas desarrolladas en reconocimiento automático de voz [17, 29, 45, 46]. Teniendo en cuenta este contexto, y basándonos en los resultados obtenidos del análisis descrito en el capítulo anterior, se propone el uso de la función BLPOC como parte de un sistema de identificación de especies de aves. Se ha de señalar que este sistema se enfoca en dos de las principales áreas de la bioacústica: la identificación de especies y, la identificación de individuos (de dichas especies). En este sentido, el sistema integral propuesto en este capitulo, mediante el análisis de señales y la identificación de las mismas, da respuesta a las siguientes preguntas:

- Considerando una especie de ave plenamente identificada acústicamente, una cierta muestra de sonido bajo análisis, ¿es de la misma especies? : la respuesta a esta pregunta se obtiene mediante la aplicación de un algoritmo de clasificación automática de especies de aves basado en Coeficientes Cepstrales Inversos en la frecuencia Mel.

- Considerando una señal acústica plenamente identificada acústicamente, una cierta muestra de sonido bajo análisis, ¿proviene del mismo individuo?: la respuesta a esta pregunta se obtiene mediante la aplicación de un algoritmo de identificación acústica individual en aves basado en la función BLPOC.

A continuación se describen las técnicas planteadas para contestar a cada una de estas preguntas.

5.1 Algoritmo de clasificación automática de especies de aves basado en Coeficientes Cepstrales Inversos en la frecuencia Mel

Justificación de Metodología: Como se mencionó anteriormente, el reconocimiento de voz es la base de las técnicas aplicadas en bioacústica. Sin embargo, dado que las vocalizaciones poseen características diferentes al habla humana, y dada la alta variabilidad de vocalizaciones entre especies de aves, aún no es bien entendido cuales características en las vocalizaciones deben ser extraídas para obtener mejores tasas de clasificación [73]. En este sentido, algunas de las investigaciones en bioacústica utilizan los Coeficientes Cepstrales en la Frecuencia Mel (MFCC) debido a que, comparada con las técnicas tradicionales de reconocimiento de voz, es una metodología relativamente sencilla [45, 74]. Por otro lado, algunos estudios recientes en reconocimiento de voz utilizan una variante de los MFCC conocida como Coeficientes Cepstrales Inversos en la Frecuencia Mel (IMFCC) para tomar en cuenta los componentes de alta frecuencia que pierden los MFCC's [14,15,51,52]. En este contexto, y dada la cercana relación entre la bioacústica y el reconocimiento automático de voz, se propone un sistema de clasificación basada en la extracción de componentes IMFCC de las vocalizaciones de aves.

Algoritmo propuesto para la clasificación de especies: Los pasos principales en el algoritmo propuesto son: pre-procesamiento, extracción de características y, sistema de clasificación de sonidos (ver Figura 5.1).

5.1.1 Pre-procesamiento

La variabilidad en las vocalizaciones (cantos y llamados) entre especies de aves dificulta la extracción automática de la información principal en las grabaciones de las vocalizaciones (especialmente cuando la información es obtenida de ambientes reales). El algoritmo de pre-procesamiento propuesto permite al usuario mediante una observación y decision, extraer la información considerada como "más importante", remover el ruido y, aplicar un umbral espectral al espectrograma resultante.

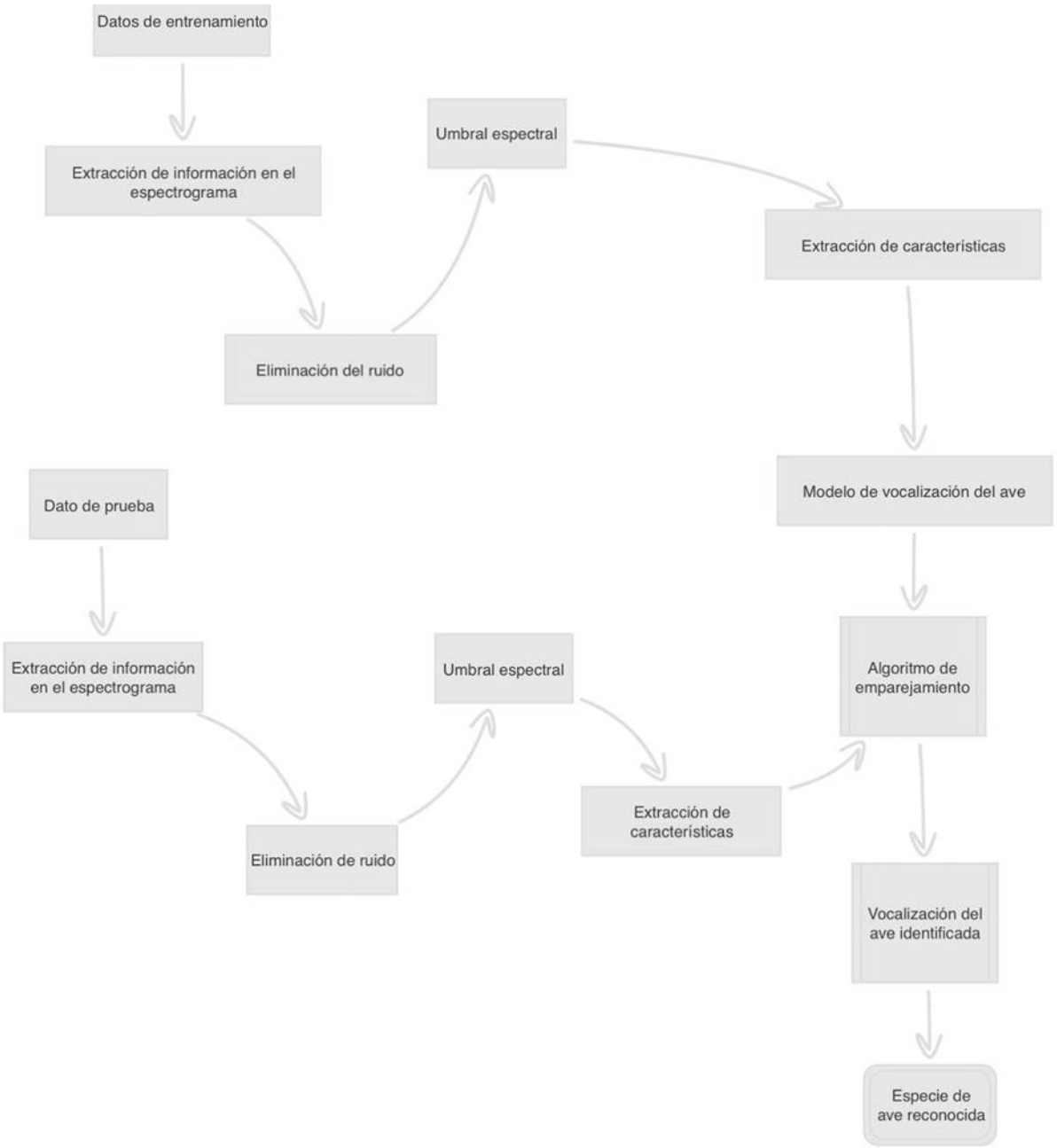


Figura 5.1 Algoritmo propuesto de clasificación de especies de aves.

5.1.1.1 Extracción de información del espectrograma

Tradicionalmente se usa la comparación de espectrogramas provenientes de las grabaciones de vocalizaciones para la clasificación de especies. Basada en esta técnica, el algoritmo propuesto, mediante la selección de una región conformada por dos puntos en el espectrograma de las vocalizaciones, extrae la información del dominio temporal y frecuencial más importante desde el punto de vista del usuario (ver Figura 5.2). Este proceso es equivalente a un filtro visual que permite la discriminación de una vocalización de ave del resto de sonidos en las grabaciones. Suponiendo la secuencia $f(n)$ como un archivo de audio (el cual contiene una vocalización de ave), la región "más importante" se extrae del espectrograma como sigue: i) calcular la Transformada Discreta de Fourier 1D (1D DFT) de la secuencia $f(n)$ como:

$$F(k) = \sum_{n=0}^{N-1} f(n) W_N^{kn} = A_F(k) \exp(j\theta_F(k)), \tag{5.1}$$

donde $k = 0, \ldots, N-1$, W_N =exp$(-j2\pi/N)$, $A_F(k)$ es un componente de amplitud y, $\theta_F(k)$ es un componente de fase. Después, ii) de la región seleccionada en el espectrograma, extraer el rango de interés en el dominio frecuencial y, convertir los puntos seleccionados que definen el rango de interés en coordenada de la 1D DFT. Después de esto, iii) calcular el valor de magnitud mínima en $F(k)$ y remplazar el valor de magnitud en las frecuencias $F(k)$ que no correspondan a la región de interés previamente seleccionada (ver Figura 5.3). Después, iv) calcular la Transformada Discreta Inversa de Fourier 1D (1D IDFT) de acuerdo con:

$$f''(n) = \frac{1}{N} \sum_{k=0}^{N-1} F(k) W_N^{-kn}, \tag{5.2}$$

donde $f''(n)$ es el archivo de audio en la frecuencia de interés, es decir, de la vocalización del ave. Luego, v) obtener el rango de interés en el dominio temporal (en coordenadas 1D DFT) y extraer el rango de interés de $f''(n)$ de la misma manera como se extrajo la información frecuencial. Finalmente, vi) después de una normalización, reproducir el audio de la secuencia resultante ($f'(n)$). Repetir el proceso hasta que el audio resultante solo contenga la información considerada como "más importante" (ver Figura 5.4).

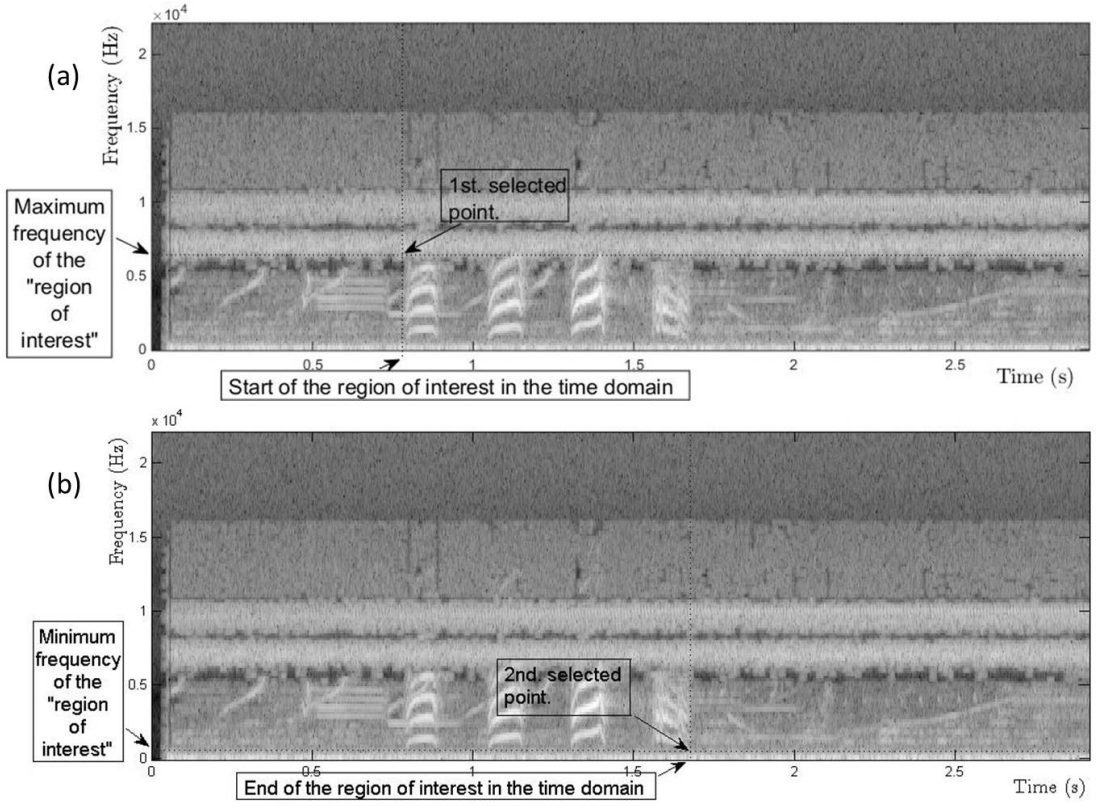


Figura 5.2 Ejemplo de la selección de una región con la información más importante en el dominio tiempo-frecuencia (en el espectrograma de una grabación de audio): a) primer y b) segundo punto seleccionado.

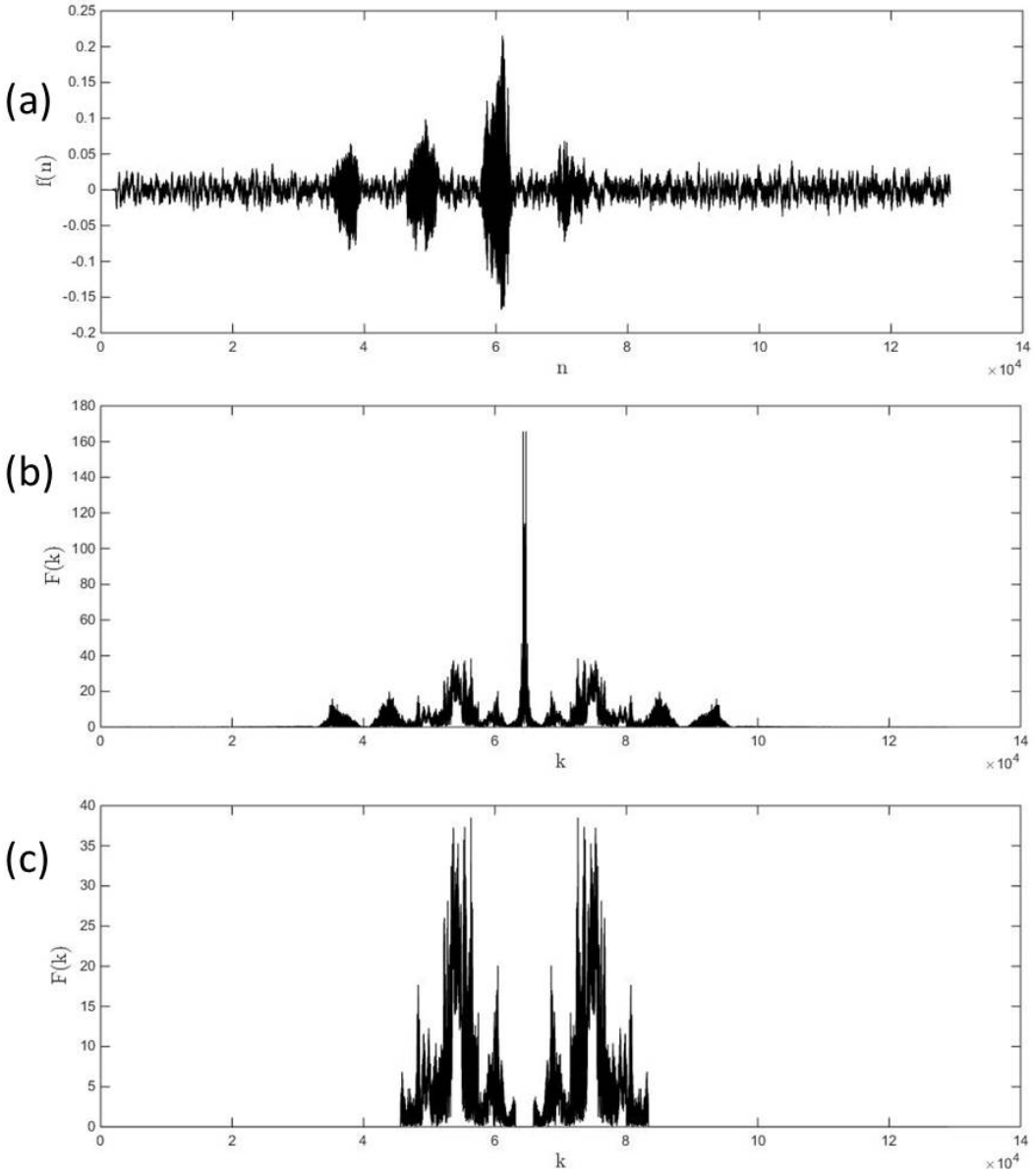


Figura 5.3 Ejemplo de la extracción de la región de interés en la frecuencia: a) secuencia de vocalización de un ave, b) 1D DFT de la secuencia y, c) región de interés en la frecuencia.

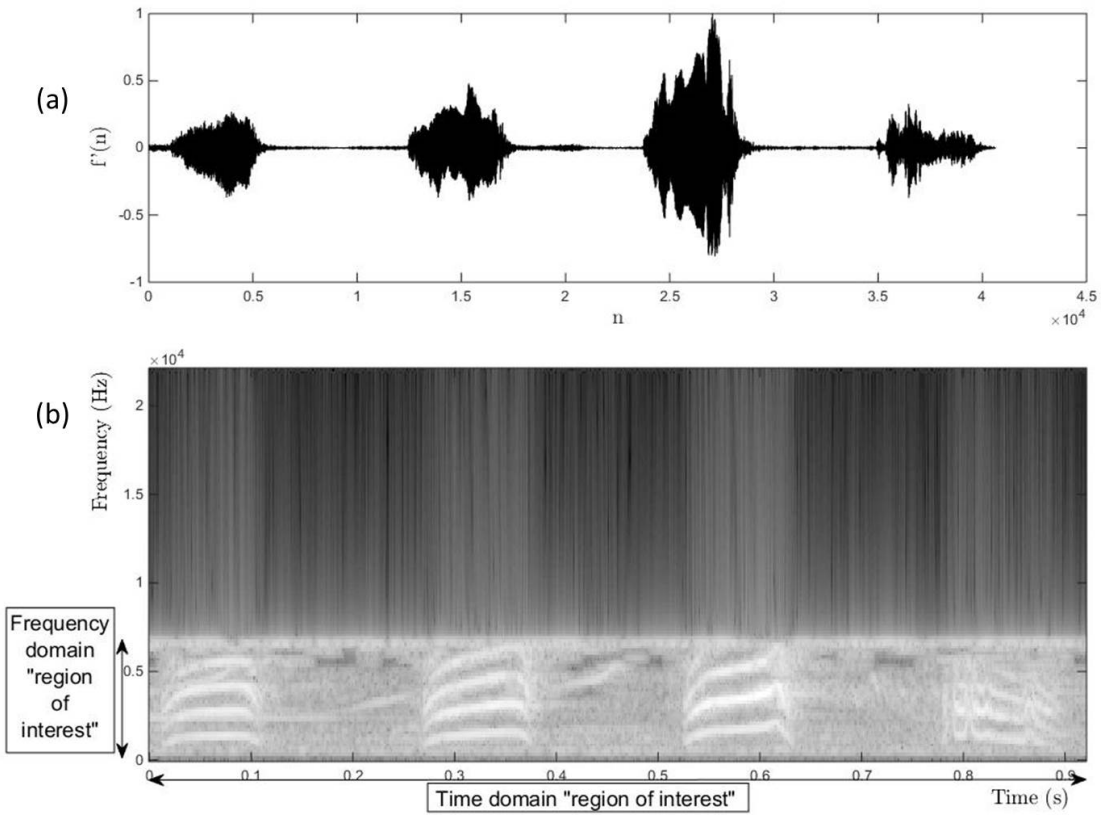


Figura 5.4 Vocalización de un ave después de la extracción de la información más importante: a) forma de onda y, b) espectrograma.

5.1.1.2 Eliminación de ruido

De grabaciones ruidosas es difícil clasificar una vocalización. El paso de eliminación de ruido consiste en la eliminación de una "región de ruido" mediante la selección de dos puntos en el espectrograma (ver Figura 5.5). En otras palabras esto es equivalente a un filtro visual que permite al usuario eliminar fragmentos de audio (previamente identificados como ruido) dentro de la grabación de una vocalizaciòn. Después de seleccionar una región de ruido ($g(n)$) en el espectrograma de la grabación, se procede de la siguiente manera: Primero, i) calcular la 1D DFT de $g(n)$ como:

$$G(k) = \sum_{n=0}^{N-1} g(n) W_N^{kn} = A_G(k) \exp(j\theta_G(k)). \tag{5.3}$$

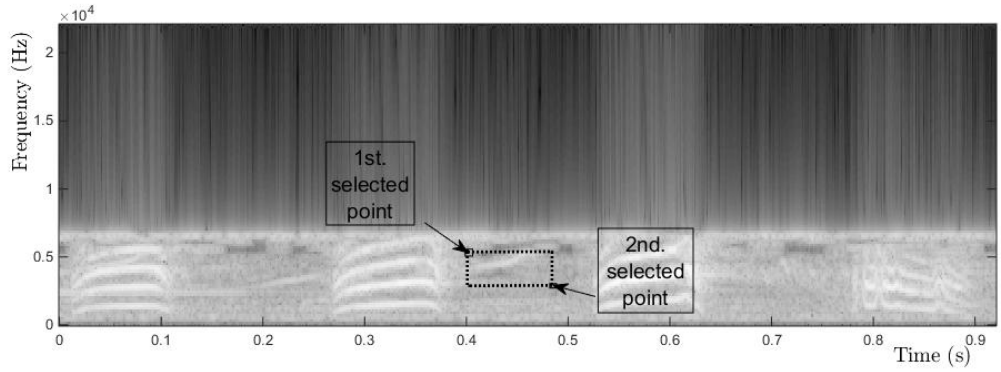


Figura 5.5 Ejemplo de una región de ruido en el espectrograma de una vocalización de ave (el rectángulo punteado representa la región seleccionada).

Después, ii) calcular el mínimo de $G(k)$ y remplazar el valor de magnitud en las frecuencias de $G(k)$. Después, calcular la IDFT de la secuencia de acuerdo con (ver Figura 5.6):

$$g'(n) = \frac{1}{N} \sum_{k=0}^{N-1} G(k) W_N^{-kn}. \tag{5.4}$$

Finalmente, iii) remplazar la región resultante ($g'(n)$) en $f'(n)$). Repetir el proceso hasta que la señal de audio resultante tenga una adecuada eliminación de ruido (ver Figura 5.7 y comparar con forma de onda de la Figura 5.3a))

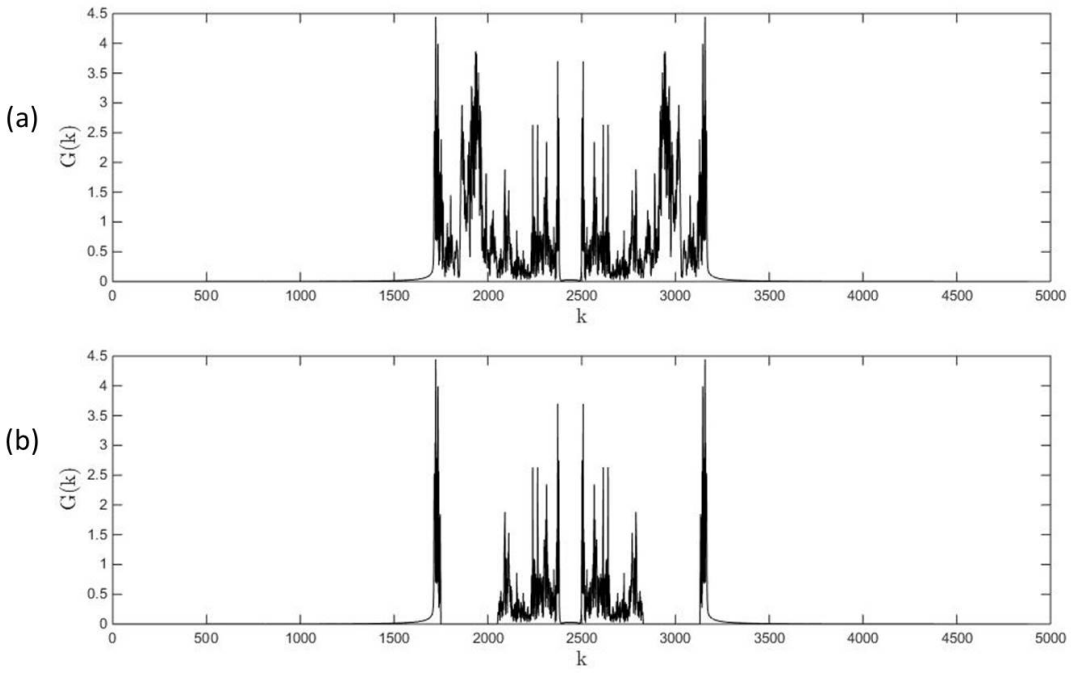


Figura 5.6 Ejemplo de la eliminación de ruido frecuencial en una vocalización de ave: 1D DFT de la región seleccionada de ruido a) antes y b) después de la eliminación.

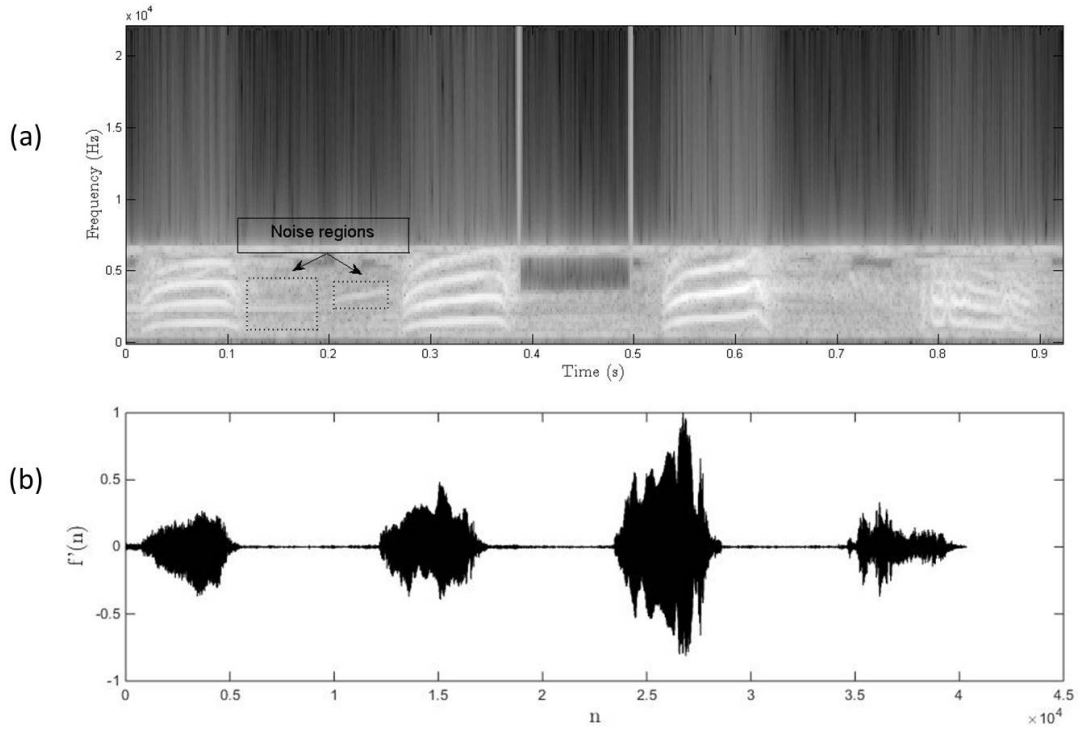


Figura 5.7 Ejemplo de: a) otras regiones propuestas como ruido y, b) forma de onda de la vocalización de ave después de la eliminación de varias regiones de ruido.

5.1.1.3 Umbral espectral

Se propone filtro de energía en el espectrograma con el fin de solo extraer (y procesar) los componentes más significativos, es decir, aquellos con la energía más alta mostrados en el respectivo espectrograma [30]. Para aplicar el umbral espectral se procede de la siguiente manera: después de segmentar y enventanar la secuencia $f'(n)$ (usando una ventana Hamming), para cada segmento ($\sum_{n=1}^{Ns} y(n)$), i) calcular la energía espectral como:

$$y(k) = \frac{1}{Ns}|\sum_{n=1}^{Ns} y(n) W_N^{kn}|^2, \tag{5.5}$$

donde Ns es el numero de puntos en la DFT, $W_N = exp(-j2\pi/Ns)$ y, $1 \leq k \leq Ns$. Luego, ii) después de una normalización, calcular un umbral de magnitud adecuado (thr) y, iii) establecer aquellos valores de magnitud de frecuencia (en $y(k)$) bajo el valor de thr, a un valor mínimo. Finalmente, iv) repetir el proceso para cada segmento en la secuencia. Después de algunos

experimentos (ver Figura 5.8), el valor de umbral establecido en los experimentos fue un valor constante de thr=0.01.

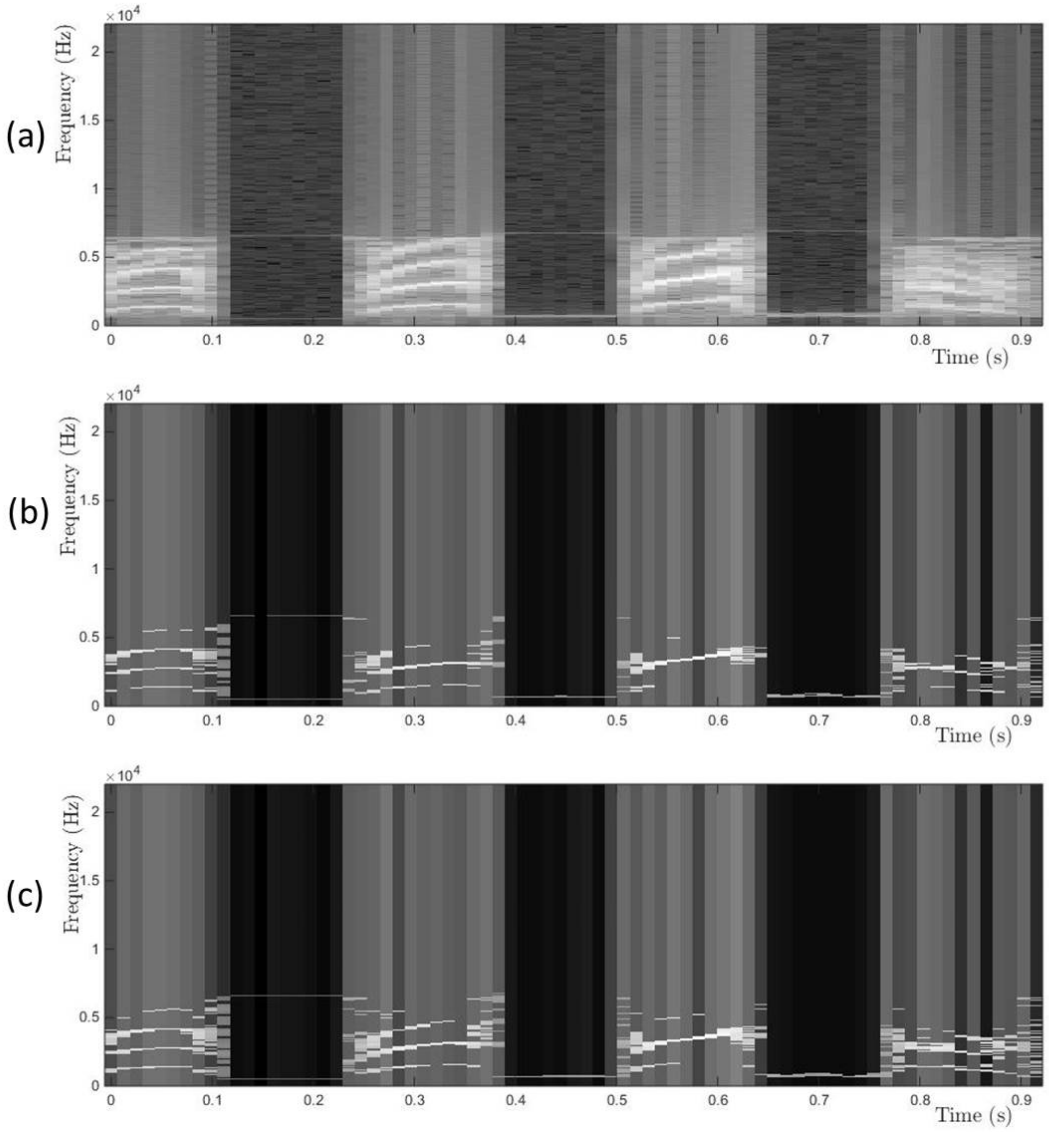


Figura 5.8 Ejemplo de umbrales espectrales en el espectrograma de una vocalización de un ave: a) thrp=0, b) thrp=0.1 y, c) thrp=0.05.

5.1.2 Extracción de características

Algunos sistemas de clasificación proponen a los Coeficientes Cepstrales en la Frecuencia Mel (MFCC) como la característica principal extraida de una vocalización de ave. En este sentido, tanto las ventajas y matemática respectiva es igual a la que se describió en la sección 3.4.

Por otro lado, los MFCC's son la base de los Coeficientes Cepstrales Inversos en la Frecuencia Mel (IMFCC). Los IMFCC's son tradicionalmente extraídos para el reconocimiento automático de hablantes como parte de modelos de fusión con otras técnicas [51, 52]. Junto con las ventajas de los MFCC's, la idea básica de los IMFCC's es el capturar la información acustica en la región de las altas frecuencias perdidas por los MFCC's (ver Figura 3.8 y 5.9).

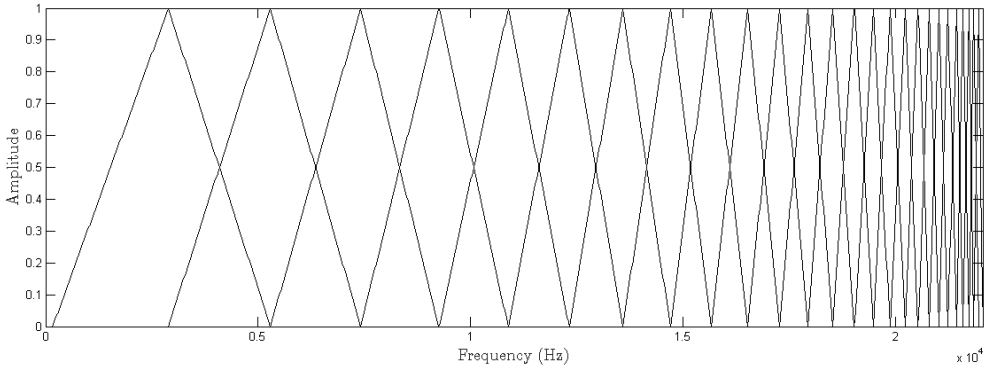


Figura 5.9 Estructura del banco de filtros IMFCC.

Para calcular el filtro Imel (inverso al banco de filtros Mel): i) determinar los limites frecuenciales del banco de filtros (f_{max} and f_{min}) y convertir las frecuencias a la escala Imel ($f_{imel}(f_{max})$ y $f_{imel}(f_{min})$ respectivamente) como:

$$f_{imel}(f_{max}) = \frac{Ns}{Fs}(f_{mel}(f_{max}) - 2595 * (\log_{10}(1 + (f_{max} - f_{max})./700)))), \tag{5.6}$$

$$f_{imel}(f_{min}) = \frac{Ns}{Fs}(f_{mel}(f_{max}) - 2595 * (\log_{10}(1 + (f_{max} - f_{min})./700)))). \tag{5.7}$$

Después, ii) calcular $p + 2$ puntos linealmente espaciados entre los puntos calculados Imel y, determinar la transformada inversa de los puntos (f_{imel}) a frecuencias (en Hz) de acuerdo

con:

$$K'_{sp}(j) = \frac{Ns}{Fs} f_{max} - 700 * (10^{(f_{mel}(f_{max}) - f_{imel}(j))/2595} - 1), \tag{5.8}$$

donde $\{K'_{sp}(j)\}_{j=0}^{p+1}$ son los puntos Imel linealmente espaciados en la escala de frecuencia. Luego, iii) la respuesta del filtro $M'_j(k)$ se calcula como:

$$M'_j(k) = \begin{cases} 0, & k \leq K'_{sp}(j-1) \\ \frac{k - K'_{sp}(j-1)}{K'_{sp}(j) - K'_{sp}(j-1)}, & K'_{sp}(j-1) \leq k \leq K'_{sp}(j) \\ \frac{K'_{sp}(j+1) - k}{K'_{sp}(j+1) - K'_{sp}(j)}, & K'_{sp}(j) \leq k \leq K'_{sp}(j+1) \\ 0, & (k \geq K'_{sp}(j+1)) \end{cases} \tag{5.9}$$

Después, para extraer los IMFCC's de las vocalizaciones de las aves: i) calcular la salida del banco de filtros Imel ($E'(j)$) mediante la multiplicación del espectro de energía de cada segmento ($y(k)$) y la ganancia del filtro ($M'_j(k)$) de acuerdo con:

$$E'(j) = \sum_{k=1}^{Ns/2-1} y(k) M'_j(k), \tag{5.10}$$

donde j es el índice de cada filtro en el banco de filtros. Después, ii) calcular la Transformada Discreta de Coseno (DCT-II) de acuerdo con:

$$C'_l(t) = \sum_{j=0}^{p-1} \log[E'(j)] \cos[t(j - \frac{1}{2})\frac{\pi}{p}], \tag{5.11}$$

donde $C_l(t)$ es el coeficiente IMFCC de t^{esimo} orden del l^{esimo} segmento y, t=1,...,13. Finalmente, iii) calcular el valor medio de los coeficientes por cada segmento para formar un vector de magnitudes. Este proceso, como en los MFCCs, se repite para cada conjunto de archivos de audio del mismo tipo de vocalización (los cuales conforman a su vez una matriz de magnitudes).

5.1.3 Sistema de clasificación de sonidos

La información extraída (MFCC's o IMFCC's) de cada conjunto de archivos de audio del mismo tipo de vocalización es usada para crear un Modelo Oculto de Markov (HMM). La matemática respectiva de los HMM es igual a la que se describió en la sección 3.4. Finalmente,

en la etapa de entrenamiento, dada secuencia (es decir una vocalización de ave), se calcula la probabilidad de haber sido generada por λ (y por tanto, generada por una especie de ave específica). Este método de clasificación se implementó usando el software gratuito descrito anteriormente [68].

5.2 Algoritmo de identificación acústica individual en aves basado en la función de correlación solo de fase limitada en banda

Justificación de Metodología: La densidad de aves en un area natural es un indicador que puede ser usado para determinar algunos indices como la supervivencia, reproducción, dispersión y crecimiento poblacional [75]. Este parámetro es medido a través de una técnica de búsqueda tradicionalmente basada en puntos de conteo, transectos lineales y redes de niebla [76–78]. Por otro lado, los puntos de conteo y los transectos lineales son usados para monitorizar cambios en la población mediante la identificación de aves por medio de la vista y el sonido. En este sentido, se debe garantizar algunas condiciones para la fiabilidad de las evaluaciones como la falta de movimiento en las aves antes de la detección y la independencia de los conteos [79]. Por otro lado, las redes de niebla son usadas para capturar individuos y determinar características como el sexo, la edad y la supervivencia [80]. Comúnmente se utilizan marcas artificiales en las aves capturadas para determinar entre otros, la dispersión de los habitats y el mapeo de territorios [81]. A pesar de que estas técnicas han mostrado buenos resultados, otras como la no estandarización de los métodos y la limitada disponibilidad de observadores expertos reducen su efectividad [78].

La identificación acústica individual en las aves involucra la extracción de características de las vocalizaciones (cantos y llamados) de un individuo que permitan su identificación entre otras vocalizaciones de individuos de la misma especies (o incluso de otras especies) [82]. Algunas de las técnicas estadísticas para el reconocimiento automático de hablantes comúnmente son aplicadas para la identificación individual de animales [83, 84]. Aunque estas técnicas poseen un buen desempeño en el reconocimiento automático de hablantes a través del uso de bases de datos de gran escala, algunas de ellas pueden proveer un desempeño inadecuado en condiciones de datos limitados (bases de datos pequeñas) [53]. En este sentido, para la identificación individual en condiciones reales, el obtener la cantidad (y calidad) de vocalizaciones de un individual para aplicar los métodos estadísticos de forma eficiente, es hoy en día una tarea compleja. En este contexto, se propone una nueva técnica para la identificación individual de aves basada en la función de correlación solo de fase limitada en banda(BLPOC). La función

BLPOC es una técnica biométrica de alta precision tradicionalmente aplicada en la identificación de personas a través de huellas digitales que, en el algoritmo propuesto, es ofrecido como un método de identificación individual de datos limitados.

Función de correlación solo de fase limitada en banda(BLPOC): La descripción de la función BLPOC quedó descrita en la sección 3.2 y solo se utiliza como referencia metodológica en esta sección.

Identificación acústica individual en aves: Una pregunta fundamental en la tarea de conservación de aves es el cómo identificar el número de individuos de una especie de ave en específico. De cualquier forma, la dificultad principal es el como extraer los mejores parámetros para describir las características inherentes de un individuo que permitan su identificación entre otros individuos de especies de aves [82].

La efectividad de la metodología tradicional para medir la población aviar depende de la experiencia de los observadores. Dado que las habilidades de los observadores varían dependiendo del nivel de audición, habilidad y experiencia, se dificulta el asegurar información confiable bajo estas condiciones [78]. En este contexto, las tecnologías computacionales modernas pueden ser aplicadas para desarrollar ayudas para los observadores [85].

En este sentido, el algoritmo propuesto, basado en la función BLPOC, identifica, a partir de la comparación de una grabación de una vocalización de ave (llamado o canto) pre-almacenada y una grabación entrante (de/o no el mismo individuo), si la grabación entrante es del mismo individuo (empate genuino) o no (empate impostor). El algoritmo propuesto posee dos variantes (dependiendo del método para analizar y extraer la vocalización de ave del archivo de audio): algoritmo de verificación automática de individuos y, algoritmo de verificación semiautomática de individuos los cuales se describen a continuación (ver Figura 5.10).

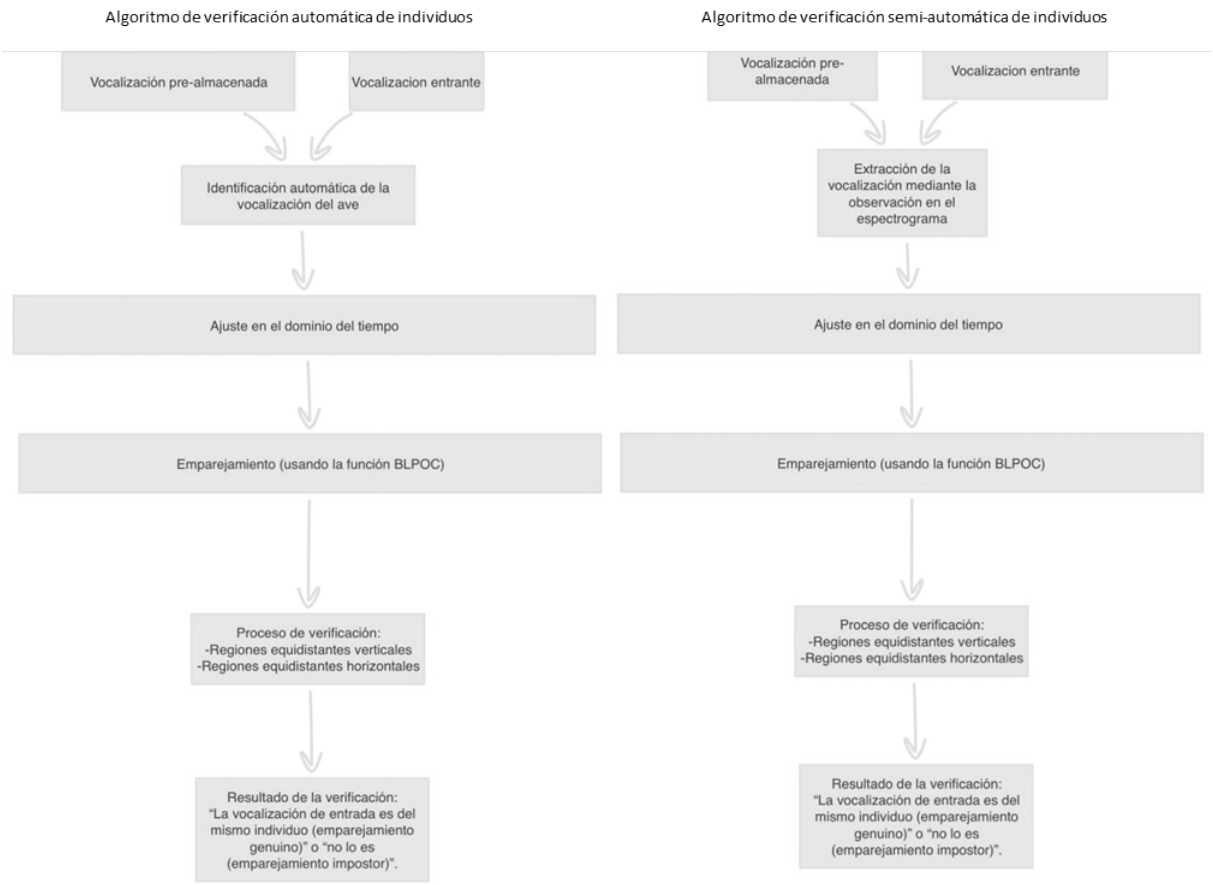


Figura 5.10 Algoritmo propuesto para la identificación acústica individual.

5.2.1 Algoritmo de verificación automática de individuos

Frecuentemente las grabaciones de audio de las vocalizaciones de aves suelen contener silencio y otros ruidos de fondo. Por tanto, el primer paso consiste en extraer los segmentos de audio (región) en la grabación, que corresponden a las vocalizaciones del ave (ver Figura 5.11). Considerando $f(n)$ como un archivo de audio de una vocalización de ave y, basado en un algoritmo para detección de actividad de voz (VAD) [86], la región de vocalización de ave (SF) se extrae de la misma manera como se indicó en la sección 3.3.1. Nuevamente, el valor de umbral de energía en los experimentos fue establecido como $thr_E = 0.001 * E$.

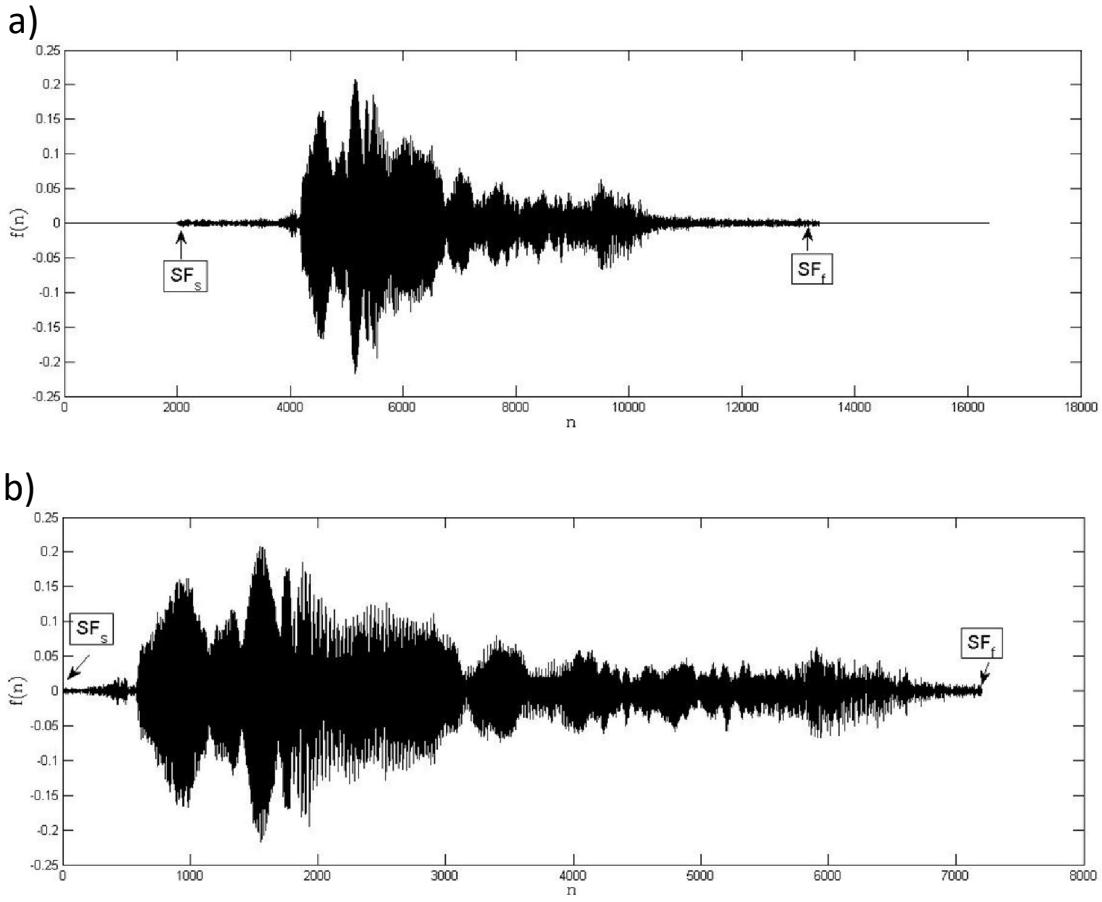


Figura 5.11 Identificación automática de la vocalización del Ave : Grabación de audio de una vocalización antes y después de la extracción automática de la vocalización.

5.2.1.1 Ajuste de tiempo y empate

Antes del empate entre grabaciones de vocalizaciones de aves, un ajuste es aplicado para reducir la variabilidad de la secuencia en el dominio temporal (ver Figura 5.12). Este proceso se realiza como se indicó en la sección 3.3.2.

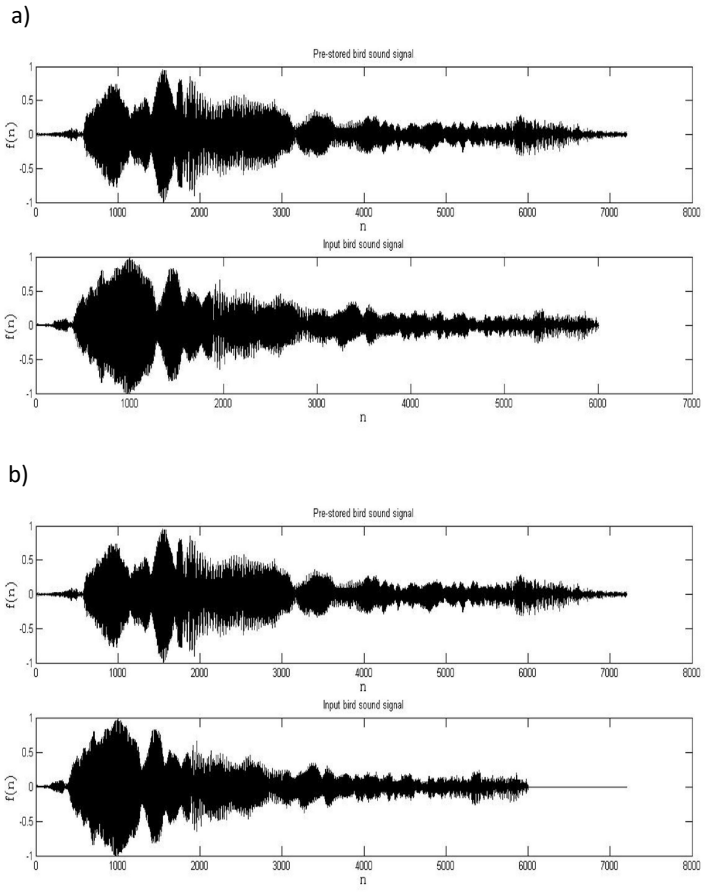


Figura 5.12 Ejemplo de una grabación de audio de una vocalización de ave a) antes y b) después del ajuste de tiempo.

La banda de frecuencia inherente se calcula siguiendo el mismo proceso que el descrito en la sección 3.3.3 (ver Figura 5.13). El valor de umbral para los experimentos se estableció como $thr_p = 0.001$. Finalmente, el empate entre las señales se calcula usando la ecuación (3.5) y el componente de frecuencia cero de la secuencia $\hat{r}_{fg}$ se desplaza al centro del espectro.

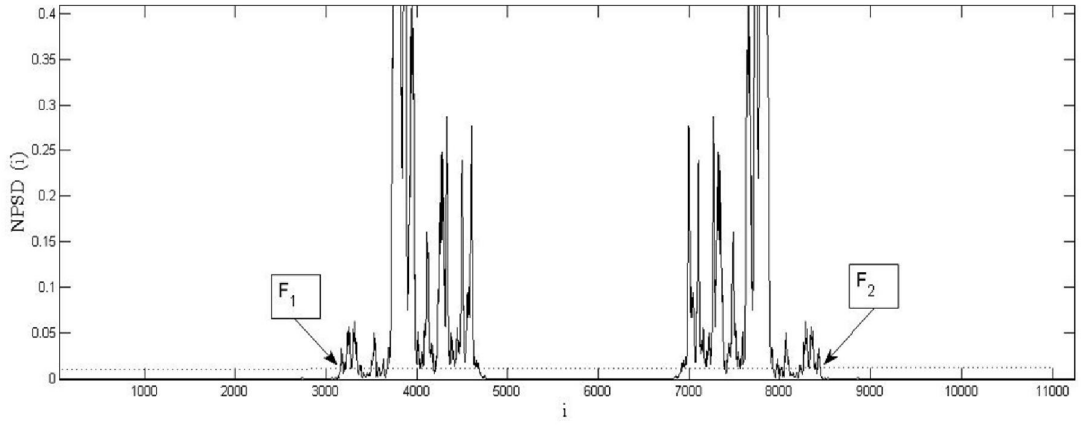


Figura 5.13 Ejemplo de la banda de frecuencia efectiva del espectro de fases cruzada de las señales comparadas. La línea punteada representa el umbral propuesto (thr_p).

5.2.2 Algoritmo de verificación semi-automática de individuos

5.2.2.1 Extracción de vocalizaciones por observación

Algunas de las habilidades de los observadores se ven afectadas por las condiciones de observación (especialmente cuando la vegetación incrementa o cuando el suelo desnudo decrementa) [77, 78]. Dado que en algunos casos la vocalización de un individuo puede ser la única fuente para la identificación, la calidad de la grabación juega un papel muy importante. Por otro lado, la variabilidad entre especies y sus vocalizaciones hace difícil el automatizar la extracción de las vocalizaciones en las grabaciones. Este proceso es equivalente al descrito en la sub sección 5.1.1.1

5.2.2.2 Ajuste de tiempo y empate

Los siguientes pasos (ajuste de tiempo y empate) son los mismos que los utilizados en el proceso del algoritmo de verificación automática de individuos que se describió anteriormente en la sub sección 5.2.1.1.

5.2.3 Decisión de verificación individual

La altura del pico máximo en la función BLPOC puede ser usada para medir similitud en la identificación humana a través del empate de imágenes; sin embargo, para la verificación de individuos de especies a través de vocalizaciones, algunas otras características (para determinar el empate genuino e impostor) necesitan ser calculadas. Dado que la función BLPOC tiene muchas formas, dependiendo en la vocalización verificada (ver Figura 5.14), el criterio propuesto para evaluar la función BLPOC es un proceso automático de verificación paramétrica basado en la forma de la función BLPOC en el emparejamiento genuino.

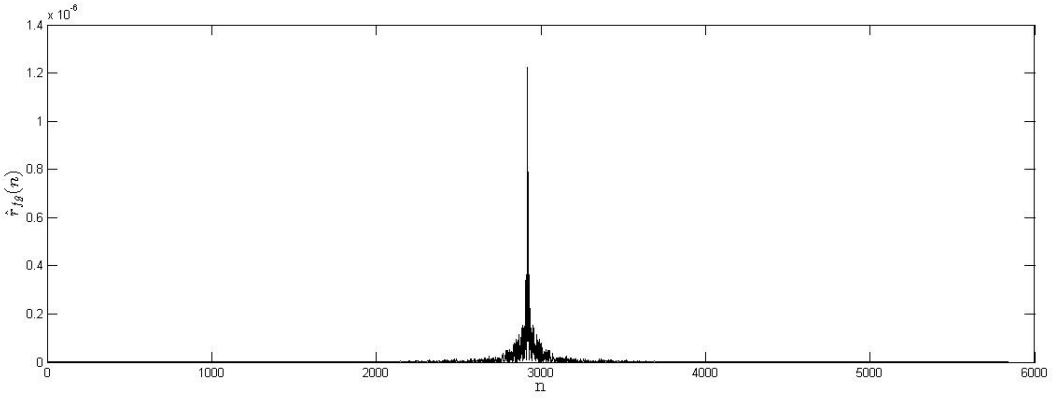
5.2.3.1 Proceso de verificación mediante regiones equidistantes verticales

El primer paso en el proceso de verificación es la evaluación de la ubicación del pico máximo en la función BLPOC mediante el proceso siguiente (ver Figura 5.15): i) determinar el numero de muestras en la función BLPOC, ii) calcular $m+1$ indices equidistantes ($ei(i)$) entre muestras y, iii) calcular la $i-esima$ región vertical equidistante (evr_i) de acuerdo con:

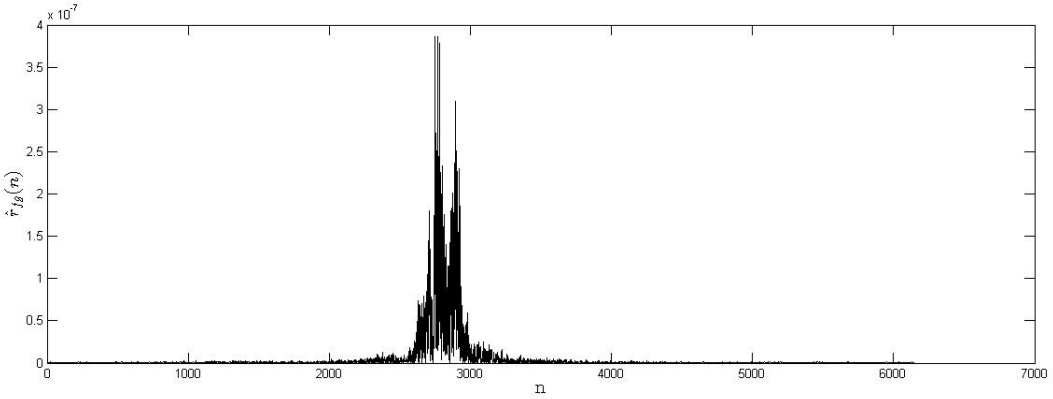
$$evr_i = [\hat{r}_{fg}(ei(i)), \ldots, \hat{r}_{fg}(ei(i+1))]. \tag{5.12}$$

Luego, iv) determinar el índice del pico máximo en la función BLPOC (P_M) y la región vertical equidistante a la que pertenece. Finalmente, v) comparar la región identificada con la región vertical equidistante de aceptación. El número de indices equidistantes en los experimentos fue un valor constante de $m = 5$ y, por otro lado, la región de aceptación fue establecida como la 3era región vertical equidistante.

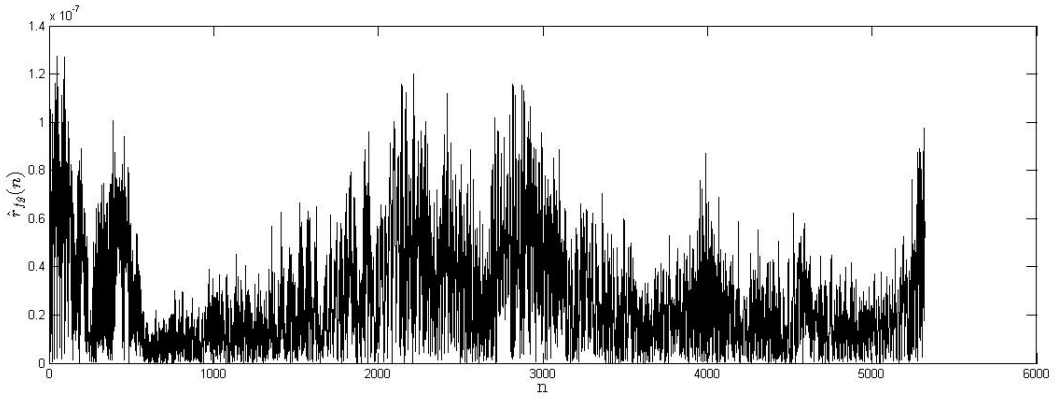
Por otro lado, dado que la función BLPOC puede tomar distintas formas, el siguiente paso es establecer la forma de aceptación de las muestras en las regiones verticales equidistantes. En este sentido, para el empate genuino, la energía se concentra en la región donde se ubica el pico máximo de la función BLPOC. Así, para identificar la forma de aceptación de la función BLPOC se procede de la siguiente manera (ver Figura 5.15): i) calcular la energía de la función BLPOC (E_i) en la $i-esima$ región vertical equidistante (evr_i) de acuerdo con:


x 10^-6
1.4
1.2
1
0.8
0.6
0.4
0.2
0
$\hat{r}_{fg}(n)$
0
1000
2000
3000
4000
5000
6000
n

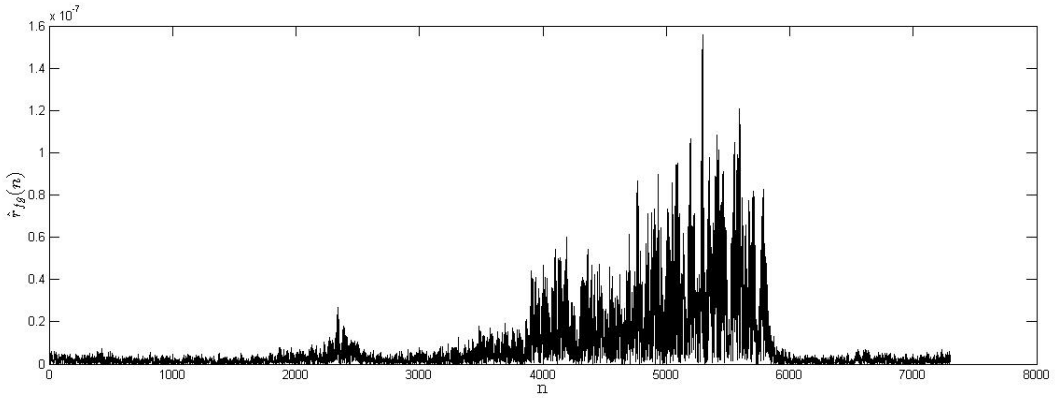
(a)


x 10^-7
4
3.5
3
2.5
2
1.5
1
0.5
0
$\hat{r}_{fg}(n)$
0
1000
2000
3000
4000
5000
6000
7000
n

(b)



(c)



(d)

Figura 5.14 Ejemplos de empate genuino e impostor: a) función BLPOC entre dos vocalizaciones idénticas del mismo individuo, b) función BLPOC entre dos vocalizaciones similares del mismo individuo, c) función BLPOC entre dos vocalizaciones similares de dos individuos diferentes de la misma especie y, d) función BLPOC entre dos diferentes individuos de diferentes especies de aves.

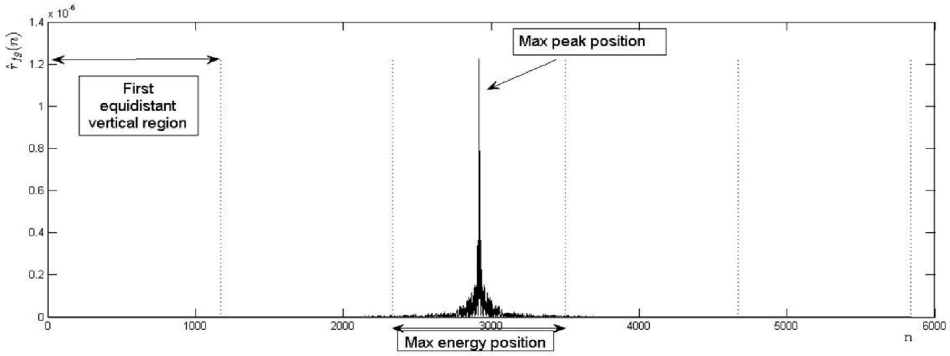


Figura 5.15 Forma aceptada de la función BLPOC a través de regiones verticales equidistantes.

$$E_i = \frac{1}{k}\sum_{j=1}^{k} |evr_i(j)|^2, \tag{5.13}$$

donde $i = 1, \ldots, m$ y k es el numero de muestras en evr_i. Después ii), identificar la región vertical equidistante con la energía máxima. Finalmente, iii) comparar la región identificada con la locación de aceptación de máxima energía. En los experimentos, la región aceptación de máxima energía en la función BLPOC se estableció como la tercera región vertical equidistante (ubicación de aceptación del pico máximo en la función BLPOC). Ver ejemplos del proceso de verificación de la función BLPOC a través de regiones verticales equidistantes en la Figura 5.16.

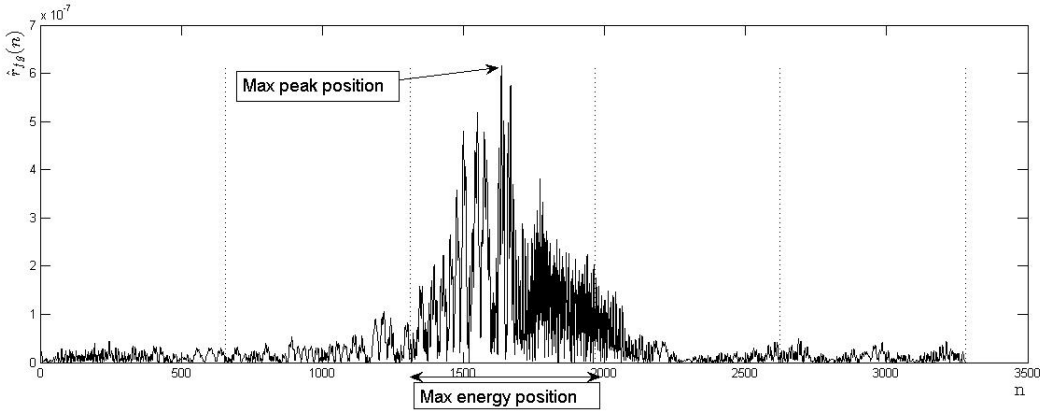
5.2.3.2 Proceso de verificación mediante regiones equidistantes horizontales

Para evitar errores de verificación (ver Figura 5.16 c)), se realiza una evaluación de la magnitud de los picos menos significativos en las regiones verticales de no aceptación (ver Figura 5.17). Dicha evaluación se realiza de la siguiente manera: i) calcular el pico máximo y mínimo de la función BLPOC y ii) calcular $m + 1$ intervalos de magnitud ($im(i)$) entre la magnitud de los picos calculados. Después, iii) calcular el pico máximo en la $i - esima$ región vertical equidistante evr_i mediante:

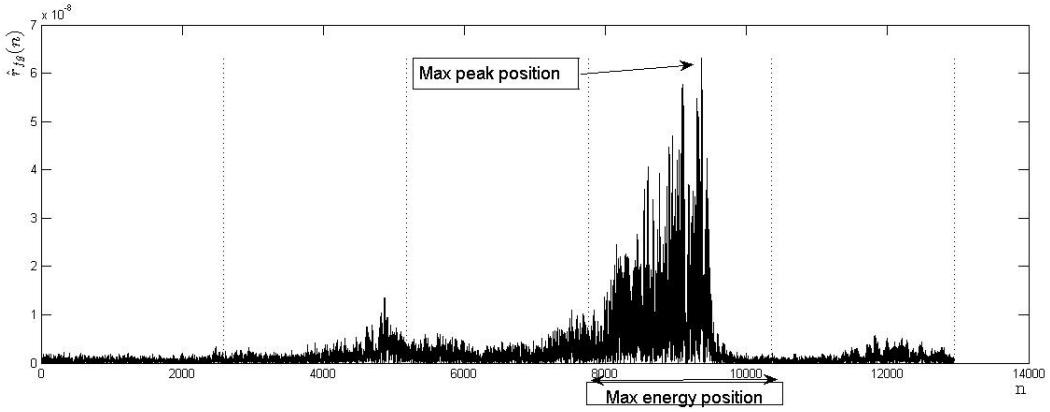
$$peak(i) = \max |segv_i|. \tag{5.14}$$

Después, iv) comparar los picos identificados ($peak(i)$) con la región horizontal de máxima aceptación. El número de intervalos de magnitud en los experimentos fue establecido como un valor constante de $n = 5$ y, por otro lado, la región de máxima aceptación se estableció como la 3era región horizontal equidistante (ver Figura 5.18).

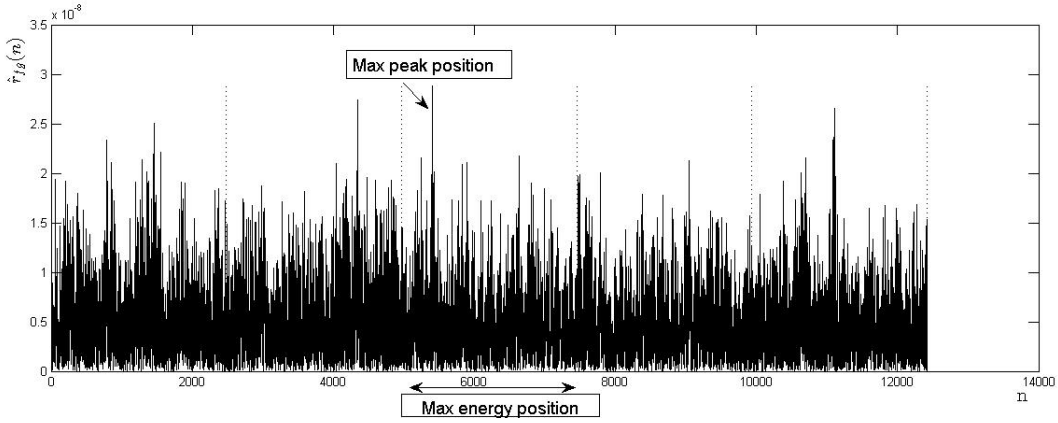
Las deformaciones por subdominios introducen posibles errores (ver Figura 5.19). En los experimentos se implementó un criterio especial basado en la evaluación del numero de picos menos significativos mayores que la región horizontal de máxima aceptación.


x 10^-7
Max peak position
Max energy position
n

(a)


x 10^-8
Max peak position
Max energy position
n

(b)



(c)

Figura 5.16 Ejemplo del proceso de verificación de la función BLPOC a través de regiones verticales equidistantes: a) empate genuino correctamente verificado como aceptado, b) empate impostor correctamente verificado como rechazado y, c) empate impostor incorrectamente verificado como aceptado.

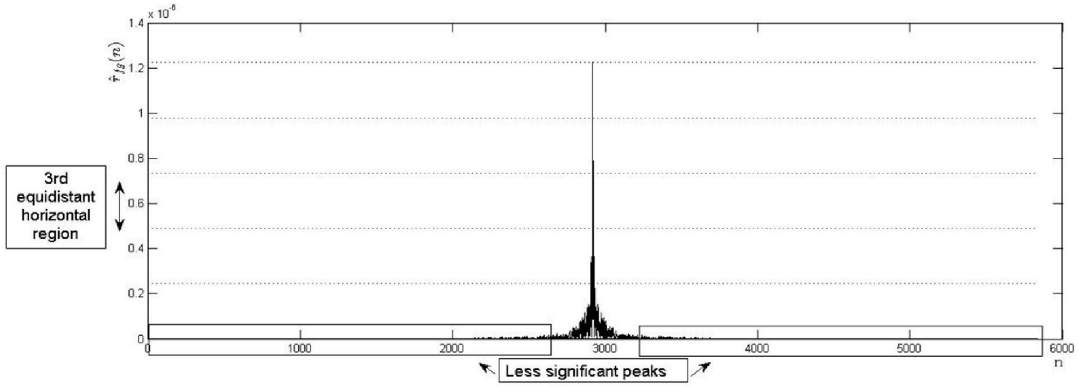


Figura 5.17 Forma aceptada de la función BLPOC a través de regiones horizontales equidistantes.

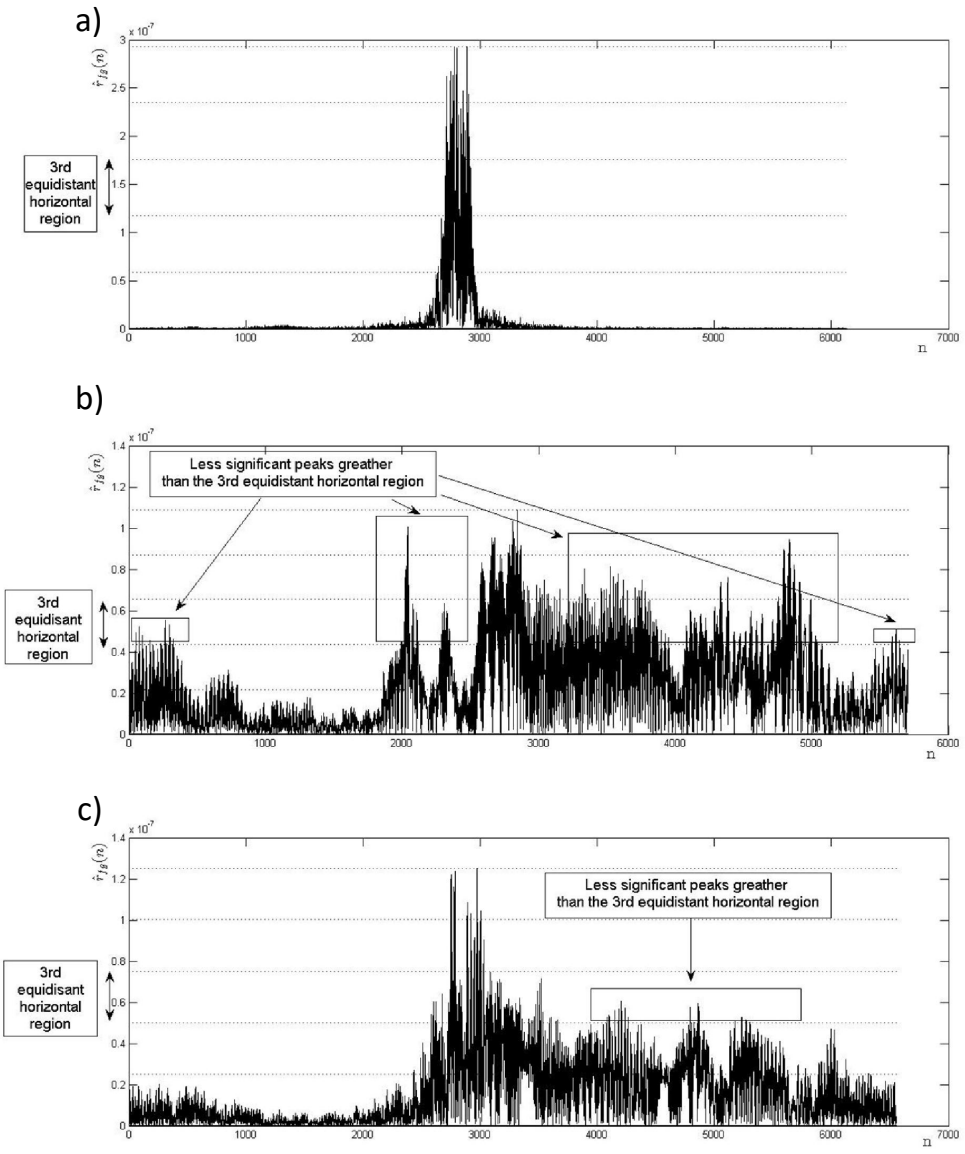


Figura 5.18 Ejemplo del proceso de verificación de la función BLPOC a través de regiones horizontales equidistantes: a) empate genuino correctamente verificado como aceptado y, b) y c), empate impostor correctamente verificado como rechazado.

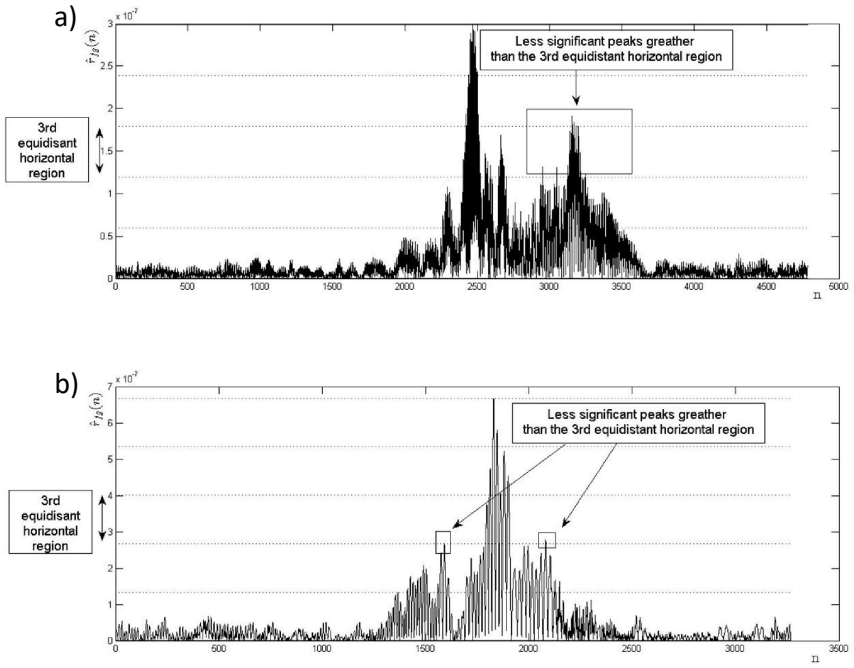


Figura 5.19 Ejemplos de posibles errores debido a deformaciones de sub-dominio: a) y b) son empates genuinos incorrectamente verificados como rechazados.

Capítulo 6

Resultados y conclusiones de reconocimiento de especies y de individuos

6.1 Resultados y discusión: Algoritmo de clasificación automática de especies de aves basado en Coeficientes Cepstrales Inversos en la frecuencia Mel

Las grabaciones de audio, para evaluar el algoritmo propuesto, se recolectaron de dos diferentes bases de datos de libre acceso: Xeno-canto y, la biblioteca de sonidos de aves de México del Instituto de Ecología, A.C (INECOL). Por un lado, Xeno-canto es una base de datos voluntaria con más de 192,000 audios de especies de aves de alrededor del mundo [87]. Por otro lado, la biblioteca de sonidos de aves de México de INECOL es una pequeña base de datos utilizada para estudiar y difundir audios de especies de aves en México [88]. Las especies de aves pueden producir diferentes vocalizaciones y, dado que la base de datos de INECOL es actualmente limitada, solo se realizó una selección parcial de los audios disponibles en las bases de datos (mediante la selección de vocalizaciones en la base de datos en INECOL con más grabaciones en la base de Xeno-canto). La Tabla 6.1 muestra el número de grabaciones por especie en la base de datos recolectada.

El algoritmo propuesto basado en IMFCC's se evaluó, comparándolo contra la extracción de MFCC's, a través de los siguientes parámetros:

Tabla 6.1 Grabaciones de las especies de aves en la base de datos recolectada.

Orden	Familia	Especie	Grabaciones
Passeriformes−Perching Birds	Furnariidae	Automolus rubiginosus	35
		Synallaxis erythrothorax	54
	Cardinalidae	Cardinalis cardinalis	41
	Thamnophilidae	Cercomacra tyrannina	15
	Tyrannidae	Myiozetetes similis	33

- Tasa de aceptación verdadera (TAR): Representa la probabilidad de que una vocalización de ave sea correctamente identificada como: "generada a partir de una especie especifica de ave".

- Tasa de rechazo verdadero (TRR): Representa la probabilidad de que una vocalización de ave sea correctamente rechazada al ser identificada como "no generada por una especie específica de ave".

Se consideró para la prueba TAR (para cada conjunto de audios del mismo tipo de vocalización) el 70% de archivos de audios para la etapa de entrenamiento; y el resto para la etapa de prueba. Por otro lado, para la prueba TRR, el total de archivos de otras especies en cada prueba TAR se utilizó en la etapa de prueba.

Tabla 6.2 Resultado de la prueba de clasificación de especies de aves (%).

	MFCC		IMFCC	
Especie de ave	TAR	TRR	TAR	TRR
Automolus rubiginosus	63.63	34.78	100	43.47
Synallaxis erythrothorax	100	28.57	100	61.90
Cardinalis cardinalis	76.92	23.80	84.61	0
Cercomacra tyrannina	0	0	0	0
Myiozetetes similis	30	4.1	10	61.90

De la Tabla 6.2 se encuentra que mediante la extracción de los IMFCC's, en comparación con los tradicionales MFCC's para las especies Automolus rubiginosus y Cardinalis cardinalis, se obtiene una mejora en la TAR (37% y 0.8% respectivamente). Además, la TAR es igual (si se extraen los MFCC's o IMFCC's) para la especie Synallaxis erythrothorax. Por otro lado, se muestra una mejora general en la TRR mediante la extracción de los IMFCC´s (especialmente para la especie Myiozetes similis. Finalmente, para la especie Cercomacra tyrannina se muestra una eficiencia aleatoria para ambos parámetros.

6.1.1 Consideraciones finales

Junto a la variabilidad de vocalizaciones entre las especies de aves, los patrones espectrales de las vocalizaciones a reconocer tienen una influencia directa en la eficiencia del algoritmo propuesto (ver Figura 6.1). Por tanto, la extracción de la región más importante se espera que tenga un desempeño mejor cuando la vocalización es fácil de identificar en el espectrograma.

(a)

(b)

(c)

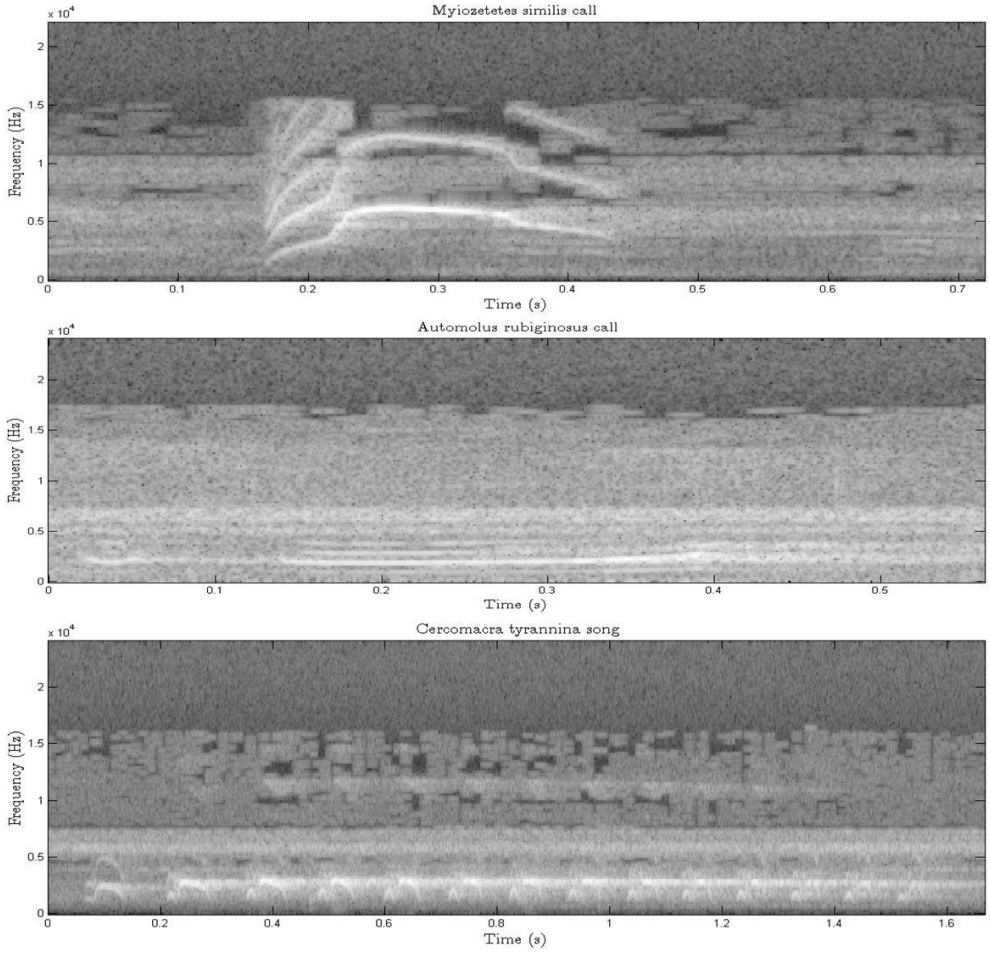


Figura 6.1 Ejemplos de espectrograma de las vocalizaciones de: a) Myiozetetes similis, b) Automolus rubiginosus y, c) Cercomacra tyrannina.

Aún que los resultados mostrados (en la Tabla 6.2) son menores que las eficiencia de MFCC's y IMFCC's para el reconocimiento de habla automática [14, 15, 51, 52], se muestra una mejora general en la TAR y TRR mediante la extracción de los IMFCC's. Por lo tanto, para un sistema de clasificación de especies basado en vocalizaciones de aves, la extracción de información acustica en las altas frecuencias (capturada por los IMFCC's) es tan significativa como aquella en las bajas frecuencias (capturada por los MFCC's).

6.2 Resultados y discusión: Algoritmo de identificación acústica individual en aves basado en la función de correlación solo de fase limitada en banda

Las grabaciones de audio utilizadas para la evaluación del algoritmo de identificación individual, como en el caso anterior, se recolectaron de las bases de datos de Xeno-canto y, la biblioteca de sonidos de aves de México de INECOL. La Tabla 6.3 muestra la lista de especies de aves y el número de grabaciones recolectadas por especie.

Tabla 6.3 Grabaciones de las especies de aves en la base de datos recolectada.

Orden	Familia	Especie	Grabaciones
Passeriformes−			
Perching Birds	Furnariidae	**a**) Automolus rubiginosus	35
		b) Synallaxis erythrothorax	78
	Cardinalidae	**c**) Cardinalis cardinalis	41
	Thamnophilidae	**d**) Cercomacra tyrannina	16
	Troglodytidae	**e**) Hylorchilus sumichrasti	18
	Tyrannidae	**f**) Myiozetetes similis	74
		g) Pitangus Sulphuratus	10
	Icteridae	**h**) Psarocolius montezuma	15
	Psittacidae	**i**) Amazona viridigenalis	9

Por otro lado, se propone una subcategoria de especies según la complejidad de las vocalizaciones (ver ejemplos en Figura 6.2). Como consecuencia, de la Tabla 6.3, las siguientes especies son consideradas, para esta investigación, como las que poseen vocalizaciones complejas:

- Cercomacra tyrannina,
- Hylorchilus sumichrasti,
- Pitangus sulphuratus,
- Psarocolius montezuma,
- Amazona viridigenalis.

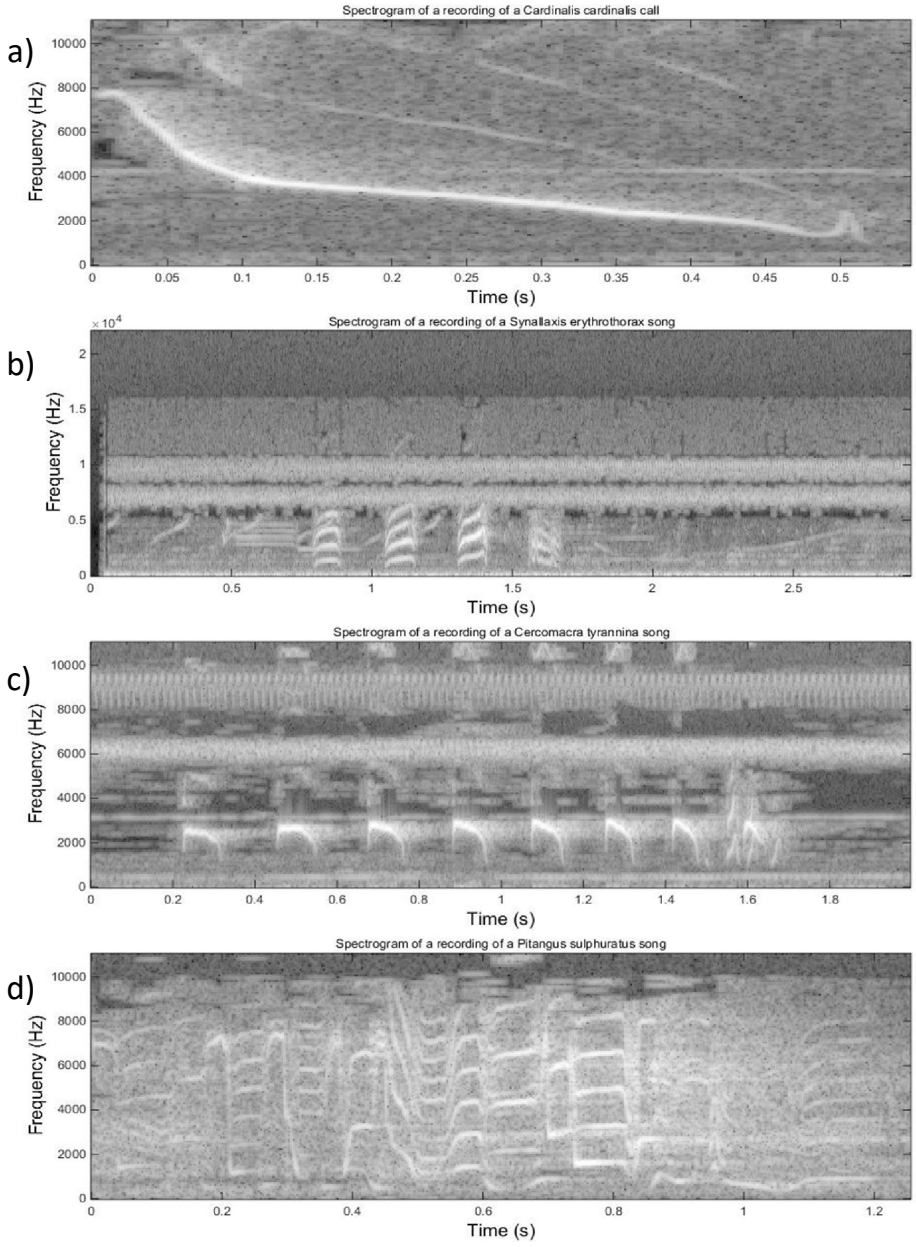


Figura 6.2 Ejemplos de complejidad en vocalizaciones: a) y b) son vocalizaciones no complejas; c) y d) son vocalizaciones complejas.

Se realizó una evaluación de la capacidad de descarte del algoritmo para garantizar una correcta identificación de individuos entre especies. En otras palabras, calcular la probabilidad

de que una vocalización sea correctamente rechazada por el algoritmo al ser identificada como "no generada por una especie de ave específica". En este sentido, la Tabla 6.4 muestra que el desempeño del algoritmo automático de verificación de individuos es superior al 90% para casi todas las comparaciones de identificación entre especies.

Tabla 6.4 Resultado del algoritmo de verificación automática en la prueba de descarte (%).

VS	**a)**	**b)**	**c)**	**d)**	**e)**	**f)**	**g)**	**h)**	**i)**
a)	.	92.5	96.5	100	100	99.22	100	97.90	100
b)	92.54	.	98.21	98.36	91.17	98.75	98.71	96.02	99.15
c)	86.4	96.08	.	100	100	69.68	94.66	100	93.6
d)	100	96.18	97.22	.	100	98.6	97.14	100	100
e)	100	91.38	100	95.5	.	98.97	92.85	100	100
f)	99.69	99.34	69.09	99.39	98.93	.	93.63	100	100
g)	100	90	84.16	100	100	99.66	.	100	100
h)	98.86	94.4	100	100	98.75	100	98.21	.	96.42
i)	100	100	98.57	97.4	100	100	100	100	.

El segundo criterio de evaluación se llevó a cabo mediante una prueba de evaluación de individuos de la misma especie. En otras palabras, calcular la probabilidad de que un individuo sea verificado como aceptado (para el empate genuino) o rechazado (para el empate impostor) por cada una de las dos variantes del algoritmo propuesto. Dado que este proceso representa un proceso combinatorio entre grabaciones, solo se utilizó una selección parcial de las grabaciones para la prueba del algoritmo semi-automático; mientras que, para el algoritmo automático, se utilizaron todas las grabaciones en la base de datos. La Tabla 6.5 muestra las eficiencia del algoritmo en la verificación de individuos. Por un lado, el algoritmo semi-automático posee aproximadamente una eficiencia del 86.53% en la verificación de individuos cuya especie posee vocalizaciones no complejas (y aproximadamente un máximo de 55% en la

verificación de individuos cuya especie posee vocalizaciones complejas). Por otro lado, el algoritmo automático posee una eficiencia de 81.93% y un máximo de 85.71%, en la verificación de individuos de especies con vocalizaciones no complejas y complejas, respectivamente.

Tabla 6.5 Resultados de la prueba de desempeño de los algoritmos en la verificación de individuos (%). Las especies con vocalizaciones están escritas en itálicas.

Especie	Algoritmo semi-automático	Algoritmo automático
a) Automolus rubiginosus	85.58	75.84
b) Synallaxis erythrothorax	81.6	76
c) Cardinalis cardinalis	96.94	92.98
d) *Cercomacra tyrannina*	18.18	82.81
e) *Hylorchilus sumichrasti*	55	56.5
f) Myiozetetes similis	82	82.9
g) *Pitangus sulphuratus*	25	75.78
h) *Psarocolius montezuma*	46.6	85.71
i) *Amazona viridigenalis*	29.4	55.55

6.2.1 Consideraciones finales

Clasificación de especies de aves: Aún que el algoritmo de verificación propuesto compara vocalizaciones de diferentes individuos de la misma especie, la Tabla 6.4 muestra que este puede discriminar incluso entre especies. En otras palabras, esto significa que la verificación automática de la función BLPOC puede ser útil también en la clasificación de especies. Por otro lado, el desempeño de descarte del algoritmo de verificación semi automática depende de la capacidad humana del usuario para discriminar vocalizaciones entre especies. En este sentido, un oído humano entrenado (en la identificación de vocalizaciones de aves) puede mejorar dicho desempeño de descarte (especialmente para las vocalizaciones de aves que tienen sonidos similares entre si).

Identificación individual de aves: A pesar de que el algoritmo automático muestra mejores eficiencias para las vocalizaciones complejas (ver Tabla 6.5), esto es principalmente debido a la capacidad de descarte del algoritmo en el empate impostor. En este sentido, un algoritmo automático de extracción de vocalizaciones más especializado puede mejorar la eficiencia del algoritmo automático propuesto en el empate genuino. Por otro lado, como se esperaba, el algoritmo semi automático muestra mejores eficiencias para vocalizaciones no complejas debido a que la información principal de las vocalizaciones en las grabaciones es fácil de extraer del espectrograma.

De las condiciones de grabación: Debido a las diferentes calidades de sonido en las grabaciones, la base de datos colectada no posee una condición de grabación especifica (incluso entre las grabaciones de individuos de la misma especie de ave). En este contexto, el algoritmo propuesto puede verificar un individuo incluso si existen diferencias en las condiciones de grabación. En cualquier caso, se recomienda para la confiabilidad en la eficiencia del algoritmo, el preferir condiciones de grabación similares para los empates de identificación.

Áreas de aplicación: Aún no es bien entendido si la captura de un individuo tiene un efecto negativo en los otros individuos; en cualquier caso, la salud de las aves es primordialmente importante. Por lo tanto, es necesario tomar especial cuidado en la identificación individual a través de las redes de niebla para evitar los problemas comunes de captura (como el estrés o la muerte de los individuos) [78, 81]. Por otro lado como se mencionó en la introducción, obtener la cantidad (y calidad) de vocalizaciones de un individuo para aplicar los métodos estadísticos tradicionales de forma eficiente, es hoy en día un desafío. En este contexto, que el algoritmo propuesto pueda proveer una evaluación de la identificación de individuos solo mediante el análisis de dos muestras de audio de los individuos (una condición de datos limitados) implica una reducción de tiempo-costo y una nueva forma sencilla de monitoreo de biodiversidad. Aún que el algoritmo propuesto no extrae las mismas características como aquellas que se pueden extraer de un individuo capturado, este es un método no invasivo para la identificación individual que, además de su aplicación en conteo, puede ser usado para determinar la dispersión del hábitat y el mapeo de territorios.

Debido a las habilidades particulares de los observadores experimentados, se recomienda que solo los observadores calificados participen en los programas de monitoreo de biodiversidad [78]. Aún que esta es una condición deseada, frecuentemente existe una falta de observadores calificados principalmente debido a la cantidad de tiempo que consumen las identificaciones [79]. Por otro lado, el entrenamiento de nuevos observadores requiere de un programa de entrenamiento intensivo que dura de dos a tres meses (y dependiendo del número de especies a identificar en el entrenamiento) [81]. Por otro lado, la identificación acustica de aves ha mostrado mayor eficiencia en reconocimiento que la identificación visual especialmente por que la identificación acústica maximiza las identificación visual [79]. Por esta razón, para incrementar la eficiencia de identificación se recomienda el combinar técnicas visuales y acústicas. En este contexto, el algoritmo propuesto se ofrece como una herramienta que puede ser aplicada junto con las técnicas tradicionales de identificación. Además, debido a que el algoritmo de identificación semi-automática se basa en la identificación de patrones en el espectrograma, su metodología puede ser usada para el entrenamiento de nuevos observadores y, para ayudar a los observadores experimentados en la confiabilidad de los conteos aviares a través de las metodologías de búsqueda.

6.3 Conclusiones

6.3.1 Algoritmo de clasificación automática de especies de aves basado en Coeficientes Cepstrales Inversos en la frecuencia Mel.

El algoritmo propuesto en esta sección propone un método de clasificación automática de especies de aves basado en la extracción de las características IMFCC de las vocalizaciones. Aunque los patrones de vocalización espectral de las aves a ser reconocidas tiene una influencia directa en la eficiencia del algoritmo, se obtuvo una mejora general en las métricas TAR y TRR mediante la extracción de los IMFCC's (comparada con la eficiencia de reconocimiento en la extracción de MFCC's). De los resultados se concluye que la información acústica de las vocalizaciones de aves en las altas frecuencias (capturadas por los IMFCC's) es tan significativa como la información acústica en las bajas frecuencias (capturadas por los MFCC´s) para la

clasificación de aves a través de vocalizaciones. Además, basado en los resultados obtenidos, se piensa que el uso de una base de datos de gran escala y de un modelo de fusión MFCC-IMFCC podría mejorar la eficiencia de reconocimiento del algoritmo propuesto.

6.3.2 Algoritmo de identificación acústica individual en aves basado en la función de correlación solo de fase limitada en banda.

Por otro lado, se propone una nueva técnica para la identificación individual de aves mediante la función BLPOC. De las pruebas de desempeño se puede concluir que las eficiencias del algoritmo dependen de la complejidad de las vocalizaciones. Por otro lado, dado que el algoritmo puede discriminar incluso entre especies y no solo entre individuos, esta es una herramienta útil para la clasificación de especies (mediante un esquema clasificación adecuada). De esta manera luego de las pruebas, el algoritmo resulta en un método efectivo para la identificación acústica de individuos de aves cuya principal ventaja es que puede proveer de una evaluación de identificación mediante el análisis de solamente un par de vocalizaciones. Además de la identificación individual, se piensa que la técnica de identificación propuesta pudiera ser usada junto con las técnicas tradicionales de conteo para determinar el mapeo de habitats y territorios.

6.3.3 Conclusiones Generales.

Mi opinión personal del trabajo desarrollado es que la investigación en bioacústica es de suma importancia para el cuidado ambiental. Desde esta perspectiva, el uso de nuevas tecnologías o la aplicación de técnicas tradicionales de otras áreas de conocimiento pueden ser útiles en temas de bioacústica siempre y cuando se comprenda de manera puntual la necesidad a resolver. Es aquí donde el procesamiento de señales (en el contexto general) es una herramienta invaluable.

En realidad, y luego de la experiencia ganada en el análisis de la literatura disponible al momento de realizar esta investigación, la bioacústica enfocada en aves es una rama de investigación aún con mucho potencial. Por ello, desde mi perspectiva, es un nicho de oportunidad para el desarrollo de nuevas aplicaciones. Específicamente, el reconocimiento de voz es una

técnica con bases bien fundamentadas con las que la ciencia ha logrado (y sigue logrando) grandes avances. Sin embargo, cuando se aplican estas técnicas en bioacústica de aves se transforma en algo novedoso. Hemos de notar por tanto que, si este nicho de oportunidad es detectado en la aplicación en aves, cuanto más no existirá para cualquier otro animal que pueda ser producto de estudio para la bioacústica. De los resultados del trabajo presentado me parece que el uso de la tecnología utilizada en la investigación presente es suficiente. Es decir, luego de analizar el uso de otras técnicas homólogas, las presentadas, aún que pueden ser perfeccionadas, pueden ser la base para el desarrollo de nuevas investigaciones. De modo particular considero que el hecho de descubrir que se puede desarrollar un reconocimiento entre individuos de aves por medio de sonido conlleva a una novedad desde el punto de vista en que se creería que todas las aves eran iguales y que no existía "individualidad". Sin embargo, luego de realizar esta identificación, se concluye y soporta la idea de que aún en la naturaleza existe la individualidad de características. Por otro lado, aún que el reconocimiento de voz ha utilizado ampliamente a los MFCC´s por su sustento en las características de audición humanas, es sorprendente considerar que la extracción de otras regiones en señales biológicas (en este caso, cantos y llamados de aves) puede tener un impacto en el reconocimiento. De este punto, nace la interrogante de si es adecuado utilizar un sistema basado en la percepción de sonido humano y para tal caso, saber si sería posible desarrollar un sistema basado en la percepción de sonido pero del organismo bajo estudio (en este caso las aves).

Por ultimo me gustaría concluir esta investigación diciendo que me encuentro satisfecho con los resultados del mismo que, más allá de un requisito para obtener el grado, abrió mi perspectiva del mundo natural y una vocación de investigación en estos temas. Para mí, y luego de concluir con esta investigación, me doy cuenta que aún existen muchas cosas que se pueden hacer en el área de bioacústica y sería excelente dedicar una vida a estos temas tan apasionantes. A continuación, y como parte de los anexos, presento los artículos desarrollados durante el periodo de la presente investigación. Entre ellos presento algunos artículos al momento de la redacción del presente documento se encuentran aún por mandar a revisión y publicación.

Apéndice
Artículos publicados sobre reconocimiento de voz

Turkish Journal of Electrical Engineering & Computer Sciences

http://journals.tubitak.gov.tr/elektrik/

Research Article

Turk J Elec Eng & Comp Sci
(2019) 27: 3150 – 3164

doi:10.3906/elk-1805-113

Limited-data automatic speaker verification algorithm using band-limited phase-only correlation function

Ángel David PEDROZA RAMÍREZ*, **José Ismael DE LA ROSA VARGAS**, **José de Jesús VILLA HERNÁNDEZ**, **Aldonso BECERRA SÁNCHEZ**
Faculty of Electrical Engineering, Autonomous University of Zacatecas, Zacatecas, Mexico

Received: 17.05.2018 • **Accepted/Published Online:** 08.05.2019 • **Final Version:** 26.07.2019

Abstract: In this paper, a new method to deal with automatic speaker verification based on band-limited phase-only correlation (BLPOC) is proposed. The aim of this study is to validate the use of the BLPOC function as a new limited-data automatic speaker verification technique. Although some speaker verification techniques have high accuracy, efficiency usually depends on the extraction of complex theoretical information from speech signals and the amount of the data for training the algorithms. The BLPOC function is a high-accuracy biometric technique traditionally implemented in human identification by fingerprints (through image-matching). When applying the BLPOC function in automatic speaker verification through the proposed algorithms (under limited-data conditions), a 98.24% true acceptance rate (TAR) and 87.17% true rejection rate (TRR) in a custom database (and 93.75% TAR and 67.05% TRR in the ELSDSR database) were obtained. The proposed algorithm is a theoretically simple method for automatic speaker verification whose main advantage is that it can provide identification under limited-data conditions. In this sense, the BLPOC function could be applicable in other limited-data biometric identifications by sound signals.

Key words: Biometrics, limited data, speaker verification, text-dependent

1. Introduction

Speech communication is an elementary process that involves some organs' coordination, word articulation, and brain processes, by which human beings evolved as a society. Automatic speech recognition is the machine recognition of words spoken by speakers. Automatic speaker identification focuses on the identification of a speaker by classifying a speech signal as being from a specific speaker among speakers in the dataset. Since the system knows the impostors, this is closed-set identification [1–4]. The speaker verification task is to confirm if a speaker is who he/she claims to be (a yes or no decision) by extracting and analyzing some speech parameters from speakers. Since the system does not know the impostors because they can not be defined, this is open-set identification [1].

A speaker verification system requires, among other steps, to extract the relevant information from speech signals and an appropriate training and test scheme. Among others [5, 6], one of the usual features extracted from voice signals are mel frequency cepstral coefficients (MFCCs). MFCC features are based on the human auditory system and are a common front-end for a speaker identification system [7]. On the other hand, to compare the extracted information from speech, statistical models are commonly used [8, 9]. Modeling requires extracting some voice distribution parameters from the speech signals uttered by the speakers in the dataset and

*Correspondence: p.a.d_16@hotmail.com

Apéndice
Artículos publicados sobre Bioacústica

A comparative between Mel Frequency Cepstral Coefficients (MFCC) and Inverse Mel Frequency Cepstral Coefficients (IMFCC) features for an Automatic Bird Species Recognition System

Angel David Pedroza Ramirez
Unidad Academica de Ingenieria Electrica
Universidad Autonoma de Zacatecas
Zacatecas, Mexico
P.A.D_16@hotmail.com

Jose Ismael de la Rosa Vargas
Unidad Academica de Ingenieria Electrica
Universidad Autonoma de Zacatecas
Zacatecas, Mexico
ismaelrv@yahoo.com

Rogelio Rosas Valdez
Unidad Academica de Ciencias Biologicas
Universidad Autonoma de Zacatecas
Zacatecas, Mexico
rogrosas@gmail.com

Aldonso Becerra
Unidad Academica de Ingenieria Electrica
Universidad Autonoma de Zacatecas
Zacatecas, Mexico
a7donso@hotmail.com

***Abstract*—In this paper a comparative between Mel Frequency Cepstral Coefficients (MFCC) and Inverse Mel Frequency Cepstral Coefficients (IMFCC) features for an automatic bird species recognition system is proposed with the aim to validate IMFCC as a feature that can also be extracted for bird species recognition. In biodiversity monitoring task there are some traditional techniques and, bioacoustics studies biodiversity by a noninvasive way based on the relationship between animal species and its sounds. Bioacoustics methodology for avian conservation are based on automatic speech recognition techniques and one of the traditional extracted features in this area are MFCC. Nevertheless some new studies uses IMFCC as a complementary frequency information. From results, it is concluded that IMFCC features have better performance than traditional MFCC features but, performance still depends on the recognized bird sound.**

***Index Terms*—bioacoustics, bird classification, HMM, IMFCC, MFCC.**

I. Introduction

Biodiversity monitoring represent an invaluable tool in conservation and climate change reduction [1], [2]. Although there are some traditional invasive monitoring techniques, a noninvasive method is preferred. Bioacoustics studies biodiversity based on the relationship between animal species and its sounds.

Bird species are very sensitive to climate change and its conservation and understanding represents and important challenge [3]. In this sense, bioacoustics traditional methodology for avian conservation are based on automatic speech recognition techniques [4], [5], [6], [7]. Although there are a huge variability of techniques and features to be extracted from sound, one of the traditional features in speech recognition, are Mel frequency cepstral coefficients (MFCC). Despite the high efficiency by using MFCC features, some new studies proposes the use of inverse mel frequency cepstral coefficients (IMFCC) as a new perspective that take into account frequency information missed by MFCC features improving the spectral modeling [8], [9], [10], [11].

Even though MFCC features have been applied in bioacoustics, the application of IMFCC features in avian conservation systems have not been applied yet. In this paper a comparative between MFCC and IMFCC features for an automatic bird species recognition system is conducted with the aim to validate the use of IMFCC features as a feature that can also be extracted for bird species recognition. From results, it can be observed that IMFCC features give better performance than traditional MFCC features. However, the performance still depends on the bird species to be recognized.

In the following: Section II explains basic mathematical background of MFCC and IMFCC features. Section III present the used automatic bird species recognition system. Section IV summarized the comparative results between MFCC and IMFCC features, and some discussion. Finally, Section V concludes this paper.

II. Feature Extraction: MFCC and IMFCC

A. MFCC

Human perception of sound can be described by the Mel scale and MFCCs are based on this configuration. The basic idea is to capture sound frequencies by a filter bank and

The authors would like to thank CONACYT and to the Programa de Doctorado en Ciencias de la Ingenieria from the Universidad Autonoma de Zacatecas.

Escuchando a la naturaleza: Del reconocimiento de voz a la bioacústica

Ángel David Pedroza Ramírez

Universidad Autónoma de Zacatecas, Unidad Académica de Ingeniería Eléctrica, Carretera a la Bufa No.1 Col. Centro Zacatecas, Zac., (492) 92 296 99.

P.A.D_16@hotmail.com

José Ismael de la Rosa Vargas

Universidad Autónoma de Zacatecas, Unidad Académica de Ingeniería Eléctrica, Av. López Velarde No. 801 Col. Centro Zacatecas, Zac., (492) 925 66 90, ext. 3979

joseismaelrv@gmail.com, ismaelrv@ieee.org

Rogelio Rosas Valdez

Universidad Autónoma de Zacatecas, Unidad Académica de Ciencias Biológicas, Av. Preparatoria s/n Col. Agronómica II, Zacatecas, Zac.,

rogrosas@gmail.com

Resumen

La bioacústica es la rama de la ciencia que, mediante información acústica, se encarga del estudio de la forma de transmisión y recepción de información biológica con el fin de alcanzar desde la identificación de especies hasta la determinación de la salud de un ecosistema. Algunos desarrollos recientes se han enfocado en la adecuación y aplicación de técnicas de reconocimiento de voz tales como el uso de Modelos Ocultos de Markov, Redes Neuronales, entre otros; con el fin de lograr el reconocimiento automatizado de especies. En esta revisión se presentan algunos avances tecnológicos en el área así como una visión global sobre las herramientas matemáticas disponibles para, mediante ellas, lograr algunos de los objetivos que la bioacústica pretende alcanzar.

Palabra(s) Clave(s): Bioacústica, Biología, Identificación de especies, Reconocimiento automático de voz.

Apéndice
Otros Artículos pendientes a publicación sobre Bioacústica

RESEARCH

In the challenge of understanding bird species biodiversity: An automatic bird species classification algorithm based on Inverted Mel Frequency Cepstral Coefficients

Ángel D. Pedroza[1*†], José I. de la Rosa[1], Rogelio Rosas[2], Aldonso Becerra[1] and José de J. Villa[1]

*Correspondence: P.A.D_16@hotmail.com
[1]Universidad Autónoma de Zacatecas,Unidad Académica de Ingeniería Eléctrica, Av. Ramón López Velarde No. 801 Col. Centro, Zacatecas, Zac., MX
Full list of author information is available at the end of the article
†Corresponding author

Abstract

Background: Biodiversity conservation is one of the most transcendent tasks in the defiance of reducing climate change. Bird species are highly sensitive to environmental changes and, therefore, their preservation and study could be invaluable tools. Bioacoustics is the science that studies biodiversity based on the relationship between animal species and their sounds. Some of the basic techniques in bioacoustics were applied in automatic speech recognition systems previously. Nevertheless, given that the bird vocalizations are different from human speech, is not yet well understood which features extracted from the bird vocalization are the most important for an automatic bird species classification system. Some recent studies in automatic speech recognition systems use the Inverted Mel Frequency Cepstral Coefficients (IMFCC) as a method that take into account high-frequency components of speech.

Results: In this paper, given the close relationship between bioacoustics and automatic speech recognition, an automatic bird species classification algorithm, based on the extraction of IMFCC from the bird vocalizations, is proposed. Although the obtained results of the efficiency test were lower than efficiencies of MFCC's and IMFCC's in automatic speech recognition, there was a general improvement in the true acceptance rate (TAR) and the true rejection rate (TRR) through the use of IMFCC's.

Conclusions: From results is concluded that the acoustic information of bird vocalizations in the high-frequencies (captured by the IMFCC's) is as significant as the acoustic information in the low-frequencies (captures by the MFCC's) for the bird species classification task through bird vocalizations. On the other hand, as was expected, the spectral pattern of the vocalizations of the birds to be recognized influences the efficiency of the proposed algorithm.

Keywords: bioacoustics; bird species classification; IMFCC

1 Background

Although climate change is related to several factors [1], biodiversity conservation is one of the main tasks in the defiance of reducing (or mitigate) this problem [2]. Recent studies demonstrate that bird species are highly sensitive to environmental

RESEARCH

Who is that bird?: Acoustic individual identification in birds based on the band-limited phase-only correlation function.

Ángel D. Pedroza[1*†], José I. de la Rosa[1], Rogelio Rosas[2], Aldonso Becerra[1] and José de J. Villa[1]

[*]Correspondence: P.A.D_16@hotmail.com
[1]Universidad Autónoma de Zacatecas,Unidad Académica de Ingeniería Eléctrica, Av. Ramón López Velarde No. 801 Col. Centro, Zacatecas, Zac., MX
Full list of author information is available at the end of the article
[†]Corresponding author

Abstract

Background: Species richness is a feature of biodiversity. The density of birds in an area is an indicator that can be used to determine some indices such as survival and, population growth. This parameter is measured through a search methodology traditionally based on point counts, linear transects, and mist nets. However, others like the no standardization of the methods and the limited availability of expert observers reduce the effectiveness of the methodology. On the other hand, bioacoustics is an alternative method of non-invasive monitoring that, based on speech recognition techniques, could be used in the individual identification of animals. However, for the acoustic identification of an individual bird at real-field conditions (where the recorded data can be limited), get the amount of vocalizations from an individual to apply in an efficient way some of the speech recognition methods, is a complex challenge.

Results: In this paper is proposed a new technique to deal with acoustic individual identification based on Band-Limited Phase-Only Correlation (BLPOC) function. The BLPOC function is a biometric technique traditionally implemented in human identification that, in this paper, is offered as a limited data individual identification method. The proposed algorithm has two variants (depending on the method to extract the bird vocalization from records): automatic and, semi-automatic verification algorithm. The evaluation of the automatic algorithm's discarding shows a performance over the 90% for almost all the identification comparatives. On the other hand, at the verification of the individuals, it is shown that the efficiencies of the algorithms depend on the complexity of the vocalizations.

Conclusions: The proposed algorithm is an effective method for the acoustic individual identification in birds whose main advantage is that it can provide an identification evaluation only by analyzing two samples from individuals. On the other hand, from the discarding efficiencies, by setting an appropriate algorithm scheme, it could be used in a bird species classification system. In addition to the individual identification, acoustic identification proposed technique could be used together with traditional counting techniques to determine the dispersion of the habitats and the mapping of the territories.

Keywords: bioacoustics; biodiversity; bird; individuals; limited data

Referencias

[1] A. M. A. P. H. G. L. G. A. Siguenza, M. Gallardo, "Biodiversidad de aves en mÃ©xico," *Revista Mexicana de Biodiversidad*, vol. 85, pp. 476–495, 2014.

[2] J. Pech, "Desarrollo de un sistema de reconocimiento de voz para el control de dispositivos utilizando mixturas gaussianas."

[3] M. B. O. d. J. M. Hagan, H. Demuth, *Neural network design*. Pws Pub. Boston, 1996, vol. 20.

[4] G. del Olmo, "Manual para principiantes en la observación de aves:pajareando," *Bruja de Monte. México, DF*, 2010.

[5] E. F. A. Salinas, F. DeClerck and N. Estrada, "Manual de técnicas para la identifiación de aves silvestres," *Programa Monitoreo de Aves (PMA) Centro Agronómico de Investigación y Enseñanza (CATIE)*, 2002.

[6] E. Olvera, "Identificación del canto de turdus migratorius (aves) utilizando un modelo acÃºstico estadístico," *Universidad Nacional Autónoma de México, Facultad de estudios superiores Zaragoza*, 2014.

[7] O. S. K. Naugolnykh, L. Ostrovsky and M.Hamilton, "Nonlinear wave processes in acoustics," *The Journal of the Acoustical Society of America*, vol. 108, no. 1, pp. 14–15, 2000.

[8] N. M. H. C. J. Barker, E. Vincent and P. Green, "The pascal chime speech separation and recognition challenge," *Computer Speech & Language*, vol. 27, no. 3, pp. 621–633, 2013.

[9] S. Z. A. Abdelaziz and D. Kolossa, "Learning dynamic stream weights for coupled-hmm-based audio-visual speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 5, pp. 863–876, 2015.

[10] Y. G. X. Cui, M. Afify and B. Zhou, "Stereo hidden markov modeling for noise robust speech recognition," *Computer Speech & Language*, vol. 27, no. 2, pp. 407–419, 2013.

[11] T. N. S. A. A. O. T. H. S. W. M. F. T. Y. T. O. e. a. M. Delcroix, K. Kinoshita, "Speech recognition in living rooms: Integrated speech enhancement and recognition system based on spatial, spectral and temporal modeling of sounds," *Computer Speech & Language*, vol. 27, no. 3, pp. 851–873, 2013.

[12] R. B. K. Frommolt and M. Clausen, "Computational bioacoustics for assessing biodiversity," in *Proceedings of the International Expert meeting on IT-based detection of bioacoustical patterns, BfN-Skripten*, vol. 234. Citeseer, 2008.

[13] P. Tubaro, "Bioacústica aplicada a la sistemática, conservación y manejo de poblaciones naturales de aves," *Etología*, vol. 7, no. 3, pp. 19–32, 1999.

[14] T. B. N. Sen and S. Chakroborty, "Comparison of features extracted using time-frequency and frequency-time analysis approach for text-independent speaker identification," in *2011 National Conference on Communications (NCC)*, 2011, pp. 1–5.

[15] M. L. S. Memon, N. Maddage and L. He, "Application of the vector quantization methods and the fused mfcc-imfcc features in the gmm based speaker recognition," in *Recent Advances in Signal Processing*, A. Zaher, Ed. Rijeka: InTech, 2009.

[16] H. L. D. L. Y. Z. Y. Li, C. Xia, "Identification of vocal individuality in male cuckoos using different analytical techniques," *Avian Research*, vol. 8, no. 1, p. 21, 2017.

[17] M. C.-C. C. M. G. V. T. Aide, C. Corrada-Bravo and R. Alvarez, "Real-time bioacoustics monitoring and automated species identification," *PeerJ*, vol. 1, p. e103, 2013.

[18] M. C. R. Carbó, A. Molero and J. Linares, "Bioacústica de fondos marinos. acústica aplicada a la pesca," *Revista de Acústica*.

[19] L. Baptista and J. Gómez, "La investigación bioacústica de las aves del archipiélago de revillagigedo: un reporte de avance," *Huitzil, Revista Mexicana de OrnitologÃa*, vol. 3, no. 2, 2015.

[20] G. Moreno-Rueda, "El comportamiento de las aves como herramienta para su identificación," *Acta Granatense*, vol. 4, no. 5, pp. 85–93, 2006.

[21] J. T. C. Duncan and N. Pettorelli, "The quest for a mechanistic understanding of biodiversity–ecosystem services relationships," *Proceedings of the Royal Society of London B: Biological Sciences*.

[22] G. H. S. W. R. Rempel, K. Hobson and J. Elliott, "Bioacoustic monitoring of forest songbirds: interpreter variability and effects of configuration and digital processing methods in the laboratory," *Journal of Field Ornithology*, vol. 76, no. 1, pp. 1–11, 2005.

[23] A. Castro, "Construcción de una red neuronal artificial para clasificar cantos de aves: una aplicación de la inteligencia artificial a la biología," *Universidad de Costa Rica*.

[24] E. Chesmore, "Application of time domain signal coding and artificial neural networks to passive acoustical identification of animals," *Applied Acoustics*, vol. 62, no. 12, pp. 1359–1374, 2001.

[25] ——, "Automated bioacoustic identification of species," *Anais da Academia Brasileira de Ciências*, vol. 76, no. 2, pp. 436–440, 2004.

[26] M. Ayob and E. Chesmore, "Probabilistic neural network for the automated identification of the harlequin ladybird (harmonia axyridis)," in *International Workshop on Multidisciplinary Trends in Artificial Intelligence*, vol. 8271, 2013, pp. 25–35.

[27] M. M. J. d. F. T. Ganchev, O. Jahn and K. Schuchmann, "Automated acoustic detection of vanellus chilensis lampronotus," *Expert systems with applications*, vol. 42, no. 15-16, pp. 6098–6111, 2015.

[28] M. V. J. Pabico, A. Gonzales and A. Mendoza, "Automatic identification of animal breeds and species using bioacoustics and artificial neural networks," *CoRR*, vol. abs/1507.05546, 2015.

[29] J. R.-M. P. Caycedo-Rosales and M. Orozco-Alzate, "Automated recognition of bioacoustic signals: a review of methods and applications," *Ingeniería y Ciencia*, vol. 9, no. 18, 2013.

[30] D. Reynolds, "An overview of automatic speaker recognition technology," in *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 4, 2002, pp. IV–4072–IV–4075.

[31] D. Sharma and I. Ali, "The effect of dc coefficient on mmfcc and mimfcc for robust speaker recognition," in *2015 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, 2015, pp. 313–317.

[32] S.Drgas and T. Virtanen, "Speaker verification using adaptive dictionaries in non-negative spectrogram deconvolution," in *Latent Variable Analysis and Signal Separation*, 2015, pp. 462–469.

[33] A. L. P. G. V. R. M. Li, J. Kim and S. Narayanan, "Speaker verification based on the fusion of speech acoustics and inverted articulatory signals," *Computer Speech & Language*, vol. 36, pp. 196 – 211, 2016.

[34] D. Sharma and I. Ali, "A modified mfcc feature extraction technique for robust speaker recognition," in *2015 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, 2015, pp. 1052–1057.

[35] H. Jayanna and S. Prasanna, "Limited data speaker identification," *Sadhana*, vol. 35, no. 5, pp. 525–546, 2010.

[36] E. S. A. S. M. Graciarena, M. Delplanche and L. Ferrer, "Acoustic front-end optimization for bird species recognition," in *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2010, pp. 293–296.

[37] P. L. C. Chou and B. Cai, "On the studies of syllable segmentation and improving mfccs for automatic birdsong recognition," in *2008 IEEE Asia-Pacific Services Computing Conference*, 2008, pp. 745–750.

[38] X. Z. Z. R. R. X. V. S. C. R. J. A. C. Kwan, G. Mei and K. Ho, "Bird classification algorithms: theory and experimental results," in *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 5, 2004, pp. V–289.

[39] S. Chakroborty and G. Saha, "Improved text-independent speaker identification using fused mfcc & imfcc feature sets based on gaussian filter," *International Journal of Signal Processing*, vol. 5, no. 1, pp. 11–19, 2009.

[40] R. Loyn, "The 20 minute search - a simple method for counting forest birds," *Corella*, vol. 10, pp. 58–60, 1986.

[41] S. G. P. Atkinson, R. Fuller and J. Vickery, "Counting birds on farmland habitats in winter," *Bird Study*, vol. 53, no. 3, pp. 303–309, 2006.

[42] F. González-García, "Métodos para contar aves terrestres," *Manual de Técnicas para el estudio de la Fauna*, vol. 1, pp. 128–147, 2011.

[43] M. M. Lambrechts and A. A. Dhondt, "Individual voice discrimination in birds," in *Current ornithology*, D. Power, Ed. Boston, MA: Springer US, 1995, pp. 115–139.

[44] E. Fox, J. Roberts, and M. Bennamoun, "Call-independent individual identification in birds," *Bioacoustics*, vol. 18, no. 1, pp. 51–67, 2008.

[45] C. T. V. Trifa, A. Kirschel and E. Vallejo, "Automated species recognition of antbirds in a mexican rainforest using hidden markov models," *The Journal of the Acoustical Society of America*, vol. 123, no. 4, pp. 2424–2431, 2008.

[46] S. G. N. Toro and T. Jiménez, "Reconocimiento de especies de anuros por sus cantos, en archivos de audio, mediante técnicas de procesamiento digital de señales," *Scientia et technica*, vol. 3, no. 32, 2006.

[47] S. F. S. C. M. Filipe, P. Branco and S. Vicente, "Affective prosody in european portuguese: Perceptual and acoustic characterization of one-word utterances," *Speech Communication*, vol. 67, pp. 58–64, 2015.

[48] S. Furui, "Recent advances in spontaneous speech recognition and understanding," in *ISCA & IEEE workshop on spontaneous speech processing and recognition*, 2003.

[49] A. Ramírez and J. de la Rosa, "El invisible y asombroso proceso de la comunicación oral: bases sobre reconocimiento de voz," *Pistas Educativas*, vol. 36, no. 112, 2018.

[50] M. Anusuya and S. Katti, "Speech recognition by machine: a review (department of computer science and engineering sri jayachamarajendra college of engineering mysore, india, 2010)," *International Journal of Computer Science and Information Security*.

[51] J. Flores, "Técnicas para el reconocimiento de voz en palabras aisladas en la lengua náhuatl," *Instituto Politécnico Nacional, Centro de Investigación en Computación, México, D.F*, 2009.

[52] A. RamÃrez, "Reconocimiento automático del locutor mediante técnicas dependientes e independientes del vocabulario para un sistema acotado por el ancho del banda telefónico y realización de us sistema experimental," *Departamento de Electrónica y Telecomunicaciones, Centro de investigación científica y de educación superior de Ensenada*, 1996.

[53] R. G. A. Buzo, A. Gray and J. Markel, "Speech coding based upon vector quantization," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 5, pp. 562–574, 1980.

[54] S. Bendezú and G. Said, "Filtro adaptivo lms y su aplicación en el reconocimiento de palabras aisladas para el control de un equipo de sonido por medio de la voz," *Pontificia Universidad Católica del Perú*.

[55] M. Aerens and H. Bourlard, "Linear and nonlinear prediction for speech recognition with hidden markov models," in *Third European Conference on Speech Communication and Technology*, 1993.

[56] D. Reddy, "Speech recognition by machine: A review," *Proceedings of the IEEE*, vol. 64, no. 4, pp. 501–531, 1976.

[57] L. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[58] D. Y. G. D. A. M. N. J. A. S. V. V. P. N. T. S. G. Hinton, L. Deng and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[59] G. JuliÃ¡n, "Las redes neuronales: quÃ© son y por quÃ© estÃ¡n volviendo," *http://goo.gl/GNFV*, Accesado por ultima vez: 26 Jun 2014.

[60] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural Networks*, vol. 61, pp. 85–117, 2015.

[61] T. Brandes, "Feature vector selection and use with hidden markov models to identify frequency-modulated bioacoustic signals amidst noise," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 6, pp. 1173–1180, 2008.

[62] R. S. P. Mowlaee and Y. Stylianou, "Advances in phase-aware signal processing in speech communication," *Speech Communication*, vol. 81, pp. 1 – 29, 2016.

[63] R. Schluter and H. Ney, "Using phase spectrum information for improved speech recognition performance," in *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.01CH37221)*, vol. 1, 2001, pp. 133–136 vol.1.

[64] I. Ito and H. Kiya, "Dct sign-only correlation with application to image matching and the relationship with phase-only correlation," in *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, vol. 1. IEEE, 2007, pp. I–1237.

[65] T. A. K. K. K. Miyazawa, K. Ito and H. Nakajima, "An efficient iris recognition algorithm using phase-based image matching," in *IEEE International Conference on Image Processing 2005*, vol. 2, 2005, pp. II–49–52.

[66] ——.

[67] T. I. N. Shabrina and H. Kunieda, "Fingerprint authentication on touch sensor using phase-only correlation method," in *2016 7th International Conference of Information and Communication Technology for Embedded Systems (IC-ICTES)*, 2016, pp. 85–89.

[68] S. A. I. Rida and A. Bouridane, "Gait recognition based on modified phase-only correlation," *Signal, Image and Video Processing*, vol. 10, no. 3, pp. 463–470, 2016.

[69] A. Teusdea and G. Gabor, "Iris recognition with phase-only correlation." *Annals of DAAAM & Proceedings*, 2009.

[70] K. K. T. A. K. Ito, H. Nakajima and T. Higuchi, "A fingerprint matching algorithm using phase-only correlation," *IEICE transactions on fundamentals of electronics, communications and computer sciences*, vol. 87, no. 3, pp. 682–691, 2004.

[71] Y. Z. S. Wu, Y. Wang and X. Chang, "Automatic microseismic event detection by band-limited phase-only correlation," *Physics of the Earth and Planetary Interiors*, vol. 261, pp. 3–16, 2016.

[72] C. F. G. G. I. M.-C. e. a. F. Bimbot, J. Bonastre, "A tutorial on text-independent speaker verification," *EURASIP Journal on Advances in Signal Processing*, vol. 2004, no. 4, p. 101962, 2004.

[73] C. Kasap and M. Arslan, "A unified approach to speech enhancement and voice activity detection," *Turkish Journal of Electrical Engineering & Computer Sciences*, vol. 21, no. 2, pp. 527–547, 2013.

[74] L. Rabiner and B. Juang, "Fundamentals of speech recognition," vol. 14, 1993.

[75] I. G. C. Dumitru and R. Vieru, "Speaker verification using hmm for romanian language," in *Proceedings ELMAR 2006*, 2006, pp. 131–134.

[76] K. Murphy, "Hidden markov model (hmm) toolbox for matlab [online]," *http://www.cs.ubc.ca/m̃urphyk/Software/HMM/hmm.html*, Accesado por ultima vez: 01 May 2017.

[77] E. G. H. G.-A. B. A. Pedroza, J. de la Rosa, "Diseño de prototipo para mejorar la dicción mediante el uso de modelos ocultos de markov." vol. 38, no. 120.

[78] M. Jayanth and B. Reddy, "Speaker identification based on gfcc using gmm-ubm," vol. 5.

[79] F. Francis and R. Vishnu, "A novel noise robust speaker identification system," vol. 10, no. 17.

[80] Y. S. Y. Lan, Z. Hu and G. Huang.

[81] E. Carl and F. Zach, "Avian habitat evaluation: should counting birds count?" *Frontiers in Ecology and the Environment*, vol. 2, no. 8, pp. 403–410, 2004.

[82] S. D. C. Ralph and J. Sauer, "Managing and monitoring birds using point counts: standards and applications," *Monitoring bird populations by point counts. Gen. Tech. Rep. PSW-GTR-149. Albany, CA: US Department of Agriculture, Forest Service, Pacific Southwest Research Station*, vol. 149, 1995.

[83] J. Karr, "Surveying birds with mist nets," *Studies in Avian Biology*, vol. 6, pp. 62–67, 1981.

[84] P. P. T. M. D. D. C. Ralph, G. Geupel and B. Milá, "Manual de métodos de campo para el monitoreo de aves terrestres," *Gen. Tech. Rep. PSW-GTR-159. Albany, CA: US Department of Agriculture, Forest Service, Pacific Southwest Research Station.*, vol. 159, 1996.

[85] J. Emlen and M. DeJong, "Counting birds: The problem of variable hearing abilities (contando aves: El problema de la variabilidad en la capacidad auditiva)," *Journal of Field Ornithology*, pp. 26–31, 1992.

[86] C. Kasap and M. Arslan, "A unified approach to speech enhancement and voice activity detection," *Turkish Journal of Electrical Engineering & Computer Sciences*, vol. 21, no. 2, pp. 527–547, 2013.

[87] "Xeno-canto," *https://www.xeno-canto.org*, Accessado por ultima vez: 07 Feb 2017.

[88] "Inecol," *http://www1.inecol.edu.mx/sonidos/menu.htm*, Accessado por ultima vez: 16 Mar 2017.