UNIVERSITAT
POLITÈCNICA
DE VALÈNCIA

Caracterización y modelado de la producción de sonidos de las ballenas beluga (*Delphinapterus Leucas*) basado en modelos de análisis / síntesis de voz

Una tesis doctoral realizada por Guillermo Fernán Lara Martínez

---

Dirigida por Ramón Miralles Ricós

Septiembre de 2016

Para Lucía

"Claro que lo entiendo.
Incluso un niño de cinco años podría entenderlo.
¡Que me traigan un niño de cinco años!"
Groucho Marx

Abstract

This thesis deals with the study of the sounds produced by beluga whales (*Delphinapterus leucas*) with a fundamental objective: its characterization and modeling. To this end, analysis / synthesis algorithms of the sounds produced by these animals are proposed. These algorithms are inspired by recent researches on how beluga whales produce the sound and in the physiognomy of the sound production organs.

This is a multidisciplinary work, and in order to achieve this goal many fields and topics have to be studied: the generation of sounds in musical instruments, time-frequency analysis techniques along with pattern recognition methods, feature selection, the potential to include algorithms that work in the cepstral domain and a quantitative analysis of the *Recurrence Plots*. All this allows to propose a sound production model capable to adapt to the peculiarities of this specie and reproduce with high fidelity its wide repertoire of sounds.

In addition, in order to validate the proposed model, different sounds from a database of beluga whale vocalisations were analysed. These sounds were compared with those generated with a generic analysis / synthesis model. Furthermore, it is proposed to use the synthesis model parameters for a new sound classification based on how the sounds are produced, achieving better results than those obtained with classifiers based on characteristics of the time-frequency diagram.

All the proposed hypotheses have been validated by doing acoustic measurements of beluga whales from the Oceanografic of Valencia (supervised by park biologists), as well as a large number of audio laboratory recordings under controlled conditions.

Finally, a passive acoustic monitoring device called SAMARUC design in the framework of the projects related to this thesis is described. The SAMARUC system has the possibility to include different signal processing algorithms for sound analysis in open water environment and is capable of recording high quality sounds. It can also provide the classification of the acoustic events as well as the noise indicators of good environmental status of our seas and oceans. These indicators include underwater noise levels as reflected in the Descriptor 11 of the Marine European Directive. For this reason we expect that this research will have a significant importance in the future years.

Resum

Aquesta tesi doctoral aborda l'estudi dels sons produïts per les ballenes beluga (*Delphinapterus Leucas*) amb un objectiu fonamental: la seua caracterització y modelat. Es proposen una sèrie d'algoritmes d'anàlisi / síntesi d'aquestos sons inspirats en recents investigacions sobre el funcionament i la fisonomia del aparell fonador de les ballenes beluga.

Es tracta d'un treball multidisciplinar, on per a arribar a aquest objectiu s'analitza la generació de sons en instruments musicales, es repassen tècniques d'anàlisi temps-freqüencia junt a mètodes de reconeixement de patrons, s'apliquen tècniques de selecció de característiques, s'analitza el potencial de la inclusió d'algoritmes que treballen en el domini cepstral i es realitza un anàlisi quantitatiu dels *Recurrence Plots* com a mètrica del determinisme. Tot aço permet plantejar un model de producción de sons capaç d'adaptar-se a les peculiaritats d'aquesta espècie i reproduir amb una alta fidelitat el seu amplíssim repertori de sons.

Ademés, amb la finalitat de donar validesa al model proposat, s'analitzen i sintetitzen diferents sons d'una base de dades de vocalitzacions de ballenes beluga per a compararles amb les generades per un model d'anàlisi / síntesi genèric. De manera addicional es proposa emprar els paràmetres del model de síntesi disenyat per a fer una nova classificació de sons basada en la seua naturalesa de producció, aconseguint millorar el resultats obtinguts amb classificadors basats en característiques del diagrama temps-freqüencia.

Les hipòtesis proposades han sigut validades amb la realització de mesures acústiques de les ballenes beluga de l'Oceanogràfic de Valéncia (supervisades pels cuidadors del parc), així com una gran quantitat de grabacions d'àudio de laboratori en condicions controlades.

Finalment es descriu el dispositiu de monitorització acústica pasiva SAMARUC, disenyat amb la funcionalitat d'incloure els algoritmes implementats en entorns d'aigües obertes, podent oferir dades de la grabació dels sons, la seua classificació, així com els indicadors del benestar mediambiental dels nostres mars i oceans. Aquestos indicadors inclueixen els nivells de so submarí tal com es diu al descriptor 11 de la Directiva Marina Europea. Es per aço que es preveu que aquesta línia d'investigació tinga un auge considerable als anys futurs.

Resumen

Esta tesis doctoral aborda el estudio de los sonidos producidos por las ballenas beluga (*Delphinapterus Leucas*) con un objetivo fundamental: su caracterización y su modelado. Para ello, se propone una serie de algoritmos de análisis / síntesis de los sonidos producidos por estos animales inspirados en las recientes investigaciones sobre el funcionamiento y la fisionomía del aparato fonador de las ballenas beluga.

Se trata de un trabajo multidisciplinar, en el que para alcanzar este objetivo, se analiza la generación de sonidos en instrumentos musicales, se repasan técnicas de análisis tiempo-frecuencia junto con métodos de reconocimiento de patrones, se aplican técnicas de selección de características, se analiza el potencial de incluir algoritmos que trabajen en el dominio cepstral y se realiza un análisis cuantitativo de los *Recurrence Plots* como medida del determinismo. Todo esto permite plantear un modelo de producción de sonidos capaz de adaptarse a las peculiaridades de ésta especie y reproducir con una alta fidelidad su amplio repertorio de sonidos.

Además, con la finalidad de dar validez al modelo propuesto, se analizan y sintetizan diferentes sonidos de una base de datos de vocalizaciones de ballenas beluga para compararlos con los generados con un modelo de análisis/ síntesis genérico. De manera adicional, se propone emplear los parámetros del modelo de síntesis para realizar una nueva clasificación de sonidos de la especie basada en su naturaleza de producción, consiguiendo mejorar los resultados obtenidos con clasificadores basados en características del diagrama tiempo-frecuencia.

Las hipótesis propuestas han sido validadas con la realización de medidas acústicas de las ballenas beluga del Oceanografic de València (supervisadas por los cuidadores del parque), así como con un gran número grabaciones de audio de laboratorio en condiciones controladas.

Finalmente se describe el dispositivo de monitorización acústica pasiva SAMARUC, diseñado con la funcionalidad de incluir los algoritmos implementados en entornos de aguas abiertas, pudiendo ofrecer datos de la grabación de los sonidos, la clasificación de éstos, así como los indicadores del buen estado medioambiental de nuestros mares y océanos. Estos indicadores incluyen los niveles de ruido submarino tal y como se recoge en el descriptor 11 de la Directiva Marina Europea y es por ello que se prevé que esta línea de investigación tenga un auge considerable en años futuros.

Agradecimientos

En primer lugar me gustaría agradecer a Ramón toda la confianza depositada en mi persona. Gracias por introducirme en el mundo de la investigación, por abrirme un abanico de posibilidades que nunca hubiera imaginado, por tu cercanía y amistad.

A todos los compañeros y profesores del Grupo de Tratamiento de Señal, desde los carismáticos Arturo y Jorge M., hasta las nuevas incorporaciones Santi, Carles y Guille, pasando por Ahmed, Beni, Patricia, Máriam, María Ángeles, Vicente, Nacho, Jorge G., Elena y David. Mención especial a Gonzalo y Alicia, con los que he compartido mucho más que un lugar de trabajo, mucho más que goteras y traslados, mucho más que cervezas y catas.

A mi familia, la política y la natural, gracias Papá y Mamá.

Gracias a Yago, por sus sonrisas y su personalidad, seguro que su hermanita también las tendrá.

Por último, y no menos importante, a Lucía, por la paciencia y apoyo en todo momento, habiendo realizado los mismos o más esfuerzos que yo para que esta tesis doctoral sea un hecho.

Índice general

Capítulo 1

Introducción y objetivos generales

1.1. Introducción

La evolución de los clasificadores en campos relacionados con el análisis de audio a lo largo de los últimos años ha sido exponencial. Un buen ejemplo es el reconocimiento de voz humana, donde existen actualmente diferentes motores de búsqueda por voz con una tasa de reconocimiento superior al 90 %. Aplicaciones como Siri o Google Now son algunos de estos ejemplos. En estos momentos las aplicaciones relacionadas con el análisis y síntesis de voz nos permiten no sólo analizar y discernir entre una determinada sílaba u otra [1], sino incluso detectar las emociones de las personas mientras están hablando [2].

A la hora de realizar una clasificación es necesario establecer tanto las familias, clases o categorías en donde clasificaremos los distintos eventos o registros, así como las características a extraer de cada una de ellos. Una buena definición de las clases y una selección de las características más relevantes será por tanto indispensable para cualquier clasificación.

A esta problemática, donde es necesario extraer características y después realizar una clasificación, se le denomina reconocimiento de patrones. El objetivo de un reconocimiento de patrones es clasificar un evento (en el caso de esta tesis doctoral serán señales audibles o sonidos) dependiendo de sus características.

Centrándonos en la selección de características, la comprensión de lo que representan y sus posibles dependencias son cualidades importantes a la hora de que dichas características sean relevantes. De esta forma, cualquier procesador, aun sin poseer gran capacidad de cálculo, tendrá la posibilidad de llevar a cabo una clasificación eficiente sin consumir una excesiva carga computacional. Cabe destacar que la evolución de los microprocesadores y el incremento en la capacidad de cálculo, unido al diseño de nuevos algoritmos de selección de características, han permitido que la cantidad de características extraídas pueda ser muy elevada, en muchas ocasiones, sin importar como son obtenidas y lo que representan.

1.2. Objetivos

Esta tesis doctoral se centra en el estudio exhaustivo de los sonidos producidos por mamíferos marinos, en concreto de la subespecie de los odontocetos denominada *delphinapterus leucas* o ballenas beluga, cetáceos que se caracterizan por la presencia de dientes, un solo orificio nasal externo y ser capaces de producir una amplia gama de sonidos. El objetivo global será diseñar un modelo de producción de sonidos que permita dar cabida a toda la gama de sonidos que estos animales son capaces de producir. Una vez realizado este modelo, se extraerán sus características más relevantes para la generación de un clasificador de sonidos.

Diversos trabajos a lo largo de los últimos años, han estado orientados a intentar esclarecer como son producidos los sonidos, canciones o vocalizaciones que realizan los cetáceos. Intentar determinar los órganos encargados de la producción de sonidos, los mecanismos empleados en su generación, así como la posibilidad de clasificar y relacionar éstos con el comportamiento de diferentes especies ha sido siempre un reto a alcanzar. La complejidad del medio marino, la presencia creciente de ruidos antropogénicos, y el cada vez más mermado número de individuos, son, entre otras, dificultades añadidas al estudio de estos sonidos y de su producción.

En todo este complejo panorama, el estudio de los sonidos generados por los cetáceos, y concretamente los producidos por las ballenas beluga, puede verse beneficiado por los avances en el procesado de señal y la aplicación de nuevos modelos de síntesis de producción. Esta tesis doctoral pretende avanzar en el modelado de los sonidos de ballenas beluga con un doble objetivo: por un lado ser capaces validar de manera indirecta algunas de las teorías sobre los mecanismos de producción de sonidos por parte de esta especie, y por otra, la extracción de las características que permitan la creación de mejores clasificadores automáticos. En esta tarea se emplearán sintetizadores similares a los empleados en voz humana pero adaptándolos a esta especie: duplicación de estructuras, diferente repertorio de sonidos, etc.

Para comenzar se realizará la extracción de un gran conjunto de características de diferente índole entre las que se elegirán las más relevantes, con el objetivo de utilizadas después en un algoritmo lo más genérico posible. De esta forma se obtendrá un primer análisis sobre las características más importantes y se detectarán las problemáticas a la hora de realizar un reconocimiento de patrones, para a partir de este momento trazar las líneas de trabajo a desarrollar combinando los conocimientos del *Data Mining* y el modelado en la producción de sonido para obtener un conocimiento más preciso de los mecanismos de producción de sonido en los odontocetos.

Por tanto, se validará a través de diferentes comparativas del modelo propuesto con los definidos por otros autores y otras especies, como por ejemplo, modelos de generación de sonido basados en LPC-Vocoder excitado, o bien por un tren de deltas, o por un ruido. Se generarán sonidos sintetizados, se mostrarán los resultados utilizando una amplia base de datos de señales de ballenas beluga y se obtendrán conclusiones de las comparativas con el resto de modelos. Esta labor se realizará mediante la comparación de espectrogramas con el método Structural SIMilarity, con el objetivo de sacar conclusiones

lo más objetivas posibles y sin tener que realizar medidas de calidad subjetivas.

En un último paso, se extraerán las características representativas de este modelo para realizar una posterior clasificación de los distintos sonidos realizados por las ballenas beluga. Estas características, al heredar a través del modelo la información relativa a la fisiología del aparato fonador de los cetáceos, deberán permitir obtener mejores clasificadores. El modelo propuesto y el clasificador, también será probado para señales producidas por delfines mulares y listados, animales con un sistema fonador muy similar al de las ballenas beluga.

El hilo argumental de la tesis será la comprensión de principio a fin del proceso de producción de sonidos, diseñando un modelo de análisis/ síntesis que refleje su funcionamiento para después extraer unas características que estarán por tanto asociadas a la fisionomía del sistema de producción de sonido de las ballenas belugas. Es decir, se realizará una selección de características de calidad, que sean relevantes, que resuman y reflejen las propiedades del modelo propuesto.

Capítulo 2

El reconocimiento de patrones. Teoría y métodos

El reconocimiento de patrones es la ciencia que se ocupa de los procesos sobre ingeniería, computación y matemáticas relacionados con objetos físicos o abstractos, con el propósito de extraer información que permita establecer propiedades entre conjuntos de eventos. Los patrones se obtienen a partir de los procesos de segmentación y extracción de características donde cada evento queda representado por una colección de descriptores o características. El sistema de reconocimiento debe asignar una clase o categoría (conjunto de entidades que comparten alguna característica que las diferencia del resto) a cada uno de los eventos. Para poder reconocer los patrones se siguen los siguientes procesos:

- Extracción y selección de características.
- Toma de decisiones.

2.1. La selección de características

El objetivo de la selección de características es evaluar con cuales de ellas un clasificador puede tener un mejor comportamiento. Se trata de un término usado habitualmente en *Data Mining* para describir las herramientas y las técnicas disponibles para reducir las características a un tamaño apropiado para su procesamiento y análisis. La selección de características no sólo implica la reducción de cardinalidad, es decir, la imposición de un límite arbitrario o predefinido en el número de características que se pueden considerar al crear un modelo, sino también la elección de características, lo que significa que el analista o la herramienta de modelado debe seleccionar o descartar activamente las características en función de su utilidad para el análisis. En resumen, la selección de características debe determinar qué y cuántas características son necesarias para minimizar la tasa de error en una clasificación.

La capacidad de aplicar la selección de características es esencial para un análisis eficiente, ya que los conjuntos de datos suelen contener mucha más información de la

necesaria para la generación de un modelo o clasificador. Por ejemplo, una base de datos de una pagina web o aplicación móvil puede contener cientos de características que describen las cualidades o acciones de cada uno de los clientes, pero si los datos de algunas de ellas están muy dispersos, no obtendrá muchas ventajas al agregarlas al modelo o clasificador. Si se mantienen las características innecesarias, se necesitará más CPU y memoria durante el proceso de entrenamiento, así como más espacio de almacenamiento para generar un buen modelo o una buena clasificación.

Aunque actualmente los recursos no sean un problema, normalmente se deben eliminar estas características porque pueden degradar la calidad de los patrones detectados debido a que algunas de ellas características son ruidosas o redundantes. Este ruido dificulta la detección de patrones significativos a partir de los datos. Además, para detectar patrones de calidad, la mayoría de los algoritmos de *Data Mining* requieren un conjunto de datos de entrenamiento mucho más grande en un conjunto de datos multidimensional. Sin embargo, en algunos casos se dispone de muy pocos datos de entrenamiento.

Si sólo unas pocas de los cientos de características de la base de datos tienen información útil para la generación de un modelo o clasificador, las demás no intervendrán en dicho proceso. Esta selección puede realizarse mediante la comprensión del experimento o modelo, extrayendo y seleccionando las características que se presuponen van a obtener más información relevante o utilizando técnicas de selección de características para detectar automáticamente las mejores características y excluir los valores estadísticamente no significativos. Este análisis automático mediante algoritmos de selección de características se aplican de manera satisfactoria en Análisis *Big Data* o *Data Mining*. En cualquier caso, la selección de características ayuda a resolver el problema de tener demasiados datos de escaso valor o muy pocos datos de mucho valor.

El Análisis *Big Data* o *Data Mining* consiste en un análisis potente donde se captan la mayoría de datos disponibles, nada centrado en la relevancia y estudio exhaustivo de las características y sí en la resolución de problemas con un desarrollo teórico complicado donde no se sabe como comenzar a proceder. En estos casos la selección de las características se hace automáticamente mediante algoritmos de selección asociados al posterior algoritmo de clasificación. Es decir, cada algoritmo tiene un conjunto de técnicas predeterminadas para aplicar de forma inteligente la reducción de características. La selección de características siempre se realiza antes del entrenamiento del modelo o clasificador y permite elegir automáticamente en un conjunto de datos las características que se usarán en el mismo. Sin embargo, también se pueden establecer parámetros manualmente para influir en el comportamiento de la selección de características.

La selección de características en *Big Data* funciona calculando una puntuación para cada características y seleccionando a continuación sólo las que han obtenido las mejores puntuaciones. También es posible ajustar el umbral para las puntuaciones más altas. Varios métodos se usan para calcular estas puntuaciones, y el método exacto que se aplica en un modelo depende de los siguientes factores:

- El algoritmo usado en el modelo.
- El tipo de datos de la característica.

- Otros parámetros que se hayan podido establecer en el modelo o clasificador.

Una vez completada la puntuación para la selección de características, sólo se incluirán en el proceso de generación del modelo, clasificador o predicción las características que seleccione el algoritmo. A continuación se enumeran los parámetros necesarios para llevar a cabo una buena selección de características.

- Número máximo de características seleccionadas: Si un modelo contiene más características en la fase de test/validación que el número especificado en el parámetro, el algoritmo pasa por alto cualquier característica que determina como no interesante.
- Máximo número de categorías de clasificación: Si un modelo contiene más categorías/clases de los especificados, los clases con menor popularidad se agrupan y se tratan como estados que faltan.

Hay muchas maneras de implementar la selección de características, dependiendo del tipo de datos con los que se esté trabajando y del algoritmo que se elija para el análisis. En los siguientes subsecciones se describen algunos de los técnicas de selección de características más utilizadas. Esta evaluación se realizará en la denominada fase de entrenamiento.

2.1.1. Selección secuencial de las características *Sequencial Features Selection* (SFS).

El algoritmo SFS tiene una manera de proceder en la búsqueda comenzando con un conjunto de características S vacío para gradualmente ir añadiendo, mediante alguna función de evaluación, la cual minimice el error cuadrático medio. En cada iteración será incluida una característica en el conjunto S, seleccionándola del resto de características que todavía no se han añadido al conjunto. Cuando la inclusión de una característica adicional suponga que el error de clasificación suba, el número de características quedará fijado y no será añadida ninguna más al conjunto S, obteniendo el mínimo error de clasificación posible para el total de características disponible.

SFS se usa comúnmente por su simplicidad y rapidez. Diferentes variantes y muchas aplicaciones han sido diseñadas basándose en dicho algoritmo [3–8].

2.1.2. Entropía de Shannon (ES)

La entropía de Shannon mide la incertidumbre de una variable aleatoria para un determinado resultado. Por ejemplo, la entropía de lanzar una moneda al aire para decidir algo a cara o cruz se puede representar como una función de la probabilidad de que salga cara. La entropía de un mensaje X, denotado por $H(X)$, es el valor medio ponderado de la cantidad de información de los diversos estados del mensaje

$$H(X) = -\sum_{i=1}^{k} P(x_i) log P(x_i) \tag{2.1}$$

donde hay k estados posibles, cada uno con una probabilidad $P(x_i)$. Esta métrica se utiliza en características discretas.

2.1.3. Grado de interés (GI)

Una característica es interesante si ofrece información útil. Dado que la definición de lo que es útil varía dependiendo del escenario, se han desarrollado diversas maneras de medir el grado de interés. Por ejemplo, medir la novedad podría ser interesante a la hora de detectar valores atípicos, pero la capacidad de diferenciar entre elementos estrechamente relacionados o, lo que es lo mismo, diferenciar su importancia, podría resultar más interesante para la clasificación.

La medida del grado de interés está basada en la entropía de Shannon, lo que significa que las características con distribuciones aleatorias tienen una entropía más alta y obtienen menos información; por tanto, esos atributos son menos interesantes. La entropía para cualquier característica se compara con la entropía de todos las demás de la manera que se muestra a continuación en la Ec. (2.2):

$$GI(c) = (H(c) - m) * (m - H(c)), \tag{2.2}$$

donde m es la entropía de todo el conjunto de características y $H(c)$ entropía de la característica c. Al restar la entropía de la característica a analizar a la entropía central, se puede evaluar cuánta información proporciona la característica en particular. Esta medida se utiliza de manera habitual cada vez que una característica contiene datos numéricos continuos no binarios.

2.1.4. Bayesiano con prioridad K2 (BPK2)

Una red bayesiana es un gráfico de estados y transiciones entre ellos; esto significa que algunos estados siempre son anteriores al estado actual y otros son posteriores, y que el gráfico no se repite ni realiza bucles.

Por definición, las redes bayesianas permiten el uso de conocimiento previo. Sin embargo, la pregunta sobre qué estados anteriores se deben utilizar para calcular las probabilidades de los estados posteriores es importante para la precisión, el rendimiento y el diseño del algoritmo.

Está basado en la optimización de una medida, la cual se usa para explorar el espacio de búsqueda formado por todas las redes que contienen las características de la base de datos. Se parte de una red inicial y ésta se va modificando obteniendo una nueva red con mejor medida. El trabajo descrito en [9] es un ejemplo de ello. Además, la medida es escalable y requiere la ordenación de las variables utilizadas como entrada. Se utiliza con características discretas.

2.1.5. Equivalente Dirichlet bayesiano con prioridad uniforme (BDEU)

La puntuación Equivalente Dirichlet Bayesiano (BDE) también utiliza el análisis bayesiano para evaluar una red dado un conjunto de datos. El método de puntuación

BDE fue desarrollado por Heckerman y está basado en la métrica Dirichet Bayesiana (BD) desarrollada por Cooper y Herskovits [9].

La distribución Dirichlet es una distribución multinomial que describe la probabilidad condicional de cada variable de la red y dispone de muchas propiedades que son útiles para el aprendizaje.

El método Equivalente Dirichlet Bayesiano con prioridad Uniforme (BDEU) considera un caso especial de la distribución Dirichlet, en el que se utiliza una constante matemática para crear una distribución fija o uniforme de estados anteriores. Esta métrica también considera la equivalencia de probabilidad; esto significa que no es de esperar que los datos diferencien estructuras equivalentes.

2.2. Teoría de los modelos de clasificación

Un sistema de clasificación se divide por tanto en dos partes: la parte de entrenamiento y test/validación o clasificación (ver Fig. 2.1). La función del módulo de preprocesado es discernir entre los diferentes patrones de interés, normalizarlos, eliminar el ruido y cualquier otra operación que contribuya a definir una representación compacta de dichos patrones.

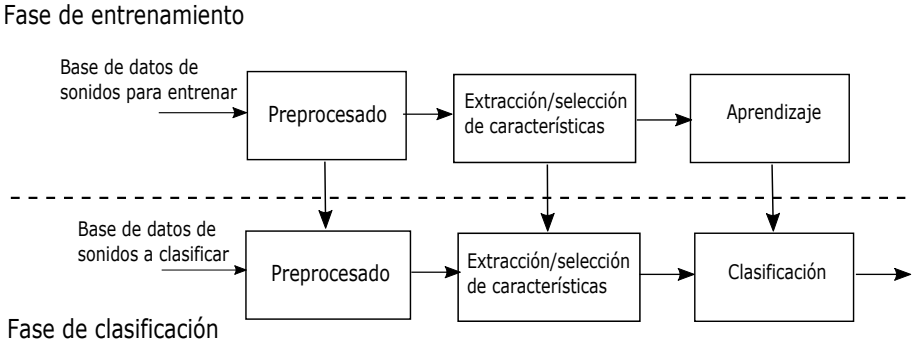


Figura 2.1: Diagrama de bloques del reconocimiento estadístico de patrones.

Tal y como se ha descrito en la Sección 2.1 de este capítulo, en la fase de entrenamiento de cualquier problema de clasificación es necesario realizar una selección de características para ordenarlas según su relevancia y así aumentar las posibilidades de obtener tasas de acierto mayores en la fase de test/validación.

En ella, el módulo de selección de características encontrará la característica apropiada para representar los patrones de entrada y el clasificador se entrenará para el espacio de la característica. La retroalimentación permitirá diseñar y optimizar las estrategias de preprocesado y de selección de características. En la fase de clasificación, el clasificador entrenado asignará una de las categorías o clases establecidas previamente basándose en las características medidas. El proceso de decisión puede ser resumido de la siguiente manera:

Un evento es asignado a una de las c categorías o clases $w_1, w_2, ..., w_c$ basándose en el

vector de d características $x = (x_1, x_2, ..., x_d)$. Las características tendrán una densidad de probabilidad. Por lo tanto el vector x de un patrón perteneciente a la clase w_i será visto como una observación aleatoria de la función de probabilidad condicionada de la clase $p(x|w_i)$. Un conjunto de buenas reglas de decisión, incluyendo la regla de decisión de Bayes, la regla de *maximum likehood* (que se verá como un caso particular de la regla de decisión de Bayes), y la regla de Neyman Pearson son adecuadas para definir los límites de cada clase o patrón. La regla de decisión óptima de Bayes que minimiza el coste se puede observar a continuación [10].

Asignamos un patrón de entrada x a la clase w_i para la cual el coste condicional $R(\alpha_i|x)$ es mínimo:

$$R(\alpha_i|x) = \sum_{j=1}^{c} \lambda(\alpha_i, w_j) P(w_j|x) \tag{2.3}$$

donde $\lambda(\alpha_i, w_j)$ es la función de pérdidas relacionada en la decisión de la clase w_i cuando la clase real es w_j y $P(w_j|x)$ es probabilidad a posteriori. En caso de que la función de pérdidas sea la siguiente:

$$\lambda(\alpha_i, w_j) = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases} \tag{2.4}$$

donde los costes condicionales se convertirán en la probabilidad condicional de errores de clasificación. Para esta función de pérdidas, la regla de decisión bayesiana se simplifica con la Regla del Máximo a Posteriori (MAP). Asignaremos por tanto el patrón de entrada x a la clase w_i si

$$P(w_i|x) > P(w_j|x) \qquad para\ \ cada\ \ j \neq i \tag{2.5}$$

Varias estrategias son utilizadas para diseñar un clasificador de reconocimiento estadístico de patrones, dependiendo del tipo de información disponible sobre las densidades condicionales de las clases [10]. Si todas las decisiones condicionales de las clases están totalmente especificadas, entonces la regla de decisión óptima de Bayes se podrá usar como clasificador.

Sin embargo, cuando las densidades condicionales de las clases no son conocidas en la mayoría de los casos prácticos, deben ser entrenadas a partir de los patrones de entrenamiento. Si la forma de las densidades condicionales de entrada es conocida (e.g. *multivariate Gaussian*), pero algunos de los parámetros de las densidades (e.g la media de los vectores y las matrices de covarianza) no lo son, tendremos un problema de decisión paramétrico.

Una estrategia común para este tipo de problemas es remplazar los parámetros no conocidos en las funciones de densidad por sus valores estimados, resultando por tanto el clasificador de Bayes integrado (*Bayes plug-in clasiffier*). La estrategia bayesiana óptima llegado a esta situación requiere información adicional como distribuciones de probabilidad a priori de los parámetros no conocidos.

Si las densidades condicionales de las clases no se conocen, entonces se utilizarán clasificadores no paramétricos. En este caso se deberán estimar las funciones de densidad o directamente construir unos límites de decisión basados en los datos de entrenamiento (e.g. regla de los k-vecinos más próximos). De hecho una percepción multicapa puede ser vista como un método supervisado no paramétrico que construye unos límites de decisión.

En el reconocimiento estadístico de patrones es necesario elegir entre realizar un aprendizaje supervisado, es decir, con eventos de entrenamiento etiquetadas, o un aprendizaje sin supervisar, es decir, con eventos sin etiquetar. La etiqueta de un patrón de entrenamiento representa la categoría a la pertenece cada patrón. En un problema de aprendizaje sin supervisar (*unsupervised learning*), a veces el número de clases deben ser aprendidas a lo largo de la estructura de cada clase.

En el árbol de la Fig. 2.2 se puede observar un breve resumen de las diferentes formas y nomenclaturas de clasificación en base al reconocimiento estadístico de patrones. Conforme se va recorriendo el árbol desde arriba hacia abajo y de izquierda a derecha, menos información está disponible para diseñar el sistema aumentando la dificultad de la clasificación.

De cualquier forma, la mayoría de los enfoques del reconocimiento estadístico de patrones tratan de implementar la regla de decisión bayesiana.

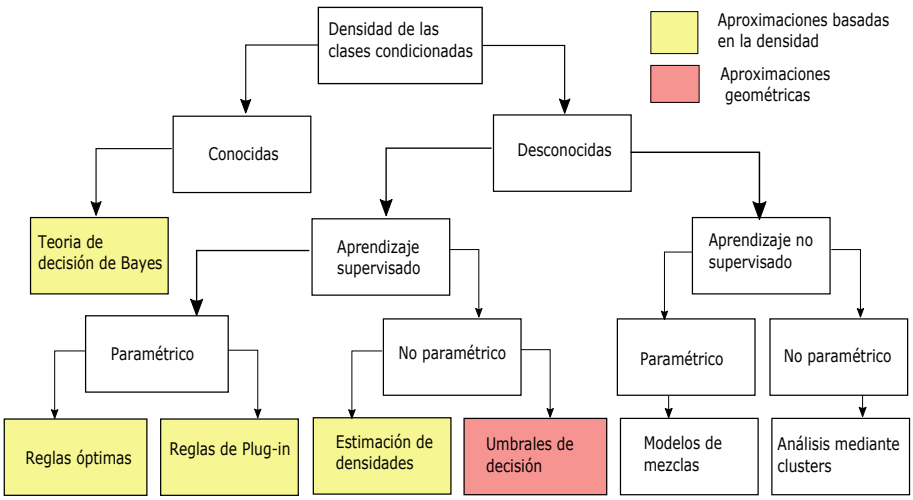


Figura 2.2: Diferentes enfoques del reconocimiento estadístico de patrones.

Cabe destacar que el análisis mediante *clusters* hace frente a los problemas de tomas de decisión de patrones en el modo no paramétrico y no supervisado. Aún así, en él no hay un número específico de categorías o clases; el objetivo es descubrir una categorización razonable de los datos (si existe). Los algoritmos de análisis de *clusters*, junto a varias técnicas de proyección y visualización multidimensional de los datos, son también contempladas en los métodos de análisis de exploración de los datos [11].

Otra decisión a tomar en el reconocimiento estadístico de patrones reside en elegir

si los límites de decisión de las clases son obtenidos directamente (enfoque geométrico) o indirectamente (enfoque basado en las densidades de probabilidad) como muestra la Fig. 2.2. El enfoque probabilístico requiere estimar primero las funciones de densidad, y después construir funciones discriminatorias que especifiquen los límites de decisión. Por otra parte, el enfoque geométrico generalmente construye los límites de decisión directamente optimizando las funciones de coste. Debemos señalar que bajo ciertas suposiciones de las funciones de densidad, los dos enfoques son equivalentes.

Sea la clasificación o la regla de decisión que se elija, se deberá entrenar a partir de los eventos de entrenamiento disponibles. Como resultado, el rendimiento de un clasificador dependerá tanto del número disponible de eventos de entrenamiento como de sus valores o características.

Al mismo tiempo, el objetivo de diseñar un sistema de reconocimiento es clasificar futuros eventos que podrán ser diferentes de los eventos de entrenamiento. Es decir, optimizar un clasificador para que maximice su rendimiento en la fase de entrenamiento no siempre permitirá obtener el rendimiento deseado en la fase de test o validación.

2.3. Modelos de clasificación más habituales

Los modelos de clasificación se utilizan para asignar elementos a un grupo discreto o una clase sobre la base de un conjunto de características específicas. Cada modelo tiene sus propias fortalezas y debilidades según el escenario donde sea aplicado. Cierto es que no hay ninguna regla teórica que pueda determinar cuál es el que debe utilizar sin simplificar las consideraciones. La elección de un modelo de clasificación de datos también está estrechamente vinculada con la comprensión del objetivo que se persigue y las características que se pretenden utilizar.

Preguntas como: ¿cuántos y cómo son los datos que se tienen?, ¿qué tipo de datos son?, ¿cuál es el objetivo?, ¿es importante visualizar el proceso?, ¿qué nivel de detalle se exige? y ¿es un factor limitante el almacenamiento? son de respuesta necesaria para saber a qué problema nos hemos de enfrentar.

Una vez entendido el tipo de datos con los cuales se ha de trabajar y teniendo claras las categorías en donde se quieren clasificar las señales, se pueden empezar a buscar los puntos fuertes de cada uno de los modelos de clasificación. Hay algunas reglas genéricas para ayudar a elegir el mejor modelo de clasificación, pero son sólo puntos de partida. Si se pretende trabajar con una gran cantidad de datos, elegir el enfoque correcto a menudo es un procedimiento empírico para lograr el equilibrio adecuado de complejidad, rendimiento y precisión. En las siguientes subsecciones se describen algunos de los clasificadores más comunes.

2.3.1. Naive Bayes

Si los datos no son muy complejos y la clasificación es relativamente sencilla se puede probar con un algoritmo *Naive Bayes*. Es un clasificador de baja varianza, más eficiente que la regresión logística y que los algoritmos de vecinos más cercanos (*Nearest Neighbor*

(kNN) cuando se trabaja con una cantidad limitada de datos disponibles para entrenar un modelo.

Naive Bayes es también una buena opción cuando los recursos de CPU y memoria son un factor limitante, dado que es muy simple, no tiende a sobreajustar los datos y puede ser entrenado muy rápidamente. También se adapta bien a nuevos datos utilizados para actualizar el clasificador.

Si los datos crecen en tamaño y varianza, y se necesita un modelo más complejo, lo más seguro es que otros clasificadores trabajen mejor, dado que su análisis simple no es una buena base para hipótesis complejas. *Naive Bayes* es a menudo el primer algoritmo a probar cuando se trabaja con el texto (filtros de spam y análisis de los sentimientos).

2.3.2. *k* Vecinos más cercanos *k Nearest Neighbor*

Categorizar los puntos de representación de los eventos en función de su distancia a otros puntos en un conjunto de datos de entrenamiento puede ser una manera simple pero eficaz de clasificar los datos. Esta es la filosofía de los algoritmos de kNN, los cuales se basan en la distancia entre un evento y los k más cercanos para obtener sus métricas.

En los métodos kNN no hay fase de entrenamiento en sí. Se insertan los datos en el modelo y se espera a utilizar el clasificador para ver su comportamiento. Cuando se necesita una nueva instancia, el modelo kNN busca el número especificado de k vecinos más cercanos; de forma que, si $k = 5$, se encontrarán las categorías de los 5 vecinos más cercanos. Si lo que se busca es aplicar una etiqueta o clase, el modelo realizará una primera búsqueda para encontrar donde debe ser clasificado.

Como bien se ha descrito anteriormente, el tiempo de entrenamiento de los algoritmos kNN es corto, pero a la hora de clasificar será más costoso que el de otros modelos, Además, el espacio de almacenamiento necesario será bastante mayor. Cuando el número de puntos de datos aumenta, al mantener todos los datos de entrenamiento a la hora de clasificar, tanto el tiempo como el espacio requerido aumentarán de manera circunstancial

El mayor inconveniente de este método es que puede ser engañado por atributos irrelevantes que oscurecen atributos importantes. Otros modelos, como los árboles de decisión son más capaces de ignorar estas distracciones. Hay maneras de corregir este problema, tales como la aplicación de pesos a los datos de manera empírica.

2.3.3. Árboles de decisión

En los algoritmos basados en árboles de decisión se pueden seguir las decisiones en el árbol desde el nodo raíz (inicio) hasta los nodos hoja (los cuales que contienen las categorías) para ver cómo se predice una respuesta. Los árboles de clasificación dan respuestas que son nominales, tales como verdadera o falsa. Poder visualizar la trayectoria completa tomada por un sonido al entrar en este tipo de clasificadores es especialmente útil si se tienen que compartir los resultados con las personas interesadas. Además, son relativamente rápidos.

La principal desventaja de los árboles de decisión es que tienden a sobreajustar, sin embargo, existen métodos conjuntos para contrarrestar y controlar esto.

2.3.4. *Support Vector Machine* (SVM)

Es posible utilizar *Support Vector Machine* (SVM) cuando los datos tienen exactamente dos clases. Este método clasifica los datos mediante la búsqueda del mejor hiperplano que separe todos los puntos de datos de una clase de los de la otra clase (el mejor hiperplano para un SVM es el que tiene el mayor margen entre las dos clases). Es posible utilizar el SVM con más de dos clases, en cuyo caso el modelo creará un conjunto de subproblemas de clasificación binaria (con un solo SVM para cada subproblema).

Hay un par de grandes ventajas por las que utilizarlo: En primer lugar, es extremadamente preciso y tiende a no sobreentrenar los datos. En segundo lugar, una vez entrenados, es una opción rápida, ya que decidirá entre una de los dos clases. Dado que los modelos SVM son muy rápidos, una vez que su modelo ha sido entrenado se pueden descartar los datos de entrenamiento si no tiene suficiente memoria disponible. También tiende a manejar clasificaciones no lineales complejas de manera satisfactoria.

Sin embargo, los SVM necesitan ser entrenados concienzudamente, por lo que se debe que invertir tiempo en el modelo antes de comenzar a utilizarlo.

2.4. Sobreentrenamiento o sobreajuste *Overfitting*

El término overfitting significa que el modelo de clasificación está tan estrechamente alineado con los datos de entrenamiento, que ante nuevos datos probablemente arrojará más errores. Una de las razones por las cuales el *overfitting* es difícil de evitar es que a menudo es el resultado de un conjunto de datos de entrenamiento insuficientes o un elevado número de características que hacen que el clasificador se especialice mucho en esos datos concretos.

Un modelo sobreajustado devuelve muy pocos errores, lo que hace que parezca atractivo a primera vista. Desafortunadamente, hay demasiados parámetros en el modelo en relación con el sistema subyacente. El algoritmo de entrenamiento sintoniza estos parámetros para reducir al mínimo la función de pérdida, pero estos resultados en el modelo sobreajustan a los datos de entrenamiento, en lugar de permitir el comportamiento deseado con nuevos datos, de hecho, cuando grandes cantidades de nuevos datos se introducen a la red, el algoritmo no puede hacer frente y pueden surgir problemas.

Idealmente, el modelo de clasificación debe ser lo más simple posible y lo suficientemente preciso para producir resultados significativos. Cuanto más complejo sea el modelo, más propenso será al overfitting. Básicamente, la mejor manera de evitarlo es asegurándose de que se están utilizando suficientes datos de entrenamiento.

2.4.1. Validación cruzada *Cross-Validation*

La validación cruzada es una técnica de evaluación del modelo utilizado para analizar el rendimiento de un algoritmo de aprendizaje automático al hacer predicciones sobre

nuevos conjuntos de datos donde no ha sido entrenado. Esto se hace particionando un conjunto de datos, usando uno de los subconjuntos para entrenar el algoritmo y los restantes para las pruebas de clasificación. La validación cruzada no utiliza todos los datos para construir un modelo, es un método común para evitar el sobreajuste durante el entrenamiento.

Un paso importante cuando se trabaja con *machine learning* es la comprobación del funcionamiento del modelo. Esta técnica consiste en que el algoritmo haga predicciones a partir de datos no utilizados durante la etapa de entrenamiento. Cada "ronda" de validación cruzada implica particionar al azar el conjunto de datos original en un conjunto de entrenamiento y establecer un test. El conjunto de entrenamiento se utiliza entonces para formar un algoritmo de aprendizaje supervisado, y el conjunto de prueba se utiliza para evaluar su rendimiento. Este proceso se repite varias veces y el promedio error de validación cruzada se usa como un indicador de rendimiento. Las técnicas de validación cruzada comunes son las siguientes:

- k-veces: consiste en particionar los datos en k subconjuntos escogidos al azar de aproximadamente el mismo tamaño. Un subconjunto se utiliza para validar el modelo entrenado utilizando los subgrupos restantes. Este proceso se repite k veces, de tal manera que cada subconjunto se utiliza exactamente una vez para su validación.
- *Holdout*: consiste en particionar los datos en exactamente dos subconjuntos de relación especificada para el entrenamiento y test.
- Submuestreo repetido al azar: consiste en realizar simulaciones de Monte Carlo de particiones aleatorias de datos y agrega de los resultados de todas ellas.
- Estratificar: consiste en particionar los datos de tal forma que los conjuntos de entrenamiento y test tienen más o menos las mismas proporciones de clase/categoría en la respuesta.
- Resustitución: en ella no se distribuyen los datos; sino que se utilizan para su validación. A menudo produce estimaciones demasiado optimistas para el rendimiento y debe ser evitado si hay datos suficientes.

2.5. Conclusiones

A lo largo de este capítulo se han explicado los principales algoritmos de selección de características y clasificadores necesarios para introducir términos que se utilizarán en los siguientes capítulos de la presente tesis doctoral.

Además, se han resumido conceptos fundamentales para que la comprensión por parte del lector sea lo más clara posible, dotando al lector de los conocimientos teóricos utilizados.

Capítulo 3

Presentación y contextualización de las señales. Enfoques y problemática

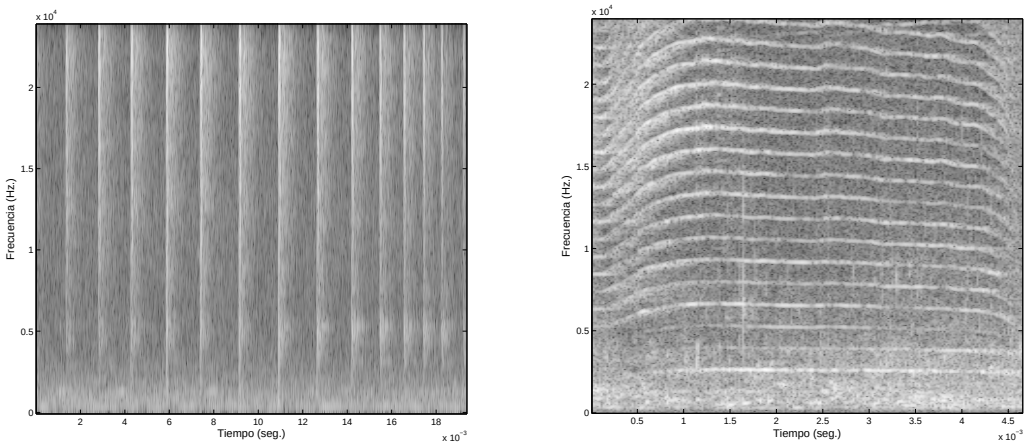
3.1. Introducción

En este capítulo se presenta una clasificación basada en la morfología del espectrograma de la señal para obtener una primera idea de las particularidades de las señales a clasificar y su entorno. Se seleccionarán las características más relevantes y a través de la elección de un clasificador óptimo, se obtendrán conclusiones acerca de si es posible realizar una buena clasificación. Además se analizará y discutirá si el enfoque de clasificación se adapta al problema de clasificación de sonidos que nos atañe. Para ello se estudiarán exhaustivamente los resultados obtenidos y se enumerarán las particularidades de las señales producidas por los odontocetos.

3.2. Los sonidos producidos por los odontocetos

En la mayoría de estudios acerca de los sonidos de los odontocetos [12, 13] se establece una clasificación teniendo en cuenta el comportamiento de las señales en el espectrograma, o lo que es lo mismo, según su morfología, atendiendo a la horizontalidad o verticalidad de los patrones obtenidos [14].

Señales como la de la Fig. 3.1a se denominaron sonidos pulsados, en referencia a que los impulsos que las componen se visualizan en el espectrograma como líneas verticales. Dentro de las señales pulsadas se encuentran las ecolocalizaciones, señales con una frecuencia de repetición de los impulsos baja (del orden de Hz), utilizadas para la localización de objetos a través del eco que se produce cuando el impulso rebota en ellos. Esta frecuencia está totalmente controlada dependiendo de lo cerca o lejos que esté el objeto en particular; cuanto más lejos, más tarda el eco en regresar al odontoceto y más tarda en producirse el siguiente impulso. Cuanto más cerca, sin embargo, incrementa la frecuencia con la que emiten impulsos, dado que el eco regresa antes [15]. No todos los sonidos pulsados son utilizados para ecolocalizar, ya que, conforme la frecuencia de



(a) (b)

Figura 3.1: a) Espectrograma de una señal pulsada. b) Espectrograma de una señal tonal.

repetición aumenta, su función pasa a ser comunicativa [16].

Por otro lado, se establece el nombre de sonidos tonales a las señales componentes horizontales en el espectrograma, como la de la Fig. 3.1b. Estas señales son utilizadas para comunicarse principalmente con individuos de su misma especie. Son usados en las interacciones sociales para identificarse como individuos [17, 18].

En la Tabla 3.1 se resumen las propiedades de estos dos tipos de sonidos teniendo en cuenta su morfología. Las señales pulsadas [19], poseen una energía alrededor de tres veces mayor que las señales tonales. Sin embargo su frecuencia fundamental es mucho menor que en las señales tonales, los rangos de frecuencia de cada una de las señales se resumen en $0{,}5 - 700$ Hz para los sonidos pulsados y de $700 - 5000$ Hz para sonidos tonales/silbido. Estos valores concretos se han obtenido a partir del análisis de los sonidos registrados por los biólogos del Oceanográfico de Valencia de las dos ballenas belugas en cautividad que posee el parque. Los valores de la frecuencia fundamental son parecidos en todos los odontocetos.

Características	**Señal pulsada**	**Señal tonal**
Morfología del espectrograma	vertical	horizontal
Rango Freq. fund.	[0.5 - 700 Hz]	[700 - 10000 Hz]
Intensidad señal	alta	baja
Propósito	ecolocalización/comunicación	comunicación

Tabla 3.1: Propiedades generales de los dos principales tipos de sonidos de los odontocetos, concretamente en las ballenas beluga.

Cabe decir que, en una gran cantidad de ocasiones, los sonidos no son puramente

tonales o puramente pulsados, sino que existen sonidos con un comportamiento mixto en el espectrograma (Fig. 3.2). Son los sonidos mixtos, de los cuales también hablaremos dada su importancia.

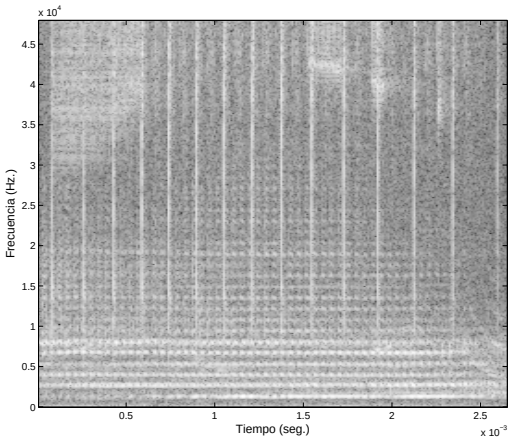


Figura 3.2: Espectrograma de una señal mixta producida por una ballena beluga.

3.3. Descripción de las características extraídas

A raíz del trabajo realizado junto con los investigadores del Oceanográfico se extrajeron características de todo tipo, pensando que algunas de ellas pudieran servir para distinguir entre las diferentes categorías de sonidos. No todas las características estuvieron basadas en las representaciones tiempo-frecuencia (espectrogramas) de las señales, sino también en las propiedades estadísticas de la señal. Además otros parámetros fueron seleccionados por la información que nos aportaban sobre los órganos internos que las ballenas belugas emplean a la hora de emitir sus sonidos (características que miden las no linealidades, el pitch, etc). En la Tabla 3.2 se muestra el conjunto de características extraídas.

Las características 1 a la 9 están definidas en la Fig. 3.3, donde f_0, f_1 y f_2 son respectivamente la frecuencia fundamental, el segundo armónico y el tercer armónico. El ancho de banda a -3dB (Δf_0, Δf_1 y Δf_2) se emplean para obtener el factor Q dadas las ecuaciones dentro de la Tabla 3.2. La densidad espectral de potencia en las tres frecuencias fundamentales también son calculadas (Fig. 3.3).

Algunas características muestran estadísticas de las señales o de otros parámetros extraídos. Si vemos estos sonidos como procesos estocásticos, la *skewness* (característica 10) mide la asimetría de la distribución de probabilidad haciendo la media aritmética tal y como se describe en la Ec. (3.1). La *kurtosis* (característica 11) mide lo picuda que es la densidad de probabilidad de la señal y se obtiene igualmente mediante el promedio de los eventos mostrado en la Ec. (3.2).

Número	Descripción
$v1$	Frecuencia fundamental f_0
$v2$	Factor Q de $f_0 = \Delta f_0 / f_0$
$v3$	Densidad espectral de potencia de la frecuencia $S_x(f_0)$
$v4$	Frecuencia fundamental f_1
$v5$	Factor Q de $f_1 = \Delta f_1 / f_1$
$v6$	Densidad espectral de potencia de la frecuencia $S_x(f_1)$
$v7$	Frecuencia fundamental f_2
$v8$	Factor Q de $f_2 = \Delta f_2 / f_2$
$v9$	Densidad espectral de potencia de la frecuencia $S_x(f_2)$
$v10$	*Skewness* de la vocalización
$v11$	*Kurtosis* de la vocalización
$v12$	Test de autocavarianza de la vocalización
$v13$	Medida de la reversibilidad temporal de la vocalización
$v14$	Sonoridad
$v15$	*Skewness* de la sonoridad
$v16$	*Kurtosis* de la sonoridad
$v17$	Test de autocavarianza de la sonoridad
$v18$	Medida de la reversibilidad temporal de la sonoridad

Tabla 3.2: Conjunto de parámetros empleados en el clasificador.

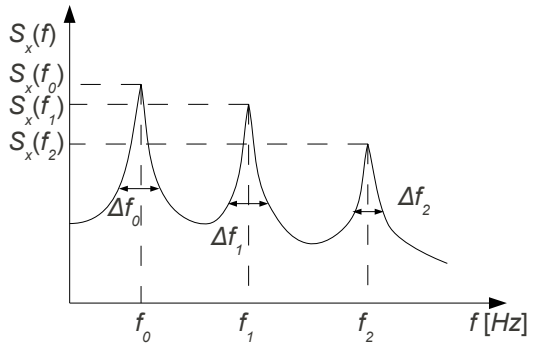


Figura 3.3: Definición de los parámetros frecuenciales.

$$v_{10} = E\left[\left(\frac{x-\mu}{\sigma}\right)^3\right] \approx \frac{1/N\sum_{n=1}^{N}\left(x(n)-\overline{x}\right)^3}{\left(1/N\sum_{n=1}^{N}\left(x(n)-\overline{x}\right)^2\right)^{3/2}} \tag{3.1}$$

$$v_{11} = E\left[\left(\frac{x-\mu}{\sigma}\right)^4\right] - 3 \approx \frac{1/N\sum_{n=1}^{N}\left(x(n)-\overline{x}\right)^4}{\left(1/N\sum_{n=1}^{N}\left(x(n)-\overline{x}\right)^2\right)^2} - 3 \tag{3.2}$$

Se incluyeron además, algunos parámetros simples que dieran información sobre la no linealidad de los sonidos: la autocovarianza de tercer orden y la reversibilidad temporal (parámetros 12 y 13 respectivamente). Se denomina autocovarianza de tercer orden a una extensión de mayor orden de la autocavarianza tradicional, tal y como se muestra en la Ec. (3.3), donde $E[\cdot]$ es un operador que calcula el valor medio de la esperanza. La estimación de este parámetro está explicada en [20].

$$v_{12} = t^{C3} = E[x(n)x(n-1)x(n-2)] \tag{3.3}$$

El parámetro que mide la reversibilidad temporal (t^{REV}) es una estimación discreta de la *skewness* o asimetria de la pendiente de la forma de onda normalizada por una estima de la desviación típica al cubo ($\widehat{\sigma}^3$) [21]. Esta medida se obtiene con la Ec. (3.4).

$$v_{13} = t^{REV} = \frac{1}{\widehat{\sigma}^3 \cdot N} \sum_{n=1}^{N} \left(\frac{x(n) - x(n-1)}{T_s} \right)^3 \tag{3.4}$$

Para series temporales la reversibilidad temporal [20] se espera que sea aproximadamente cero. Sin embargo, los procesos que son temporalmente irreversibles tienen un valor de $t^{REV} > 0$ ó $t^{REV} < 0$.

La caracterización de la sonoridad o no sonoridad (parámetro 14) se inspira en los algoritmos lineales típicamente empleados para la síntesis de la voz humana [22]. Para obtener este parámetro se estima la autocorrelación y el periodo de pitch T_p y se normaliza por la energía del sonido y el factor $1 - \frac{T_p}{N}$ para compensar la atenuación triangular de la función de autocorrelación. Este parámetro es capaz de discriminar las vocalizaciones sonoras de las no sonoras (ver Ec. (3.5)).

$$v_{14} = \frac{1}{1 - T_p/N} \cdot \frac{R_{ee}(T_p)}{R_{ee}(0)} \tag{3.5}$$

3.4. Fase de entrenamiento: selección de características y clasificadores

Se eligió un algoritmo de selección de características secuencial (SFS) [10] que permitió obtener resultados característica a característica, seleccionando y ordenando las características más representativas. De esta manera cada una de las características se priorizaron según la importancia que tenían a la hora de decidir entre una categoría u otra.

Se utilizó un enfoque basado en las densidades de probabilidad dado su carácter generalista necesario a la hora de dar cabida a un gran número de características. Para ello se eligió un clasificador gaussiano, en concreto el algoritmo de *Naive Bayes*. Se trata un clasificador probabilístico basado en la estadística bayesiana con una suposición de independencia entre las características extraídas el cuál consta de dos fases bien diferenciadas a la hora de realizar la clasificación:

I. Se utilizan los eventos de entrenamiento para estimar los parámetros de las funciones de densidad de probabilidad.

II. Se predice la probabilidad de que los eventos pertenezcan a cada una de las clases. La suposición de independencia de clase condicionada simplifica la fase de entrenamiento. Esto permite una mejor estimación de los parámetros de *Naive Bayes* que se requieren para la clasificación y permite utilizar menos eventos de entrenamiento que otros clasificadores.

Dado que a priori se desconocía el tipo de distribución más apropiada se probó con varias de ellas. A continuación se enumeran junto a una breve descripción:

- La distribución normal es apropiada para características que tienen una distribución normal en cada clase. Para cada característica modelada con una distribución normal, el clasificador *Naive Bayes* estima una distribución normal separada para cada clase calculando la media y la desviación típica del conjunto de entrenamiento.
- La distribución *kernel* es apropiada para características que tienen una distribución continua. No asume una condición tan restrictiva como la distribución normal y se puede usar en casos donde la distribución de cada característica puede ser sesgada o tener múltiples modos. Necesita más tiempo de computación y más memoria que la distribución normal. El clasificador *Naive Bayes* calculará para cada característica una estimación de la función densidad de probabilidad *kernel* para cada clase del conjunto de entrenamiento.
- La distribución *Multinomial* es apropiada cuando todas las características representan conjuntos de eventos. El clasificador cuenta las probabilidades relativas de cada conjunto para cada clase y define la distribución *Multinomial* para cada fila dada por el vector de probabilidades de la clase correspondiente.
- La distribución *Multivariate Multinomial* es apropiada para características que definen categorías. Para cada característica que se modela con esta distribución, el clasificador de *Naive Bayes* calcula un conjunto separado de probabilidades para el conjunto de características por cada clase.

Además de estos cuatro clasificadores Gaussianos *Naive Bayes*, se incluyeron dos clasificadores discriminantes (diaglinear y diagcuadratico), debido que están basados en el cálculo de las matrices de covarianza, característica adecuada para una cantidad de características como las que se necesitan evaluar.

Con este objetivo y con el de comprobar la importancia de cada una de las características con los clasificadores propuestos, se eligieron para el conjunto de entrenamiento una serie de sonidos realizados por las ballenas belugas del Oceanográfico de Valencia, en concreto se utilizaron 20 sonidos de ruido antropogénico, 20 sonidos tonales, 20 sonidos pulsados y 20 sonidos de mandíbula *Jawclap*, todos ellos clasificados por los biólogos creando un vector de categorías de referencia, para poder evaluar los resultados de los clasificadores, comparándolo con las categorías predichas.

Cabe destacar que los sonidos mixtos no han sido incluidos en este modelo de clasificación dada su comportamiento combinado (horizontal y vertical) en su diagrama tiempo-frecuencia.

Clasificador	Orden de relevancia de las características	Opt. #	Tasa de Error
Diaglinear	6,15,2,10,5,17,12,7,9,14,11,1,4,16,18,3,8,13	10	21.73
Diagquadratic	9,15,3,10,2,18,11,7,1,17,14,16,4,13,8,5,16,12	11	20.77
N.B. Gaussian	1,15,6,11,8,17,9,14,18,10,2,7,3,4,13,5,16,12	11	19.49
N.B. Kernel	12,16,6,2,10,13,7,15,1,4,18,8,17,3,11,14,9,5	11	5.11
N.B. Multinomial	9,15,6,3,10,12,16,18,13,1,14,17,4,7,11,2,8,5	12	26.52
N.B. Multivariate Multinomial	1,4,3,6,7,8,5,9,10,11,12,13,14,2,15,16,17,18	4	0.96

Tabla 3.3: Número óptimo de características y características más significativas ordenadas de más a menos significativas. (N.B.= *Naive Bayes*)

Para cada uno de los clasificadores, según el algoritmo SFS, se calculó el orden óptimo y el número óptimo de características para minimizar la tasa de error en la fase de entrenamiento. Los resultados del SFS con los clasificadores *Naive Bayes* se pueden observar en la Tabla 3.3.

Dados los resultados obtenidos en la Tabla 3.3 para la fase de entrenamiento se puede concluir que:

- El mejor clasificador durante esta etapa de entrenamiento fue el basado en *Naive Bayes* con la configuración *Multivariate Multinomial*. Este lograba menos de 1 % de tasa de error al clasificar con únicamente cuatro características (1, 4 ,3 y 6).
- El clasificador *Naive Bayes* basado en la distribución *kernel* fue el segundo mejor, ya que usando 11 características conseguía menos de un 5 % de tasa de error al clasificar.
- Los demás clasificadores tenían un comportamiento bastante similar con alrededor de un 20 % de tasa de error, siendo el clasificador basado en la configuración *Multinomial* el que peor prestaciones arrojaba.

Se seleccionaron las características más importantes de los sonidos (obtenidas por el SFS), diseñándose los dos clasificadores para la fase de test/validación (*Multivariate Multinomial* y *kernel*) con el número de características óptimas de cada uno de ellos.

3.5. Resultados de la fase de test/validación

En esta fase de test se utilizaron 45 sonidos de ruido antropogénico, 58 sonidos tonales, 47 pulsados y 13 *Jawclaps*, en total 163 sonidos. Se instaló en los tanques de agua de las ballenas belugas el hidrófono Bruel Kjaer modelo 8103 (sensibilidad: -211 dB re 1V/uPa +- 2 dB, rango de respuesta: 0.1 Hz hasta 100.0 kHz +1.0/-6.0 dB), conectado a un amplificador y acondicionador de carga Brüel Kjaer modelo Nexus 2690 (filtros de paso: 10.0 Hz - 80.0 kHz, ganancia: 0 a 80 dB). La señal amplificada y acondicionada se digitalizó en tiempo real a 24 bits /192 kHz (filtro Nyquist: 80.0 kHz) por medio de una

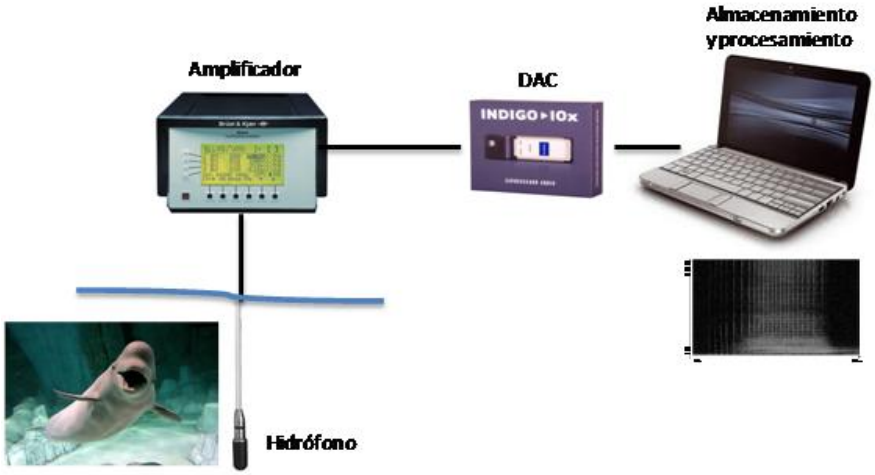


Figura 3.4: Configuración del sistema de medida.

tarjeta de adquisición de sonido iOX- INDIGO la cual estaba conectada a un ordenador portátil (procesador 4 GHz Intel Core-Duo, 1GB RAM).

Las Tablas 3.4 y 3.5 presenta las matrices de confusión para los dos mejores clasificadores de *Naive Bayes* obtenidos en la fase de entrenamiento, es decir, con distribución *Multivariate Multinnormal* y con distribución kernel, obtenidas cada una de las dichas matrices eligiendo las características óptimas seleccionadas en la fase de entrenamiento, es decir, 11 y 4 respectivamente.

Se introduce el parámetro TPR como la tasa de aciertos positivos o *True Positive Rate* que como su nombre indica representa un coeficiente entre los sonidos clasificados correctamente y el total de los sonidos para una categoría en concreto.

Clasificador *N. B. Multivariate Multinomial* (4 carac.)

Clasificado\Real	Ruido	Tonal	Pulsado	*Jawclap*
Ruido	24	0	0	0
Tonal	0	33	1	0
Pulsado	0	1	29	0
Jawclap	0	0	0	13
No clasificado	21	24	18	0
TPR	53.33 %	58.62 %	61.70 %	100 %

Tabla 3.4: Matriz de confusión obtenida con el clasificador *Naive Bayes Multivariate Multinomial*, referente a la fase de test (163 sonidos).

Se resumen a continuación los resultados más relevantes:

- El clasificador *Naive Bayes* Multivariado Multinomial logra una tasa de acierto global entorno a un 60 %, ya que de los 163 sonidos a clasificar, sólo lo hace bien

Clasificador *N. B. Kernel* (11 carac.)

Clasificado\Real	Ruido	Tonal	Pulsado	*Jawclap*
Ruido	41	5	4	3
Tonal	0	47	9	0
Pulsado	2	3	32	1
Jawclap	1	1	2	9
No clasificado	1	2	0	0
TPR	91.11 %	81.03 %	68.08 %	69.23 %

Tabla 3.5: Matriz de confusión obtenida con el clasificador *Naive Bayes kernel*, referente a la fase de test (163 sonidos).

en 99 ocasiones. Realizando un análisis más en detalle llama la atención que esta modalidad de *Naive Bayes* tiende a no clasificar los sonidos que no tiene claro a que categoría corresponden. Este comportamiento puede resultar ventajoso ya que cuando el clasificador realiza una predicción no suele equivocarse, a costa de que el porcentaje de perdidas o *missing rate* es cercano al 40 % de las señales.

- Respecto al clasificador *Naive Bayes* en configuración *kernel* usando el número y orden óptimo de características tiene un comportamiento mucho mejor que el algoritmo analizado anteriormente, presentando una tasa de acierto global en entorno a un 79 %. Esto se refleja en los diferentes TPR de cada una de las categorías de sonidos, destacando para bien la obtenida para los ruidos antropogénicos.

Haciendo hincapié en una de las características más relevantes obtenidas mediante el algoritmo SFS se obtiene la Fig. 3.5, donde se puede observar el clustering obtenido para las diferentes categorías para la característica v_3 (densidad espectral de potencia de la frecuencia fundamental $S_x(f_0)$), característica v_{14} (sonoridad) y característica v_{12} (test de autocavarianza). Es interesante observar como con sólo tres características el clasificador no permite distinguir de manera satisfactoria entre las diferentes categorías.

3.6. Discusiones, problemas y detalles acerca de la clasificación basada en la morfología del espectrograma

Esta primera experiencia muestra como la extracción de características de forma amplia, es decir, extrayendo un buen número de ellas, unido a la utilización de un algoritmo adecuado de selección de características puede ser una buena alternativa si no se tiene un conocimiento muy extenso del entorno donde se está trabajando. Algoritmos de clasificación Bayesianos basados en la densidad de probabilidad permiten obtener y conocer información importante.

Acorde con la clasificación proporcionada por los biólogos en tres categorías: sonidos tonales (comunicativos), sonidos pulsados (comunicativos y agresivos) y palmadas

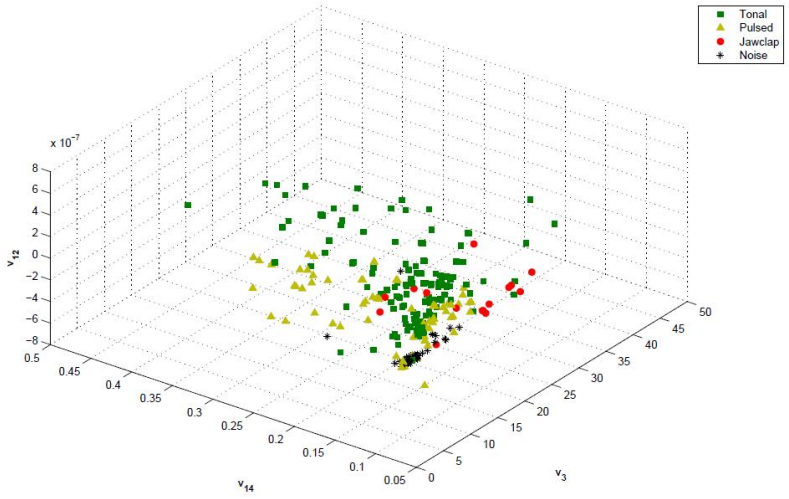


Figura 3.5: Clustering de las diferentes categorias con las características v_3 (Densidad espectral de potencia de la frecuencia fundamental $S_x(f_0)$), característica v_{14} (sonoridad) y característica v_{12} (test de autocovarianza).

(agresivos), la selección de características SFS permite mostrar como las características relacionadas con la sonoridad del habla y los estadísticos de tercer orden dan una información muy valiosa a la hora de clasificar.

Varias lineas de investigación se abren a partir de este trabajo, como el análisis exhaustivo de la sonoridad de las vocalizaciones. Los resultados indican que las ballenas belugas podrían tener la habilidad controlar la vibración de sus órganos internos que pudieran actuar como las cuerdas vocales en los humanos.

3.6.1. Propiedades de las vocalizaciones: Medida de la no linealidad de sonidos de ballenas belugas

En [23] se muestra que el mecanismo de producción de no linealidades permite a los individuos generar sonidos impredecibles y altamente complejos sin necesidad de mecanismos neuronales complejos equivalentes. En [24] se observó y se midió la presencia de no linealidades para ballenas jorobadas. De una manera similar, se observa la presencia de no linealidades en los sonidos producidos por las ballenas beluga. Para poder mostrarlo se presenta el *clustering* obtenido con el detector y clasificador automático diseñado donde las características v_{14} (sonoridad) y $\mid v_{13} \mid$ (medida de la reversibilidad temporal) están relacionadas con la existencia de dinámicas no lineales en los mecanismos de producción del sonido.

Tal y como y como se puede ver en la Fig. 3.6, casi todas las palmadas así como un

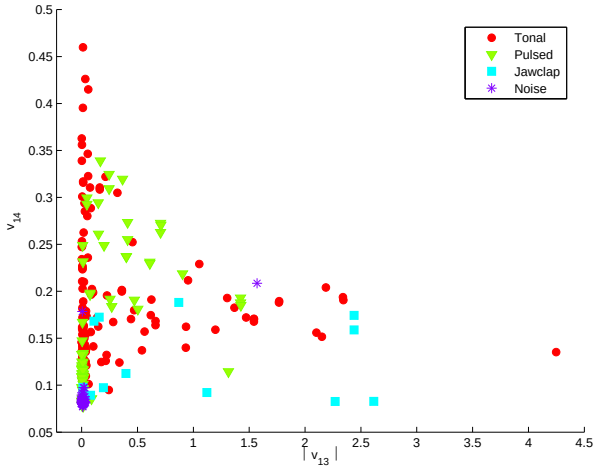


Figura 3.6: Clustering de las diferentes categorias con las características | v_{13} | (medida de la reversibilidad temporal) y v_{14} (sonoridad).

porcentaje considerable de las vocalizaciones pulsadas y tonales parecen ser producidas por mecanismos no lineales (valores de la reversibilidad temporal no nulos). Es por ello interesante realizar un estudio con más detalle para tener la seguridad que los valores altos obtenidos para la reversibilidad temporal (valores indicadores de la no linealidad) son realmente producidos por no linealidades.

El estudio de las no linealidades en las ballenas belugas podría ser un factor clave en el estudio de la anatomía de las belugas. Por ejemplo, la presencia de no linealidades en las vocalizaciones, podría indicar que algunos de los sonidos tonales complejos producidos por la belugas podrían generados por una vibración irregular parecida a la que poseen los mamíferos terrestres [23].

Es importante resaltar que no en todas las vocalizaciones que aparecen en la Fig. 3.6 y que muestran un alto v_{13} son debidas a una dinámica no lineal. Esta característica se calcula en el análisis de clasificación para toda la duración de la vocalización y podría indicar la no estacionaridad.

3.6.2. La importancia de la ventana de análisis en el cálculo del diagrama tiempo-frecuencia

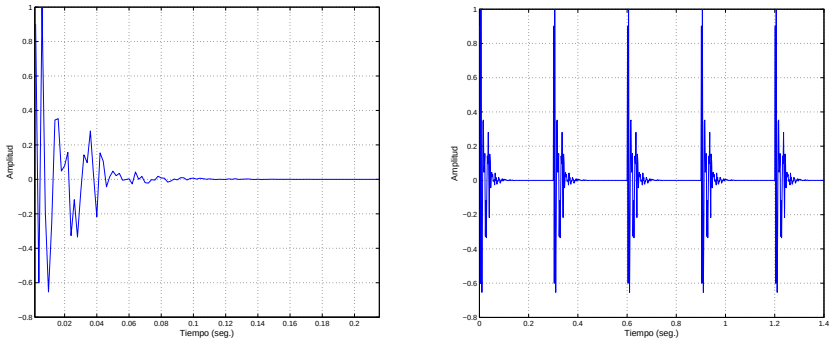
A continuación, se estudia cómo sonidos que provienen de una misma naturaleza de producción, pueden interpretarse de una manera diferente. Aunque el aspecto del espectrograma de señales pulsadas y las señales tonales vibratorias es totalmente diferente, esto no implica que este tipo de sonidos hayan sido producidos por órganos totalmente diferentes. Ambas están provocadas por un elemento vibrante, cada uno especializado en un rango de frecuencias y con una energía diferente.

Se pretende reflexionar sobre un parámetro utilizado para la obtención del diagrama

tiempo-frecuencia o espectrograma: la Longitud de la Ventana de Análisis (LVA). Esta variable es una de las razones por las cuales una clasificación de sonidos basada en el diagrama tiempo-frecuencia o espectrograma de la señal puede resultar muchas veces engañosa.

En la Ec. (3.6) modelamos el movimiento de una membrana en todas direcciones $g(n)$ a partir de una señal de ruido blanco gaussiano de media nula con varianza unidad $\eta(n)$ atenuada por una función exponencial (Fig. 3.7a).

$$g(n) = \eta(n)exp(-\sigma n) \quad \text{donde} \quad \sigma \leq 0{,}1 \tag{3.6}$$



(a) (b)

Figura 3.7: a) $g(n)$ Simulación del impulso u onda provocado por el movimiento de una lengüeta en todas direcciones. b) $x(n)$ Representación de la repetición periódica de $g(n)$.

Si se convoluciona $g(n)$ por un tren de deltas con separación T_0, se simulará la vibración de una lengüeta (Fig. 3.7b) con una frecuencia de repetición $f_0 = 1/T_0$, obteniendo $x(n)$ (ver Ec. (3.7)). Se ha observado empíricamente que este modelo es apropiado para simular dicha vibración.

$$x(n) = g(n) * \sum_{s=-\infty}^{\infty} \delta(n - sT_0) \tag{3.7}$$

A continuación se modela el movimiento de la ventana de análisis para la obtención del espectrograma mediante la Ec. (3.8)

$$w_{\tau_l} = w(n - \tau_l) = \begin{cases} a_0 + a_1 cos(\frac{2\pi(n-\tau_l)}{\text{LVA}-1}) & \tau_l \leq n \leq \text{LVA} + \tau_l \\ 0 & 0 \leq n \leq \tau_l \text{ y LVA} + \tau_l \leq n \leq \text{TAM} \end{cases} \tag{3.8}$$

donde w_{τ_l} es la ventana de análisis utilizada, $\tau_l = l \cdot d$ con $0 \leq l \leq N-1$, siendo $d = \text{LVA}/2$ el desplazamiento con el que se mueve la ventana de análisis y $N = \text{TAM}/d - 1$ el número de ventanas w_{τ_l} que necesitan hasta cubrir todo $x(n)$ y TAM el tamaño de la señal $x(n)$.

La variable w_{τ_l} corresponde a una ventana Hamming con $a_0 = 0{,}538$ y $a_1 = 0{,}461$. Para cada τ_l se obtiene su w_{τ_l}, consiguiendo N vectores x_{τ_l} guardados en X_w.

$$\left[X_w\right] = \begin{bmatrix} x_{\tau 0} \\ x_{\tau 1} \\ \cdots \\ \cdots \\ x_{\tau(N-2)} \\ x_{\tau(N-1)} \end{bmatrix} = \begin{bmatrix} w_{\tau 0} \\ w_{\tau 1} \\ \cdots \\ \cdots \\ w_{\tau(N-2)} \\ w_{\tau(N-1)} \end{bmatrix} \left[x(n)\right] \tag{3.9}$$

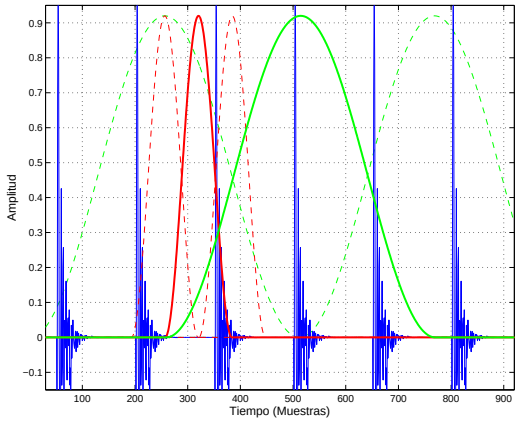


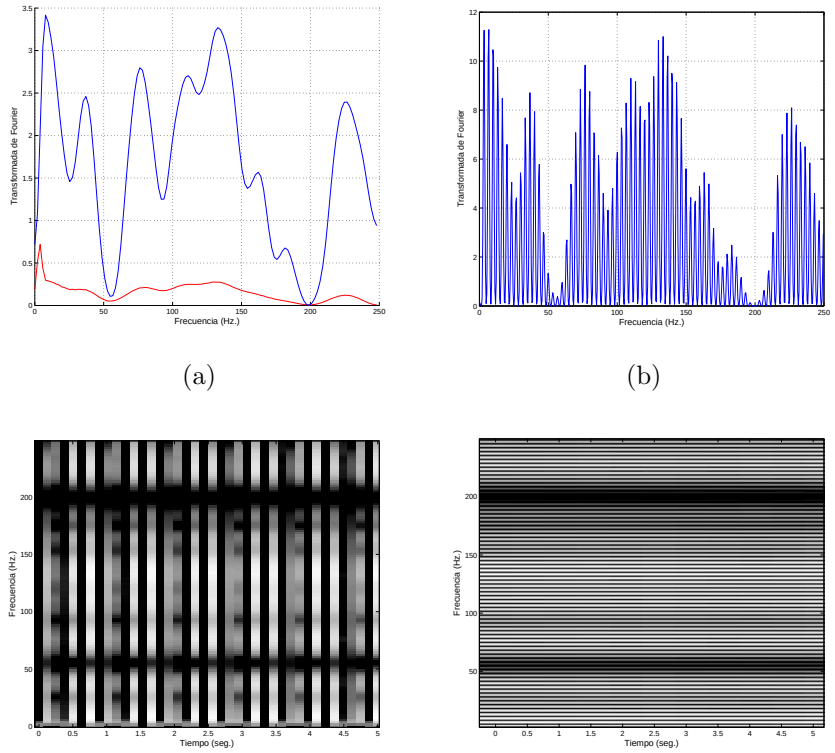
Figura 3.8: Detalle de $y(n)$ (curva azul) junto a dos ventanas de análisis (curva roja LVA= 128 y curva verde LVA= 512). Las curvas discontinuas rojas y verdes muestran como se desplazan las ventanas de análisis para obtener los espectrogramas de las Fig. 3.9c y 3.9d respectivamente.

En la Fig. 3.8 se observa como la ventana de análisis se mueve a lo largo de $x(n)$ obteniendo finalmente X_w.

Se define la matriz X_{wLVA} como X_w según el tamaño de la LVA. A la hora de obtener el diagrama tiempo-frecuencia del sonido, se calcula la Transformada de Fourier de X_{wLVA} (particularizando con LVA= 512 y LVA= 128), calculando dos espectrograma según los dos LVAs utilizados. En las curvas rojas de la Fig. 3.8 se representa el movimiento de una ventana de análisis con LVA= 128 y en las curvas verdes el de una ventana de análisis con LVA= 512.

Cuando LVA= 512, los trozos x_{wl} se consideran estacionarios al abarcar varios periodos de la señal $x(n)$ (T_0), es decir, cuando LVA$> 2T_0$. Se obtiene entonces $TF[x_{wl}] \approx TF[x_{w(l+1)}]$.

Debido a las propiedades de la Transformada de Fourier, cuando se aplica a señales periódicas que contienen varios pulsos como son x_{wl} cuando LVA=512, se consiguen unas curvas como las de la Fig. 3.9b, donde la periodicidad T_0 de la señal se traslada a una



(a) (b)

(c) (d)

Figura 3.9: a) Transformada de Fourier de x_{wl} para LVA= 128 muestras; Curva azul cuando la ventana de análisis selecciona un impulso y la curva roja cuando no. b) Transformada de Fourier de $x(n)$ para LVA= 512 muestas. c) Espectrograma de $x(n)$ con LVA= 128 muestras. d) Espectrograma de $x(n)$ con LVA= 512 muestras.

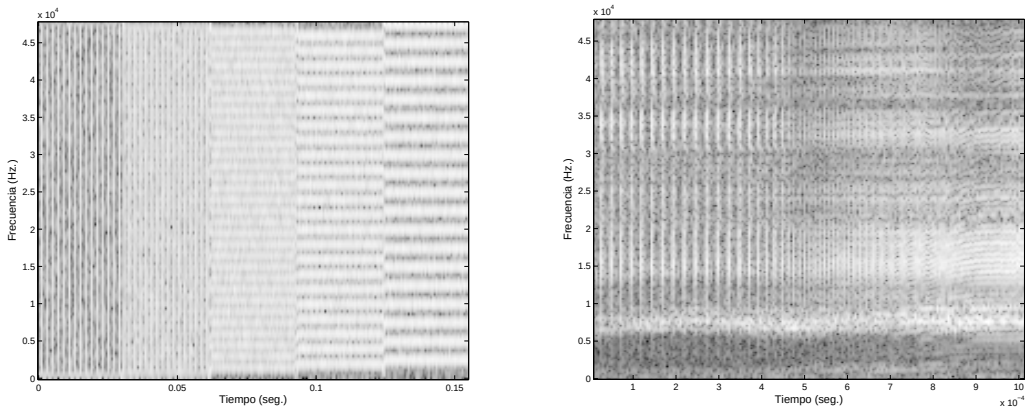
separación de picos de f_0. El espectrograma resultante de realizar $TF[X_{w512}]$ es como el de la Fig. 3.9d donde se pueden observar multitud de componentes horizontales.

En el caso de la ventana de análisis de LVA=128 representada por la curva roja en la Fig. 3.8, la matriz X_{w128} contiene trozos diferentes entre si (con o sin pulso), reflejados en su Transformada de Fourier de esta manera:

- Si el trozo seleccionado (x_{wl}) tiene contenido un impulso obtendremos una transforma de Fourier como la curva azul de la Fig. 3.9a.

- Si, por el contrario, el trozo seleccionado ($x_{w(l+1)}$) no contiene impulso, su transformada de Fourier será como la curva roja de la Fig. 3.9a. Se obtendrá un espectrograma como el de la Fig. 3.9c, donde se podrán ver las diferencias entre las distintas transformadas de Fourier, visualizadas en componentes horizontales que corresponderán a los trozos que tengan contenido un impulso.

Se observa como, dependiendo de la LVA, el espectrograma que se obtiene es totalmente diferente. Con LVA= 128 muestras el espectrograma tiene componentes verticales (Fig. 3.9c) y, sin embargo, cuando LVA= 512 muestras se obtiene un espectrograma con componentes horizontales como el de la Fig. 3.9d.



(a) (b)

Figura 3.10: a) Espectrograma de un sonido sintético. b) Ejemplo de un sonido real producido por una ballena beluga.

En la Fig. 3.10a se ha fijado LVA= 128 y se ha simulado una señal que aumenta su frecuencia fundamental o de repetición (f_0). Conforme aumenta f_0, disminuimos su periodo de repetición hasta que LVA$> 2T_0$ se obtiene una representación tiempo-frecuencia con componentes horizontales. Como ejemplo real se muestra la Fig. 3.10b donde vemos el espectrograma de una señal real producida por una de las ballena beluga del Oceanográfico de Valencia. Fijando LVA= 128, el aspecto del espectrograma cambia según aumenta su f_0 pasando de componentes verticales a tener componentes horizontales.

A la vista de lo aquí expuesto, queda patente la gran dependencia que existe entre la forma del espectrograma y la LVA. Por tanto, habrá que tenerla en cuenta a lo hora de interpretar de manera correcta un espectrograma. Tanto la obtención de la frecuencia fundamental (f_0) como la separación previa de las señales mixtas se realizará a través de transformar la señal al dominio cepstral, lo que será explicado en profundidad en el Capítulo 5 de esta tesis doctoral.

3.6.3. La existencia de sonidos resonantes

Existe una tendencia errónea por la cual a todos los sonidos tonales producidos por la mayoría de aves y mamíferos se les suele denominar silbidos. Sólo a los que son producidos sin necesidad de un elemento vibrante, como los silbidos humanos o los sonidos producidos por instrumentos musicales de viento como la flauta, se les debería llamar así [25]. De hecho, en el campo de los odontocetos se denomina silbido a cualquier señal

tonal, sin importar su naturaleza de producción y simplemente por tener un diagrama tiempo-frecuencia con componentes horizontales.

En esta tesis se considera importante no realizar la suposición de que las señales tonales se consideren silbidos. Un ejemplo de estos sonidos tonales se puede ver en la Fig. 3.11 donde se muestra un espectrograma donde es posible diferenciar claramente entre los dos sonidos. Se propone denominar a los silbidos como sonidos resonantes, la razón, explicada exhaustivamente en el Capítulo 4, será debida a la naturaleza de producción de estos sonidos.

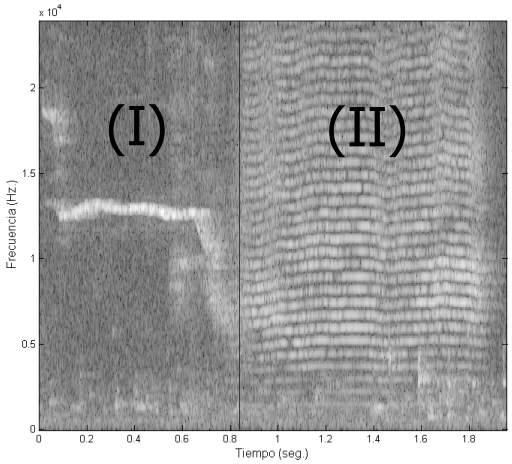


Figura 3.11: Silbido y sonido tonal de una ballena beluga.

3.6.4. Los sonidos mixtos

Finalmente y para terminar de presentar el repertorio habitual de sonidos de las ballenas beluga en la Fig. 3.12 se puede ver un espectrograma de una señal mixta. Este tipo de señales aparecen en numerosas ocasiones en las grabaciones realizadas. Se denominan sonidos mixtos debido a su morfología en el espectrograma, donde se puede observar tanto un comportamiento vertical como horizontal. De ahí su nomenclatura.

Por tanto, en una clasificación como la realizada es una tarea complicada llegar a identificar a qué categoría asignar este tipo de sonidos, ya que los valores de las características extraídas obtenidos son confusos, debido a la existencia de ambos comportamientos.

3.7. Conclusiones

A la vista de los resultados de este primer estudio, se propone, además de realizar una nueva clasificación sin tener en cuenta la morfología de su espectrograma, el diseño de un modelo de análisis/ síntesis que permita incluir y esclarecer las problemáticas de la inclusión de los sonidos mixtos y resonantes, donde todas las señales tonales no sean

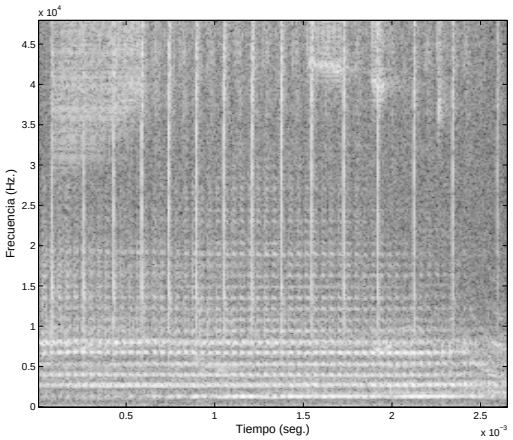


Figura 3.12: Espectrograma de una señal mixta producida por una ballena beluga.

tratadas de la misma manera, dada la capacidad de los odontocetos de emitir sonidos con una naturaleza de producción diferente.

Se dedicará un capítulo completo de esta tesis al estudio de algoritmos de medida del determinismo para diferenciar entre los dos tipos de naturaleza de producción del sonido trabajando con la señal temporal y no mediante las propiedades del espectrograma.

Además, el modelo que se planteará en el Capítulo 6, manteniendo una asociación clara con la fisionomía del aparato fonador de los odontocetos, deberá dar cabida a los sonidos mixtos, proponiendo una solución clara y elegante. Tal y como se describirá en el Capítulo 5, trabajar con el dominio cepstral nos permitirá conseguir dicho objetivo.

El contenido mostrado en este capítulo ha generado la publicación de los siguientes artículos en revista:

- Jorge Moragues Escrivá, Arturo Serrano Cartagena, Guillermo Lara Martínez, Jorge Gosalbez Castillo y Luis Vergara Dominguez. “Detection of acoustic events with application to environment monitoring”. Waves. Volumen 4. Páginas 25-33. Año 2012.

- Ramón Miralles Ricós, Guillermo Lara Martínez, Jose Antonio Esteban Simón y Alberto Rodriguez Martínez. “The Pulsed To Tonal Strength Parameter And Its Importance In Characterizing And Classifying Beluga Whale Sounds”. Journal of the Acoustical Society of America. Volumen 131. Número 3. Páginas 141-147. Año 2012.

- Ramón Miralles Ricós, Guillermo Lara Martínez, Alicia Carrión Garcia, Jose Antonio Esteban Simón. “Automatic Detection and Classification of Beluga Whale Vocalizations. Advances in Applied Acoustics”. Volumen 2. Número 2. Páginas 61-70. Año 2013.

y la participación en los siguientes congresos:

- Guillermo Lara Martínez y Ramón Miralles Ricós. “Naive Bayes Classifier for Automatic analysis of Beluga Whale songs”. 4th International Workshop on Marine Technology (MARTECH). Año 2011.
- Ramón Miralles Ricós y Guillermo Lara Martínez. “Detección automática de sonidos de rorcual en presencia de explosiones de cañones de aire sísmicos”. TECNIACUSTICA. 43º Congreso Español de Acústica. 7º Congreso Ibérico de Acústica. 8º Congreso Iberoamericano de Acústica. Año 2012.
- Ramón Miralles Ricós y Guillermo Lara Martínez. “An automatic system for detection and classification of beluga whale sounds”. 42th Annual Symposium of the European Association for Aquatic Mammals. Tenenife, España. Año 2014.

Capítulo 4

La naturaleza de la producción de los sonidos en los odontocetos

4.1. Introducción

En este capítulo se estudiarán y caracterizarán las diferentes maneras de producir sonidos. Se analizará la teoría de tubos y los principios de las vibraciones en distintos instrumentos musicales de viento y se extraerán conclusiones para intentar comprender el sistema de producción de sonidos en mamíferos marinos e identificar los diferentes sonidos que son capaces de realizar teniendo en cuenta su naturaleza.

En primer lugar se estudiarán diversos instrumentos musicales, observando de una manera detallada cómo y por qué se producen sonidos dentro de ellos. El siguiente paso será estudiar los sonidos producidos por los seres humanos, dado el control y conocimiento que se tiene en la actualidad. A través de una serie de señales realizadas por varios sujetos en un entorno controlado, se analizarán, no sólo sonidos vibratorios, sino también sonidos ruidosos y silbidos. Se obtendrán las conclusiones oportunas asociando los sonidos producidos por los seres humanos con los producidos por los instrumentos musicales.

La última parte del capítulo mostrará el estado actual de los mecanismos de producción de sonidos que poseen los odontocetos, en concreto las ballenas beluga, para intentar establecer similitudes con los sonidos estudiados anteriormente. El conocimiento y análisis de los sonidos producidos por los cetáceos podría suponer un método novedoso a la hora de conocer y controlar los comportamientos de los cetáceos de una manera no invasiva. Dada la creatividad mostrada por estos animales en algunas estudios [26, 27], no parece atrevido pensar que un mejor conocimiento su sistema de generación de sonidos pueda llevarnos a comprender, en un futuro a medio-largo plazo, el nivel de inteligencia que poseen estos animales.

4.2. Teoría de tubos y creación de vibraciones

Para poder entender la naturaleza de producción en instrumentos musicales y en los mamíferos, es necesario explicar una serie de propiedades sobre las diferentes maneras con las que se puede producir sonido.

4.2.1. Teoría de tubos

Se denominan tubos sonoros a cavidades que contienen una columna gaseosa (columna de aire) capaz de producir ondas acústicas al ser convenientemente excitada. La onda acústica se produce en la columna gaseosa, y no en el tubo que la contiene; las dimensiones del tubo provocarán la aparición mediante resonancia de ondas acústicas a frecuencias asociadas a estas dimensiones. Los tubos sonoros pueden ser cerrados, es decir, que poseen una sola abertura y tubos abiertos, que poseen dos o más (ver Fig. 4.1).

Una columnas de aire contenida en un tubo sonoro se comporta, desde ciertos puntos de vista, como cuerdas musicales, por lo tanto una columna de aire vibrante poseen nodos, o sea puntos donde la vibración es nula, y vientres, equidistantes de los anteriores, donde la vibración alcanza su máxima amplitud. La vibración de una columnas de aire es longitudinal; los nodos serán por tanto, puntos de condensación y los vientres, puntos de dilatación; en los extremos cerrados siempre se producen nodos (máximo de presión) y en los extremos abiertos se producen vientres (máximo de vibración o movimiento).

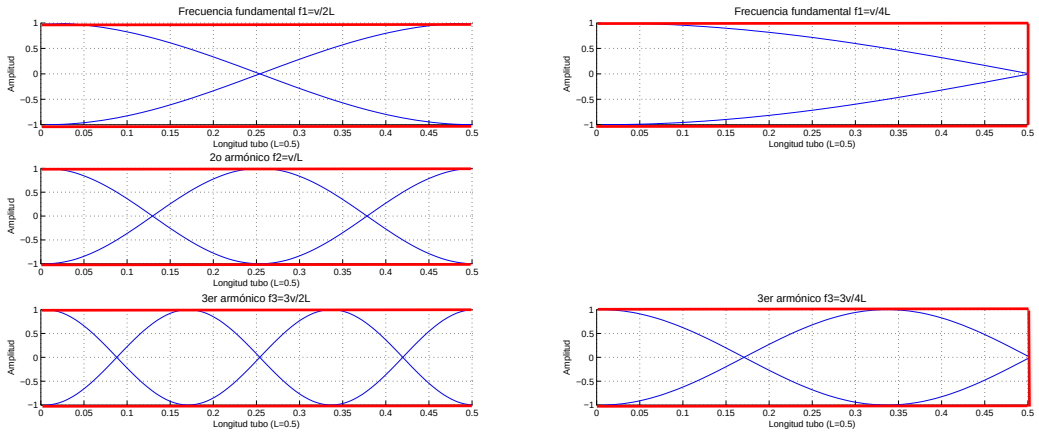
A destacar que no es necesario que las aberturas de un tubo coincidan con los extremos, pudiendo éstos estar cerrados y haber una o más aberturas en otras partes del tubo. Una columna de aire puede vibrar con toda su longitud o dividida en segmentos iguales, en el primer caso se obtiene un sonido a la frecuencia fundamental, y en los otros los armónicos: segundo, si la columna vibra dividida en mitades; tercero, si vibra en tercios, etc.

Tomando como punto de partida que en los extremos de un tubo abierto, sólo pueden haber vientres de vibración y el tubo produce un sonido a la frecuencia fundamental cuando vibre con un nodo único en su centro. Siguiendo con el razonamiento, cuando el tubo produce su segundo armónico, posee dos nodos y tres vientres y así sucesivamente.

En los tubos cerrados, la onda se forma con un nodo en el extremo cerrado y un vientre (ver Fig. 4.1a) en el extremo abierto. A igualdad de longitud de tubo, el tubo abierto produce un sonido de frecuencia doble que el cerrado. Los tubos abiertos emiten la serie completa de armónicos correspondientes a su longitud, mientras que los cerrados, emiten sólo los armónicos de orden impar (Fig. 4.1b).

Los principios de Bernoulli [28] aplicables tanto a los tubos abiertos como a los tubos cerrados, las cuales tendremos en cuenta a la hora de analizar los sonidos son los siguientes:

- La frecuencia del sonido producido por un tubo, tanto abierto como cerrado, es directamente proporcional a la velocidad de propagación. Un ejemplo claro de esto se da cuando una persona inspira helio en lugar de aire y su voz se vuelve más



(a) Tubo abierto (b) Tubo cerrado

Figura 4.1: Detalle los nodos y vientres de los modos que se propagan dentro de los tubos sonoros. En la parte izquierda corresponden a un tubo abierto y en la parte derecha a un tubo cerrado. Las curvas corresponden a la amplitud de la onda estacionaria que se propaga en el interior del tubo sonoro.

aguda mientras quede Helio en sus pulmones, dado que la velocidad de propagación del helio es igual a 965 m/s, mucho más alta que los 330 m/s del aire.

- La frecuencia del sonido producido por un tubo, tanto abierto como cerrado, es inversamente proporcional a la longitud del tubo. A mayor longitud del tubo, más grave es el sonido, es decir, posee una frecuencia fundamental menor.
- A igualdad de longitud entre un tubo abierto y otro cerrado, el abierto produce un sonido de frecuencia doble que el cerrado, es decir, el abierto produce un sonido a la octava del cerrado.
- Los tubos abiertos producen la serie completa de armónicos, mientras que los cerrados sólo los armónicos de frecuencia impar de la fundamental.

4.2.2. Creación de vibraciones

A continuación se explica brevemente la creación de las vibraciones. Se denomina vibración a un movimiento repetitivo alrededor de una posición de equilibrio. Este movimiento es debido a una energía cinética que provoque la deformación de una lengüeta o el choque de una membrana con otro material hasta volver a su posición de equilibrio. Esta deformación suele producirse en todas direcciones, lo que provoca una señal impulsiva con un ancho de banda muy grande. Si este movimiento tiene una periodicidad, se produce una vibración y la frecuencia a la cual se repite el movimiento será la frecuencia fundamental f_0, denominada frecuencia de vibración. Un ejemplo claro de esta

Tipo excitación	Tipo de tubo	Instrumento
Sin elemento vibrante	Tubos de Embocadura	Flauta Travesera
		Flauta de pico y Tubos órgano
Con elemento vibrante	Tubos de Lengüeta	Acordeón y Armónica
		Clarinete, Saxofón y Tubos órgano
		Oboe y Fagot
	Tubos de Lengüeta Labial	Trompetas, Trombones y Tuba

Tabla 4.1: Clasificación de los tubos sonoros según el modo de excitación de la columna de aire.

explicación son los latidos del corazón, los cuales están formados por el movimiento de las paredes del órgano, las cuáles se mueven en todas direcciones y con una f_0 de algo más de un segundo.

Una vez el material está vibrando, el gas o liquido que está en contacto con él también vibra de la misma forma, y dependiendo de si está o no contenido en el tubo sonoro, provoca que los modos excitados con la vibración se propaguen o no por ella, coloreando la señal. Si la frecuencia de la vibración es una frecuencia entre 100 Hz y 10 KHz se denomina onda acústica [29], ya que será audible.

4.3. La naturaleza de los sonidos producidos en instrumentos musicales de viento y seres humanos

En esta sección se mostrará como es posible aplicar las teorías vistas anteriormente para el estudio de los diferentes sonidos de los instrumentos musicales y de los seres humanos. El objetivo final será clasificarlos y asociarlos a una naturaleza de producción concreta.

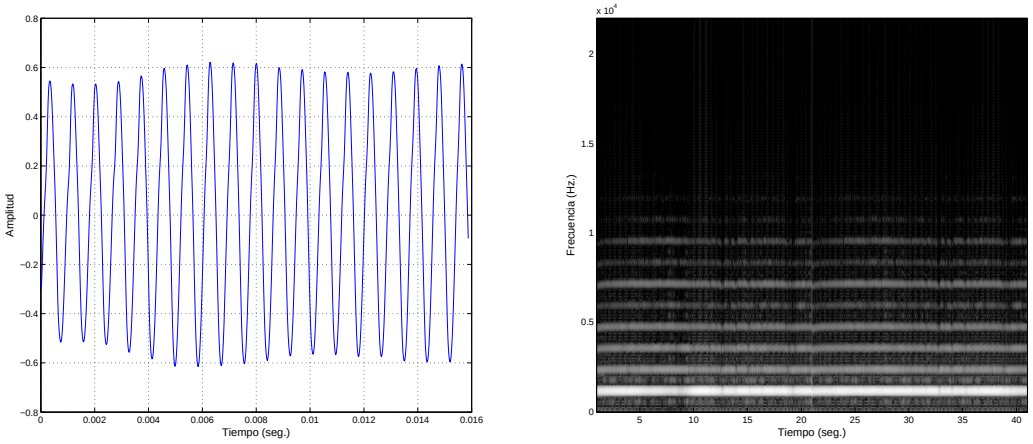
4.3.1. Instrumentos musicales de viento

Los instrumentos musicales de viento han sido utilizados desde hace varios siglos y han ido evolucionando en su fabricación para conseguir unos sonidos muy característicos, con una facilidad de uso y aprendizaje elevada. La Tabla 4.1 resume la clasificación de los instrumentos musicales basada en el modo de la excitación del tubo sonoro. Los dos grupos principales son los instrumentos que realizan sonidos mediante la resonancia de los tubos sonoros (sin elemento vibrante) y los que los realizan mediante una vibración física que excita de los tubos sonoros (con elemento vibrante).

I. **Sin elemento vibrante.** Este tipo de sonidos se caracterizan por permitir la propagación de pocos armónicos o modos en su interior, ya que la onda acústica se produce por la resonancia del tubo sonoro sin la necesidad de que ningún elemento vibrante la provoque. Dentro de esta familia se incluyen los tubos de embocadura

que son tubos sonoros que poseen una abertura convenientemente dispuesta llamada embocadura, uno de cuyos bordes es biselado. Contra este borde incide una corriente de aire que se divide en dos ramas; la rama que penetra en el tubo origina pequeñas vibraciones que a su vez excitan por resonancia la columna aérea contenida en el tubo.

En la Fig. 4.2a se puede ver un ejemplo de una señal en tiempo producida por una flauta. Es apreciable la señal tiene el comportamiento habitual de una resonancia en los primeros modos, que son los que sufren una menor atenuación dentro del tubo sonoro. En la Fig. 4.2b se muestra la representación tiempo-frecuencia donde se puede ver como el sonido que se ha propagado dentro del tubo sonoro de la flauta tiene pocos armónicos debido que el sonido no ha sido provocado por ninguna excitación. Se observa además como el nivel del primer armónico es mucho mayor a todos los demás, por ello la forma temporal de la señal es casi sinusoidal.



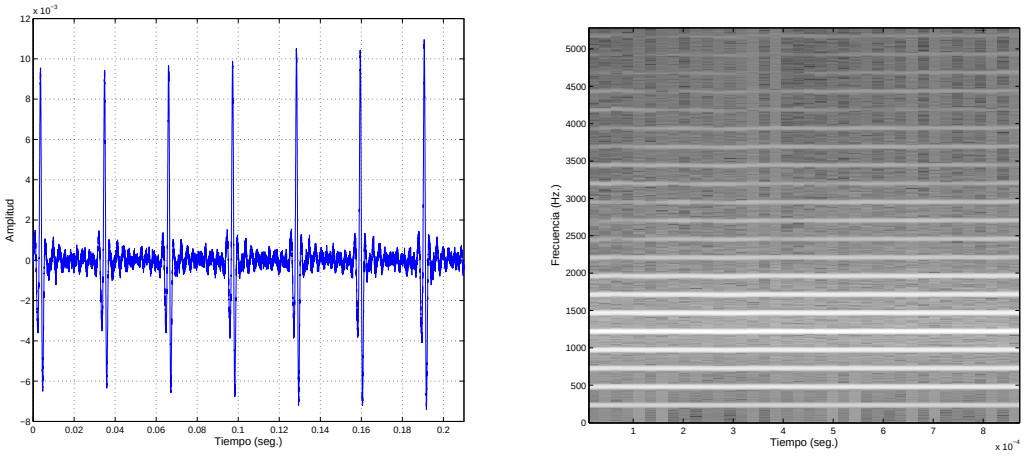
(a) (b)

Figura 4.2: Sonido de una flauta. a) Señal temporal. b) Espectrograma de la señal.

II. **Con elemento vibrante.** Existen dos familias de instrumentos que tienen esta naturaleza. En primer lugar se tratarán los instrumentos de tubos de lengüeta, los cuales están formados por pequeñas láminas elásticas, generalmente de metal o de madera (caña) que, sujetas a un soporte de manera conveniente, vibran al paso de una corriente aérea excitando el tubo sonoro y produciendo sonido. La otra gran familia dentro de los instrumentos excitados mediante vibración física es la de los tubos de lengüeta labial.

Un ejemplo de la naturaleza vibrante de un sonido se encuentra en la Fig. 4.3a. se puede apreciar como la señal temporal tiene un aspecto donde un impulso periódico. Esto es debido al contacto de los labios con la apertura de la trompeta, lo que provoca una vibración de la boquilla, transfiriendo esta vibración al aire contiguo.

En la Fig. 4.3b se puede ver como el diagrama tiempo-frecuencia o espectrograma de la señal está formado por muchos más armónicos que en el caso de la flauta y que además, el primer armónico no tiene por qué ser el más importante, ya que todos los modos excitados pasarán a través del tubo sonoro y se colorearán cambiando su intensidad según su frecuencia.



(a) (b)

Figura 4.3: Sonido de una trompeta. a) Señal temporal. b) Espectrograma de la señal.

4.3.2. La producción de sonido en los seres humanos

En el caso de los seres humanos y de forma resumida, se describen a continuación los dos sonidos más frecuentes, relacionándose con los instrumentos de viento descritos anteriormente: la voz humana y los silbidos, sonidos realizados en diferentes puntos del sistema de producción de sonido que posee el ser humano.

Uno de los órganos más importantes en dicho sistema son las cuerdas vocales, las cuales se organizan en cuatro pliegues vocales, dos superiores, que no participan en la articulación de la voz y dos inferiores, las verdaderas cuerdas vocales, responsables de la producción de la voz. Si se abren y se recogen a los lados, el aire pasa libremente sin hacer presión y respiramos. Si por el contrario se juntan, el aire pasa entre ellas, iniciando un movimiento muy rápido de vibración con el cual se produce el sonido que denominamos voz. Al cerrarse más las cuerdas vocales comienzan a vibrar a modo de lengüetas, produciéndose un sonido cuya frecuencia de repetición o frecuencia fundamental varía en forma inversa al tamaño de las cuerdas y la tensión en estas. De forma resumida se observa en la Fig. 4.4 los órganos más importantes que intervienen tanto en la producción de la voz humana como en la producción de silbidos.

En la Fig. 4.5a se muestra el sonido de la letra “a” pronunciado por una persona. Se trata de una señal periódica, donde se aprecia la vibración de las cuerdas vocales

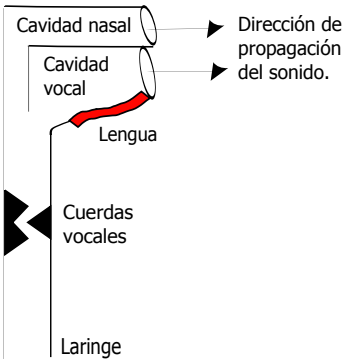
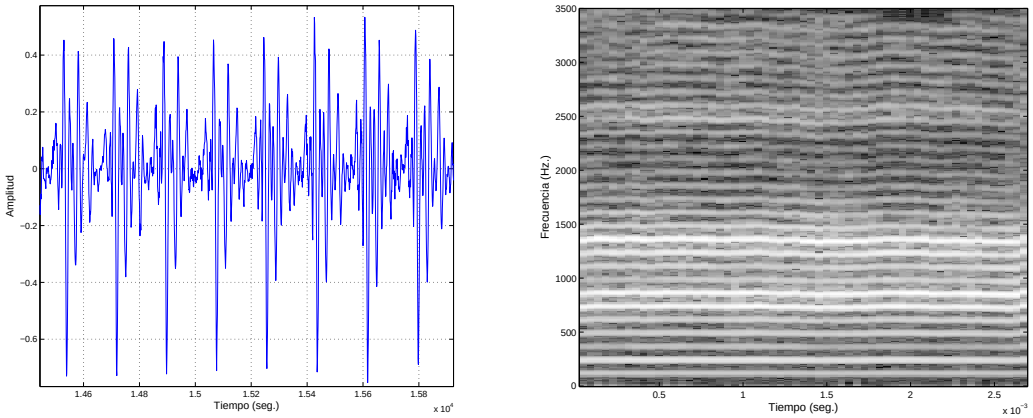


Figura 4.4: Morfología y órganos involucrados en la producción de la voz humana.



(a) (b)

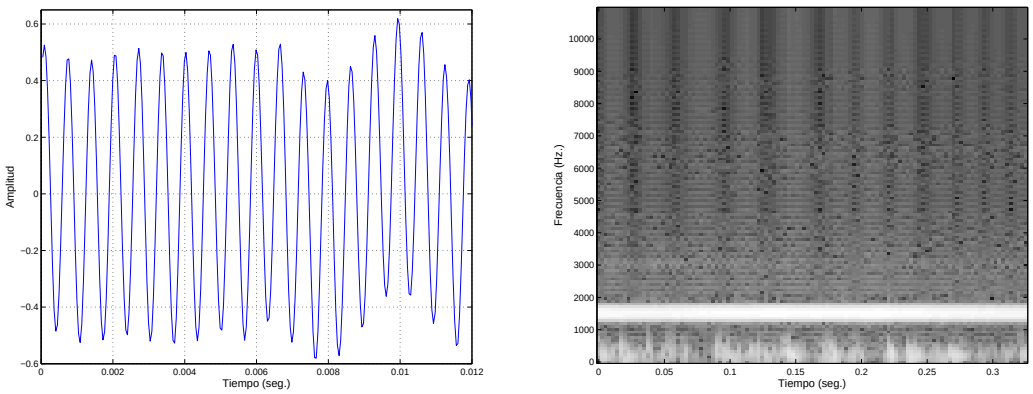
Figura 4.5: Sonido de una persona pronunciado la letra “a”. a) Señal temporal. b) Espectrograma de la señal.

en todas las direcciones, lo que provoca, tal y como se puede ver en la Fig. 4.5b un espectrograma con un ancho de banda muy grande y por ello con numerosos armónicos.

Este ejemplo en particular, extensible a la producción de las vocales, tiene semejanzas con los instrumentos con un elemento vibrante, como los de lengüeta labial o membranosa donde la vibración de los labios a la boquilla. En los seres humanos, las membranas vibrantes son las cuerdas vocales. Estas traspasan la vibración al aire contenido en las cavidades en que actúan como tubos sonoros. Dichas cavidades actúan como resonadores acústicos.

Si se realiza un análisis espectral del sonido, una vez atravesado estas cavidades (mostradas en la Fig. 4.4), el efecto de la resonancia produce un énfasis en determinadas frecuencias del espectro obtenido, a las que se les denominará formantes. Existen tres formantes fundamentales debidos respectivamente a las cavidades oral, bucal y nasal.

Los silbidos se generan de forma totalmente diferente, ya que son producidos en la parte final del sistema de producción de sonido, más concretamente en la parte final de la cavidad vocal. Un silbido es un sonido resultante de hacer pasar un soplido a través de los labios ayudándose, o no, con los dedos. La frecuencia fundamental de este sonido varía con la posición de los labios, la lengua o los dientes. La cavidad bocal actúa a modo de caja de resonancia. Estos sonidos por tanto son producidos por la resonancia que se produce entre los labios, sin necesidad de ningún elemento vibrante.



(a) (b)

Figura 4.6: Sonido de un silbido de una persona. a) Señal temporal. b) Espectrograma de la señal.

En la Fig. 4.6a se puede ver la señal temporal de un silbido realizado por un ser humano. Se puede apreciar como es mucho más suave y sin los impulsos provocados por la vibración. Además, si nos fijamos en su espectrograma (Fig. 4.6b), se puede observar la similitud con otros tipos de sonido creados sin elemento vibrante, por ejemplo con sonidos creados por instrumentos musicales como la flauta y el clarinete (ver Fig. 4.2b).

4.4. La producción de sonidos en los odontocetos

Ya en el campo de los cetáceos, concretamente en los odontocetos, se mostrará el estado del arte en la temática, explicando los órganos relacionados con el sonido que producen estos mamíferos marinos y los diferentes experimentos e hipótesis que se han llevado a cabo hasta el momento.

Varios autores comenzaron a estudiar la fisiología y órganos que utilizan para la producción del sonido [30–33], mostrando cómo evolucionaron a partir de los mamíferos terrestres teniendo que adaptarse a otro medio de transmisión, el agua.

Estos animales emplean un amplio repertorio de sonidos para la ecolocalización y la comunicación, sonidos que no se producen en la laringe, como en los mamíferos terrestres, sino más bien en un complejo sistema de sacos de aire nasal, tejido conectivo

y compartimentos de grasa [34, 35]. La anatomía de estos órganos y la amplia gama de sonidos con propiedades diferentes pero controladas [17, 36, 37] han llamado la atención científica al problema de cómo producen sonidos los odontocetos.

Los primeros trabajos de Norris y sus colaboradores [38] mostraron que los sonidos se generaban a través de una fuente por encima del rostro de los delfines, hallazgo que fue corroborado por Diercks [39], el cual utilizó un conjunto de hidrófonos ventosa para localizar acústicamente la fuente de sonido en una ubicación dentro del orificio nasal. Usando técnicas cinoradiográficas [40], electrodos de electromiografía y catéteres de presión [30, 32], se demostró que la producción de sonido en los odontocetos es realizada neumáticamente por una acumulación de presión en los sacos de aire vestibulares desplazando dorsalmente el aire más allá de los dos orificios nasales.

Era por tanto evidente que la fuente de sonido tenía que encontrarse por encima de los orificios nasales, y se propusieron varios candidatos: la membrana diagonal [41], las fosas nasales [42], los sacos de aire [43], el espiráculo o los labios fónicos *monkey lips* [44]. Un problema común, sin embargo, era que no podía comprobarse con mediciones fisiológicas, problema que se resolvió con la publicación de un artículo de Cranford [35] que supone la identificación de una estructura anatómica homóloga en una amplia gama de especies de odontocetos. Esta estructura fue denominada por primera vez como *Monkey-Lips-Dorsal-Bursa* (MLDB), formada por dos pares de labios fónicos y sus correspondientes membranas *dorsal bursa*, sistema que se encuentra duplicada en todas las especies de odontocetos, excepto en los cachalotes (*Physeter macrocephalus*). En [31, 33] se explica como los sonidos se generan por las vibraciones de una membrana o tejido, cambiando la presión entre los sacos de aire colocados a los extremos de las cavidades nasales para mover el aire de un lado a otro del sistema, haciendo vibrar el sistema MLDB situado dentro de él.

Todos los órganos nombrados se pueden visualizar en la Fig. 4.7. Estudiándolos en detalle se puede entender como se propaga el sonido hasta que sale del odontoceto hacia el exterior, sea en un medio acuático o aéreo. Los sacos de aire permiten el cambio de presión del aire haciéndolo circular por los dos tubos pasando a través del sistema MLDB de cada uno de ellos. Este sistema está conectado con el melón, órgano gelatinoso situado en la parte superior de la boca del odontoceto. Otro elemento importante para entender la forma en que los sonidos se propagan hacia el exterior es el espiráculo, por donde expulsan el aire al respirar. Este órgano tendrá una importancia relativamente grande cuando los odontocetos realicen sonidos aéreos.

Más en detalle, Madsen demostró el carácter vibratorio de los sonidos producidos a través del sistema MLDB en [45]. Introduciendo helio en la cavidad aérea resonante, comprobó que ciertos sonidos no cambian su frecuencia fundamental y por tanto eran debidos a vibraciones de los labios fónicos. Dichas vibraciones serían propagadas por el melón, al estar en contacto con él. Es decir, las vibraciones no se desplazan por ninguna cavidad resonante, en tal caso la sustitución de helio en lugar de aire, hubiera provocado el aumento de la frecuencia fundamental, haciendo el sonido más agudo, dada que la densidad del helio ($0{,}166 Kg/m^3$) es mucho menor a la del aire ($1{,}205 Kg/m^3$). Este valor de las densidades corresponde las métricas NTP (métricas medidas en condiciones

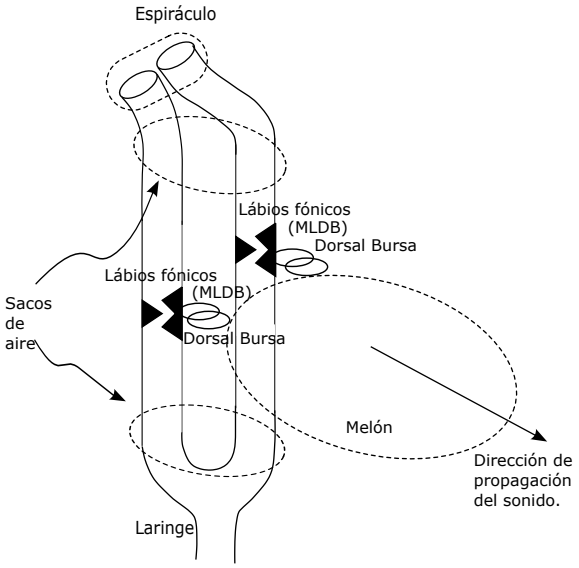


Figura 4.7: Detalle de los órganos involucrados en la generación del sonido en los odontocetos.

de temperatura de 25 grados centígrados y presión de 101.325 kPa).

4.4.1. Efecto de la resonancia en cavidades en la producción de sonidos vibratorios: Experimento del globo con helio

Al igual que realizó Madsen [45] en 2012 y con objeto de ver cuales son los órganos involucrados en las ballenas beluga se realizó el siguiente experimento. Se grabaron sonidos producidos por un globo al deshincharse cuando se mantenía la tensión en la boquilla del globo. Dichos sonidos se consideran de naturaleza vibratoria, ya que la membrana creada en la boquilla del globo vibra en el momento que se expulsa el aire contenido al ejercer presión sobre ésta.

Varias propiedades son parecidas al sistema de producción de sonido los odontocetos. En primer lugar, la membrana colocada en el extremo vibra a una frecuencia de repetición parecida a la de los labios fónicos. En segundo lugar, esta vibración no se propaga dentro de ninguna cavidad, lo que impide se provoquen formantes que coloreen de los armónicos de la señal. Es posible visualizar un ejemplo de la señal temporal que produce esta vibración en la Fig. 4.8a y de la representación de esta señal en el diagrama tiempo-frecuencia en la Fig. 4.8b.

Al igual que realiza Madsen en el aparato fonador de las belugas, se completa el experimento introduciendo helio en lugar de aire en el interior del globo. En las Fig. 4.8b y 4.8d se puede observar como la frecuencia fundamental provocada por la vibración de la membrana de la boca del globo es la misma en los sonidos producidos deshincharse el globo, sea con aire o con con helio dentro de él. Se utilizaron algoritmos de detección del pitch convencionales para la comprobación y medición del pitch en los dos sonidos,

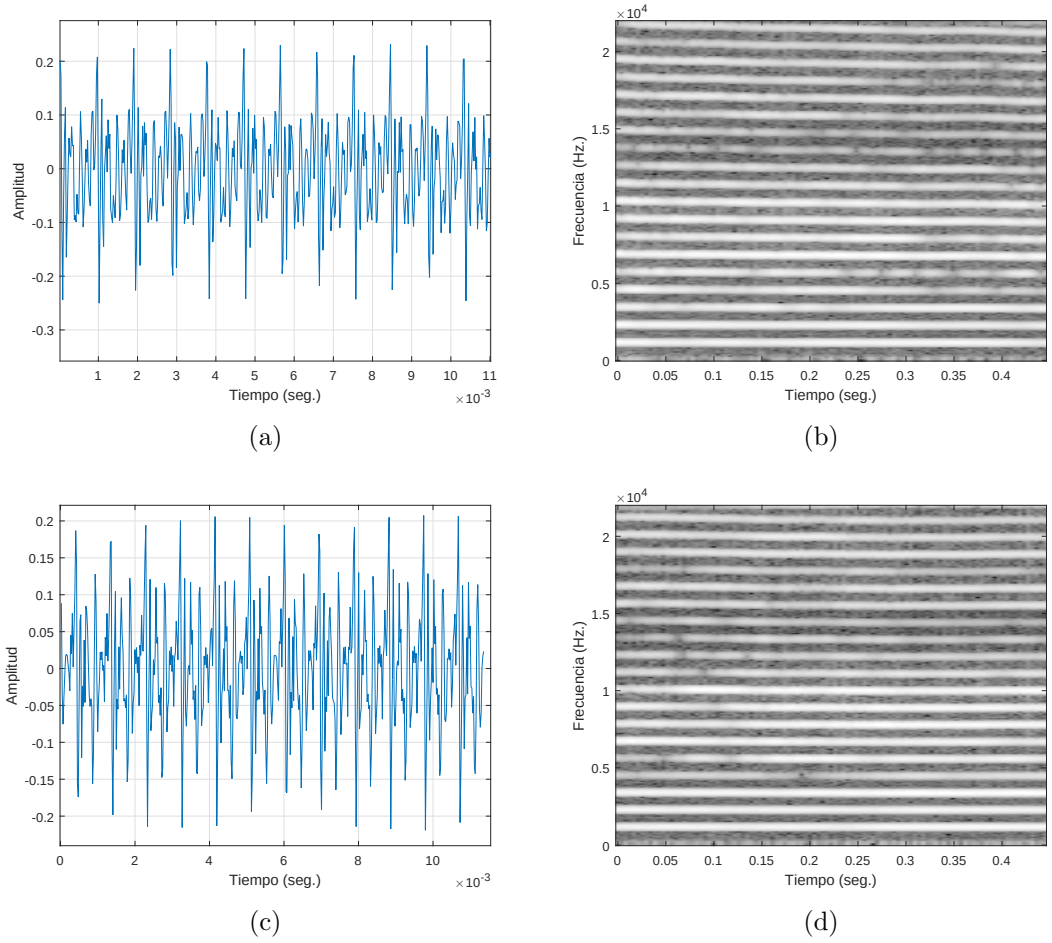


Figura 4.8: Ejemplo de la vibración de una membrana o lengüeta. Globo desinflándose. a) Señal temporal utilizando aire. b) Espectrograma de la señal utilizando aire. c) Señal temporal utilizando helio. d) Espectrograma de la señal utilizando helio.

estos algoritmos se explicarán en detalle en el Capítulo 6. Se representa un dibujo en la Fig. 4.9 donde se puede ver las dos realizaciones.

Esto es debido a que todos los gases tienen el mismo número de partículas por unidad de volumen, para una misma temperatura y presión, independientemente del gas del que estemos hablando. Pero no todas las moléculas de gas tienen la misma masa, por lo que la densidad del gas cambia. El nitrógeno (principal componente del aire que respiramos) tiene una masa unas siete veces superior a la del helio, por que la densidad del helio es mucho menor y las ondas sonoras pueden viajar más rápido. A 20°C, el sonido viaja a 927 m/s a través del helio y a 344 m/s a través del aire.

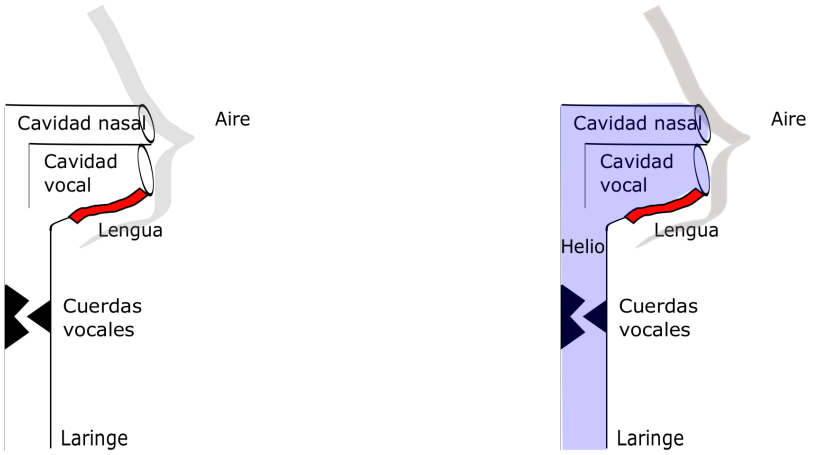
$$f = \frac{v}{\lambda} \tag{4.1}$$

(a) (b)

Figura 4.9: Dibujo de un globo. a) Con aire en su interior. b) Con helio (en azul) en su interior.

Cuando la onda viaja más rápido, aumenta la frecuencia, sin necesidad de variar la longitud de onda (λ) intrinseca del elemento de vibración. Es necesario remarcar este punto porque las membranas o sistemas vibrantes siempre lo hacen a la misma λ, es decir, emiten el mismo número de vibraciones por segundo, tanto si respiramos aire como si respiramos helio, ya que este último rasgo viene determinado por la longitud y grosor de las membranas.

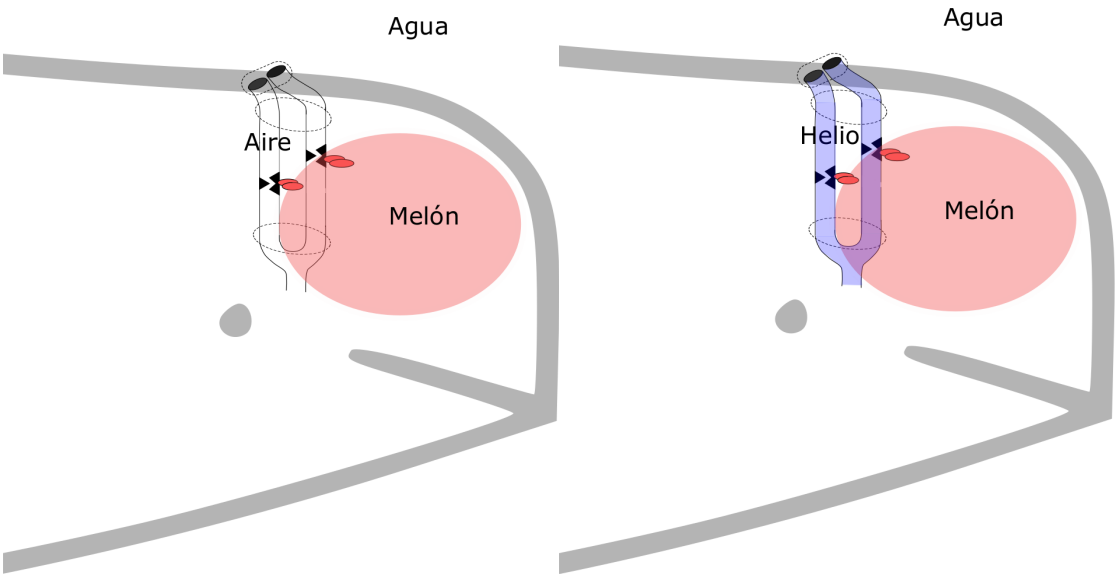
Una vez vibran las membranas del globo, las ondas mecánicas provocadas salen al exterior (aire) sin aumentar la frecuencia fundamental, es decir, la frecuencia fundamental del sonido captado es independiente del elemento que rellene el globo.



(a) (b)

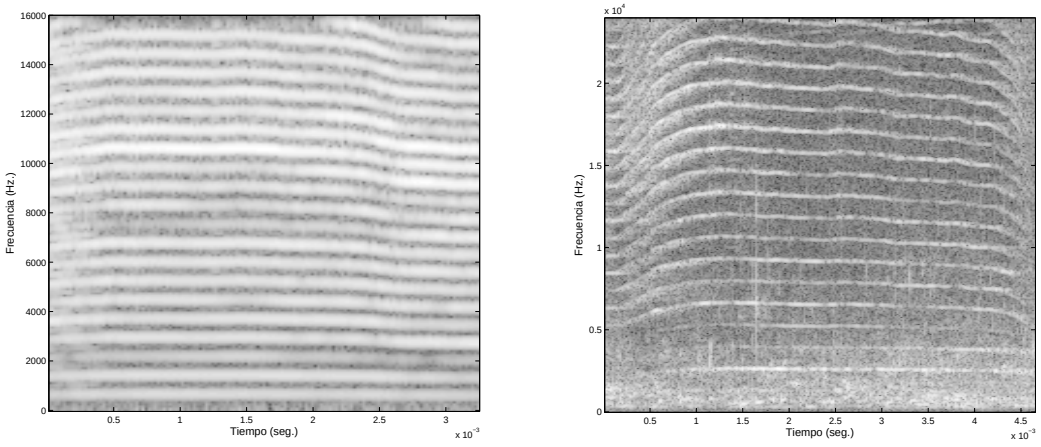
Figura 4.10: Dibujo del sistema producción de sonido de los seres humanos. a) Con aire en su laringe. b) Con helio (en azul) en su laringe.

Sin embargo, en los seres humanos, aunque las cuerdas vocales vibren a la misma longitud de onda independientemente del helio, al ser un gas formado por moléculas de un solo átomo, las ondas sonoras viajan mucho más rápido, aumentando la frecuencia cuando resuena en una cavidad como puede ser la laringe.



(a) (b)

Figura 4.11: Dibujo del aparato fonador de las ballenas belugas. a) Con aire en su laringe. b) Con helio (en azul) en su laringe.



(a) (b)

Figura 4.12: a) Espectrograma de un sonido producido por un globo. b) Espectrograma de un sonido producido por una ballena beluga.

En el caso de las ballenas beluga (ver Fig. 4.11) si se compara el espectrograma de un sonido generado por un globo al deshincharse (ver Fig. 4.12a) con el obtenido por el de un sonido vibratorio producido por una ballena beluga (ver Fig. 4.12b), la similitud es evidente, dando a entender una similitud en el mecanismo de producción del sonido,

en este caso mediante la vibración de membranas . Esta similitud, unida al experimento del helio realizado por Madsen [45], concluye que:

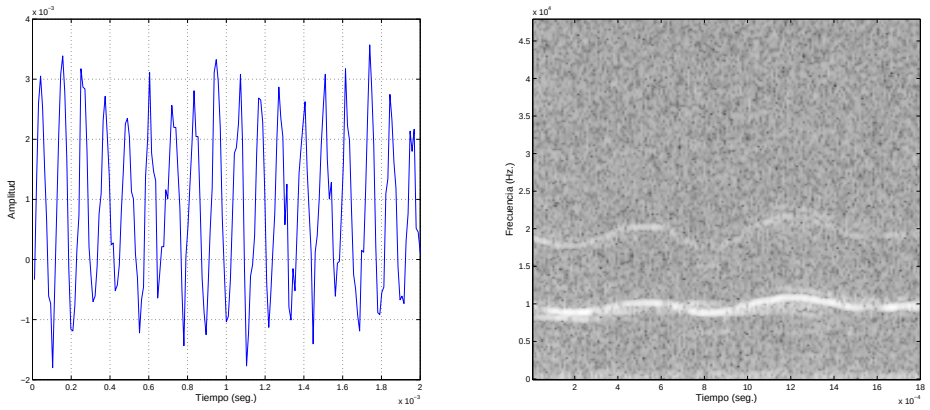
I. Los tubos donde se encuentran los labios fónicos (sistema MLDB) no intervienen en la transmisión de la vibración dado que no aparecen formantes que colorearían el sonido producido. La vibración por tanto se transmite al melón por medio de las *dorsal bursa* que lo conectan a los labios fónicos. Algo parecido pero en los sonidos pulsados o clicks, se demostró anteriormente en [46].

II. La creación de los sonidos se produce por vibración de los labios fónicos y las membranas *dorsal bursa* y no por la resonancia en los tubos que las contienen.

4.4.2. La producción de los sonidos resonantes

En algunos animales terrestres se producen de dos maneras diferentes; silbando [47], donde un flujo de aire crea fluctuaciones en la presión dentro de las cavidades vocales, provocando resonancias dentro de él, o por la vibración llevada a cabo por corrientes de aire en las cuerdas vocales donde su masa y su tensión determinan su frecuencia de repetición o frecuencia fundamental [48, 49]. El aumento o disminución de la frecuencia se produce o bien cambiando la frecuencia de resonancia de los espacios de aire en caso de los sonidos resonantes, o por el contrario, cambiando la presión del aire o la tensión de las cuerdas vocales [50].

Estas dos maneras de producir sonidos fueron caracterizadas por varios investigadores referentes en este ámbito. En [25, 43] explicó como son generados los sonidos al resonar volúmenes de aire en las cavidades nasales, sonidos producidos sin necesidad de un elemento vibrante. Un ejemplo de estos sonidos resonantes se puede ver en la Fig. 4.13.



(a) (b)

Figura 4.13: Sonido resonante de una ballena beluga. a) Señal temporal. b) Diagrama tiempo-frecuencia.

Como ejemplo de la diferente naturaleza de producción de las señales, en la Fig. 4.14c se muestra un espectrograma donde se puede diferenciar claramente entre los dos tipos de naturaleza. La primera parte de la señal (Fig. 4.14a) es debida a una resonancia y no existe excitación mediante vibración. Se puede ver que es similar a los silbidos producidos por el ser humano. La segunda parte de la señal (Fig. 4.14b), producida mediante la vibración de los labios fónicos, es similar a la señal de voz humana.

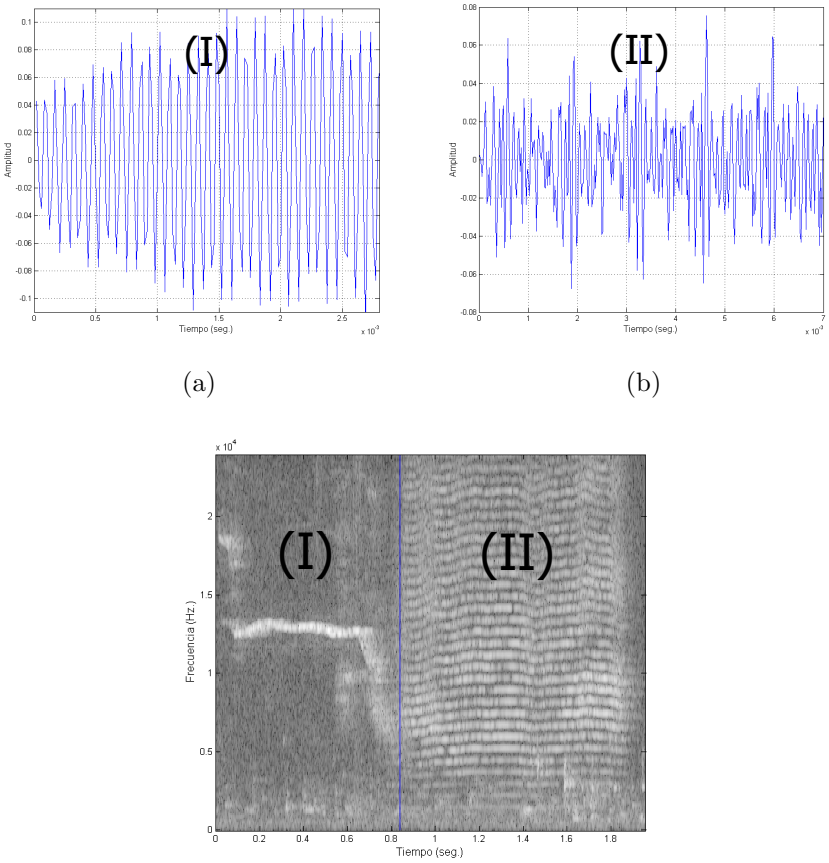


Figura 4.14: Sonido concatenado resonante-vibratorio de una ballena beluga. a) Detalle de la señal temporal de la parte resonante de la señal. b) Detalle de la señal temporal de la parte vibratoria de la señal. c) Espectrograma de la señal.

A la vista de esto parece que los odontocetos son capaces de realizar varios tipos de sonidos tonales dentro de su aparato fonador. Se definen por tanto dos tipos de sonidos tonales: vibratorios y resonantes. Se propone, a la vista de los resultados presentados y de acuerdo con los últimos estudios y medidas empíricas realizadas, que los odontocetos son capaces de realizar tanto sonidos vibratorios, producidos por los labios fónicos, como sonidos debidos a resonancias, producidos sin la necesidad de que los labios fónicos

vibren, es decir, mediante la resonancia de los tubos por donde circula en aire, provocando señales con pocos armónicos y una importancia clara del primero de ellos.

4.4.3. La combinación de los dos pares de labios fónicos

Conforme a la manera de combinar los dos pares de labios fónicos, Lammers propuso en [51] la forma en la cual los cetáceos dirigían sus sonidos dependiendo del retardo entre los impulsos producidos por ellos, consiguiendo así direccionar el haz y ecolocalizar en una dirección o en otra. Esto supondría que para generar sonidos pulsados, los dos labios fónicos deberían estar funcionando simultáneamente (ver Fig. 4.15), lo que imposibilitaría la creación de señales mixtas (sonidos compuestos por componentes frecuenciales y pulsados simultáneamente).

Sin embargo, posteriormente, Madsen [52] demostró de una manera práctica que los cetáceos consiguen direccionar su ecolocalización mediante el melón, indicando que la producción de estos sonidos depende únicamente de unos labios fónicos, lo que permitiría dedicar el otro par de labios fónicos para realizar sonidos tonales y posibilitaría la creación de las señales mixtas (ver Fig. 4.16).

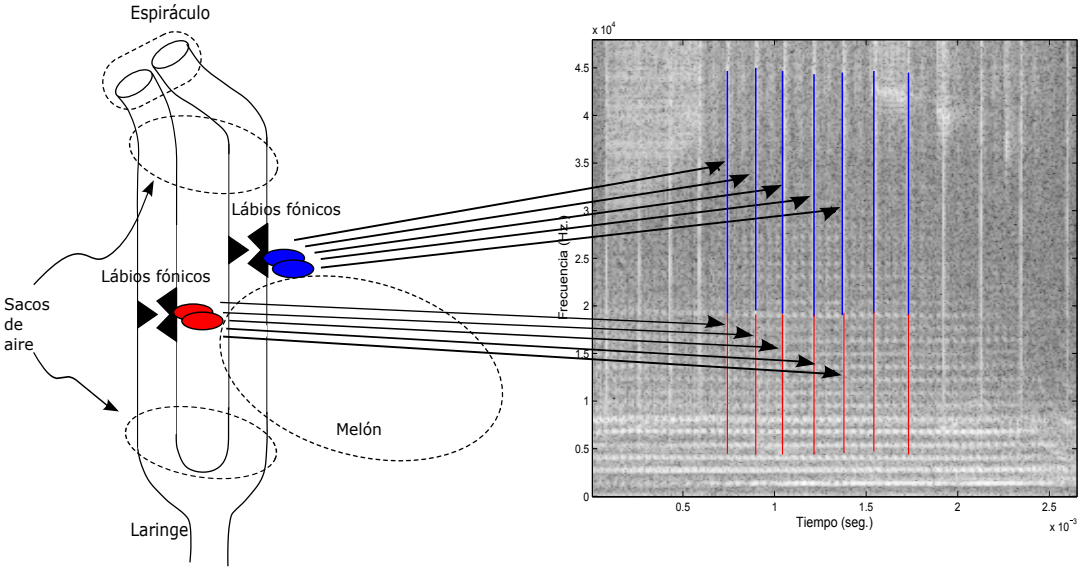


Figura 4.15: Asociación entre el espectrograma de un sonido mixto y su producción según Lammers.

En recientes estudios Lammers se retractó de su hipótesis, y junto a Madsen [19], concluyeron de modo experimental, que cada uno de los labios fónicos estaban especializados en la realización de sonidos vibratorios con rangos de f_0 diferentes. (ver Fig. 4.16. La identificación de dos fuentes de sonido potenciales en la mayoría de los odontocetos ayudó a explicar que estos animales son capaces de producir varios sonidos al mismo tiempo, denominados sonidos mixtos [53–56].

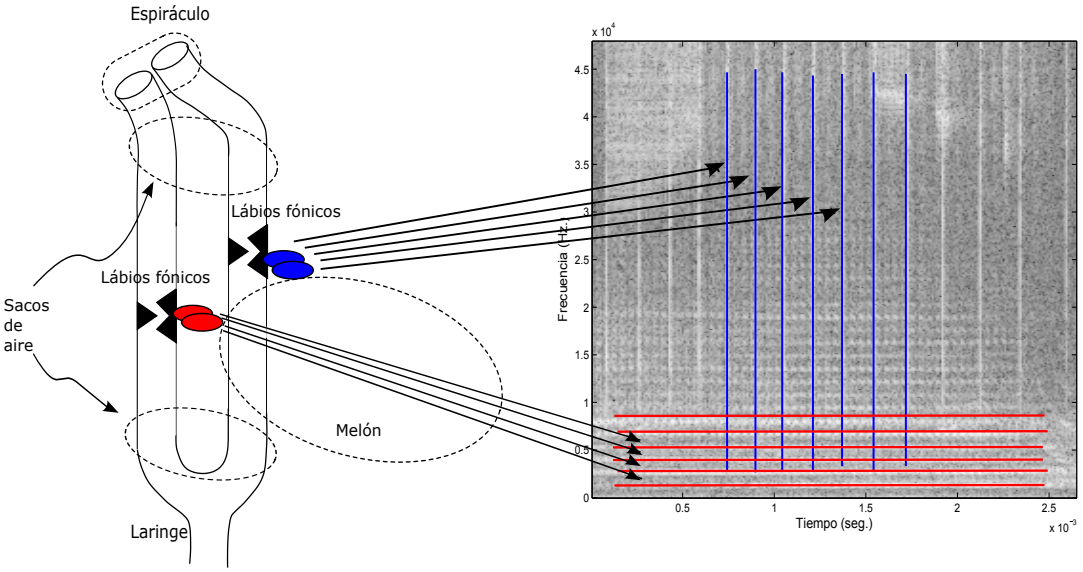


Figura 4.16: Asociación entre el espectrograma de un sonido mixto y su producción según Madsen.

Esta simultaneidad es relevante a la hora de explicar que los odontocetos son capaces de producir dos sonidos a la vez y por tanto, que poseen dos fuentes de sonido dentro de su estructura de producción. Finalmente, Madsen [19] concluye que los labios fónicos del lado derecho del animal están especializados en impulsos o *clicks* y los labios fónicos situados al lado izquierdo producen sonidos tonales, lo que traducido a la hipótesis de esta tesis doctoral indica que los labios fónicos realizan sonidos iguales pero cada uno de ellos los realiza con diferente frecuencia de vibración.

A al vista de lo explicado, se considera a los sonidos mixtos como la combinación de dos sonidos vibratorios con diferente f_0. Esto será importante a la hora de realizar un modelo del cual se extraigan características a la hora de clasificar adecuadamente la mayoría de sonidos realizados por los odontocetos.

4.5. Propagación del sonido a través de los órganos presentados

A la hora de llegar a una conclusión concreta sobre como se propagan los sonidos una vez creados en el sistema MLDB se realizaron una serie de grabaciones con la colaboración de los biólogos y entrenadores del Oceanográfico de Valencia. Debido a la capacidad que poseen sus ballenas beluga para la producción de sonidos no sólo en el medio subacuático (medio por el cuál se comunican por naturaleza) sino también por el medio aéreo, es interesante realizar un estudio exhaustivo relacionado con la dependencia entre la producción de los sonidos con la propagación de ellos previa a su salida al exterior.

4.5.1. Identificación de los órganos encargados de la propagación mediante acelerómetros

Con la ayuda de los entrenadores del Oceanografico de Valencia se diseñaron una serie de experimentos partiendo de la comunicación entrenador-beluga y gracias a la inteligencia que estos animales poseen. En 3 tandas repetidas de 10 minutos el entrenador realizó una serie de gestos a los cuales, las dos ballenas beluga respondieron con el sonido correspondiente, es decir, cada gesto correspondía a un sonido.

La Fig. 4.17 muestra el entrenador realizando uno de los gestos a la ballena beluga, a la cual ésta respondía siempre con un mismo sonido. En concreto se colocaron dos acelerómetros iguales de la marca PCB Piezoelectrics, modelo 353B17, con una sensibilidad de $0{,}956 mV/m/s^2$, uno en el melón y otro en el espiráculo con el propósito de capturar las vibraciones mecánicas en estos puntos, vibraciones provocadas por la salida del sonido al exterior.

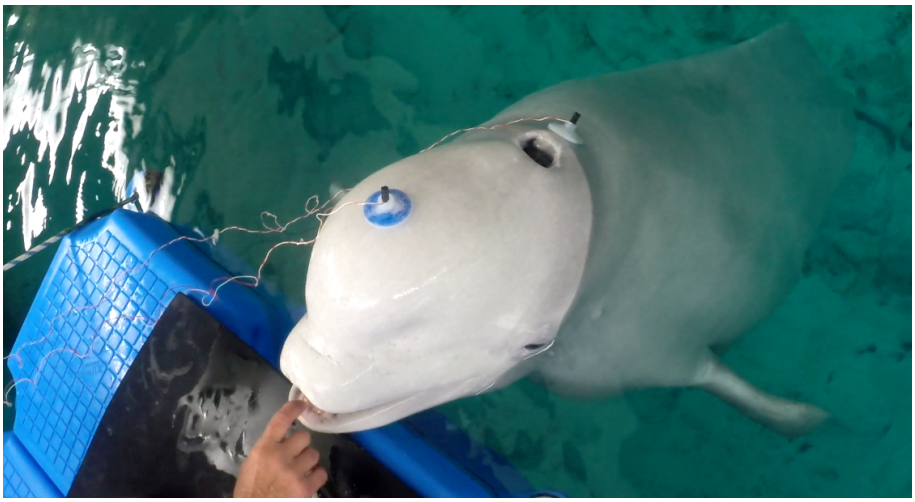


Figura 4.17: Muestra de la ballena beluga en el momento de la interacción con el entrenador.

Con el objetivo de obtener más información acerca de las diferencias entre los sonidos realizados dentro y fuera del agua, se realizarán varias hipótesis respecto al lugar por el cuál el sonido producido por el aparato fonador de las ballenas belugas sale al exterior.

Tal y como se puede observar en detalle en la Fig. 4.18, el acelerómetro colocado en el melón se situó dentro de una ventosa azul y el colocado en el espiráculo se situó en una ventosa blanca. Gracias a la colocación de los acelerómetros dentro de las ventosas, se consiguió que los acelerómetros quedaran totalmente fijados al cuerpo del animal resistiendo cualquier tipo de movimiento de su cabeza.

Para familiarizarse con el tipo de datos extraídos de los acelerómetros se muestra la Fig. 4.19 como ejemplo. En ella se puede visualizar que una de las dos gráficas de las que se compone está coloreada en gris, indicando que la amplitud máxima de dicha señal es mayor que la otra. Se define la amplitud máxima de las dos señales como $A_{espiraculo}$

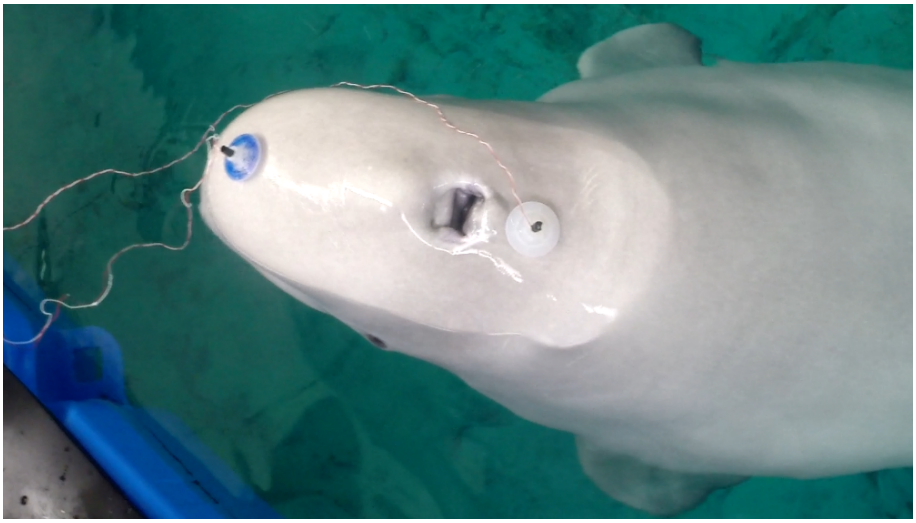


Figura 4.18: Muestra de la colocación de los dos acelerómetros en la ballena beluga.

y A_{melon}. Además, en cada una de las dos curvas aparece marcada una línea roja y otra verde. La línea roja ($T_{espiraculo}$) indica el momento en el que llega la vibración al espiráculo y la línea verde (T_{melon}) indica el momento que llega al melón. A continuación se realiza un análisis cualitativo de estos parámetros.

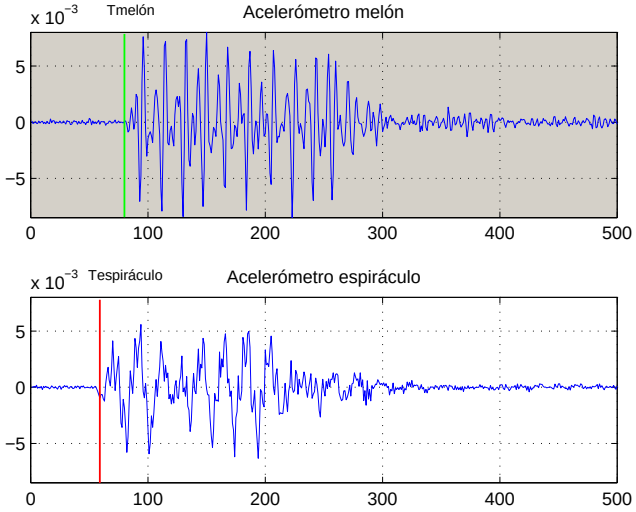


Figura 4.19: Sonido de ejemplo.

Al igual que en los seres humanos a la hora del posicionamiento y síntesis de sonidos en estéreo se computará de una manera sencilla la diferencia del tiempo de llegada (DT) de llegada y la diferencia de amplitudes máximas (DA) de las señales captadas por los

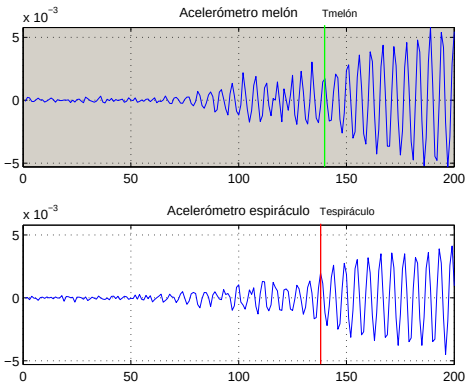
2 acelerómetros. En las Ec. (6.4) y (6.5) se detalla como se calculan ambos valores, basándonos en los anteriormente descritos.

$$DT = T_{melon} - T_{espiraculo} \tag{4.2}$$

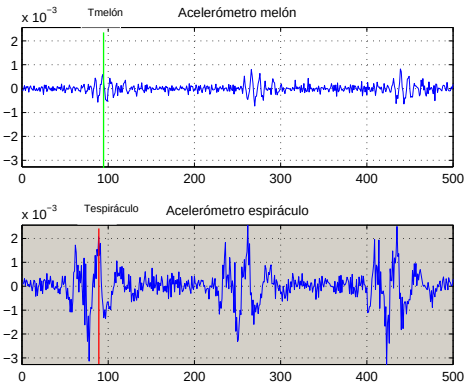
donde T_{melon} y $T_{espiraculo}$, tal y como se puede ver en la Fig. 4.19, son respectivamente el tiempo que tarda la señal en llegar al acelerómetro azul situado en el melón y el tiempo que tarda la señal en llegar al acelerometro blanco situado en el espiráculo.

$$DA = A_{melon} - A_{espiraculo} \tag{4.3}$$

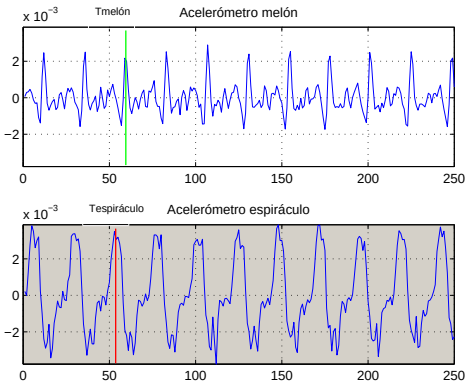
donde A_{melon} es la amplitud máxima con la que la señal llega al acelerómetro azul situado en el melón y $A_{espiraculo}$ la amplitud máxima con la que la señal llega al acelerómetro blanco situado en el espiráculo, tal y como se puede ver en la Fig. 4.19.



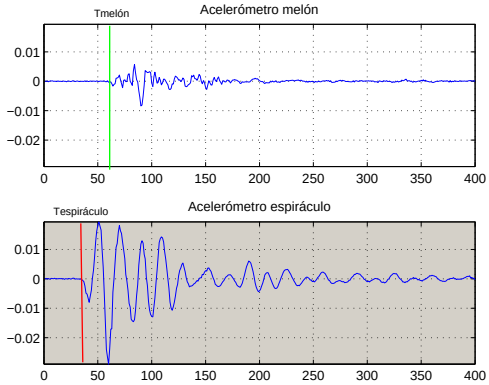
(a) Sonido resonante.



(b) Sonido vibratorio (frec. fund. baja)



(c) Sonido vibratorio (frec. fund. alta)



(d) Sonido espiráculo

Figura 4.20: Curvas de las señales recogidas por los acelerómetros.

Sonidos producidos	Nº sonidos	Diferencia temporal (DT)	Diferencia amplitud (DA)
Sonidos resonantes	42	Negativa	Positiva
Sonidos vibratorios	43	Negativa	Negativa
Espiráculo	67	Negativa	Negativa
Sonidos mixtos	8	Negativa	Depende
Sonidos concatenados	4	Negativa	Depende

Tabla 4.2: Diferencias temporales y de amplitud en los acelerómetros por cada tipo de sonido.

En la Fig. 4.20 se puede ver una muestra de 4 sonidos diferentes, por orden de aparición: un sonido resonante, un sonido vibratorio con frecuencia fundamental baja, un sonido vibratorio con frecuencia fundamental alta y un sonido producido por el movimiento del espiráculo.

Cada uno de los sonidos constan de dos curvas, una primera captada por el acelerómetro situado en el melón y una segunda captada por el acelerómetro situado en el espiráculo. Se comprueba que todas las ondas acústicas salen antes por el espiráculo que podría dar a entender que una vez creado el sonido en el sistema MLDB el paso por el melón como transductor de vibraciones del medio aéreo al medio acuoso retrasa la señal creada respecto a las vibraciones que salen directamente por el espiráculo sin ninguna transición aparente. Es por tanto que la diferencia temporal (DT) es siempre negativa en cualquier de los 4 casos.

Respecto a la diferencia de amplitud (DA) en el caso de los sonidos vibratorios suponiendo el efecto del melón en la adaptación de los sonidos vibratorios, provoca que las vibraciones lleguen con más amplitud por el espiráculo. Un caso especial son los sonidos provocados por el paso del aire a través del espiráculo y su movimiento en forma de impulsos. En esta ocasión el sonido se crea en el mismo espiráculo y las vibraciones producidas llegan al melón mediante la capa de grasa que estos animales tienen como piel y no a través del melón. Sin embargo en los sonidos resonantes (Fig. 4.20a), la DA es positiva, lo que indica una naturaleza diferente de producción respecto a los sonidos vibratorios.

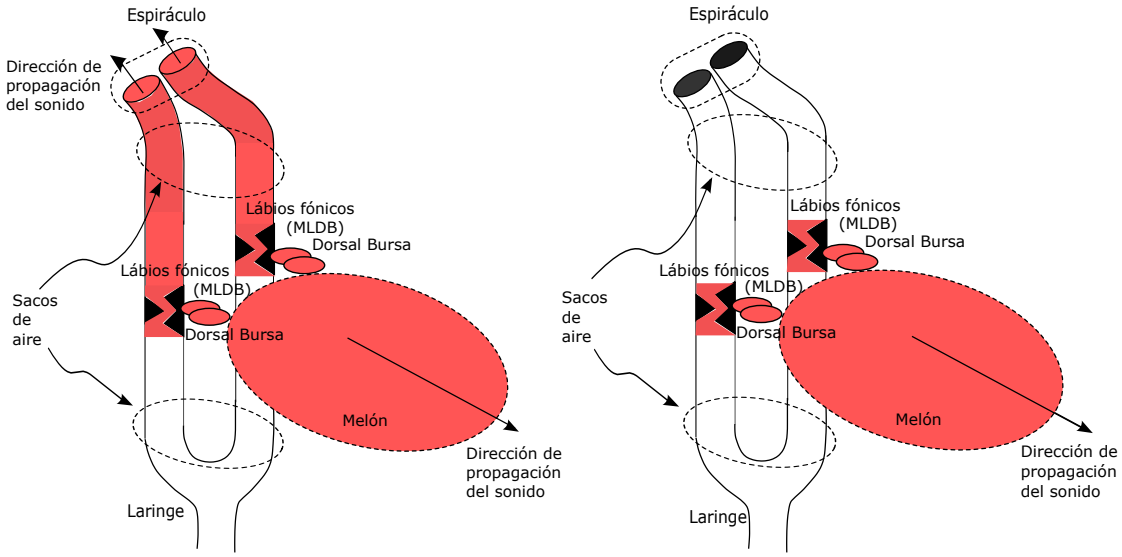
La Tabla 4.2 resume los valores DA y DT de cada uno de los sonidos. En ella se recogen todos los sonidos grabados, clasificados en sonidos resonantes, como el de la Fig. 4.20a, sonidos vibratorios como los de las Fig. 4.20b y 4.20c, así como sonidos mixtos, sonidos concatenados y sonidos producidos por el espiráculo como el de la Fig. 4.20d. Destacar que en los sonidos concatenados, que consisten en un tipo de sonido seguido de otro diferente, dependiendo del sonido que se sitúe en primer lugar, la DA será positiva o negativa.

En el caso de los sonidos mixtos, compuestos o bien por dos sonidos vibratorios producidos simultáneamente (uno con una frecuencia fundamental más baja que el otro), o bien por un sonido vibratorio de frecuencia fundamental baja unido a un sonido resonante, resulta complicado identificar la DA dado que se debe separar previamente las señales para realizar cualquier tipo de análisis. Teniendo en cuenta el objetivo del expe-

rimento, tanto los sonidos mixtos como los concatenados no son relevantes a la hora de establecer cualquier hipótesis asociada a él.

4.5.2. Propagación e independencia con la producción

Una vez explicados y analizados los resultados obtenidos mediante la captación de vibraciones a través de los dos acelerémetros, se proponen unas conclusiones en consecuencia a ellos. En la Fig. 4.21, se pueden observar los órganos formantes del sistema de producción de sonidos en un odontoceto. Respectivamente, la Fig. 4.21a se puede observar en rojo como los sonidos creados en el sistema MLDB, es decir, los vibratorios, se propagan a través del aparato fonador tanto por el melón por las membranas *dorsal bursa* como por el espiráculo a través de la laringe. Sin embargo, en la Fig. 4.21b explica como cuando los odontocetos realizan los sonidos debajo del agua, es decir, en un medio acuoso, el espiráculo permanece cerrado y los sonidos que realizan se transmiten a través del melón hacia el exterior.



(a) Propagación del sonido en un medio aéreo. (b) Propagación del sonido en un medio acuoso.

Figura 4.21: Órganos involucrados del sistema de producción de sonido de los odontocetos.

El melón hace de transductor de las ondas acústicas aire-agua, además de una labor de direccionamiento de señales de ecolocalización, emitiendo las ondas acústicas mecánicas provocadas por el sistema MLDB en el caso de que el medio donde se encuentre sea el agua. La forma en que los sonidos se propagan hacia el medio aéreo es el espiráculo. La hipótesis que se realiza es que es posible emplear sonidos aéreos y extrapolar que los sonidos submarinos serán producidos por los mismos órganos, con la salvedad de que el espiráculo se encuentra cerrado al estar el animal dentro del agua.

Tiene sentido recalcar que, al igual que ocurre en los seres humanos a través de la faringe y boca, cualquier vibración producida dentro de un tubo, se propaga por él hasta que encuentra un obstáculo para ello. En los odontocetos este obstáculo es el espiráculo, el cual se contrae, cerrando el tubo cuando el individuo está dentro del agua y se abre para respirar al salir a la superficie.

Las diferencias en la amplitud entre todos los sonidos resonantes y los sonidos vibratorios refuerzan la hipótesis de que ambos tipos de sonidos se producen de forma diferente. La solución sobre donde son producidos los sonidos resonantes es inabordable y escapa al alcance de esta tesis doctoral.

Se comprueba que el espiráculo colorea la señal original (que se propaga directamente a través de las membranas *dorsal bursa* hacia el melón) como si de un filtro lineal $h(n)$ se tratara. Tal y como ocurre con la boca de los seres humanos, en los sonidos aéreos el espiráculo toma una importancia relevante, ya que los sonidos son filtrados por el espiráculo, modificando su comportamiento temporal. Cabe decir que en los sonidos subacuáticos, al permanecer cerrado el espiráculo son transmitidos o a través del melón. En definitiva, la diferencia entre los sonidos subacuáticos con los sonidos aéreos recae en los siguientes aspectos:

- En los sonidos aéreos, el sonido se transmite de dos formas, a través de las membranas *dorsal bursa* al melón y por el al medio acúatico, además de por los tubos que comunican el sistema MLDB con el espiráculo, por donde sale el sonido al medio aéreo.

- En los sonidos realizados con la cabeza fuera del agua, la amplitud del sonido al salir del melón es menor que al salir por el espiráculo, debido a que los sonidos creados en dicho medio aéreo en el interior del cuerpo del animal y la adaptación al medio aéreo no es necesaria.

- Los sonidos subacuáticos, al estar cerrado el espiráculo, son sólo transmitidos a través del melón.

- Tanto los sonidos aéreos como los subacuáticos comparten la misma producción y parte de la propagación, añadiendo en el caso aéreo la propagación por la laringe hasta el espiráculo.

4.6. Propuesta de nuevas categorías de clasificación

En la Tabla 4.3 se puede ver resumida la clasificación propuesta para los sonidos de los odontocetos. En la clasificación basada en la morfología del espectrograma existen dos tipos de señales: los sonidos pulsados y los sonidos tonales. Los primeros caracterizados por tener asociado diagrama tiempo-frecuencia con componentes verticales y los segundos por una representación en forma de líneas horizontales. Sin embargo, en la clasificación basada en la naturaleza de producción de los sonidos, los sonidos tonales se dividen en sonidos vibratorios con f_0 alta y sonidos resonantes, y después de la exposición efectuada

conforme a la longitud de la ventana de análisis, se equipara las señales pulsadas a los sonidos vibratorios con f_0 baja.

Clasificación	**Tipos de sonidos**		
Según espectrograma	Pulsados	Tonales	
Según excitación	Vibratorios f_0 baja	Vibratorios f_0 alta	Resonantes

Tabla 4.3: Tipos de sonido según clasificación.

4.7. Conclusiones

Este capítulo sigue con en la línea de la investigación de Madsen [19], la cual especializa cada sonido para unos labios fónicos. En concreto los de la parte derecha del animal en sonidos vibratorios, sonidos con frecuencias de vibración medio-bajas (reflejado con componentes verticales en el espectrograma), y los de la parte izquierda en sonidos vibratorios con frecuencias de vibración altas (reflejado con componentes horizontales en el espectrograma), normalmente utilizadas para la comunicación entre individuos [45].

A lo largo de todo el capítulo se habla sobre los distintos sonidos según su naturaleza de producción (Tabla 4.4), mostrando como los instrumentos de viento como la trompeta y el trombón, es decir, instrumentos de viento con una membrana vibratoria, la voz humana y los sonidos vibratorios que producen los odontocetos, guardan similitudes en cuanto a la naturaleza de producción del sonido, reflejada tanto en la forma de los espectrogramas, como en su forma temporal.

Tipos de sonidos	**Humanos**	**Instrumentos**	**Odontocetos**
Con elemento vibrante	Voz	Trompeta, trombón	Sonidos vibratorios
Sin elemento vibrante	Silbidos	Flauta, clarinete	Sonidos resonantes

Tabla 4.4: Resumen de la clasificación de los sonidos.

A su vez, entre los instrumentos de viento sin elemento vibrante, los silbidos humanos y los sonidos resonantes de los odontocetos se comparte la forma de producción del sonido, reflejada en las señal temporal y en la forma de espectrogramas que estos sonidos tienen asociados.

Tal y como se argumenta a lo largo de todo el capítulo, los sonidos provocados por los labios fónicos en ambas ramas del sistema de producción de sonidos se producen de la misma manera. Es decir, que se comporten de manera diferente en el espectrograma no quiere decir que sean creados de manera distinta, sino que su f_0 es distinta y es el efecto de la LVA lo que provoca que se vea de una manera o de otra. Estos sonidos basados en la excitación de un elemento vibrante se caracterizan por tener un elevado ancho de banda y una frecuencia fundamental con menor nivel que los demás armónicos.

Capítulo 5

El dominio cepstral aplicado a sonidos subacúaticos

5.1. Introducción

En el ámbito del tratamiento de señal, la transformación del dominio temporal al dominio frecuencial es un recurso muy potente y útil para extraer información que de otra forma es complicada obtener [57]. Tanto en ultrasonidos, para visualizar los ecos [58, 59] como en la detección de no linealidades, en el tratamiento de imagen y filtrado [60] y en la detección de bordes [61], se desarrollaron y se siguen desarrollando multitud de algoritmos con el inconveniente habitual del coste computacional asociado al cálculo de la transformada de Fourier necesaria para la obtención del dominio frecuencial. Tales son los avances al trabajar en el dominio frecuencial que los posibles inconvenientes computacionales son asumibles dada la importancia de la información que se obtiene.

Las ondas acústicas son señales que se adecuan perfectamente al análisis de Fourier dada su naturaleza sinusoidal. El desarrollo de algoritmos de reconocimiento de audio [62], la obtención del pitch de la voz humana o la identificación de los armónicos [63] son algunas de las aplicaciones en las que actualmente la transformada de Fourier sigue siendo clave.

Sin embargo, dado el potencial que supuso la aplicación de la transformada de Fourier en la voz humana, se siguió investigando proponiendo un modelo de generación de voz basado en sistemas lineales invariantes, donde, asociándose perfectamente con la manera de producción de sonido de voz humana, era posible definirla mediante una fuente de sonido $x(n)$, localizada en la vibración de las cuerdas vocales y un filtro lineal invariante $h(n)$, localizado básicamente en el tracto vocal y nasal, los cuales daban como resultado la voz humana $y(n)$.

El filtro $h(n)$ es fácilmente calculable mediante los Coeficientes de Predicción Lineales (LPC) obtenidos a partir del dominio frecuencial, es decir, mediante la transformada de Fourier. Pero para conseguir adecuadamente la frecuencia de vibración de las cuerdas vocales es necesario volver a transformar de la señal del dominio frecuencial a lo que se denomina el dominio cepstral [64,65]. Con ello se consiguió obtener un modelo de síntesis

fácilmente asociable al aparato fonador humano.

Varios estudios mostraron la importancia del dominio cesptral en señales acústicas en ámbitos que sobrepasan la voz humana, pero que tienen mucha relación con ella, como por ejemplo los sonidos provocados por otros mamíferos [66, 67]. En estos estudios se demostró que la obtención de una señal acústica en el dominio cepstral puede utilizarse a la hora de clasificar los diferentes sonidos de una forma sencilla y compacta.

Dado que muchos de los sonidos producidos por los odontocetos son sonidos mixtos, es decir, producidos simultáneamente cada uno con distinto pitch, una de las partes más importantes a la hora de realizar un modelado que se adapte a la fisionomia del aparato de producción de sonidos de estos animales, será realizar una buena separación entre dichos sonidos para poder caracterizarlos por separado [68].

El dominio cepstral es una transformación usada para convertir señales combinadas con la convolución, en sumas de su espectro. Los picos que aparecen indican el período fundamental del que dependen los armónicos del espectro. Para que aparezcan picos se deben usar señales que sí contengan armónicos, como la suma de dos senos, en la cuál el segundo tenga un armónico de la primera [68]. Tener combinación de dos señales que poseen armónicos permite utilizar las ventajas de realizar una transformación al dominio cepstral para poder separar la información armónica mediante una máscara adecuada.

En este capítulo se muestra como el dominio cesptral puede utilizarse no sólo como herramienta para clasificar sonidos de odontocetos, sino también para poder separar los sonidos producidos en su aparato fonador, el cual posee una estructura de producción en paralelo capaz de realizar dos sonidos simultáneamente.

5.2. La utilización del dominio cepstral para la clasificación de sonidos de mamíferos marinos mediante los Mel-Frequency Cepstral Coefficients

Una de las maneras más comunes de utilizar el dominio cepstral es a partir de los *Mel-Frecuency Cepstral Coefficients* (MFCC). En el ámbito del audio subacuático se han aplicado los MFCC a sonidos producidos por los cetáceos demostrando que es posible utilizarlos para obtención de la frecuencia fundamental de dichos sonidos y su posterior clasificación [69, 70]. Los sonidos producidos por la mayoría de odontocetos están contenidos en las bandas de frecuencia de Mel, debido a ello, la utilización de los MFCC es exitosa.

La obtención del dominio cepstral se realiza aplicando dos transformaciones a cada trozo de 25 ms $x_t(n)$ de la señal de entrada $x(n)$. En el caso de los MFCC, la segunda transformación es una transformada Discreta del Coseno (DCT). A continuación se muestran los pasos a la hora de obtener los MFCC:

I. Troceamos la señal $x(n)$ en trozos de 25 ms $x_t(n)$

II. Realizamos la transformada discreta de Fourier de cada una de los trozos $X_t(\Omega) = DFT[x_t(n)]$

III. Tomamos el valor absoluto

IV. Filtrado en bandas de Mel usando una ventana triangular $S_K = K_{mel-filters}|X_t(\Omega)|$ obteniéndose un vector de K=20 componentes $S_K = [S_1, ..., S_{20}]$

V. Tomamos logaritmos: $log(S_K)$

VI. Realizamos la DCT: $c_l = DCT[log(S_K)]$ donde $1 \leq l \leq L$ consiguiendo $L = 14$ coeficientes c_l, creando el vector $c_{1:L} = [c_1, ..., c_L]$, vector que corresponde a los denominados MFCC.

5.2.1. Comportamiento y propiedades de los MFCCs

El comportamiento de los MFCC tiene varias cualidades que lo hacen adecuado para el análisis de audio. En la Fig. 5.1 se observa como señales débiles en el tiempo, representadas en forma de espectrograma de $K = 20$ coeficientes (S_K), son transformadas al dominio cepstral. El carácter frecuencial que pueden poseer las señales en un espectrograma no se pierde en el dominio cepstral (Fig. 5.1a y Fig. 5.1b).

Además en las Fig. 5.1c y 5.1d se observa como la pendiente de cualquier señal en un espectrograma está reflejada en los MFCC. Por último y como cualidad más importantes de las MFCC, se representa en las Fig. 5.1e y 5.1f una señal a la cual se le añade un armónico en un momento determinado. Vemos que la presencia de este armónico provoca que el nivel en algunos MFCC aumente considerablemente.

Los MFCC permiten por tanto obtener información frecuencial compactada en $L = 14$ coeficientes por trozo $x_t(n)$ gracias a la utilización de la DCT y las bandas de frecuencia de Mel, con el ahorro computacional que esto supone a la hora de realizar clasificadores, ofreciendo una fácil solución tanto a problemas de detección como de clasificación de señales en las cuales el espectrograma no es del todo identificable. La capacidad de los MFCC de identificar los armónicos de una señal, el reflejo en forma de repeticiones de cualquier señal tonal y el mantenimiento del comportamiento morfológico de dicha señal hacen del dominio cepstral un dominio adecuado para conseguir información no visible en el espectrograma, estableciendo mayores diferencias entre señales y así poder clasificarlas y discriminarlas de una manera más sencilla.

Velocidad y aceleración MFCCs

En el caso de los odontocetos, las vocalizaciones, al contrario de los seres humanos, tienen un comportamiento muy diferente en el tiempo. El lenguaje de los seres humanos conlleva una cantidad de información por segundo mucho mayor que los demás mamíferos, ya que cada palabra contiene sílabas, que son pronunciadas en un periodo muy corto de tiempo. Los odontocetos suelen mantener durante más de medio segundo los sonidos que emiten, simplemente son capaces de modificar el pitch, subiendo o bajando la frecuencia fundamental.

Dado que la información se mantiene durante mayor cantidad de tiempo, es posible utilizar la información previa y posterior de cada trozo obteniendo características tem-

porales que se utilizarán en la detección o clasificación. Los MFCC obtenidos hasta este punto se denominarán MFCC estáticos. Se introducirán además los MFCC de velocidad y de aceleración. Los MFCC de velocidad (d_l) se obtendrán a partir de la cantidad $P = 2$ de vectores previos y posteriores de MFCC estáticos (c_l) tal y como muestra la Ec. (5.1) y la Fig. (5.2a):

$$d_{1:L,n} = \frac{\sum_{p=1}^{P} p(c_{1:L,n+p} - c_{1:L,n-p})}{2\sum_{p=1}^{P} p^2} \tag{5.1}$$

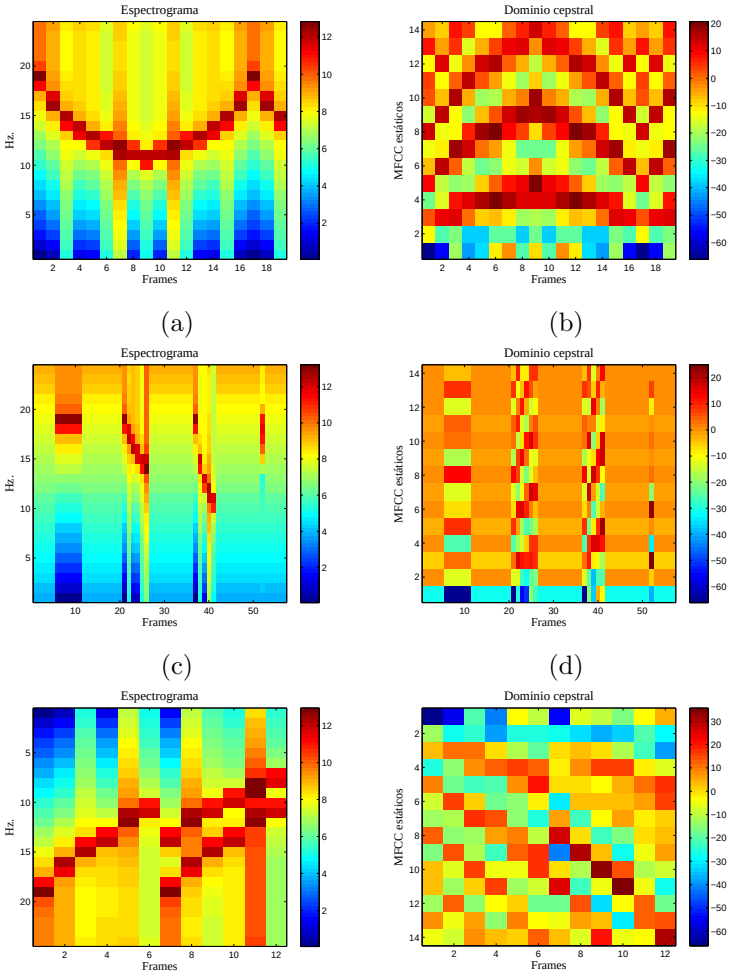
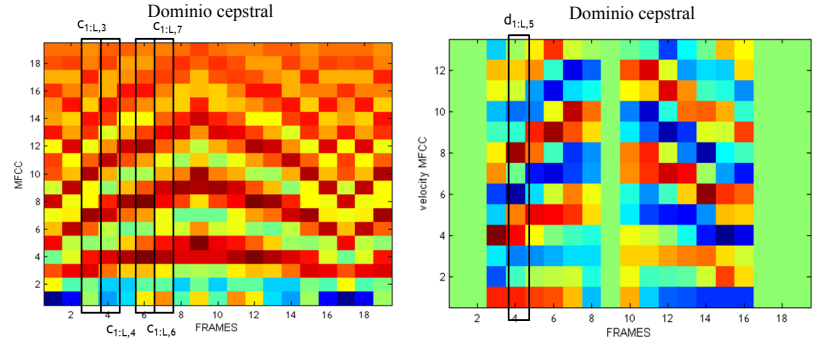
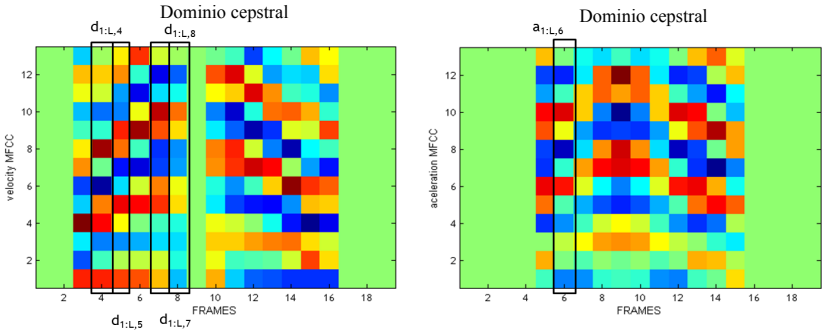


Figura 5.1: a) Espectrograma de una señal tonal. b) MFCC de dicha señal. c) Espectrograma de una serie de señales con diferentes pendientes. d) MFCC de la serie de señales. e) Espectrograma de una señal con un armónico a partir de un momento determinado. f) MFCC de dicha señal.



(a) MFCC estáticos (izquierda) a partir de los cuales se obtienen los MFCC de velocidad (derecha). Los rectángulos negros seleccionan los vectores MFCC estáticos necesarios a la hora de obtener los MFCC de velocidad.



(b) MFCC de velocidad (izquierda) a partir de los cuales se obtienen los MFCC de aceleración (derecha). Los rectángulos negros seleccionan los vectores MFCC de velocidad necesarios a la hora de obtener los MFCC de aceleración.

Figura 5.2: Muestra del cálculo de los MFCC de velocidad y aceleración.

Por ejemplo, del trozo x_f número 5, se obtiene el vector de MFCC estáticos $c_{1:L,5}$. Establecidos $P = 2$ y $L = 14$, se calcula $d_{1:L,5}$ a partir de los MFCC estáticos previos $c_{1:L,3}$ y $c_{1:L,4}$, y posteriores $c_{1:L,6}$ y $c_{1:L,7}$. Del mismo modo, a través de los $P = 2$ vectores de MFCC de velocidad $(d_{1:L})$ anteriores y posteriores, se conseguirán los coeficientes MFCC de aceleración $(a_{1:L})$ como se puede observar en la Fig. 5.2b y en la Ec. (5.2):

$$a_{1:L,n} = \frac{\sum_{p=1}^{P} p(d_{1:L,n+p} - d_{1:L,n-p})}{2\sum_{p=1}^{P} p^2} \tag{5.2}$$

De esta manera se obtendrá el vector mostrado en la Ec. (5.3), a partir de los MFCC estáticos (c_l), los MFCC de velocidad (d_l) y los MFCC de aceleración (a_l).

$$MFCC_{total} = [c_1, ..., c_{14}, d_1, ..., d_{14}, a_1, ..., a_{14}] \tag{5.3}$$

Este vector de MFCCs ($MFCC_{total}$) compuesto de 42 componentes, resume la información frecuencial (c_l) del trozo $x_t(n)$ analizado, además de la evolución temporal (d_l y a_l). De esta manera se compacta en un vector toda la información necesaria para realizar una clasificación o detección con muy pocos coeficientes, adecuada para realizar unas matrices de entrenamiento lo suficientemente asequibles computacionalmente.

5.3. Aplicación del análisis en el dominio cepstral al modelado de señales mixtas: Separación de fuentes de Sonido y Estimación de la longitud de la Ventana de Análisis (SSEVA)

En las sección anterior se ilustra como las cualidades de las señales producidas por los odontocetos son más diferenciables en el dominio cepstral. Siguiendo los pasos que se observan en la Fig. 5.3, el dominio cesptral permite identificar por separado el pitch de cada una de las dos señales para después realizar un filtrado en el dominio cepstral (*liftering*) para después volver al dominio temporal con las dos señales separadas.

Cabe decir que para realizar de manera correcta esta separación de fuentes se debe desplazar a través de la señal una ventana de análisis de una longitud de ventana de análisis (LVA) la cual incluya dos o más periodos de ambas señales. Además, se necesita que la LVA sea lo suficientemente pequeña para que el trozo a analizar pueda ser considerado estacionario. Con una LVA= 4096 muestras se cumplen las dos premisas, ya que comparado con la frecuencia de muestreo ($f_s = 96000$ Hz) a la que están grabadas las señales la LVA es pequeño.

A continuación se explica el procesado propuesto en la Fig. 5.3 para realizar la separación de fuentes y estimar la LVA:

I. Se seleccionan 4096 muestras de la señal $y(n)$

II. Se realiza la transformada de discreta Fourier de cada una de los trozos $Y(\Omega) = DFT[y(n)]$

III. Se toman logaritmos: $log(Y(\Omega))$

IV. Se realiza la transformada de discreta Fourier Inversa, es decir, se llega al dominio cepstral: $y(c) = DFT^{-1}log(Y(\Omega))$

V. Ya en el dominio cepstral, se realiza un filtrado con dos filtros complementarios, obteniendo dos señales $y_{lenta}(c)$ y $y_{rapida}(c)$

VI. Se realiza la transformada de discreta Fourier de cada una de las señales: $Y_{lenta}(\Omega) = DFT[y_{lenta}(c)]$, $Y_{rapida}(\Omega) = DFT[y_{rapida}(c)]$

VII. Se toman logaritmos de las dos señales.

VIII. Se realiza la transformada de discreta Fourier Inversa de las dos señales.

IX. Se mueve la ventana de análisis 2058 muestras y se vuelve al paso 1.

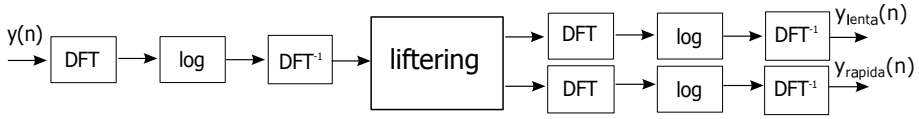


Figura 5.3: Separador de Señales y Estimador de longitud de la Ventana de Análisis (SSEVA).

Para demostrar el funcionamiento del algoritmo propuesto se modelan los sonidos mixtos $y(n)$ como la suma de dos señales vibratorias de distinto periodo, es decir, una señal vibratoria de f_0 baja $y_{lenta}(n)$ y una señal vibratoria de f_0 baja $y_{rapida}(n)$.

$$\begin{aligned} y(n) &= y_{lenta}(n) + y_{rapida}(n) \\ &= x_{lenta}(n) * h_{lenta}(n) + x_{rapida}(n) * h_{rapida}(n) \end{aligned} \tag{5.4}$$

Tal y como se verá en el Capítulo 7, ambas señales vibratorias tienen como excitación ($x_{rapida}(n)$ y $x_{lenta}(n)$) un tren de deltas (ver Ec. (6.4) y Ec. (6.5)), coloreado por los filtros $h_{rapida}(n)$ y $h_{lenta}(n)$

$$x_{lenta}(n) = \sum_{k_p=-\infty}^{\infty} \delta\left(n - k_p T_{Long}\right), \tag{5.5}$$

$$x_{rapida}(n) = \sum_{k_t=-\infty}^{\infty} \delta\left(n - k_t T_{Short}\right), \tag{5.6}$$

donde T_{Long} y T_{Short} son respectivamente los periodos correspondientes a las excitaciones.

Cabe decir que la literatura referente al modelado de señales mediante Vocoders como el LPC-V (ver Capítulo 7) ofrecen una nomenclatura amplia a la hora de denominar a la excitación correspondiente a la creación de sonidos con pitch, como es el caso de las señales mixtas. Tren de pulsos o tren de impulsos son expresiones alternativas usadas para referirse a dicha excitación.

En esta tesis doctoral se empleará la denominación tren de deltas. Todas las nomenclaturas son utilizadas para referirse a una señal discreta formada por un impulso unidad (ver Ec. (5.7)) repetido periodicamente, obteniendo excitaciones como las de las Ec. (6.4) y (6.5).

$$\delta(n) = \begin{cases} 1 & n = 0 \\ 0 & n \neq 0 \end{cases} \tag{5.7}$$

Aplicando la transformada de Fourier a la Ec. (5.4) se obtiene lo siguiente:

$$\begin{aligned} Y(\Omega) = DFT\left(y(n)\right) &= DFT\left(y_{rapida}(n) + y_{lenta}(n)\right) \\ &= DFT(y_{rapida}(n)) + DFT(y_{lenta}(n)) \\ &= Y_{rapida}(\Omega) + Y_{lenta}(\Omega) \\ &= X_{lenta}(\Omega)H_{lenta}(\Omega) + X_{rapida}(\Omega)H_{rapida}(\Omega) \end{aligned} \tag{5.8}$$

donde $X_{lenta}(\Omega)$ y $X_{rapida}(\Omega)$ son otro tren de deltas con periodo igual a la frecuencia fundamental de la excitación ($1/T_{Long}$ y $1/T_{Short}$) como vemos en las Ec. (5.9) y (5.10)

$$X_{lenta}(\Omega) = \frac{\pi}{2}\sum_{k_p=-\infty}^{\infty}\delta\left(\Omega - \frac{k_p}{T_{Long}}\right) \tag{5.9}$$

$$X_{rapida}(\Omega) = \frac{\pi}{2}\sum_{k_t=-\infty}^{\infty}\delta\left(\Omega - \frac{k_t}{T_{Short}}\right) \tag{5.10}$$

tomando logaritmos, se obtiene lo siguiente:

$$\log Y(\Omega) = \log\left\{X_{lenta}(\Omega)H_{lenta}(\Omega) + X_{rapida}(\Omega)H_{rapida}(\Omega)\right\} \tag{5.11}$$

A partir de este momento se tienen dos casos posibles: cuando $k_t\Omega_{Short} \neq k_p\Omega_{Long}$ y cuando $k_t\Omega_{Short} = k_p\Omega_{Long}$. En el primer caso y por las propiedades de la $\delta(\Omega)$ se tiene que:

$$\begin{aligned} \log Y(\Omega) &= \log\left\{X_{lenta}(\Omega)H_{lenta}(\Omega)\right\} + \log\left\{X_{rapida}(\Omega)H_{rapida}(\Omega)\right\} \\ &= \log X_{lenta}(\Omega) + \log H_{lenta}(\Omega) + \log X_{rapida}(\Omega) + \log H_{rapida}(\Omega) \end{aligned} \tag{5.12}$$

para finalmente, mediante la Transformada de Discreta Fourier Inversa, obtener la señal en el dominio cepstral.

$$\begin{aligned} y(c) &= DFT^{-1}\left\{\log Y(\Omega)\right\} \\ &= DFT^{-1}\left\{\log X_{lenta}(\Omega) + \log H_{lenta}(\Omega) + \log X_{rapida}(\Omega) + \log H_{rapida}(\Omega)\right\} \\ &= DFT^{-1}\log X_{lenta}(\Omega) + DFT^{-1}\log H_{lenta}(\Omega) \\ &+ DFT^{-1}\log X_{rapida}(\Omega) + DFT^{-1}\log H_{rapida}(\Omega) \\ &= x_{lenta}(c) + h_{lenta}(c) + x_{rapida}(c) + h_{rapida}(c) \end{aligned} \tag{5.13}$$

En el segundo caso $k_t\Omega_{Short} = k_p\Omega_{Long}$, dado que empíricamente se ha comprobado que la potencia de $Y_{lenta}(\Omega) \geq 5Y_{rapida}(\Omega)$:

$$\begin{aligned}\log Y(\Omega) &= \log\{Y_{lenta}(\Omega) + Y_{rapida}(\Omega)\} \\ &= \log\left\{Y_{lenta}(\Omega)\left(1 + \frac{Y_{rapida}(\Omega)}{Y_{lenta}(\Omega)}\right)\right\} \\ &= \log Y_{lenta}(\Omega) + \log\left\{1 + \frac{Y_{rapida}(\Omega)}{Y_{lenta}(\Omega)}\right\} \\ &\approx \log Y_{lenta}(\Omega) \end{aligned} \tag{5.14}$$

la señal obtenida en el dominio cesptral será:

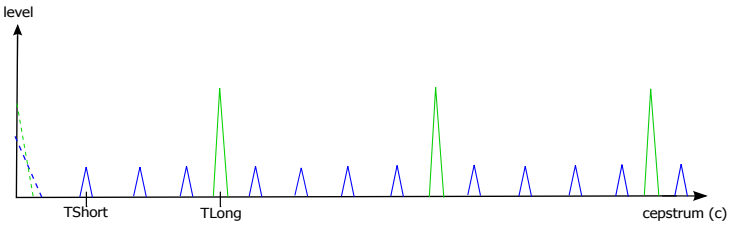
$$\begin{aligned} y(c) &= DFT^{-1}\{\log Y(\Omega)\} \\ &\approx DFT^{-1}\{\log Y_{lenta}(\Omega)\} \\ &= DFT^{-1}\{\log\left(X_{lenta}(\Omega)H_{lenta}(\Omega)\right)\} \\ &= DFT^{-1}\{\log X_{lenta}(\Omega) + \log H_{lenta}(\Omega)\} \\ &= DFT^{-1}\log X_{lenta}(\Omega) + DFT^{-1}\log H_{lenta}(\Omega) \\ &= x_{lenta}(c) + h_{lenta}(c) \end{aligned} \tag{5.15}$$

Tal y como se demuestra, en ambos casos (ver las Ec. (5.13) y (5.15)), se ha separado la información relevante de la señal lenta, teniendo en cuenta además su mayor potencia en comparación con la señal rápida.

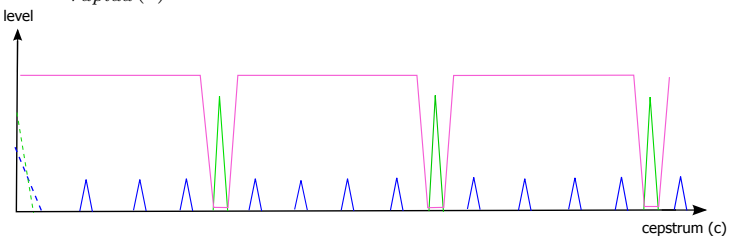
En la Fig. 5.4 se observa como queda distribuida esta información a lo largo del dominio cepstral. La excitación $x_{lenta}(c)$ será un tren de deltas centradas en $T_{Long} = n2\pi/\Omega_{Long}$ y $x_{rapida}(c)$ será un tren de deltas centradas en $T_{Short} = p2\pi/\Omega_{Short}$, siendo $n = 1, 2, ..., \infty$ y $p = 1, 2, ..., \infty$ respectivamente. En líneas discontinuas se obtiene la información relativa a $h_{lenta}(c)$ en verde y a $h_{rapida}(c)$ en azul, ambas se sitúan en valores bajos del dominio cepstral. En línea continua verde muestra los picos correspondientes a los periodos T_{Long} de la excitación de $y_{lenta}(c)$. En línea continua azul se puede ver el periodo de la señal vibratoria rápida $y_{rapida}(c)$ centrado en T_{Short}.

Una vez situados en el dominio cepstral y mediante un filtro peine, se seleccionan los tramos en los cuales se tiene la información armónica de la señal vibratoria lenta $x_{lenta}(n)$. En la Fig. 5.4 se puede observar en rosa el filtro utilizado, se trata de un filtro peine con polos en nT_{Long} siendo $n = 1, 2, ...\infty$ con el cual se obtiene la información armónica de $x_{lenta}(n)$. Con el filtro inverso (curva naranja) se consigue la información restante, es decir la que concierne a $x_{rapida}(n)$. Una vez separadas las distintas partes de las dos señales que componen la señal mixta y siguiendo el esquema de la Fig. 5.3, se realiza la transformación inversa, situándonos en el dominio del tiempo, donde se obtiene la señal vibratoria de f_0 baja $y_{lenta}(n)$ y la señal vibratoria de f_0 alta $y_{rapida}(n)$ por separado.

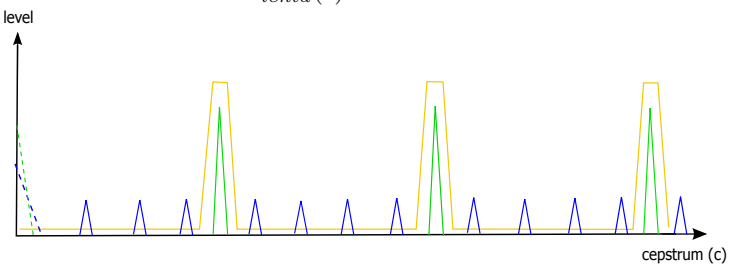
Además, el algoritmo consigue una elección apropiada de la LVA con el objetivo de analizar cada una de las señales en un posterior modelado de forma que $\text{LVA}_{\text{lenta}} >$



(a) Curva verde discontinua $h_{lenta}(c)$, curva verde continua $x_{lenta}(c)$, curva azul discontinua $h_{rapida}(c)$, curva azul continua $x_{rapida}(c)$



(b) Curva rosa máscara aplicada para la obtención de la información armónica de $x_{lenta}(c)$.



(c) Curva naranja (inversa a la curva rosa) máscara aplicada para la obtención de la información restante.

Figura 5.4: Información cepstral de una señal mixta y visualización del *liftering* propuesto.

$2T_{Long}$ y $LVA_{rapida} > 2T_{Short}$. Esta condición, analizada en el Capítulo 3 permitirá un análisis correcto mediante coeficientes de predicción lineal (LPC) (ver Capítulo 7).

En la Fig. 5.5, se puede observar el comportamiento en frecuencia del algoritmo de Separación de señales y Estimación de la longitud de la Ventana de Análisis (SSEVA), es decir, aplicando el filtro peine en el dominio cepstral. En la Fig. 5.5a se muestra el espectrograma de una señal sintética, que se asemeja a un sonido mixto de una ballena beluga, se puede ver como efectivamente está compuesta por dos señales de diferentes pitch o frecuencia de repetición. En las Fig.5.5b y 5.5c se muestra el espectrograma de cada una de las dos señales a la salida del algoritmo SSEVA. Respectivamente se observa el buen comportamiento del filtrado cepstral a la hora de discriminar la señal con más f_0 de la que tiene menos. En el caso de que la señal esté compuesta por solo un sonido, el algoritmo SSEVA simplemente encaminará el sonido por una de sus dos salidas, según

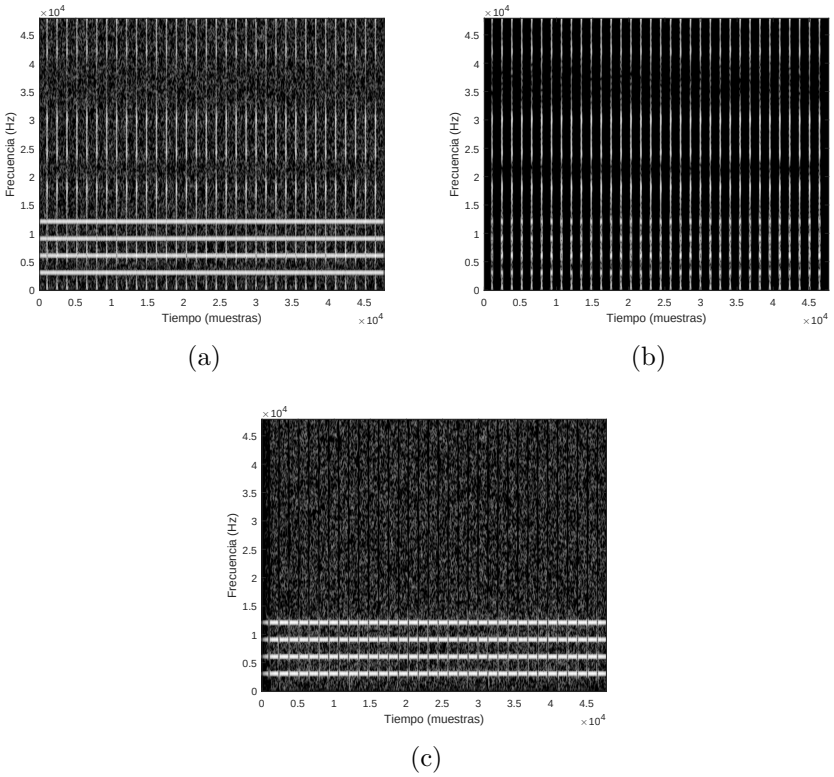


Figura 5.5: a) Diagrama tiempo-frecuencia de una señal sintética compuesta de dos señales con diferente pitch. b) Diagrama tiempo-frecuencia de la señal sintética de una de las salidas del algoritmo SSEVA, concretamente de la señal lenta. c) Diagrama tiempo-frecuencia de la señal sintética de una de las salidas del algoritmo SSEVA, la señal rápida.

su frecuencia fundamental.

El potencial del domimio cepstral nos permite, además de detectar el pitch de los sonidos de las ballenas beluga y con ello estimar que LVA se deberá utilizar en su posterior modelado mediante coeficientes LPC-V, separar las señales mixtas mediante el algoritmo SSEVA explicado.

El contenido mostrado en este capítulo ha generado la participación en el siguiente congreso:

- Guillermo Lara, Ramón Miralles y Alicia Carrión. "Right Whale activity detector and sound classifier using Mel-Frequency Cepstral Coefficients". Expuesto por Guillermo Lara en el 6th International Workshop on Detection, Classification, Localization and Density Estimation of Marine Mammals using Passive Acoustics. University of St. Andrews. Escocia. Año 2013.

Capítulo 6

El análisis cuantitativo de los *Recurrence Plots* como algoritmo de caracterización de la naturaleza de producción del sonido

6.1. Introducción

En este capítulo se estudiará una de las características con más importancia de los sonidos de los odontocetos, el determinismo. En el Capítulo 3 se explicó la clasificación tradicional de este tipo de señales en función de su representación tiempo-frecuencia, es decir, como sonidos tonales o pulsados. Para ello no sólo se utilizaron características obtenidas a partir del espectrograma, sino también otras obtenidas en el dominio temporal o por ejemplo, características estadísticas donde se puede obtener otro tipo de información, más cercana a la caracterización de la naturaleza de la señal.

Más concretamente, se obtuvieron características relacionadas con la linealidad, como por ejemplo la reversibilidad temporal, las cuales tenían una importancia fundamental para la clasificación de los sonidos. A raíz de estos resultados se comenzó a plantear la realización de un estudio de la linealidad y determinismo de las señales de los cetáceos con el objetivo de estudiar a fondo los buenos resultados que ofrecían estas características a la hora de diferenciar entre los dos tipos de señales fundamentales.

Además, y teniendo en cuenta la propuesta de esta tesis referente al cambio en el enfoque de la clasificación de los sonidos donde se plantea una clasificación basada en la naturaleza de producción del sonido y no una clasificación teniendo el cuenta el comportamiento del espectrograma, parece apropiado emplear este tipo de características en la clasificación.

Este capítulo trata por tanto de analizar los sonidos de los odontocetos bajo el punto de vista de características como el determinismo, con el objetivo de incluir en el modelo presentado un procesado que permita distinguir si los sonidos son creados de una manera u otra en el aparato fonador de estos animales.

6.2. Conceptualización del problema

La voz humana puede tener fragmentos sordos o sonoros. La clasificación de estas regiones permite una segmentación acústica preliminar a aplicaciones con procesado de audio tales como la síntesis de voz, el reconocimiento de la voz o aumento de volumen de la voz para personas con disfunciones auditivas [71].

Los denominados algoritmos de detección de pitch o *Pitch Detection Algorithms* (PDA) son algoritmos diseñados para estimar la frecuencia fundamental de una señal quasi periódica en voz humana, música y tonos. Estos algoritmos pueden estar diseñados para trabajar tanto en el dominio temporal, como el frecuencial o una combinación de ellos. La detección de pitch conlleva que los sonidos que no lo posean sean clasificados como sonidos sordos, dentro de estos sonidos se puede encontrar sonidos periódicos (resonantes) o sonidos no periódicos (consonantes). La mayoría de los PDA también ofrecen la capacidad de clasificar los dos tipos de sonidos, aunque muchos de ellos se centren más en una estimación más concreta del pitch que no en la distinción comentada.

En el caso que nos ocupa en esta tesis doctoral dejaremos a un lado los sonidos sordos no periódicos y nos centraremos en distinguir entre sonidos con pitch (vibratorios) y sonidos resonantes, ambos periódicos. Se tendrá en cuenta esta consideración a la hora de elegir los algoritmos adecuados que permitan no sólo averiguar la frecuencia fundamental de un sonido sino averiguar si esta frecuencia es causada por una vibración o una resonancia. Es decir, caracterizar la naturaleza de producción de los sonidos de las ballenas beluga. Realizar una correcta diferenciación de este tipo de sonidos en odontocetos es crucial para un correcto modelado.

Tal y como se explicó en el Capítulo 4, la frecuencia fundamental de una señal vibratoria se denomina pitch o frecuencia de repetición de la vibración. En el caso de los sonidos resonantes se denomina frecuencia de resonancia y son producidos, como su nombre indica, por resonancias y no por la vibración de ningún elemento. En resumen y para facilitar la lectura, sonidos vibratorios poseen pitch y los sonidos resonantes poseen frecuencia de resonancia.

Tendremos que tener en cuenta diferentes consideraciones para poder aplicar los algoritmos para la detección del pitch a los sonidos producidos por los mamíferos marinos. La manera en que los mamíferos marinos, en concreto los cetáceos, producen sus sonidos o vocalizaciones es un campo muy complejo de estudio. La evolución les ha dotado de unos órganos encargados de la producción de sonido extremadamente sofisticados y adaptados para poder ser capaces de realizar sonidos subacuáticos. El conocimiento sobre estos órganos y su funcionamiento es un factor clave para entender el repertorio de sonidos que son capaces de realizar además de conocer para qué los utilizan. Este conocimiento ha sido logrado sólo para un número limitado de cetáceos, los odontocetos.

Cabe destacar que, al tratarse de sonidos producidos por un sistema de producción duplicado al del aparato fonador humano, se puede llegar a tener dos sonidos simultáneos con dos pitches diferentes. En el Capitulo 5 se describió la manera de separar estos sonidos.

En la Fig. 6.1 se puede observar un ejemplo de los dos tipos de producción de sonidos

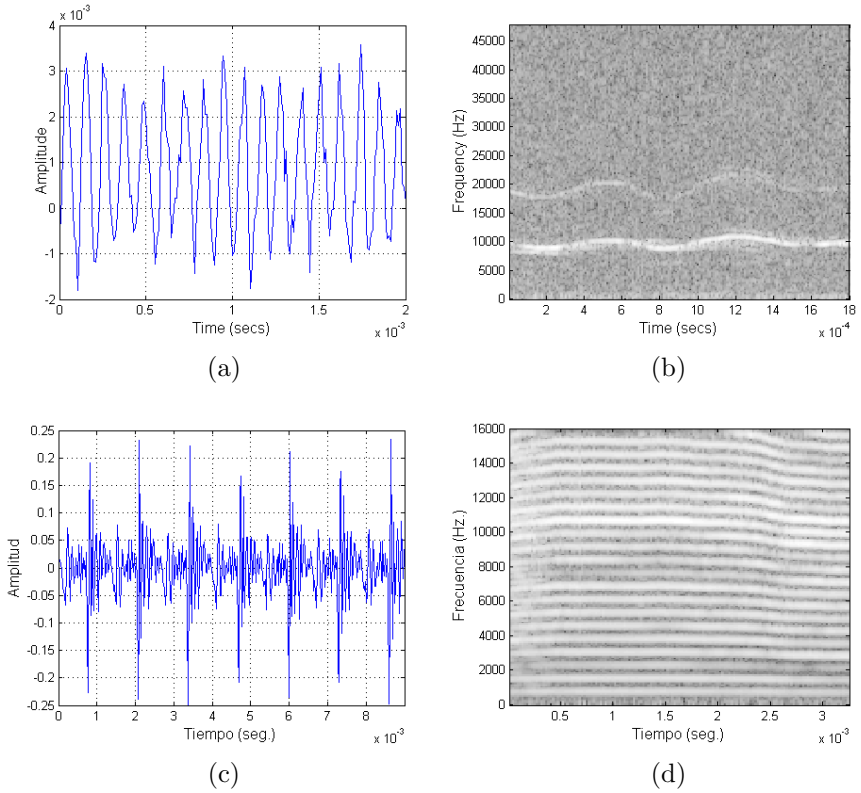


Figura 6.1: a) Segmento de un sonido resonante en el dominio temporal y b) en el dominio frecuencial en forma de espectrograma. c) Segmento de un sonido vibratorio en el dominio temporal y d) en el dominio frecuencial en forma de espectrograma.

en el dominio temporal y frecuencial. El sonido de la Fig. 6.1a está provocado por resonancias donde únicamente los primeros armónicos son los propagados (ver Fig. 6.1b), por tanto tiene forma sinusoidal en el dominio temporal.

Por otra parte, al sonido 6.1c está provocado por vibraciones, tal y como se puede observar en los impulsos concatenados en el dominio temporal. Esto queda reflejado en el dominio frecuencial en una representación con una gran cantidad de armónicos (ver Fig. 6.1d).

Características como el nivel del primer armónico o fundamental, la relación entre su intensidad y las de los demás armónicos o simplemente la cantidad de armónicos que se pueden observar ofrecen la posibilidad de distinguir sobre el tipo de sonido sin más que visualizar el espectrograma. Además permiten conocer si el sonido ha sido realizado mediante una excitación de un elemento vibrante o simplemente mediante la resonancia del tubo sonoro. En la Tabla 6.1 se pueden observar las diferencias en los espectrogramas de los dos tipos de mecanismos de producción.

Se recuerda que aunque trabajar en el dominio frecuencial ofrece ventajas como la de poder visualizar y detectar diferentes patrones de forma satisfactoria, también posee

Características	Sin elemento vibrante	Con elemento vibrante
Nivel primer armónico (NPA)	El mayor	$\leq$ que resto armónicos
Nivel armónico principal / NPA	≈ 1	≥ 2
Número de armónicos	≈ 10	≈ 2

Tabla 6.1: Características frecuenciales del espectrograma de los sonidos según el modo de excitación de la columna de aire.

inconvenientes. La elección del parámetro LVA es uno de ellos, explicándose con anterioridad en esta tesis en la Sección 3.6.2 del Capítulo 3. Por tanto, será conveniente buscar otras alternativas para realizar la diferenciación entre sonidos resonantes y vibratorios.

El análisis de la modalidad de la señal a la hora de detectar algún parámetro o característica no es común en el ámbito del tratamiento y detección de audio. Sin embargo el estudio del determinismo a la hora de caracterizar la naturaleza de la producción del sonido, identificando la presencia de un elemento vibrante, es una técnica novedosa a tener en cuenta como alternativa a los algoritmos basados en el dominio frecuencial.

6.3. Algoritmos de detección de pitch más habituales

Los PDA son utilizados en varios contextos (fonética, codificación de la voz, etc) y por tanto poseen varias aplicaciones donde su importancia es considerable. Existen una gran variedad de algoritmos entre los que, dependiendo de la aplicación en concreto, se podrá elegir.

6.3.1. Métodos basados en el dominio temporal

Algunos de los PDA funcionan en el dominio temporal extrayendo características como la autocorrelación, la similitud entre la señal y ésta misma desplazada en el tiempo, etc. Entre todos ellos destacaremos el siguiente:

Zero-Crossing Rate (ZCR)

Uno de los algoritmos más simples consiste en medir la distancia entre los puntos que cruzan por cero de la señal. Este método no funciona bien con formas de onda complejas compuestas de múltiples ondas sinusoidales con diferentes periodos. Sí existen casos en los que el algoritmo ZCR puede ser una medida útil, por ejemplo en aplicaciones de audio donde se asume que existe una sola fuente. La simplicidad del algoritmo lo hace muy fácil de implementar.

6.3.2. Métodos basados en el dominio de la frecuencia

Otra familia de métodos para la detección de pitch operan en el dominio de la frecuencia, localizando picos en este dominio mediante la transformada de Fourier. En el

dominio de la frecuencia es posible la detección polifónica, utilizando el espectrograma. Esto requiere un mayor tiempo de procesado cuanto más exhaustivo queramos el resultado. Cabe decir que la eficiencia de la transformada de Fourier permite que actualmente el cálculo del espectrograma sea totalmente abordable para sonidos de duración media-corta.

Algoritmos basados en el dominio de la frecuencia como el *harmonic product spectrum* [72] que intenta emparejar las características del dominio frecuencial con mapas predefinidos (útil para la detección del pitch en instrumentos de notas fijas) y la detección de picos debidos a series armónicas son muy utilizados.

Para mejorar la estimación del pitch a partir del espectrograma, técnicas como el reasignamiento espectral (basado en la fase) o la interpolación de Grandke (basada en la magnitud) pueden ser útiles en caso de no requerir una precisión como la aportada por la transformada de Fourier. En el caso que atañe a este tesis doctoral, haremos hincapié en el algoritmo de *Subharmonic to harmonic ratio* ya que tiene un buen equilibrio entre complejidad y calidad en la separación/ detección del pitch.

Subharmonic to harmonic ratio (SHR)

En 2002, Sun propuso un algoritmo de detección de pitch [73] que consistía en analizar las modulaciones frecuenciales y de amplitud producidas en la voz humana debido a la vibración de las cuerdas vocales. Este efecto se manifiesta en el dominio de la frecuencia por la presencia de subarmónicos.

más en detalle, la magnitud de los subarmónicos respecto a los armónicos refleja el grado de modulación en un sonido, permitiendo la detección de un sonido provocado por una vibración (vibratorio) cuando esta diferencia es pequeña y de un sonido resonante cuanto es alta. Por tanto, el parámetro denominado *Subharmonic to harmonic Ratio* (SHR) se propuso para medir la diferencia entre los armónicos y los subarmónicos comparando su intensidad. Será calculado por tanto en el dominio frecuencial.

Además de determinar el pitch o frecuencia de vibración de un sonido, el SHR puede usarse como un parámetro de medida de calidad de voz. Es sabido que diferentes grados de irregularidad pueden provocar diferentes grados de sensación de aspereza en el sonido y esto refleja el estado del aparato fisiológico subyacente. Por lo tanto, es deseable tener una medida objetiva para cuantificar esta relación, la cual puede ser utilizada como un índice para clasificar el modo de producción de la voz de una persona o para comparar la calidad de la voz de diferentes sujetos.

En la actualidad el algoritmo en lugar de buscar un único pico, el cual representa la suma de los armónicos y los subarmónicos, trata de descomponer sus efectos y determinar si los subarmónicos son los suficientemente potentes como para ser considerados candidatos a pitch. Se describe en [73].

Se define como $A(f)$ cada trozo del espectrograma obtenido mediante la Transformada de Fourier con una LVA pequeña (128 muestras por ejemplo) y se obtiene la frecuencia fundamental f_0, siendo el parámetro SH la suma de la amplitud de los armónicos, tal y como muestra la Ec. (6.1):

$$SH = \sum_{n=1}^{N} A(nf_0), \tag{6.1}$$

donde N es el número máximo de armónicos considerados. Si se define la frecuencia de los subarmónicos como un medio de la frecuencia fundamental, se obtiene el parámetro SS sumando la amplitud de los subarmónicos (ver Ec. (6.2)):

$$SS = \sum_{n=1}^{N} A((n - 1/2)f_0) \tag{6.2}$$

Como consecuencia, el ratio SHR (ver Ec. (6.3)) se obtiene dividiendo SS entre SH:

$$SHR = \frac{SS}{SH} \tag{6.3}$$

Para su aplicación dentro del marco de esta tesis, el algoritmo de SHR ha sido modificado para poder calcular más armónicos de los habituales, es decir, se ha aumentado N, y se ha permitido la detección de f_0 mucho mayores que las utilizadas en voz humana.

6.4. La utilización de los *Recurrence plots* para la caracterización de la señal

Analizar la complejidad del sistema dinámico que produce una señal para realizar algoritmos de detección de pitch se puede lograr estudiando los diagramas del espacio de fase o sus representaciones en forma de *Recurrence Plots* (RP). El hecho de que los sonidos resonantes son más predecibles que los sonidos vibratorios es un buen punto de partida a la hora de intentar realizar este tipo de algoritmos mediante el determinismo. La aplicación de los *Recurrence Plots* y los *Recurrence Quantification Analysis* (RQA) en sonidos de ballenas beluga es algo novedoso y presenta la ventaja de la utilización del dominio temporal al no perder la información de fase. Esto es interesante dado que en el dominio temporal se refleja perfectamente el mecanismo físico subyacente de la naturaleza de producción de un sonido. Al final de esta sección se muestran algunas figuras y ejemplos que ilustran este comportamiento.

La caracterización de la modalidad de señal es un campo emergente e interdisciplinar que trata de abordar el problema de la detección de la presencia de los mecanismos subyacentes de generación no lineal en una señal dada. El estudio de estos fenómenos se ha evitado durante muchos años y, sin embargo, es una práctica común para modelar estos procesos utilizando modelos subóptimos pero matemáticamente manejables. Sin embargo, una detección y caracterización de la naturaleza no lineal y determinista de la señal adecuada pueden transmitir información importante en un gran número de situaciones, como por ejemplo en disfunciones de voz [74].

El reciente enfoque para la caracterización de la modalidad de la señal consiste en el uso de los *Recurrence Plots*, así como su análisis cuantitativo, llamado *Recurrence*

Quantification Analysis. Los RP han demostrado ser una valiosa herramienta de visualización de datos y análisis en el estudio de complejos sistemas dinámicos en un gran número de disciplinas tales como: biología, neurociencia, ingeniería, finanzas, ciencias de la tierra, etc. Recientemente, Miralles y Lara en [75] publicaron la posibilidad de ralizar una caracterización de la modalidad de la señal usando conceptos referentes a los RP. Este enfoque abre nuevos horizontes para lograr un mejor análisis de la modalidad de la señal y el desarrollo de nuevas pruebas de no linealidad en función de la RP y RQA.

La representación RP de una serie temporal $x(n)$ fue introducida por Zbilut et al. (1991). Varias modificaciones de los RP han sido propuestas, a continuación formularemos la más utilizada. Siendo $x(n)$ una serie temporal de duración N, es posible obtener el *Delay Vector* (DV) en el instante i para una *embedding dimension* m y un *time lag* τ como:

$$x(i) = [x(i), x(i+\tau), x(i+2\cdot\tau), ..., x(i+(m-1)\cdot\tau)] \tag{6.4}$$

La *Distance Plot* (DP) puede ser calculada tomando la norma de todas las posibles combinaciones de los DVs.

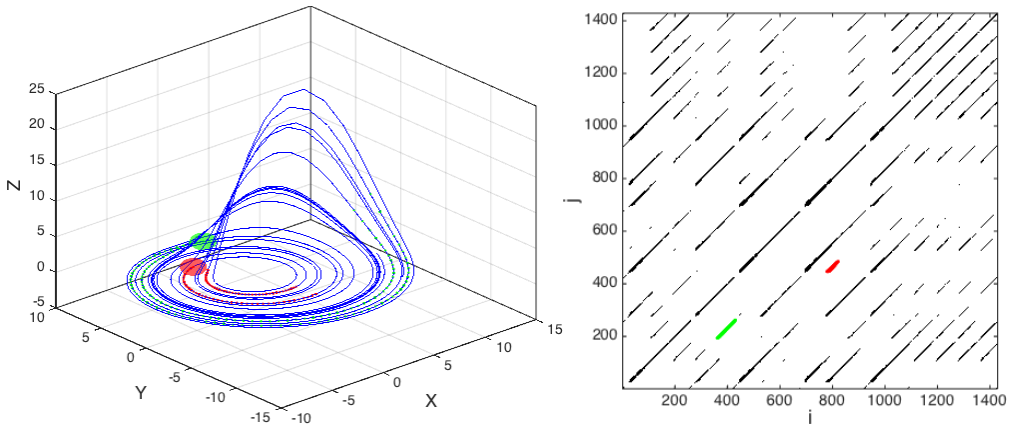
$$DP_{ij} = DP(i,j) = \|x(i) - x(j)\|, \qquad x(i) \in \mathcal{R}^m, i, j = 1, ..., N - m\tau \tag{6.5}$$

En ésta ocasión hemos usado la L2-norm como $\|\cdot\|$. Algunos autores se refieren a esta gráfica como *Global Recurrence Plot*. A partir de $DP(i,j)$ el RP puede ser calculado como:

$$RP(i,j) = \Theta(\epsilon - DP_{ij}) \tag{6.6}$$

donde $\Theta(\cdot)$ es la función *Heaviside Step* y ϵ es un umbral de recurrencia a fijar. Además de muchas otras ventajas, la representación RP convierte el diagrama del *phase space* de un espacio $\mathcal{R}^m$ hacia un espacio $\mathcal{R}^2$. Esto ofrece una manera más fácil de analizar sistemas complejos independientemente de la *embedding dimension*. La recurrencia de los estados, o las veces en las que la trayectoria del *phase space* se sitúa más o menos en la misma área es indicada como puntos negros en la representación RP. Como consecuencia de esto, el sistema dinámico subyacente puede ser analizado o caracterizado midiendo el número y duración de las recurrencias.

La Fig. 6.2a muestra un fragmento de trayectoria del *phase space* de un atractor de Rössler (en azul). La Fig. 6.2b muestra la representación RP del mismo fragmento. Los puntos negros indican que dos estados están dentro de una distancia/umbral ϵ (usando la norma nombrada anteriormente). Cuando dos trayectorias discurren paralelas una a otra durante un número determinado de estados provoca líneas diagonales en el RP. La región roja y verde de la Fig. 6.2 muestran esta idea. De este modo, estudiando la longitud y distribución de las diagonales se puede obtener información del sistema que lo generó.



(a) (b)

Figura 6.2: a) Segmento de la trayectoria del phase space del sistema de Rossler (para parámetros estándar $b = 0{,}2$) usando sus tres componentes y b) su RP correspondiente. Los puntos en la trayectoria en a) que discurren paralelos dentro de una distancia ϵ son mapeados como lineas diagonales en b).

6.4.1. Estructuras en los *Recurrence Plots* y *Recurrence Quantification Analysis*

El propósito inicial de los RPs es la inspección visual de las trayectorias multidimensionales del espacio de fases o *phase space*. La visualización mediante RPs nos aporta pistas sobre la evolución temporal de estas trayectorias. La ventaja de los RPs es que también se pueden aplicar a datos de menor tamaño e incluso no estacionarios.

Los RPs muestran características con patrones a gran y pequeña escala. Los patrones a gran escala pueden revelar información acerca de homogeneidad, periodicidad, etc. Los patrones a pequeña escala (la textura) están formados por puntos, lineas diagonales, además de lineas verticales y horizontales. La presencia de puntos es debida a estados raros, es decir a estados que no son persistentes en ningún momento o con grandes fluctuaciones. Sin embargo, la presencia de líneas diagonales corresponde cuando un segmento de la trayectoria del *phase space* discurre paralelo a otro segmento. La longitud de esta diagonal es determinada por la duración de la evolución de la trayectoria de estos segmentos. Finalmente, la presencia de líneas horizontales (o verticales) señalan un periodo de tiempo donde un estado no cambia o cambia muy despacio. Estas estructuras o patrones a pequeña escala son la base para el análisis cuantitativo de los RPs (RQA).

Entre todas las medidas de los RPs que componen el RQA, existe una de especial interés: el porcentaje de los puntos que forman líneas diagonales. Como ya se ha comentado anteriormente, la aparición de lineas diagonales implica una evolución similar de estados de la señal en momentos diferentes, lo cual puede indicar que el proceso es determinista. Si las lineas diagonales se muestran al lado de puntos independientes, el proceso puede ser caótico. Esto se puede cuantificar como el porcentaje de puntos recurrentes que forman líneas diagonales con el parámetro DET definido en la Ec. (6.7).

$$DET = \frac{\sum_{l=l_{min}}^{N} l \cdot P(l)}{\sum_{i,j}^{N} RP(i,j)} \tag{6.7}$$

donde $P(l)$ es el histograma de longitud l de las lineas diagonales y l_{min} es la longitud mínima por la cual una diagonal es considerada (típicamente $l_{min} = 2$).

Se propone la métrica RQA_{DET} como un promedio de esta cuantificación a la hora de obtener una medida que permita detectar el existencia o no de pitch en un sonido. A partir de un número fijo de trozos del sonido se obtiene DET_t, donde el subíndice t hace referencia a cada uno de los trozos analizados. A continuación se promedia este vector para conseguir la métrica propuesta (ver Ec. (6.8)):

$$RQA_{DET} = \frac{\sum_{t=1}^{Ntrozos} DET_t}{Ntrozos} \tag{6.8}$$

Este parámetro RQA_{DET} hace referencia al grado de determinismo, basado en el *phase space* de una señal.

6.4.2. Aplicación del análisis de la modalidad de la señal a la caracterización de vocalizaciones de mamíferos marinos

En las instalaciones del Oceanográfico de Valencia se grabaron sonidos producidos por las ballenas beluga en un experimento controlado y repetible. En concreto, 5 sonidos resonantes y 5 vibratorios fueron recogidos, es decir, 5 sonidos de cada tipo, con la intención de demostrar que la identificación de la manera en que se produce cada sonido puede realizarse a través de la caracterización de la modalidad de la señal mediante el estudio de la morfología de los RPs.

En las Fig. 6.3a y 6.3c se puede ver como el *phase space* de dos de los sonidos grabados son totalmente diferentes, lo cual, se refleja al obtener los RPs correspondientes (Fig. 6.3b y 6.3d). Los RPs han sido obtenidos usando un ϵ de 10 % del máximo diámetro del *phase space* [76].

La Tabla 6.2 muestra los resultados de la medida RQA_{DET} para los sonidos grabados, donde se puede observar la media y la desviación típica obtenidas para cada uno de los dos grupos de sonidos.

	Sonidos de beluga whale	
	Resonantes	Vibratorios
RQA_{DET}	0.95 ± 0.01	0.71 ± 0.11

Tabla 6.2: Resultados del parámetro del determinismo obtenido para sonidos resonantes y vibratorios.

Tal y como muestran los resultados de la Tabla 6.2 es posible diferenciar y clasificar los diferentes sonidos por su naturaleza de producción a través del determinismo basado en RPs. Se puede ver que la media del determinismo de los 5 sonidos resonantes es de

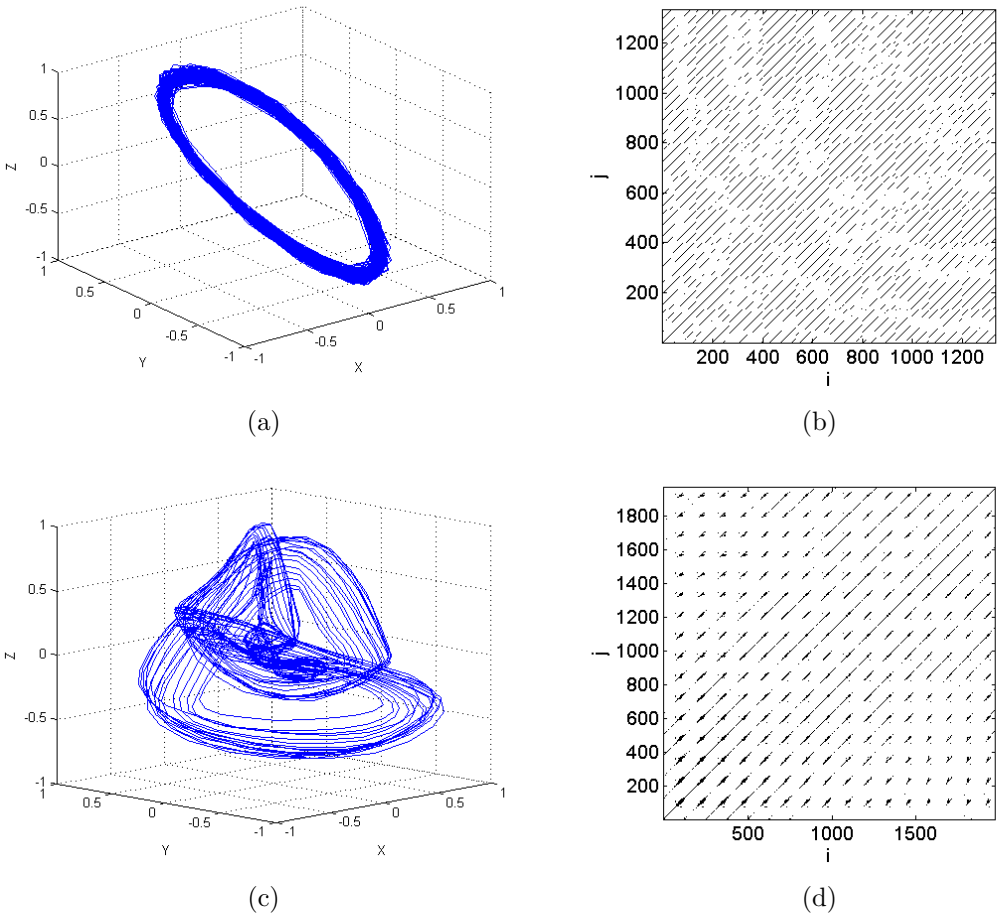


Figura 6.3: a) Sonido resonante en el *phase space* y b) visto como *Recurrence plot*. c) Sonido vibratorio en el *phase space* y d) visto como *Recurrence plot*.

0.95 con una desviación típica de 0.01, valores diferentes a los obtenidos para los sonidos vibratorios, con una media de 0.71 y una desviación típica de 0.11. Estos resultados muestran que la medida RQA_{DET} puede ser una herramienta interesante para distinguir entre los sonidos producidos mediante resonancias o vibraciones. Mediante un umbral en un rango de valores comprendidos entre [0.85-0.9] se podrá clasificar de forma satisfactoria entre los dos tipos de sonidos. En la sección siguiente, con el objetivo de profundizar y comparar con otros métodos, se elegirá el umbral óptimo, utilizando una base de datos de más de 150 sonidos.

6.5. Comparativa con otros métodos

Dado que la medida del determinismo a través de los *Recurrence Plots* puede ser utilizada para la caracterización de sonidos, en concreto para la identificación de sonidos

resonantes y vibratorios, se procederá a realizar una comparativa con el fin de poner en contexto la solución propuesta. En este caso se capturaron 168 sonidos durante 30 minutos de grabación producidos por las ballenas beluga del Oceanográfico. Supervisados por sus biólogos y científicos, se clasificaron entre vibratorios y resonantes considerando esta clasificación como base con la cual comprobar los resultados obtenidos por cada una de los métodos a comparar.

Para la comparación se seleccionaron los dos algoritmos más robustos utilizados a la hora de decidir si un sonido es sordo o sonoro, dejando a un lado los algoritmos que mejor identifican la frecuencia fundamental de un sonido. El primero de ellos realizado en el dominio temporal, en concreto se trata del algoritmo de *Zero-Crossing Rate*. El segundo, basado en en dominio frecuencial será el algoritmo de SHR. Ambos introducidos en la Sección 6.3 de este capítulo.

Como simplemente se trata de obtener un orden de magnitud para cada uno de los métodos, se trata de buscar el umbral óptimo en cada uno de ellos comprobando cuantos sonidos no son correctamente clasificados en función del total de sonidos de cada clase: los sonidos resonantes y sonidos vibratorios. En la Tablas 6.3, 6.4 y 6.5 se pueden ver los resultados obtenidos de estos dos métodos junto al propuesto en este capítulo. En cada una de ellas se obtiene la Tasa de Error (TE) en función de un umbral seleccionado, como:

$$TE = \frac{\text{nº de errores}}{\text{nº de señales de la clase}} \tag{6.9}$$

Están formadas por 4 columnas, la primera correspondiente a los valores del umbral, la segunda y tercera correspondientes a la TE de la clase resonante y la clase vibratorio respectivamente. En la cuarta se obtiene la TE total obtenida a partir de la suma de las dos anteriores.

Método ZCR

Umbral	TE Res. (%)	TE Vib. (%)	TE Tot. (%)
0.060	13.33	31.58	44.91
0.065	15.00	27.37	42.37
0.070	15.00	23.16	38.16
0.075	**15.00**	**18.95**	**33.95**
0.080	21.67	15.79	37.46
0.085	25.00	14.74	39.74
0.090	28.33	14.74	43.07
0.095	28.33	11.58	39.91
0.100	31.67	8.42	40.09
0.105	31.67	6.32	37.98

Tabla 6.3: Tasa de Error (TE) obtenida para el método ZCR en función del umbral.

En la Tabla 6.3 correspondiente al algoritmo ZCR se puede ver como las TE más bajas se obtienen con el umbral situado en el valor 0.075. Los valores del umbral en el

Método SHR

Umbral	TE Res. (%)	TE Vib. (%)	TE Tot. (%)
0.30	25.00	5.26	30.26
0.32	20.00	5.26	25.26
0.34	**16.67**	**5.26**	**22.93**
0.36	16.67	8.42	25.09
0.38	16.67	10.53	27.19
0.40	16.67	10.53	27.19
0.42	15.00	11.58	26.58
0.44	15.00	17.89	32.89
0.46	15.00	23.16	38.16
0.48	13.33	29.47	42.81

Tabla 6.4: Tasa de Error (TE) obtenida para el método SHR en función del umbral.

método de ZCR corresponden al número de cruces por cero dividido por el numero de total de muestras del sonido. La TE para la clase silbido se sitúa en 15.00 %, la TE para la clase vibratorio en un 18.95 % y la TE total con un valor de 33.95 %.

En el caso de la Tabla 6.4 correspondiente al algoritmo SHR se puede ver como las TE más bajas se obtienen con el umbral situado en el valor 0.34. Los valores del umbral corresponden a la métrica SHR. La TE para los sonidos resonantes se sitúa en 16.61 %, la TE para la clase vibratorio en un 5.26 % y la TE total con un valor de 22.93 %.

Método RQA_{DET}

Umbral	TE Res. (%)	TE Vib. (%)	TE Tot. (%)
0.78	0.00	51.58	51.58
0.80	1.67	48.42	50.09
0.82	3.33	41.05	44.39
0.84	5.00	33.68	38.68
0.86	10.00	21.05	31.05
0.88	**13.33**	**13.68**	**27.02**
0.90	25.00	7.37	32.37
0.92	36.67	7.37	44.04
0.94	46.67	5.26	51.93
0.96	75.00	1.05	76.05

Tabla 6.5: Tasa de Error (TE) obtenida para el método del determinismo en función del umbral.

Éste método, aunque con mejores resultados que el algoritmo ZCR, tiene un coste computacional mucho mayor, ya que trabajar en el dominio frecuencial siempre es más costoso computacionalmente que trabajar en el tiempo. Se observa como ventaja que a la hora de clasificar los sonidos vibratorios, la TE obtenida es muy baja (5.26 %) menor

todavía que la obtenida para los resonantes (16.67 %). Gracias a ello se obtiene una TE conjunta del 22.93 %.

En la Tabla 6.5 correspondiente al algoritmo propuesto con la medida RQA_{DET} se puede ver como las TE más bajas se obtienen con el umbral situado en el valor 0.88. Los valores del umbral corresponden a la medida RQA_{DET}. La TE para la clase silbido se sitúa en 13.33 %, la TE para la clase vibratorio en un 13.68 % y la TE total de 27.02 %

Además, en la Tabla 6.6 se puede comprobar como la media y la desviación típica de la medida RQA_{DET} para sonidos resonantes y para sonidos vibratorios se parece bastante a las obtenidas en la Sección 6.4.2 (ver Tabla 6.2).

	Sonidos de beluga whale	
	Resonantes	Vibratorios
RQA_{DET}	0.92 ± 0.05	0.73 ± 0.16

Tabla 6.6: Resumen del parámetro del determinismo obtenido para sonidos resonantes y vibratorios.

6.6. Conclusiones

A lo largo de este capítulo se comprueba como la caracterización de la señal a través del determinismo puede ser una herramienta útil para utilizar algoritmo de detección de pitch, así como una buena característica a la hora de realizar una clasificación de los sonidos. La utilización de métricas asociadas a la cuantificación de los *Recurrence Plots* en sonidos de ballenas belugas es algo novedoso, pero sin embargo son señales ideales para sacar todo el partido a este tipo de análisis.

Se puede decir que la medida propuesta RQA_{DET} permite la detección del pitch de un sonido aprovechando toda la información que nos permite obtener el dominio temporal, como por ejemplo, la fase de la señal, diferenciando entre los sonidos vibratorios y sonidos resonantes, adaptándose al modelo subyacente de la naturaleza de producción del sonido.

Una de las ventajas del algoritmo propuesto es la independencia que trabajar en el dominio del tiempo nos aporta sobre el dominio frecuencial, siempre más costoso computacionalmente y con mayores problemas asociados como el ocasionado por longitud de la ventana de análisis, para obtener con una buena resolución de la frecuencia fundamental del sonido y de sus armónicos.

A la hora de comentar las conclusiones de la comparativa se tendrá en cuenta que se van a comparar dos métodos que trabajan en el dominio temporal: el ZCR y el RQA_{DET} y uno que lo hace en el frecuencial como es el SHR. De hecho, el coste computacional del algoritmo en este último dominio siempre es bastante mayor que en el dominio temporal, propiedad importante si en un futuro se plantea la utilización del alguno de estos PDA en un clasificador que trabaje en tiempo real (ver Apéndice A).

Las diferencias entre los valores TE obtenidos por el método que utiliza el dominio temporal se inclinan a favor del método propuesto RQA_{DET} consiguiendo 27 % de TE

con el umbral óptimo, 6 puntos menos que los 34 % del método ZCR para su mejor umbral. Sin embargo, cuando comparamos la métrica RQA_{DET} con el valor obtenido por el método del dominio frecuencial SHR en su mejor caso, un 23 %, se tiene que es 4 puntos peor.

Cabe destacar que con el incremento del ruido de fondo en las grabaciones todos los algoritmos bajarían sus prestaciones de la misma manera. En el SHR, las amplitudes de los subarmónicos que se utilizan para la obtención del ratio podrán ser enmascaradas mediante el ruido de fondo y confundiendo al algoritmo al perder la información causada por la vibración de los sonidos vibratorios. En este caso se obtendrá el mismo ratio que en los sonidos resonantes. En el caso de los métodos temporales, trabajar en este dominio hace que cualquier cambio en la señal afecte de manera directa a las métricas, debiendo mover el umbral utilizado para discernir entre una clase u otra conforme se detecte el posible aumento del ruido de fondo.

Dada las ventajas e inconvenientes descritos en las conclusiones se ha decidido que la métrica RQA_{DET} se adapta perfectamente al problema a resolver y será interesante, por tanto, dotar al modelo de análisis / síntesis que se propondrá en el Capítulo 7 de esta métrica como PDA. El coste computacional menor, la clara asociación con la naturaleza de la producción del sonido, el problema de la LVA (en el caso de trabajar en frecuencia) y la necesidad que obtener características independientes unas de otras (al extraerlas de dominios diferentes) en un futuro clasificador, han sido considerados propiedades prioritarias.

El contenido mostrado en este capítulo ha generado la publicación de los siguientes artículos en revista:

- Ramon Miralles, Alicia Carrión, David Looney, Guillermo Lara, y Danilo Mandic. "Characterization of the complexity in short oscillating time series: An application to seismic airgun detonations". Journal of the Acoustical Society of America. Volumen 138. Número 3. Páginas 1595-1603. Año 2015.

- Alicia Carrión, Guillermo Lara, Ramón Miralles, Jorge Gosalvez e Ignacio Bosch. "On the use of Recurrence Quantification Analysis for Signal Modality Characterization: two applications". Waves. Volumen 7. Páginas 5-14. Año 2015.

y las participaciones en los siguientes congresos:

- Alicia Carrión Garcia, Ramón Miralles y Guillermo Lara Martínez. "Scattering materials characterization based on Recurrence Plots Quantification Analysis (RQA)". Sixth International Symposium on Recurrence Plots. Grenoble. Año 2015.

- Ramón Miralles Ricós, Alicia Carrión García y Guillermo Lara Martínez. "Computing the Delay Vector Variance using the Recurrence Plots". Sixth International Symposium on Recurrence Plots. Grenoble. Año 2015.

Capítulo 7

Modelado de señales bioacústicas de ballenas beluga

7.1. Introducción

Una vez estudiada la naturaleza de los sonidos de los odontocetos y la fisionomía de su aparato fonador, se plantea el reto de implementar un modelo simplificado de este, debiéndose diseñar para mantener un equilibrio entre reproducir de la forma más fiel algunas de las características del aparato de producción de sonido de estos animales y facilitar el análisis desde el punto de vista de procesado de señal. Esta es una aproximación similar a la que se realiza en los seres humanos donde se extraen características adaptadas al análisis, síntesis y codificación de voz [77–79].

Los trabajos de Parmentier en [80] son ejemplos de como, en otras especies, la generación de los sonidos está relacionada con la fisionomía de sus aparatos fonadores y de como su estudio en conjunto ayuda a progresar en el conocimiento. En concreto, se compararon las estructuras involucradas en la comunicación acústica en tres especies de peces anguila los cuales producen sonidos a través de la vejiga natatoria relacionando los sonidos a la acción de los músculos de vejiga natatoria que los provocan.

Otro ejemplo se puede leer en trabajos de Thorpe, Warner y Gant [81–83] donde se muestran detalles sobre el mecanismo de producción de sonido en las aves. Thorpe en [81] describió el aparato vocal de las aves y su comparación con el aparato vocal humano. Tanto la vocalización de aves y voz humana se producen por la vibración de membranas durante la fase de exhalación de la respiración. Warner en [82] comparó la anatomía del órgano vocal de las aves (siringe) en pájaros cantores. Su principal conclusión fue que las únicas áreas vibrantes, por lo tanto, las únicas fuentes de sonido son dos membranas ubicadas una en cada bronquio, las pueden hacer vibrar independientemente unas de otras, lo que produce dos tonos armónicamente no relacionados al mismo tiempo.

La mayoría de los murciélagos producen señales de ecolocalización en su laringe, que van de 20 a 200 kHz. Mientras que los murciélagos de herradura y nariz de hoja emiten estos sonidos a través de la nariz, por una estructura en forma de hoja carnosa, algunas especies realizan el sonido con vibraciones en su lengua. Skinner en [84, 85] investigó esta

estructura revelando que su anchura está relacionada con la longitud de onda de la señal.

Sissom en [86] investigó el ronroneo en gatos. La frecuencia fundamental del ronroneo, sobre todo en los gatos domésticos (alrededor de 25 Hz) sigue mecanismos alternativos de producción de sonido de la vibración de las cuerdas vocales introducida por el flujo respiratorio. Esto ha sido sugerido también por Remmers en [87], donde mostró que el ronroneo es producido por contracciones de músculos de la laringe que modulan el flujo de aire respiratorio que pasa a través de las cuerdas vocales.

Marler en [88] trabajó en la asociación entre modelos análisis / síntesis y los órganos encargados de la producción de sonidos. Se demostró cómo la teoría de Sistemas Lineales Invariantes $y(n) = x(n) * h(n)$, donde una fuente y un filtro adecuados son capaces de modelar cualquier tipo de sonido, se podía aplicar a las aves con el fin de explicar la evolución de la elongación tráquea. Más de 60 especies de aves poseen una tráquea alargada. Las frecuencias formantes dependen de la longitud del tracto vocal y forma, con las frecuencias más bajas que indica tractos vocales más largos. Los formantes son por lo tanto un buen indicador de tamaño del cuerpo en muchas especies. Sin embargo, algunas especies poseen ya sea una laringe móvil que pueden retraer para alargar el tracto vocal durante vocalizaciones. Además se sugiere que estas características evolucionaron a través de la selección natural, es decir a medida de que el tamaño del cuerpo aumenta, los animales pueden producir vocalizaciones con formantes con una frecuencia menor.

Diversos autores, entre ellos Castellote, comenzaron a estudiar características relacionadas con la producción del sonido en mamíferos marinos [89–92]. Características como la frecuencia fundamental, la cantidad de armónicos, la duración, la intensidad, el orden o la gramática de los sonidos concatenados se utilizaban a la hora de caracterizar y clasificar distintos sonidos que seres humanos y mamíferos marinos son capaces de realizar. Es decir, comenzaron a estudiar características relacionadas con los sistemas físicos reales de producción de sonidos a la hora de diseñar un modelo de análisis / síntesis.

Además, Cranford y Thomas [93, 94] intentaron esclarecer como son producidos los sonidos, canciones o vocalizaciones que realizan los cetáceos. Siguiendo en la línea de estos últimos trabajos y al igual que se ha realizado con cada uno de los animales anteriormente nombrados, ha sido necesario realizar un estudio minucioso sobre los órganos, fisiología y funcionamiento del sistema de producción de los mamíferos marinos, en nuestro caso, de los odontocetos.

En el Capítulo 3 de esta tesis se estudió la naturaleza que poseen sus diferentes sonidos producidos, haciendo hincapié en la generación del sonido dentro del aparato fonador de la ballena beluga. Además, se propuso una clasificación de los sonidos conforme a dicha naturaleza con el objetivo de aportar nuevos enfoques sobre la producción del sonido. Es importante resaltar que los odontocetos son capaces de crear señales mixtas, es decir, generar dos señales vibrantes a la vez, una por cada una de las dos estructuras *Monkey-Lips-Dorsal-Bursa* (MLDB) existentes en su sistema de producción de sonido.

Tal y como Madsen explica en [19], cada uno de los dos tipos de los sonidos mostrados en la Tabla 7.1 está especializado en un rango de frecuencias diferente y complementario. Se puede observar, por ejemplo, que el número de armónicos en los sonidos resonantes es menor que en los sonidos vibratorios. Esto es inherente a la naturaleza de la vibración,

es decir, cuando se produce una vibración se generan muchos más armónicos que cuando no la hay. Los armónicos se propagan por los tubos del sistema de producción de los odontocetos, que actuan a modo de filtro.

Características	**Señal vibratoria f_0 baja**	**Señal vibratoria f_0 alta**	**Señal resonante**
MLDB utilizado	Izquierda	Derecha	Derecha
Rango Freq. fund.	[0.5 - 700 Hz]	[700 - 5000 Hz]	[700 - 5000 Hz]
Vibración Labios fónicos	Si	Si	No
Nº armónicos	Muchos	Muchos	Pocos
Intensidad señal	alta	baja	baja
Propósito	ecolocalización o comunicación	comunicación	comunicación
Espectrograma	vertical	horizontal	horizontal
Tipo de excitación	Tren de deltas	Tren de deltas	Ruido blanco

Tabla 7.1: Propiedades generales de los sonidos de los odontocetos, concretamente en las ballenas beluga según en la rama MLDB donde son producidos.

A la vista de todo esto, se propone emplear una estructura de producción de sonido basada en el modelo LPC-Vocoder (LPC-V) [95, 96], el cual dará cabida a todos los sonidos que los odontocetos son capaces de generar. Tanto el modelo referencia como el modelo propuesto se emplearán para analizar varios sonidos producidos por las ballenas beluga que posteriormente serán sintetizados. Estos sonidos generados serán comparados con el sonido original. En la Sección 7.5 se mostrarán los resultados de la comparativa y en la Sección 7.6 se obtendrán las conclusiones.

7.2. LPC-Vocoder en voz humana

En el modelado y síntesis de la voz humana se utilizan soluciones que se relacionan claramente con los órganos encargados de la producción del sonido. Por ejemplo, los modelos basados en los *Linear Prediction Coefficients* (LPC) [96] consisten en un modelo típico de sistemas lineales invariantes tal que $x(n) * h(n) = y(n)$ y es comúnmente denominado LPC-Vocoder.

Dada la cantidad de variables que se van a presentar en el presente capítulo, en la Tabla 7.2 se realiza un resumen de los parámetros utilizados en el modelo básico LPC-V para que el lector se familiarice con la notación y acrónimos empleados. En concreto se introducen los conceptos del Orden del Filtro (N) y la Longitud de Ventana de Análisis (LVA), parámetros de los cuales dependerán los sonidos sintéticos que generen los modelos.

Modelo	**Variables**	**Acrónimo**
LPC-V	Orden Filtro LPC	N_{LPCV}
LPC-V	Longitud Ventana de Análisis	LVA_{LPCV}

Tabla 7.2: Parámetros del modelo LPC-V y sus acrónimos.

Al analizar el sonido $y(n)$, se obtienen los LPC correspondientes al filtro $h(n)$ y la excitación $x(n)$, los cuales son codificados para ser utilizados posteriormente en la etapa de síntesis como $x'(n)$ y $h'(n)$ y así generar un sonido parecido al original (ver Fig. 7.1). Este modelo es capaz de sintetizar de forma aproximada cualquier sonido producido por el ser humano a partir de una excitación, y un filtro que la coloree en frecuencia. Se modela $h(n)$ como un filtro autoregresivo (AR) asociado con el tracto vocal donde los polos del filtro cambian según el ser humano mueve la boca para articular sonidos.

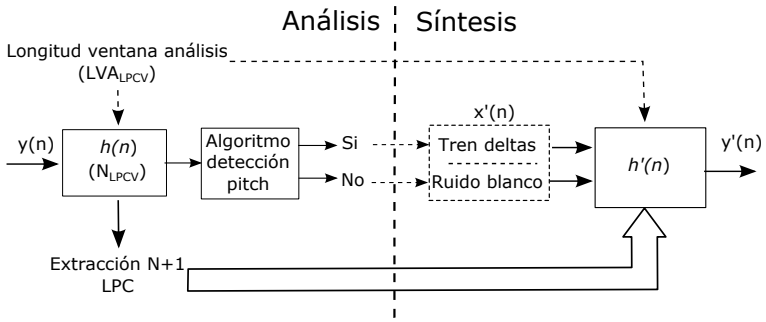
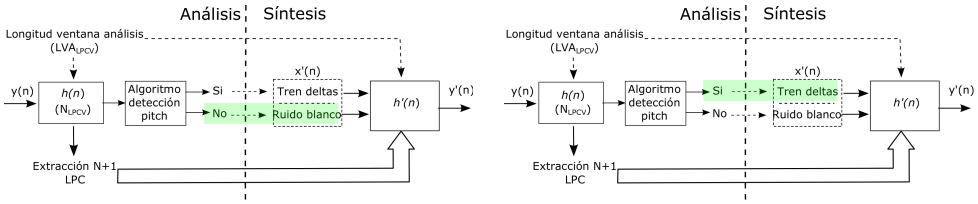


Figura 7.1: Modelo básico de análisis / síntesis utilizado habitualmente en voz humana: LPC Vocoder (LPC-V).

En el modelo LPC-V es importante la elección del tipo de excitación ($x'(n)$) se utilizará en la etapa de síntesis posterior. Para ello, el algoritmo detección de pitch realiza la detección del pitch de la señal. Según su resultado, se elegirá entre dos tipos de excitación. Cuando el sonido es producido por un elemento vibrante se elegirá un tren de deltas como forma de excitación y cuando el sonido se forma a través de resonancias sin necesidad de vibración, $x'(n)$ será ruido blanco. En el Capítulo 5 se estudiaron varios algoritmos para realizar esta tarea y se propuso la variante referente al determinismo que es la que emplearemos en el resto de la tesis.

En la Fig. 7.2a se observa como el algoritmo detección de pitch no detecta pitch en los los sonidos resonantes y por tanto selecciona ruido blanco como excitación en la parte de síntesis $x'(n)$. Sin embargo, con los sonidos vibratorios, el análisis de detección de pitch es positivo y por tanto, es necesaria la información del pitch a la hora proceder a la síntesis del sonido. Para ello se introduce un tren de deltas como excitación (Fig. 7.2b). Por tanto, en la etapa de síntesis, a partir de los LPC de $h'(n)$ y la excitación $x'(n)$ elegida, se podrán modelar de manera correcta estos dos tipos de sonidos.

Sin embargo el LPC-V no se adapta a la estructura de producción de sonidos que poseen los odontocetos. Estos poseen un par de órganos que pueden funcionar de forma simultanea, con capacidad para generar sonidos de diversa naturaleza y diferentes frecuencias. Será necesario diseñar un modelo adaptado a la fisionomía del sistema de producción de sonidos de los odontocetos capaz caracterizar todos los sonidos que estos animales son capaces de producir.



(a) Señal resonante (b) Señal vibratoria

Figura 7.2: Funcionamiento del LPC-V según las señales a tratar.

7.3. Modelo de producción de sonido propuesto: Doble Excitación LPC Vocoder

Partiendo de este modelo básico, algunos autores siguieron investigando en generar modelos más avanzados que permitan codificar de una manera más eficiente y con mejor calidad de reproducción [79]. Estos modelos tienen el inconveniente de perder la clara asociación entre parámetros del LPC-V y los órganos encargados de la producción del sonido. Se propone en este capítulo seguir empleando un modelo basado en el LPC-V, donde no se pierda esa relación con la fisionomía del animal. Duplicando la estructura de una rama del LPC-V vista en la Fig. 7.1, se propone un modelo de Doble Excitación LPC-Vocoder (DELPC-V) como el de la Fig. 7.3 que sintetice los sonidos que los odoncetos son capaces de realizar. El modelo tiene una asociación clara con su aparato fonador, formado por dos estructuras MLDB donde cada una de las dos fuentes se generan, se propagan y se suman.

Al igual que en la sección anterior, en la Tabla 7.3 se realiza un resumen de los parámetros utilizados en el modelo propuesto DELPC-V, de los cuales dependerán los sonidos sintéticos que generen los modelos.

Modelo	**Variables**	**Acrónimo**
DELPC-V	Orden Filtro LPC Rama Rápida	$N_{\text{RR-DELPCV}}$
DELPC-V	Orden Filtro LPC Rama Lenta	$N_{\text{RL-DELPCV}}$
DELPC-V	Longitud Ventana de Análisis Rama Rápida	$LVA_{\text{RR-DELPCV}}$
DELPC-V	Longitud Ventana de Análisis Rama Lenta	$LVA_{\text{RL-DELPCV}}$

Tabla 7.3: Parámetros del modelo DELPC-V y sus acrónimos.

En el modelo propuesto, siguiendo con la línea de investigación de [19], cada una de las ramas está especializada en la generación de sonidos en rangos de frecuencia fundamental complementarios como muestra la Tabla 7.1. La rama del DELPC-V que se encarga de las señales con una frecuencia fundamental más baja se denomina rama lenta, por el contrario, la rama encargada de las señales vibratorias rápidas o con una frecuencia fundamental más alta se denomina rama rápida. Además, al igual que el modelo LPC-V, el nuevo modelo puede adaptarse a sonidos sean vibratorios o sean resonantes.

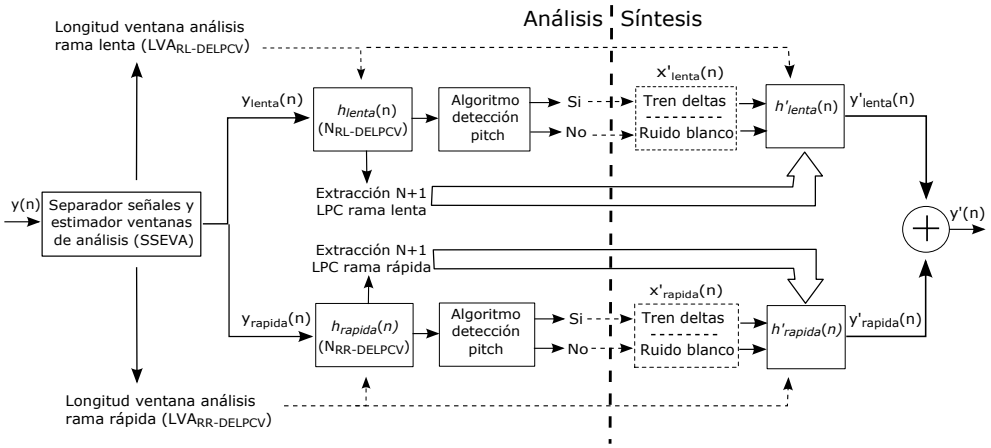


Figura 7.3: Modelo de análisis / síntesis propuesto por los autores: *Doble Excitación LPC Vocoder*.

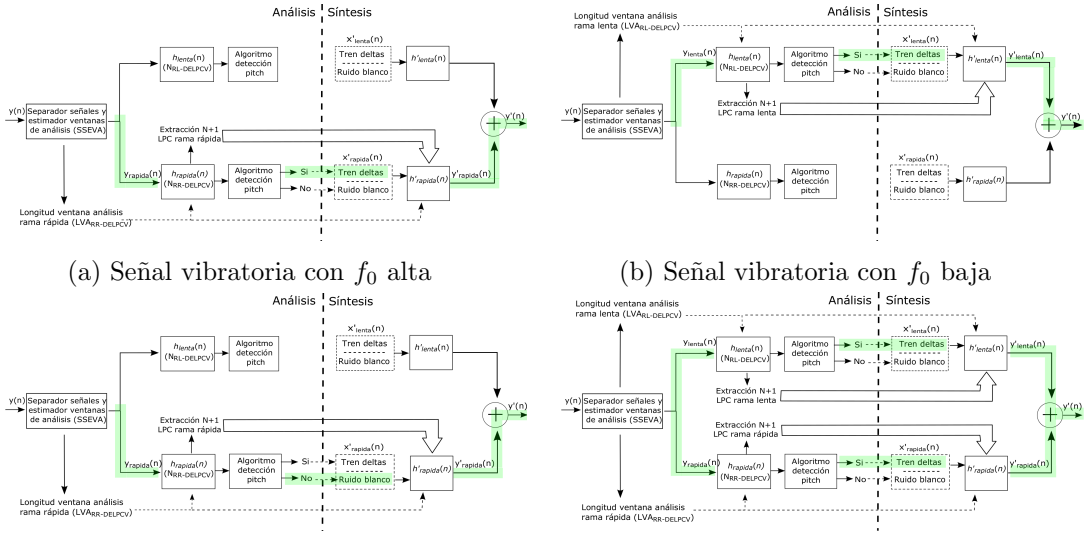
En la parte de análisis, una vez separados los sonidos por el algoritmo Separador de Señales y Estimador de longitud de la Ventana de Análisis (SSEVA) explicado en el Capítulo 5, se extraen los LPC y se elige la excitación de cada una de las señales. Ya en la etapa de síntesis, a partir de las características obtenidas, el modelo será capaz de generar dos señales sintéticas y'_{lenta} e y'_{rapida}, las cuales se promediarán para obtener la señal $y'(n)$.

En la Fig. 7.4 se puede observar como el modelo DELPC-V puede adaptarse tanto señales simples (resonantes o vibratorias) como a señales mixtas. En las señales vibratorias simples (Fig. 7.4a y Fig. 7.4b), el algoritmo SSEVA permite encaminar la señal por la rama adecuada según la frecuencia fundamental que posea, consiguiendo así el espectrograma con componentes horizontales necesario para la extracción correcta de los LPC.

En el caso de que la señal simple a analizar sea resonante (Fig. 7.4c), dado que las frecuencias fundamentales de esta señal están en el rango de la rama rápida del modelo, el algoritmo SSEVA encamina la señal por dicha rama, el bloque *Algoritmo detección pitch* analiza que no existe pitch en esta señal y en la parte de síntesis del modelo se elige ruido blanco como excitación x'_{rapida}. Por último, en las señales mixtas (Fig. 7.4d), el algoritmo SSEVA es necesario para separar cada una de las dos señales vibratorias que las componen por la rama correspondiente del DELPC-V y así poder extraer los LPC de manera correcta, gracias a la elección de la LVA para cada una de las ramas.

7.4. Métodos para el análisis y comparativa entre los modelos LPC-V y DELPC-V

Una vez explicados los elementos y consideraciones más importantes del DELPC-V, se evaluará el parecido entre una serie de sonidos producidos por las ballenas belugas



(a) Señal vibratoria con f_0 alta

(b) Señal vibratoria con f_0 baja

(c) Señal resonante

(d) Señal mixta

Figura 7.4: Funcionamiento del DELPC-V según las señales a tratar.

del Oceanográfico de Valencia y las señales sintéticas obtenidas por los modelos LPC-V y DELPC-V.

Existe una problemática importante a la hora de medir la calidad de los sonidos generados ya que no se trabaja con voz humana. En numerosos trabajos se han realizado medidas de calidad de sonidos sintéticos para voz humana, por ejemplo en aplicaciones *Text-To-Speech* (TTS), donde no existe una referencia con la cual comparar el sonido generado, o para medir la calidad de un enlace de telecomunicaciones. Se distinguen dos maneras diferentes a la hora de evaluar la calidad un sonido sintético.

La primera consiste en métodos de evaluación subjetivos, obteniendo una medida de opinión media o *Mean Opinion Score* (MOS) [97]. La segunda usa evaluaciones objetivas, comúnmente basadas en modelado perceptual. Generalmente, las medidas objetivas necesitan algún sonido de referencia para poder ser comparado y tenerlo no es un problema trivial. Sin embargo, las medidas objetivas han evolucionado en los últimos años y los resultados obtenidos muestran una alta correlación con los obtenidos por los métodos subjetivos. Además, han aparecido nuevas métricas objetivas no intrusivas que no necesitan una referencia acústica, es decir, el sonido original. Es el caso de la evaluación de la calidad vocal por percepción o *Perceptual evaluation of speech quality* (PESQ).

7.4.1. *Perceptual evaluation of speech quality* (PESQ)

Dependiendo de la información disponible, los algoritmos que miden la calidad de la voz pueden ser divididos en dos categorías principales:

I. Los algoritmos referenciados, los cuales hacen uso de la señal de referencia para realizar la comparación. Estos puede comparar cada muestra de la señal de referen-

cia con las correspondientes de las señales sintéticas. Las medidas proporcionadas por estos algoritmos nos dan una buena exactitud y repetibilidad pero sólo pueden ser aplicados a pruebas con unas condiciones muy controladas y definidas.

II. Los algoritmos no referenciados. Estos no necesitan el sonido de referencia para estimar la calidad de la señal sintética. Dichos algoritmos, como el P.563 [98] no proporcionan medidas con una gran exactitud pero proporcionan características sobre la fuente de emisión del sonido.

PESQ [99] es un algoritmo referenciado que analiza la señal de audio muestra a muestra después de un alineamiento temporal entre el sonido referencia y el sintético. PESQ puede aplicarse para proporcionar una medida de calidad entre los sonidos originales del principio de una red de comunicación y los resultantes a la salida, así como para caracterizar sus componentes. Los resultados del algoritmo modelan la puntuación de la opinión media o *mean opinion score* (MOS), cuya escala va de 1 (malo) hasta 5 (excelente).

Este método es el resultado de varios años de trabajos de desarrollo y es también aplicable los codificadores vocales. Los sistemas reales pueden incluir filtrado y retardo variable, así como distorsiones debidas a errores de canal y a codificadores de baja velocidad binaria. El método de la medida de la calidad vocal por percepción o *Perceptual Speech Quality Measure* (PSQM), descrito en el UIT-T P.861, sólo se recomienda para uso en la evaluación de codificadores vocales debido a no poder tener en cuenta efectos como el filtrado, el retardo variable y las distorsiones cortas localizadas. El método PESQ trata estos efectos mediante la ecualización de la función de transferencia, la alineación de tiempo y un nuevo algoritmo para promediar distorsiones en función del tiempo. Se destaca que existe una versión del método adaptado a un ancho de banda mayor (50-7000 Hz) el cual permite la comparación de sonidos no centrados en la voz humana (300-3400 Hz). Este mayor ancho de banda se quedará igualmente corto con sonidos muestreados a 96 kHz.

En caso de esta tesis doctoral, conseguir que varios expertos en acústica submarina obtengan unas medidas subjetivas es una labor complicada, dado los pocos expertos que existen y la dificultad a la hora de poder reunirlos para realizar la sesión. Se debe tener en cuenta que el oído humano esta adaptado a las frecuencias que la voz es capaz de realizar, pero no a rangos de frecuencias de los sonidos de los odontocetos, que son mayores. Estos expertos, en una mayoría de ocasiones acaban mirando el espectrograma a la hora de decidir si un sonido se parece a otro.

Con todo ello se descarta una métrica subjetiva para la evaluación de las sonidos generados por los dos modelos. Además y dado que los métodos objetivos de análisis perceptual realizados en el dominio temporal sólo llegan a medir un rango de frecuencias entre 50-7000 Hz, sería necesario una medida subjetiva para que pudiera ser evaluado su validez en frecuencias con una ancho de banda mayor, es decir, desde 50 Hz hasta 96 kHz. Otros métodos de comparación objetivos de la calidad del sonido han sido propuestos en [100, 101], pero no se adaptan a la necesidad de esta tesis ya que están centrados en medir anomalías o disfunciones de la voz humana.

Dado que el método de comparación PESQ no está optimizado para comparar señales con un ancho de banda mayor que sobrepasen a 7 kHz como frecuencia máxima, las métricas obtenidas serán válidas hasta esa frecuencia, dejando fuera de la comparación toda la información frecuencial que queda por encima. Por todas estas razones, se ha creído conveniente realizar la comparación basándonos en la similitud de los espectrogramas, tratando a estos como una imagen. El método empleado será el *Structural SIMilarity*, explicado en [102].

7.4.2. *Structural Similarity* (SSIM)

Este método es utilizado a la hora de comparar imágenes de una manera objetiva, teniendo en cuenta el contraste, la energía, y la estructura, características con las cuales los seres humanos evalúan la calidad de una imagen, siempre teniendo como referencia la imagen original.

La diferencia respecto a otras técnicas como MSE o PSNR, las cuales estiman el error absoluto, es que SSIM es un modelo basado en la percepción, que considera la degradación de la imagen como un cambio percibido en su información estructural, además de incorporar fenómenos perceptibles importantes, incluido el enmascaramiento de la luminancia y del contraste.

Se define la información estructural como la dependencia de los píxeles de una imagen cuando están espacialmente cerca. Estas dependencias llevan información importante sobre la estructura de los objetos en la escena visual. El enmascaramiento de la luminancia es un fenómeno por el cual la imagen se distorsiona tendiendo a ser menos visible en regiones brillantes mientras que el enmascaramiento por contraste es el fenómeno donde las distorsiones se vuelven menos visibles donde no hay una actividad significativa o textura en la imagen.

Este método ha sido utilizado previamente en varias ocasiones para un cometido parecido al que concierne esta tesis. Kandadai en [103], Cooper en [104] y Perovic en [105] consideraron que el SSIM, originalmente desarrollado como un método objetivo para medir la calidad de una imagen, podía ser utilizado para medir la calidad de un audio.

En particular, Kandadai estudió dos realizaciones diferentes del índice SSIM: la primera aplicada a cortos instantes de tiempo de una secuencia de audio mientras que la segunda obtuvo el espectrograma para después realizar comparaciones con el SSIM en dicho. Además se evaluó la exactitud de las dos realizaciones en base a otras métricas de calidad de audio objetivas ya contrastadas con evaluaciones subjetivas, obteniendo resultados coherentes.

7.5. Análisis

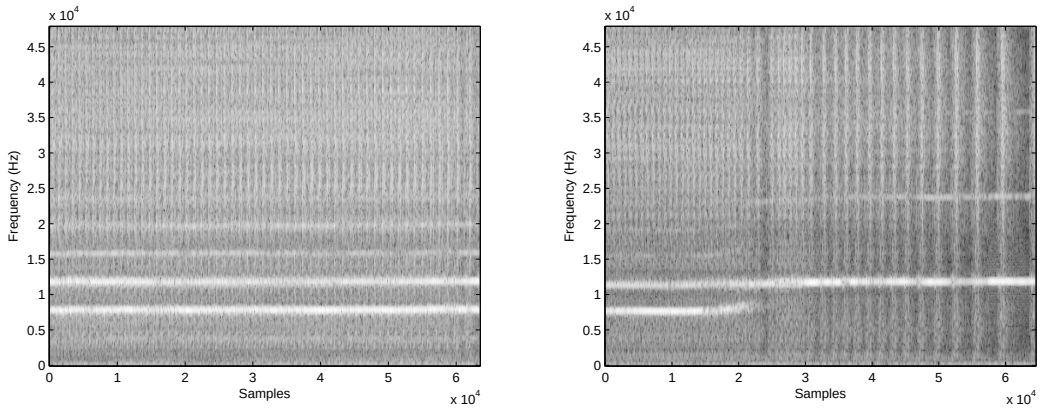
A la vista de todo lo expuesto anteriormente y con el objeto de comparar los sonidos sintetizados por los modelos con los sonidos originales se empleará el método SSIM. Dicho método proporcionará una medida objetiva sobre la imagen del espectrograma que nos permitirá evaluar la calidad de los sonidos sintéticos generados por los modelos

LPC-V y DELPC-V.

7.5.1. Señales mixtas

Se elige comenzar con el estudio de las señales mixtas, las cuales utilizan las dos ramas del modelo DELPC-V, cuyos resultados resumirán si es realmente conveniente la existencia de estas dos ramas en paralelo. Para ello se analizarán y sintetizarán 54 sonidos mixtos producidos por las ballenas beluga del Oceanográfico de Valencia, dos de los cuales corresponden a los espectrogramas de la Fig. 7.5. Todos los sonidos son de la misma duración 2.046 segundos, grabados con una frecuencia de muestreo de 96000 Hz (196416 muestras). Se utilizará un solape del 25 % a la hora de analizar las señales en el diagrama tiempo-frecuencia.

Se recuerda, tal y como se explicó en el Capítulo 5, que el algoritmo SSEVA calcula, además de la separación de los dos sonidos que componen el sonido mixto, las dos LVAs a utilizar en cada una de las ramas del DELPC-V. En una mayoría de las señales se obtiene $\mathrm{LVA_{RR\text{-}DELPCV}} = 128$ y $\mathrm{LVA_{RL\text{-}DELPCV}} = 4096$. Las Fig. 7.5 muestran los espectrogramas de dos de las señales mixtas a analizar y sintetizar con los modelos LPC-V y DELPC-V.



(a) (b)

Figura 7.5: Dos sonidos mixtos reales realizados por una de las ballenas belugas del Oceanográfico de Valencia.

Las evaluaciones se realizan generando un par de sonidos sintéticos (uno por modelo) para cada orden de filtro, es decir se obtiene un sonido para cada $\mathrm{N_{LPCV}}$ utilizado en el modelo LPC-V y otro sonido para cada $\mathrm{N_{RR\text{-}DELPCV}}$ utilizado en el modelo DELPC-V. Cabe destacar que para el modelo DELPC-V se ha fijado $\mathrm{N_{RL\text{-}DELPCV}} = 10$, siendo éste un valor empírico que puede codificar perfectamente la señal de la rama lenta. Estos sonidos generados se comparan con el sonido original. Se obtienen por tanto dos curvas donde se puede apreciar como evoluciona la similitud entre los sonidos sintéticos generados y el sonido original, donde 0 es nada similiar y 1 es totalmente similar, conforme

se aumenta N_{LPCV} y $N_{RR\text{-}DELPCV}$. Es necesario tener en cuenta que a mayor N, menor codificación y mayor similitud entre la señal original y la señal sintética.

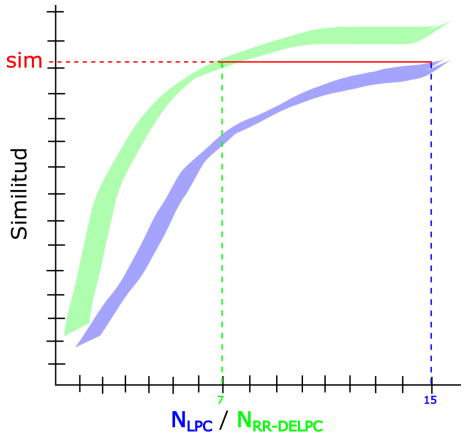


Figura 7.6: Ejemplo de la obtención de la métrica Diferencia de Orden (DO).

A la hora de compararlos, tal y como se muestra en 7.6, se obtiene la curva azul cuando se generan los sonidos en el modelo LPC-V aumentando N_{LPCV} (eje x) y la curva verde cuando se generan los sonidos en el modelo DELPC-V aumentando $N_{RR\text{-}DELPCV}$ (eje x). Dado que la longitud de la ventana de análisis es la misma en la rama rápida del modelo DELPC-V y en la rama única del modelo LPC-V, posible comparar el par de sonidos generados para cada N.

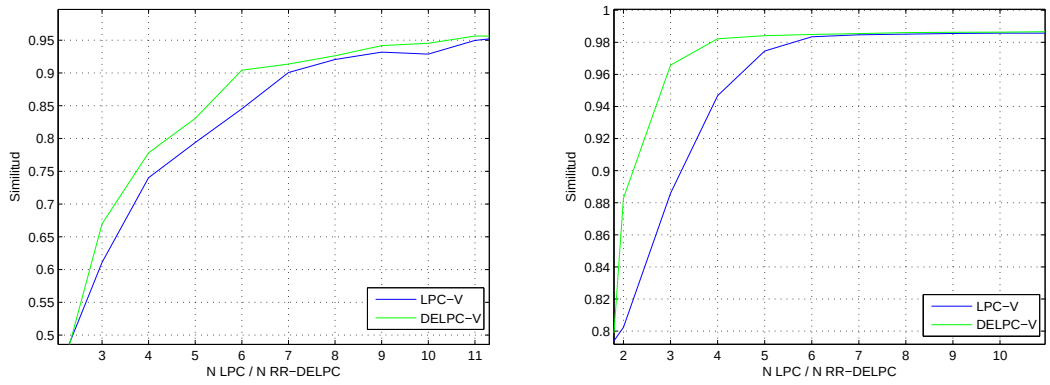
En la Fig. 7.6 se puede observar un ejemplo sobre la obtención de la métrica que denominaremos Diferencia de Orden (DO), la cual refleja de manera simple la reducción de la codificación del modelo DELPC-V respecto al modelo LPC-V. En definitiva, refleja como el modelo DELPC-V consigue generar una señal sintética más parecida a un sonido original utilizando filtros con un orden menor en contra del modelo LPC-V, el cual necesita filtros con un orden mayor para poder llegar a generar esa misma similitud.

La métrica DO consiste en encontrar para cada $N_{RR\text{-}DELPCV}$ el N_{LPCV} que iguale el valor de similitud (sim) obtenido ($N_{LPCV}{}_{sim} - N_{RR\text{-}DELPCV}{}_{sim}$). Se calcula la diferencia entre estos dos valores, obteniendo un vector de diferencias de 15 componentes, los mismos que número de $N_{RR\text{-}DELPCV}$ a evaluar, es decir, todo el eje x. Se toma el máximo de este vector, consiguiendo una medida que muestra cuantitativamente la diferencia de orden para codificar una señal mediante un modelo u otro.

Concretamente en el ejemplo de la Fig 7.6 el máximo DO se obtiene en $N_{RR\text{-}DELPCV} = 7$ (línea verde discontinua), dado que la similitud lograda (sim) para este orden (línea roja con curva verde) es igualada en la curva azul con $N_{LPCV} = 15$ (línea azul discontinua). Por tanto $DO = 8$, valor que refleja que con el modelo propuesto (DELPC-V) se consigue una similitud determinada con valor= sim equivalente a la obtenida con el modelo LPC-V habitual con un filtro de 8 coeficientes LPC menos.

En la Fig. 7.7 se muestran varios ejemplos reales sobre como evoluciona la similitud respecto al espectrograma de la señal original a medida que aumenta N_{LPCV} o

$N_{\text{RR-DELPCV}}$. Las curvas azules muestran las similitudes entre la señal original y la señal sintética del modelo LPC-V, aumentando el N_{LPCV}. La curva verde, correspondiente al modelo DELPC-V se obtiene fijando $N_{\text{RL-DELPCV}} = 10$ y aumentando $N_{\text{RR-DELPCV}}$.



(a) (b)

Figura 7.7: Similitud de la señal original con las señales sintetizadas por el LPC-V (curva azul) y por el DELPC-V (curva verde) aumentando $N_{\text{RR-DELPCV}}$ y N_{LPCV}. a) Sonido de la Fig. 7.5a. b) Sonido de la Fig. 7.5b.

Se puede observar como el valor de la similitud obtenido respecto al sonido original es cada vez más cercano a 1 conforme se aumenta N. Además, también se puede apreciar (en estas figuras) que con el modelo DELPC-V se obtienen métricas de similitud mayores para un mismo N, lo que es importante a la hora de realizar una codificación con menos datos.

Se presentan las Tablas 7.4 y 7.5 donde se resume la información de la métrica explicada anteriormente para los 54 sonidos mixtos. En ellas también se muestran los coeficientes totales que se obtienen para la codificación con el modelo DELPC-V así como con el modelo LPC-V para cada uno de los sonidos. Para su cálculo se multiplica la cantidad de trozos en los cuales hemos troceado la señal por los coeficientes (Coef. = N+1) con los cuales codificamos cada uno de ellos.

Las LVA proporcionadas por el algoritmo SSEVA son $\text{LVA}_{\text{RR-DELPCV}} = \text{LVA}_{\text{LPCV}} = 128$ y $\text{LVA}_{\text{RL-DELPCV}} = 4096$, por tanto se trocean en 60 partes los sonidos en la rama lenta del modelo DELPC-V y en 1920 partes en la rama rápida del DELPC-V y en el modelo LPC-V.

Por ejemplo, en el sonido mixto nº 2 se obtiene una $DO = N_{\text{LPCV}} - N_{\text{RR-DELPCV}} = 10$, es decir, el caso de mayor reducción de coeficientes entre un modelo y el otro es de 10 coeficientes menos y se consigue con $N_{\text{LPCV}} = 5$. Por tanto, en la rama rápida del DELPC-V, dado los 1920 trozos en los que se divide cada sonido y los 6 coeficientes utilizamos para codificar cada trozo, se obtienen 11520 coeficientes. En la rama lenta fijado el $N_{\text{RL-DELPCV}}$ a 10, es decir, con 11 coeficientes LPC y multiplicados por 60 trozos hacen 660 coeficientes. En total, el modelo DELPC-V utiliza 12180 coeficientes.

Sonido	Método	Modelo DELPC-V					Modelo LPC-V		Reducción (%)
		N_{RR}	Coef.	N_{RL}	Coef.	Total Coef.	N	Total Coef.	
Mixta1	SSIM	6	13440	10	660	14100	7	15360	8.20
Mixta2	SSIM	5	11520	10	660	12180	15	30720	60.35
Mixta3	SSIM	6	13340	10	1408	14100	8	17280	18.40
Mixta4	SSIM	12	24960	10	660	25620	15	30720	16.60
Mixta5	SSIM	7	15360	10	660	16020	10	21120	24.14
Mixta6	SSIM	12	24960	10	660	25620	15	30720	16.60
Mixta7	SSIM	6	13440	10	660	14100	15	30720	54.01
Mixta8	SSIM	6	13440	10	660	14100	8	17280	18.40
Mixta9	SSIM	6	13440	10	660	14100	15	30720	54.01
Mixta10	SSIM	6	13440	10	660	14100	15	30720	54.01
Mixta11	SSIM	6	13440	10	660	14100	15	30720	54.01
Mixta12	SSIM	6	23440	10	660	14100	8	17280	18.40
Mixta13	SSIM	7	15360	10	660	16020	10	21120	24.14
Mixta14	SSIM	12	24960	10	660	25620	15	30720	16.60
Mixta15	SSIM	7	16216	10	660	16020	10	21120	24.14
Mixta16	SSIM	12	24960	10	660	25620	15	30720	16.60
Mixta17	SSIM	6	13440	10	660	14100	15	30720	54.10
Mixta18	SSIM	6	13440	10	660	14100	15	30720	54.10
Mixta19	SSIM	8	17280	10	660	17940	15	30720	41.60
Mixta20	SSIM	6	13440	10	660	14100	14	29460	52.13
Mixta21	SSIM	6	13440	10	660	14100	15	30720	54.10
Mixta22	SSIM	7	15360	10	660	16020	15	30720	47.85
Mixta23	SSIM	7	15360	10	660	16020	15	30720	47.85
Mixta24	SSIM	5	11520	10	660	12180	15	30720	60.35
Mixta25	SSIM	11	23040	10	660	23700	15	30720	22.85
Mixta26	SSIM	5	11520	10	660	12180	15	30720	60.35
Mixta27	SSIM	5	11520	10	660	12180	15	30720	60.35

Tabla 7.4: Resumen codificación según modelo y método de comparación. Señales mixtas desde la 1 hasta la 27.

En el modelo LPC-V, para obtener la simulación conseguida con $N_{LPCV} = 5$, se necesita un $N_{LPCV} = 15$, es decir 30720 coeficientes. Por tanto, la reducción de coeficientes entre modelos será del 60.35 %.

En la Fig. 7.8 se resume la reducción de los coeficientes en la codificación (eje y, derecha, en naranja) gracias al modelo DELPC-V propuesto, además de la diferencia de orden (eje y, izquierda, en azul) correspondiente a los 54 sonidos mixtos. Es obvio que a más DO, menos coeficientes tengo que utilizar en la codificación.

Destacar que se consigue reducir como máximo un 60.35 % (en más de 10 señales mixtas) y un mínimo de 8.2 % (señal mixta nº 1) el número de coeficientes utilizados

Sonido	Método	Modelo DELPC-V					Modelo LPC-V		Reducción (%)
		N_{RR}	Coef.	N_{RL}	Coef.	Total Coef.	N	Total Coef.	
Mixta28	SSIM	8	17280	10	660	17940	15	30720	41.60
Mixta29	SSIM	5	11520	10	660	12180	15	30720	60.35
Mixta30	SSIM	11	23040	10	660	23700	15	30720	22.85
Mixta31	SSIM	7	15360	10	660	16020	15	30720	47.85
Mixta32	SSIM	6	13440	10	660	14100	10	21120	33.23
Mixta33	SSIM	7	16216	10	660	16020	15	30720	47.85
Mixta34	SSIM	9	19200	10	660	19860	15	30720	35.35
Mixta35	SSIM	13	26880	10	660	27540	15	30720	10.35
Mixta36	SSIM	7	15360	10	660	16020	15	30720	47.85
Mixta37	SSIM	8	17280	10	660	17940	15	30720	41.60
Mixta38	SSIM	6	13440	10	660	14100	15	30720	54.10
Mixta39	SSIM	5	11520	10	660	12180	15	30720	60.35
Mixta40	SSIM	9	19200	10	660	19860	15	30720	35.35
Mixta41	SSIM	8	17280	10	660	17940	15	30720	41.60
Mixta42	SSIM	12	24960	10	660	25620	15	30720	16.60
Mixta43	SSIM	8	17280	10	660	17940	15	30720	41.60
Mixta44	SSIM	7	15360	10	660	16020	15	30720	47.85
Mixta45	SSIM	7	15360	10	660	16020	15	30720	47.85
Mixta46	SSIM	10	21120	10	660	21780	15	30720	29.10
Mixta47	SSIM	9	19200	10	660	19860	15	30720	35.35
Mixta48	SSIM	5	11520	10	660	12180	14	29460	58.35
Mixta49	SSIM	5	11520	10	660	12180	15	30720	60.35
Mixta50	SSIM	7	15360	10	660	16020	15	30720	47.85
Mixta51	SSIM	5	11520	10	660	12180	14	29460	59.94
Mixta52	SSIM	5	11520	10	660	12180	15	30720	60.35
Mixta53	SSIM	5	11520	10	660	12180	15	30720	60.35
Mixta54	SSIM	5	11520	10	660	12180	15	30720	60.35

Tabla 7.5: Resumen codificación según modelo. Señales mixtas desde la 28 hasta la 54.

en la codificación. En todas las señales se reduce, por tanto, media del 39.25 % de los coeficientes, dando validez al modelo DELPC-V propuesto.

7.5.2. Señales vibratorias de frecuencia fundamental baja

En este caso se analizarán y sintetizarán 25 sonidos vibratorios de f_0 baja producidos por las ballenas beluga del Oceanográfico de Valencia. Al igual que con los sonidos mixtos, todos son de la misma duración 2.046 segundos, grabados con una frecuencia de muestreo de 96000 Hz (196416 muestras). Se utilizará un solape del 25 % a la hora de obtener el diagrama tiempo-frecuencia de las señales.

A la hora de codificar ty sintetizar señales vibratorias lentas como la de la Fig.

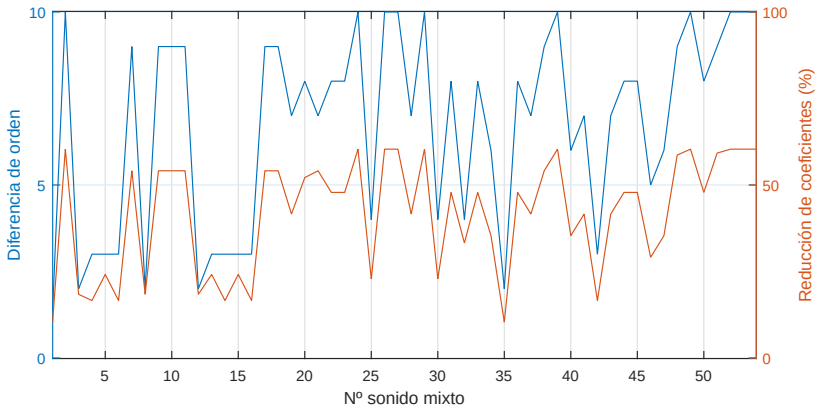


Figura 7.8: Curvas correspondientes a la reducción de los coeficientes de codificación en los sonidos mixtos (en naranja) y su métrica DO correspondiente (curva azul).

7.9, el modelo DELPC-V (ver Fig. 7.4b) actua de la siguiente forma: el algoritmo SSE-VA encamina toda la señal por la rama lenta, estimando la $\mathrm{LVA}_{\mathrm{RL\text{-}DELPCV}}$ en 4096 muestras, quedando la rama rápida inactiva. Por otra parte, el modelo LPC-V utiliza un $\mathrm{LVA}_{\mathrm{LPCV}} = 128$ muestras, su configuración habitual. A continuación se compara en términos de similitud y de eficiencia como se comportan los modelos DELPC-V y LPC-V con este tipo de sonidos.

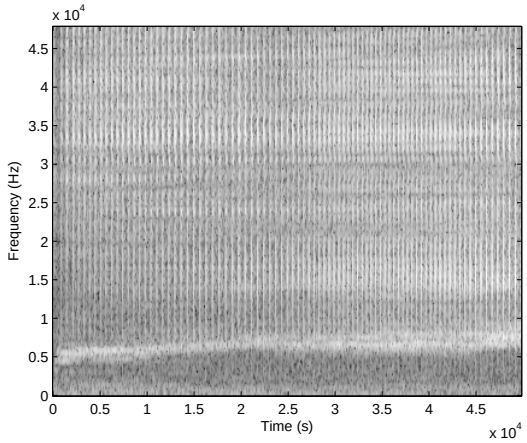


Figura 7.9: Señal vibratoria lenta realizada por una de las ballenas beluga del Oceanográfico de Valencia.

A la hora de comparar la calidad de los sonidos generados por los dos modelos, es necesario identificar el orden del filtro del modelo LPC-V $\mathrm{N}_{\mathrm{LPCV}}$ para el cual se obtiene el máximo de similitud. Dado un sonido en particular se obtiene un valor de similitud para cada $\mathrm{N}_{\mathrm{LPCV}}$ (eje x), dando lugar a la curva azul de la Fig. 7.10.

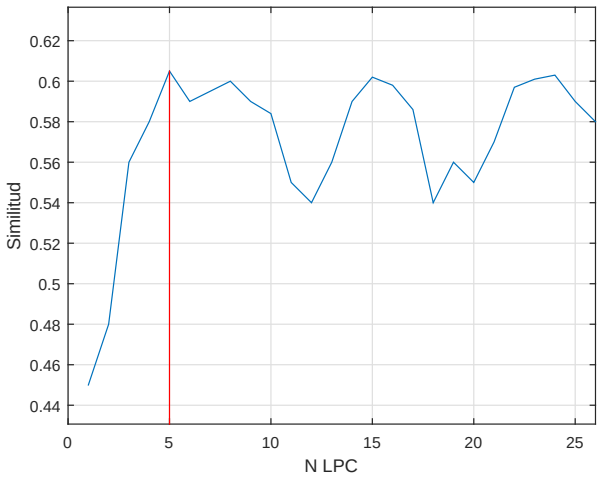


Figura 7.10: Similitud del sonido vibratorio de frecuencia fundamental baja nº 1 con los sonidos generados por el modelo LPC-V aumentando N_{LPCV}.

La Fig. 7.10 muestra como la utilización de la rama lenta en el modelo DELPC-V proporciona unos resultados parecidos a los aportados por el modelo LPC-V sea cual sea el valor N_{LPCV} utilizado. Se puede ver como el incremento de N_{LPCV} (eje x), no tiene una relación dependiente con la obtención de señales que se parezcan más a la original.

Esto es así debido a que la elección de una LVA_{LPCV} con tanta resolución temporal consigue que no haga falta un N_{LPCV} muy grande para codificar la información frecuencial. Para cada uno de los sonidos vibratorios con f_0 baja se fijará el N_{LPCV} según el máximo de similitud obtenido en la curva azul de la Fig. 7.10.

Se procederá entonces a la evaluación del modelo DELPC-V. En la Fig. 7.11 se obtiene la curva verde donde se observa como la similitud de los sonidos producidos por el modelo DELPC-V va siendo cada vez más alta a medida que el $N_{RL\text{-}DELPCV}$ aumenta, superando a la similitud obtenida (curva azul) por el modelo LPC-V con $N_{LPCV} = 5$. El punto de corte entre las dos curvas permitirá obtener el $N_{RL\text{-}DELPCV}$ que se utilizará para obtener la Tabla 7.6. Para sonido vibratorio de frecuencia fundamental baja nº 1 representado, se obtiene $N_{RL\text{-}DELPCV} = 10$.

En la Tabla 7.6 se muestran los coeficientes utilizados en la codificación para los modelos LPC-V y DELPC-V. Para el modelo LPC-V se ha seleccionado el N_{LPCV} que maximiza la similitud (ver Fig. 7.10), y para el modelo DELPC-V el $N_{RL\text{-}DELPCV}$ que iguala dicha similitud (ver Fig. 7.11).

Los resultados son los siguientes: en el DELPC-V, al utilizar una $LVA_{RL\text{-}DELPCV}$ más larga, o lo que es lo mismo, al codificar menos trozos, consigue una reducción respecto al modelo LPC-V. Podemos afirmar que codificar las señales vibratorias lentas de la base de datos con el modelo DELPC-V contempla reducciones de coeficientes en la codificación cercanas al 90 % respecto al modelo LPC-V, manteniendo la similitud con la señal original.

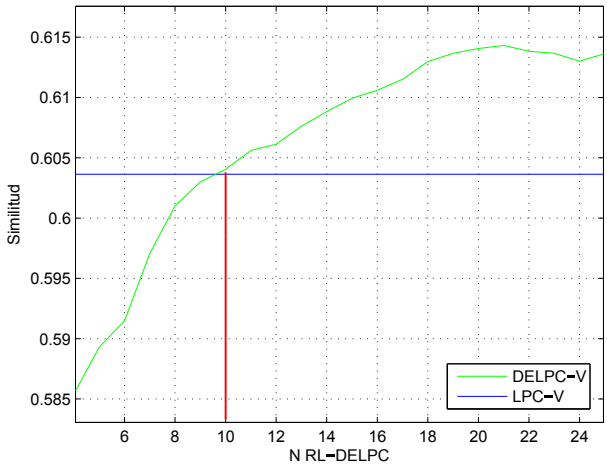


Figura 7.11: Similitud del sonido vibratorio de frecuencia fundamental baja nº1 con los sonidos por el DELPC-V (curva verde) aumentando $N_{\text{RL-DELPCV}}$. La curva azul corresponde a la similitud obtenida con $N_{\text{LPCV}} = 5$ por el modelo LPC-V.

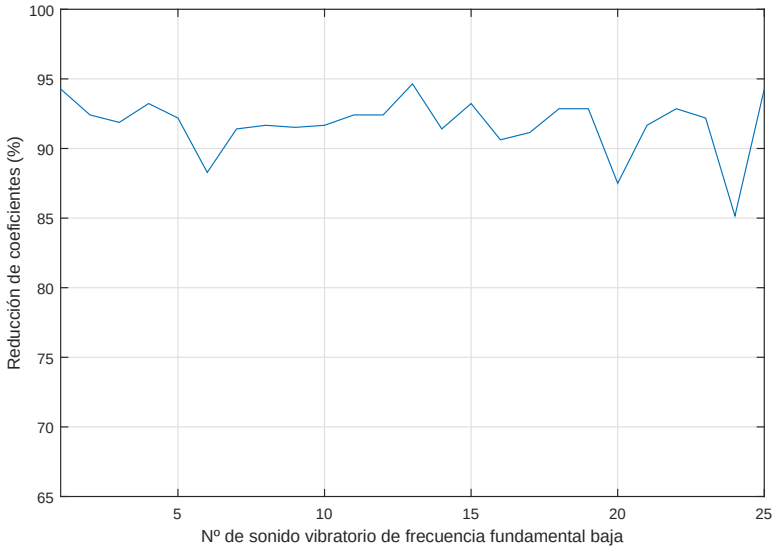


Figura 7.12: Curvas correspondientes a la reducción de los coeficientes de codificación en los sonidos vibratorios de frecuencia fundamental baja.

La Fig. 7.12 muestra un resumen de los resultados. El porcentaje de reducción obtenido es posible gracias a que los LPC funcionan bien al convertir los diagramas tiempo-frecuencia con información vertical (obtenidos con $\text{LVA}_{\text{LPCV}} = 128$) en diagramas tiempo-frecuencia con componentes horizontales, cosa que se logra al aplicar una

Sonido	Método	Modelo DELPC-V					Modelo LPC-V		Reducción (%)
		N_{RR}	Coef.	N_{RL}	Coef.	Total Coef.	N	Total Coef.	
Vibraf0baja1	SSIM	-	-	10	660	660	5	11520	94.79
Vibralenta2	SSIM	-	-	16	1020	1020	6	13440	92.41
Vibraf0baja3	SSIM	-	-	12	780	780	4	9600	91.87
Vibraf0baja4	SSIM	-	-	12	780	780	5	11520	93.23
Vibraf0baja5	SSIM	-	-	14	900	900	5	11520	92.19
Vibraf0baja6	SSIM	-	-	14	900	900	3	7680	88.28
Vibraf0baja7	SSIM	-	-	10	660	660	3	7680	91.41
Vibraf0baja8	SSIM	-	-	15	960	960	5	11520	91.67
Vibraf0baja9	SSIM	-	-	18	1140	1140	6	13440	91.52
Vibraf0baja10	SSIM	-	-	15	960	960	5	11520	91.67
Vibraf0baja11	SSIM	-	-	16	1020	1020	6	13440	92.41
Vibraf0baja12	SSIM	-	-	16	1020	1020	6	13440	92.41
Vibraf0baja13	SSIM	-	-	11	720	720	6	13440	94.64
Vibraf0baja14	SSIM	-	-	10	660	660	4	7680	91.41
Vibraf0baja15	SSIM	-	-	12	780	780	5	11520	93.23
Vibraf0baja16	SSIM	-	-	17	1080	1080	5	11520	90.62
Vibraf0baja17	SSIM	-	-	16	1020	1020	5	11520	91.15
Vibraf0baja18	SSIM	-	-	15	960	960	6	13440	92.86
Vibraf0baja19	SSIM	-	-	15	960	960	6	13440	92.86
Vibraf0baja20	SSIM	-	-	15	960	960	4	7680	87.50
Vibraf0baja21	SSIM	-	-	15	960	960	5	11520	91.67
Vibraf0baja22	SSIM	-	-	15	960	960	6	13440	92.86
Vibraf0baja23	SSIM	-	-	14	900	900	5	11520	92.19
Vibraf0baja24	SSIM	-	-	18	1140	1140	4	7680	85.16
Vibraf0baja25	SSIM	-	-	10	660	660	5	11520	94.27

Tabla 7.6: Coeficientes utilizados en la codificación para 25 los sonidos vibratorios de f_0 baja, incluido el mostrado en la Fig. 7.9.

$N_{\text{RL-DELPCV}}$ lo suficientemente grande (4096). Se recuerda que esta explicación está realizada en detalle en la Sección 3.6.2 de esta tesis doctoral.

Al contrario que en las señales mixtas, la Fig. 7.12 no muestra la métrica DO. Esto es debido a que $\text{LVA}_{\text{RL-DELPCV}} \neq \text{LVA}_{\text{LPCV}}$ y por tanto los filtros no son equivalentes, como sí lo eran en el caso de la utilización de la rama rápida del modelo DELPC-V.

7.5.3. Señales vibratorias de frecuencia fundamental alta y señales resonantes

Para el caso de las señales vibratorias rápidas o señales resonantes, la rama lenta se encuentra inactiva tal y como muestra la Fig. 7.4a, debido a que el separador de señal encamina toda señal por la rama rápida del DELPC-V. Por tanto, y dado que el modelado en esa rama del DELPC-V es el mismo que en el LPC-V, los resultados son

los mismos y no es necesario su análisis.

7.6. Conclusiones

Extraer las características más relevantes para realizar un clasificador especializado en las señales de las ballenas beluga es uno de los objetivos primordiales de este capítulo. Conseguir analizar los sonidos producidos por dichas ballenas a partir de la obtención de características para un buen modelado y la posterior síntesis, provocará que dichas características sean también las adecuadas a la hora de realizar una buena clasificación de las señales.

En este capítulo se demuestra que el modelo DELPC-V que se propone se adapta de manera satisfactoria a la fisionomía del sistema de producción de sonido de los odontocetos, reflejándose esta asociación en los resultados obtenidos al codificar cualquier tipo de sonido, obteniendo sonidos sintéticos más parecidos a los originales que los conseguidos mediante el modelo LPC-V.

Todos los algoritmos que componen el modelo DELPC-V se podrán utilizar a la hora de obtener características para un clasificador que ofrezca mejores porcentajes que los habitualmente utilizados en el ámbito de la monitorización acústica pasiva.

El modelo presentado DELPC-V pese a ser más complejo consigue con menos coeficientes modelar mejor las señales acústicas de las ballenas beluga que el modelo LPC-V. La explicación de esto es obvia ya que este modelo imita mejor la fisionomía del animal. De forma inversa ese mayor parecido de las señales se puede emplear para reforzar las teorías de producción de sonido propuestas por los biólogos.

Capítulo 8

Extracción de parámetros del modelo propuesto para su aplicación en un clasificador de sonidos

8.1. Introducción

En este último capítulo se comprobará que es posible realizar una clasificación utilizando características extraídas del modelo propuesto en el Capítulo 7, características basadas en lo tratado en el Capítulo 5 y el Capítulo 6, es decir, en métricas obtenidas en el dominio cepstral y en el determinismo de la señal.

Todo el estudio realizado en los capítulos anteriores hace posible la comprensión de los sonidos y su naturaleza. Se estudiaron las capacidades que tienen estos mamíferos a la hora de generar sonido, mucho mayores que las de un humano, a la hora de diseñar un modelo asociado a la fisionomía del aparato fonador de los odontocetos. Todos y cada uno de los elementos del modelo tienen relación con alguna función del sistema de producción de sonidos, de forma que el modelo propuesto presenta un claro parecido.

Se aprovechará la extracción de características para ir repasando las principales temáticas estudiadas y explicadas a lo largo de los diferentes capítulos de la presente tesis doctoral, pudiendo realizar unas conclusiones al final de este capítulo que resuman de manera global los resultados e ideas a los cuales se ha llegado.

8.2. Características extraídas del modelo DELPC-V

En esta sección se realiza una de las partes más importantes de la tesis doctoral, la extracción de las características del modelo propuesto para la posterior comprobación de que son adecuadas y válidas para la realización de un clasificador de sonidos de la ballena beluga. De hecho la proposición de un modelo de análisis / síntesis lo más adaptado posible se ha realizado para poder realizar una selección de características de calidad y no por cantidad (como se hizo en el Capítulo 3). Se ha intentado extraer

pocas características pero relevantes, todas relacionadas con el modelo DELPC-V, el cual resume la fisionomía del aparato fonador de estos odontocetos.

Para ello, se realizará una categorización de sonidos que se adapten perfectamente a las características a seleccionar. Tal y como se indica en el Capítulo 4, fueron cuatro las clases de sonidos a identificar:

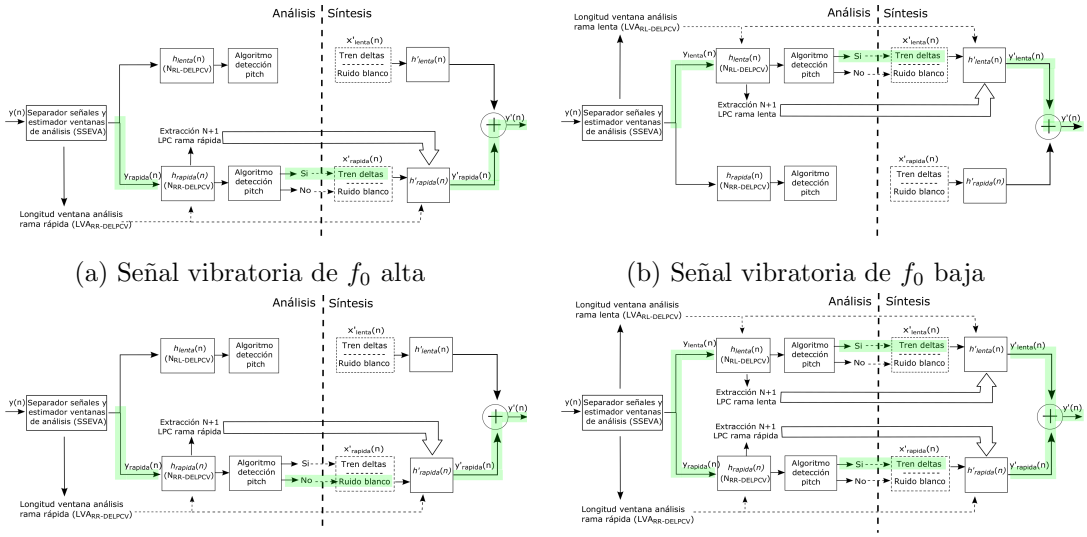
I. Sonidos vibratorios de f_0 alta.

II. Sonidos vibratorios de f_0 baja.

III. Sonidos resonantes.

IV. Sonidos mixtos.

Estos 4 tipos de sonidos engloban a cualquier sonido más específico, y dado que las ballenas belugas son denominadas como los canarios del mar, esta elección es básica para poder llevar a cabo una clasificación exitosa.



(a) Señal vibratoria de f_0 alta

(b) Señal vibratoria de f_0 baja

(c) Señal resonante

(d) Señal mixta

Figura 8.1: Funcionamiento del DELPC-V según las señales a tratar.

Recordamos la Fig. 8.1 donde se muestra de una manera global el funcionamiento del modelo propuesto según el tipo de sonido a clasificar. Básicamente, cuando la señal no es mixta se podrán dar tres casos básicos:

I. En el caso de que sea un sonido vibratorio con una f_0 dentro del rango [700 - 10000 Hz], el algoritmo SSEVA encaminará la señal por la rama rápida. El Algoritmo detección pitch determinará que existe pitch y en la parte de síntesis del modelo elegirá un tren de deltas como excitación (Fig. 8.1a).

II. En el caso de que sea un sonido vibratorio con una f_0 dentro del rango [0.5 - 700 Hz], el algoritmo SSEVA direccionará la señal por la rama lenta. El Algoritmo detección pitch determinará que existe pitch y en la parte de síntesis del modelo elegirá un tren de deltas como excitación (Fig. 8.1b).

III. En el caso de que sea resonante, dado que las frecuencias fundamentales de esta señal están en el rango de la rama rápida del modelo, el algoritmo SSEVA encaminará la señal por dicha rama. El Algoritmo detección pitch determinará que no existe pitch y en la parte de síntesis del modelo elegirá ruido blanco como excitación x'_{rapida} (Fig. 8.1c).

Sin embargo, cuando se analiza una señal mixta (Fig. 8.1d), el algoritmo SSEVA separará cada una de las dos señales vibratorias que la componen y después las direccionará por la rama correspondiente del DELPC-V. De esta manera se podrán extraer los LPC de manera correcta gracias a la elección de la LVA para cada una de las ramas.

Las dos partes más importantes de este modelo DELPC-V son las explicadas con detenimiento en el Capítulo 5 y el Capítulo 6, referentes al dominio cepstral y al determinismo respectivamente. Tal y como se puede leer en los párrafos anteriores, mediante el algoritmo SSEVA y el Algoritmo detección pitch, es posible discernir entre las cuatro clases o categorías definidas, categorías basadas en la naturaleza del sonido.

En el Capítulo 5 se estudió el dominio cepstral y se diseño un algoritmo capaz de encaminar el sonido por una de las dos ramas que componen el modelo. Recordemos que cada una de ellas estaba especializada en un rango de frecuencias, al igual que la fisionomía real aparato fonador de las ballenas belugas. A su vez, cuando el sonido a analizar era mixto dicho algoritmo permitía separar las dos señales y encaminarlas por la rama correcta para el aprovechamiento de las ventajas de los LPC.

En el Capítulo 6 se propuso un algoritmo de detección de pitch basado en el nivel de determinismo de la señal, que pudiera competir con otros algoritmos que realizan esta labor, como por ejemplo utilizando el dominio frecuencial o detectando los pasos por cero de la señal. La rapidez de cálculo, unida a la independencia que aporta, al ser calculada en un dominio diferente al frecuencial, hace que la métrica sea la elegida para la detección de pitch. Se recuerda que el objetivo no es encontrar específicamente la frecuencia de resonancia, sino simplemente determinar si existe o no pitch, es decir, si la frecuencia fundamental está causada por resonancias o por vibraciones.

Se elegirán cinco características como representativas. En la Fig. 8.2 se puede observar el modelo DELPC-V y en color azul las características extraídas, así como su procedencia. La primera de las cinco características procede de la métrica de determinismo propuesta y tendrá como función evaluar si un sonido tiene o no pitch. Se extraerá de la rama rápida, ya que las señales resonantes sólo aparecen en el rango de esta rama. Esta primera característica será RQA_{DET} (el subíndice r identifica la rama de donde se extrae la característica) y se obtendrá mediante el algoritmo de medida de diagonales descrito en el Capítulo 6 de la misma forma que se utiliza en el modelo DELPC-V, haciendo un análisis de un número fijado de trozos del sonido, obteniendo un valor para cada uno de ellos y calculando el promedio.

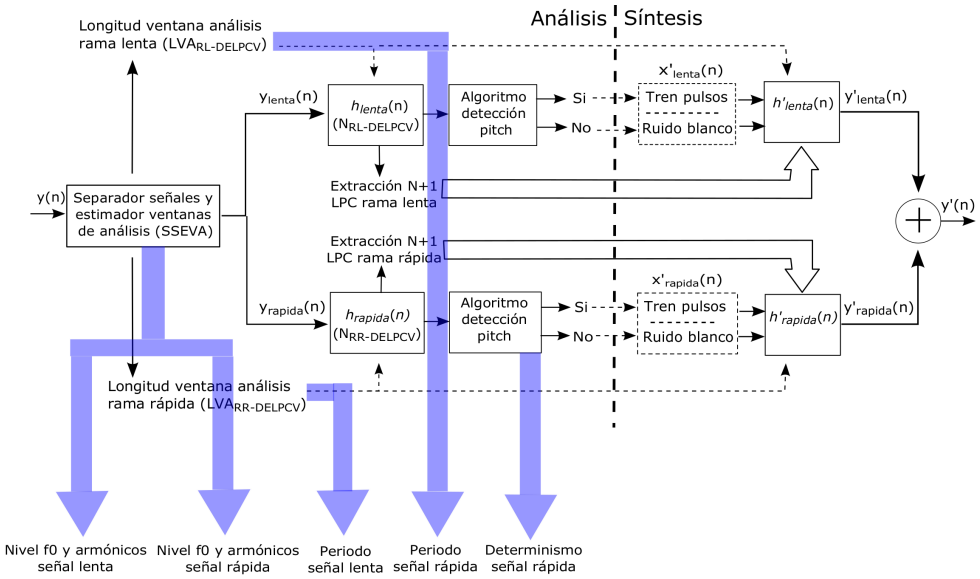


Figura 8.2: Modelo de análisis / síntesis Doble Excitación LPC-Vocoder. Las flechas azules muestran las características extraídas y su procedencia.

Las dos siguientes características que se han elegido tienen que ver con el algoritmo SSEVA y el dominio cepstral, ya que para decidir si un sonido está compuesto por dos (es decir, un sonido mixto) se necesita tener una medida que nos indique el nivel de señal (armónicos incluidos) de cada uno de ellos, una vez lleguen a la rama correspondiente. Se recuerda que el dominio cepstral resume en sus coeficientes información tanto de la frecuencia fundamental como de la cantidad de armónicos que posee y su energía.

Se obtendrá el diagrama cepstral-temporal un sonido y se calculará el nivel máximo de cada uno de los *frames* obteniendo un vector de máximos de tantos componentes como *frames* tenga el diagrama. Se promediará este vector de máximos vec_{max} obteniéndose LEV_r y LEV_l (ver Ec. (8.1)). Véase que los subíndices r y l identifican la rama de donde se extrae la característica.

$$LEV = P[vec_{max}] \tag{8.1}$$

donde $P[\cdot]$ es un operador que calcula el valor medio y vec_{max} el vector de máximos.

La Longitud de la Ventana de Análisis (LVA) extraída del algoritmo SSEVA para cada una de las ramas será utilizada también como característica. A través de ella tendremos información sobre la frecuencia fundamental que hay en cada una de las ramas. La LVA será nula cuando no se encuentre dicha frecuencia en su rango de frecuencias establecido. Esta característica estará asociada a la capacidad de encaminar una señal por la rama correcta del modelo DELPC-V.

Se utiliza el cálculo realizado para la obtención de las características LEV_l y LEV_r, y de cada uno de los máximos obtenidos se obtendrá la frecuencia cepstral donde se sitúa, obteniendo un vector de las mismas componentes que el vector de máximos. Se

Nº Carac	Nombre	Objetivo
1	RQA_{DETr}	Medida del determinismo en la rama rápida para detección de pitch
2	LEV_r	Medida relacionada con el nivel de f_0 y armónicos en el sonido de la rama rápida
3	LEV_l	Medida relacionada con el nivel de f_0 y armónicos en el sonido de la rama lenta
4	IFF_r	Medida relacionada con la f_0 en el sonido de la rama rápida
5	IFF_l	Medida relacionada con la f_0 en el sonido de la rama lenta

Tabla 8.1: Resumen de las características seleccionadas a partir del modelo DELPC-V

promediará este vector de frecuencias cepstrales vec_{ceps} obteniendo la característica IFF o información de la Frecuencia Fundamental (ver Ec. (8.2))

$$IFF = P[vec_{ceps}] \tag{8.2}$$

Se denominarán IFF_r y IFF_l, donde $P[\cdot]$ es un operador que calcula el valor medio y vec_{ceps} el vector de frecuencias cepstrales. En la Tabla 8.1 se resumen las cinco características extraídas.

8.3. Elección y aplicación en clasificadores según el tipo de características

En esta sección se comprobará lo relevantes que son las características extraídas. Para ello se necesitará una buena base de datos con sonidos de todos los tipos para lograr averiguar de manera inequívoca y amplia si clasificador y por tanto las características, cumplen su cometido y la selección de características ha sido satisfactoria.

Tipo de sonido	Fase entrenamiento	Fase test/validación	Total
Mixto	20	27	47
Vibratorio de f_0 baja	20	73	93
Vibratorio de f_0 alta	20	64	84
Resonante	20	20	40
Total	80	184	264

Tabla 8.2: Cantidad de sonidos de la base de datos.

Se han utilizado para la fase de entrenamiento 20 sonidos de cada tipo. Para la fase de test o clasificación para validar el clasificador entrenado se han utilizado 27 sonidos mixtos, 73 sonidos vibratorios lentos, 64 vibratorios rápidos y 20 sonidos resonantes. La Tabla 8.2 resume estos números.

Cabe destacar que ningún sonido ha sido repetido en ambas fases y que la base de datos de sonidos es la misma que la utilizada en la clasificación del Capítulo 3, incluyendo unos pocos más en el presente capítulo. De esta manera se podrán comparar algunos de los resultados.

8.3.1. Fase de entrenamiento

Para la correcta elección del tipo de clasificador en función de las características extraídas, se han evaluado las métricas obtenidas utilizando los clasificadores descritos anteriormente en la fase de entrenamiento. En la Tabla 8.3 se puede ver la tasa de éxito para cada clasificador.

Se ha utilizado la validación cruzada en todos ellos para la comprobación del rendimiento del algoritmo de aprendizaje, mediante la partición k veces del conjunto de entrenamiento ($k = 5$).

Tipo de Clasificador	Tasa de acierto global (%)	Tasa de error global (%)
SVM Lineal	90	10
SVM Cuadrático	90	10
Árbol de decisiones simple	86.3	13.7
Árbol de decisiones complejo	83.8	16.2
k vecinos más cercanos con distancia euclídea	83.8	16.2

Tabla 8.3: Resultados de los clasificadores en la fase de entrenamiento con las 5 características

Se obtiene que el clasificador con una mayor tasa de acierto es el *Support Vector Machine* lineal. En la Tabla 8.4 se presenta su matriz de confusión.

SVM lineal (5 características)

Clasificado\ Real	Mixta	Vibratorio de f_0 baja	Vibratorio de f_0 alta	Resonante
Mixto	18	0	1	2
Vibratorio de f_0 baja	0	20	0	1
Vibratorio de f_0 alta	0	0	17	1
Resonante	2	1	2	17
TPR	90 %	100 %	85 %	85 %

Tabla 8.4: Matriz de confusión obtenida con el clasificador SVM lineal, referente a la fase de entrenamiento (80 sonidos)

A continuación estudiaremos si la incorporación de todas las características mejora o empeora los resultados. Para ello se evaluarán otra vez los clasificadores sustrayendo del entrenamiento varias de las características aleatoriamente. Los mejores resultados se darán cuando no utilicemos la característica relacionada con el nivel de los armónicos y frecuencia fundamental de la rama lenta IFF_l.

Como se puede ver en la Tabla 8.5 utilizando únicamente cuatro características se obtienen mejores resultados en algunos de los clasificadores. En una mayoría de ocasiones, la utilización de un número elevado de características no implica la obtención de mejores resultados y es otra de las variables a tener en cuenta junto con el *overfitting*. Se puede ver que en este caso, el clasificador con mayor porcentaje de fiabilidad es el árbol de decisiones

Tipo de Clasificador	Tasa de acierto global (%)	Tasa de error (%)
SVM Lineal	91.3	8.8
SVM Cuadrático	88.8	11.3
Árbol de decisiones simple	**95**	**5**
Árbol de decisiones complejo	95	5
k vecinos más cercanos con distancia euclídea	85	15

Tabla 8.5: Resultados de los clasificadores en la fase de entrenamiento con 4 características

Árbol de decisiones simple (4 características)

Clasificado\ Real	Mixta	Vibratorio de f_0 baja	Vibratorio de f_0 alta	Resonante
Mixto	19	0	0	0
Vibratorio de f_0 baja	0	20	0	0
Vibratorio de f_0 alta	0	0	17	0
Resonante	1	0	3	20
TPR	95 %	100 %	85 %	100 %

Tabla 8.6: Matriz de confusión obtenida con el clasificador de árbol de decisiones simple, referente a la fase de entrenamiento (80 sonidos)

simple. Se elegirá por tanto este clasificador trabajando con cuatro características para la fase de test/validación de la siguiente sección.

En la Tabla 8.6 se puede observar la matriz de confusión para dicho clasificador referente a la fase de entrenamiento para los 80 sonidos (20 de cada tipo).

Para que el clasificador elegido no tenga *overfitting* se incrementarán los sonidos en la fase de entrenamiento y no el número de características extraídas. Finalmente se eligió el árbol de decisiones simple con 4 características: RQA_{DETr}, LVA_r, LVA_l y LEV_r.

8.3.2. Fase de test, validación o clasificación

Una vez elegido el clasificador adecuado, teniendo en cuenta el número de características, su comportamiento y el posible *overfitting*, se procederá a predecir/clasificar los sonidos preparados para ello.

En la Tabla 8.7 se muestra la matriz de confusión de los sonidos en la fase de test/validación. Se puede observar como la tasa de acierto global es del 91 %, un muy buen resultado dada la dificultad que conlleva realizar una clasificación de sonidos de ballenas beluga.

Las señales mixtas son las que peor porcentaje de clasificación correcta tienen, confundiéndose con señales vibratorias de f_0 baja o con señales vibratorias de f_0 alta. La razones pueden ser las siguientes:

I. Cuando el clasificador se confunde e indica que son vibratorias lentas es debido a que el nivel de la frecuencia fundamental y armónicos en la rama rápida (carac-

Árbol de decisiones simple (4 características)

Clasificado\ Real	Mixta	Vibratorio de f_0 baja	Vibratorio de f_0 alta	Resonante
Mixta	19	0	1	1
Vibratorio de f_0 baja	3	72	0	0
Vibratorio de f_0 alta	4	0	59	0
Resonante	1	1	4	19
TPR	70.37 %	98.63 %	92.19 %	95 %

Tabla 8.7: Matriz de confusión con el clasificador de árbol de decisiones simple, referente a la fase de validación (184 sonidos)

terística LEV_r) es lo suficientemente bajo como para pensar que no existe señal en dicha rama. En ese momento, el clasificador pensará que no es un sonido compuesto o mixto, decidiendo por tanto que se trata de un sonido vibratorio de f_0 baja.

II. Cuando el clasificador indica que son vibratorias rápidas es debido a que el nivel de la frecuencia fundamental y armónicos (característica LEV_l) es lo suficientemente bajo como para pensar que no existe señal en la rama lenta del modelo. En ese momento, el clasificador pensará que no es un sonido compuesto o mixto. Una vez llegado a este punto, la característica relacionada con la detección del pitch determinará que el sonido es vibratorio de f_0 alta.

Una de las ventajas de la elección de un clasificador tipo árbol de decisión es la rapidez de cálculo, muy adecuado cuando se incluya este clasificador en aplicaciones como la descrita en el apéndice A.

8.4. Conclusiones

En el presente capítulo se ha comprobado el potencial del modelo DELPC-V propuesto, mediante la utilización de sus características en un modelo de clasificación seleccionado entre varias opciones recomendadas.

Mediante la fase de entrenamiento y la búsqueda del clasificador que maximice los porcentajes de clasificación correcta, se comprueba que con las características que se han extraído del modelo es posible realizar un clasificador con buenos porcentajes, demostrando así que todas ellas son relevantes y que el estudio realizado en todos los capítulos de la presente tesis doctoral han tenido una contribución notable a la hora de obtener los resultados mostrados.

Es posible establecer vínculos con los resultados del Capítulo 3 para obtener alguna conclusión, pero se debe tener en cuenta que tanto las categorías de clasificación como las características utilizadas son diferentes.

Por ejemplo, la tasa de acierto global es más de 10 puntos mejor en la clasificación realizada con las características extraídas a partir del modelo propuesto en esta tesis

doctoral que en la anterior. Esto es más llamativo ya que se han incluido las señales mixtas, y los sonidos tonales se han separado en dos categorías (vibratorios de f_0 alta y resonantes).

De hecho, analizando exhaustivamente la tasa de verdaderos positivos o *True Positive Rate* (TPR) en los sonidos pulsados, se obtiene un aumento de casi 30 puntos más dado el 98.63 % conseguido en la clasificación de este capítulo. Algo parecido sucede con el TPR obtenido para los sonidos tonales del Capítulo 3, donde aun con un valor significativo (81.03 %) es más de 10 puntos menor que cualquiera de los TPRs obtenidos por los sonidos resonantes y vibratorios de f_0 alta.

Cabe destacar que en las categorías definidas para la clasificación propuesta en este capítulo, no se han incluido los ruidos antropogénicos dado que, al utilizar las características extraídas sobre un modelo de análisis / síntesis de sonido, el comportamiento sería inesperado. Algo parecido ocurre con los sonidos provocados por la mandíbula de estos animales, los *Jawclaps*, dado que se producen con otra parte del cuerpo no presente en el aparato fonador de los odontocetos.

Capítulo 9

Conclusiones generales

A continuación se presentan las conclusiones derivadas de la tesis doctoral. Se englobarán dentro de tres temáticas diferentes, las obtenidas gracias al estudio y comprensión de sistema de producción del sonido y sus problemáticas asociadas, las asociadas al nuevo planteamiento de clasificación y diseño de un modelo adaptado a él, y las concernientes a la clasificación de los sonidos.

COMPRENSIÓN DEL SISTEMA DE PRODUCCIÓN DE SONIDO Y PROBLEMÁTICAS ASOCIADAS A LAS CARACTERÍSTICAS DE LOS SONIDOS

La comprensión del aparato fonador de los odontocetos mediante el estudio de sus sonidos y la comparación con los realizados por los seres humanos y por instrumentos musicales fue clave para entender su comportamiento. Gracias a un paso previo consistente en la extracción de todo tipo de características, se pudo resolver cuales de ellas eran más relevantes, además de esclarecer en qué cualidades de los sonidos se debía de hacer hincapíe. Se llegaron a las siguientes conclusiones:

I. Los sonidos pulsados y un grupo de los sonidos tonales son producidos de la misma manera (vibración sistema MLDB) y pueden coincidir en el tiempo de manera simultánea. Es por tanto necesario la existencia de dos fuentes de sonidos en paralelo, dos MLDB, cada uno de ellos especializado en la realización de sonidos vibratorios con diferente rango de frecuencia fundamental.

   La utilización del algoritmo propuesto SSEVA basado en el dominio cepstral para la separación de fuentes y estimación de la longitud de la ventana de análisis permite realizar dicho cometido.

II. La existencia de sonidos con diferente comportamiento en el dominio temporal sugiere la definición de una clase de sonidos parecidos a los típicos silbidos producidos por los seres humanos, donde no es necesaria la vibración de ningún elemento para producirlos. Su naturaleza de producción es el detonante para la utilización de algoritmos basados en el dominio temporal. Se recuerda que estos sonidos están

agrupados dentro de la categoría de sonidos tonales por la morfología de su diagrama tiempo-frecuencia.

El diseño del algoritmo basado en la cuantificacion de los *Recurrence Plot* como medida del determinismo de la señal permite diferenciar entre los sonidos resonantes y vibratorios.

III. La importancia de la LVA en los diagramas tiempo-frecuencia. La utilización de una LVA con distinto tamaño afecta a la manera de visualizar el diagrama tiempo-frecuencia de un sonido, pudiendo tomar una forma horizontal o una forma vertical. Esta conclusión es imprescindible a la hora de proponer un cambio de las categorías de clasificación, basadas esta vez en la naturaleza de producción del sonido.

NUEVAS CATEGORIAS DE CLASIFICACION Y MODELO DISEÑADO

Con todo lo visto anteriormente es necesario establecer nuevas normas. En paralelo se decidió utilizar una nueva categorización de los sonidos y un nuevo modelo de análisís / síntesis para la codificación y generación de estos. Estos dos pasos supusieron las siguientes conclusiones:

I. El modelo denominado DELPC-V, basado en el modelo LPC-V para voz humana, puede dar cabida a toda la gama de sonidos que son capaces de producir los odontocetos. Este modelo tiene claras asociaciones con la fisionomía del aparato fonador de las ballenas belugas.

Se comprueba el funcionamiento de los diferentes algoritmos incluidos en el modelo DELPC-V al analizar el ahorro de coeficientes para la codificación. De hecho, genera sonidos sintéticos con una similitud mayor con el sonido original que otros modelos menos adaptados al sistema de producción de sonido del animal como el LPC-V utilizando el mismo orden de filtro LPC.

II. La extracción de características a partir del modelo DELPC-V y su utilización en un clasificador demuestra que el modelo se adapta bien a la gran cantidad de sonidos que son capaces de producir.

Se comprueba que el estudio y compresión en detenimiento de la fisionomía y los problemas intrínsecos de las ballenas beluga facilita el desarrollo de algoritmos para el análisis de los sonidos. Mediante la utilización de clasificadores habituales se demuestra que tanto las características seleccionadas del modelo como la nueva categorización permiten el diseño de un clasificador capaz de discriminar de una manera satisfactoria entre la mayoría de los sonidos provocados por las ballenas beluga.

III. La categorización de los sonidos propuesta, asociada a la naturaleza de su producción, refleja mejores resultados a la hora de clasificar los sonidos que el enfoque basado en la morfología del diagrama tiempo-frecuencia.

ENFOQUE DE CLASIFICACIÓN

I. Conforme a la utilización de las técnicas de clasificación, una de las conclusiones más interesantes del trabajo realizado por esta tesis doctoral es la comprobación de que una selección de características basada en la comprensión del problema permite obtener un número pequeño y relevante para su utilización en diferentes clasificadores de manera rápida y eficaz a la hora de hacer una clasificación de sonidos.

   Cabe destacar que tanto la clasificación utilizada en el Capítulo 8 como la realizada en el Capítulo 3, pueden ser complementarias en entornos más complejos. La resolución de un problema de amplio espectro y con poco conocimiento sobre él puede iniciarse extrayendo el mayor número de características diferentes posible, para después, una vez averiguadas las propiedades y peculiaridades del problema realizar el diseño de un nuevo modelo de clasificación con las características más relevantes identificadas.

9.1. Líneas Futuras

Se plantea la realización de otros modelos de análisis / síntesis adaptados a otro tipo de cetáceos que posean un mecanismo de producción de sonido diferente. Con la ayuda del modelo planteado en la presente tesis doctoral se podrán obtener ideas de como realizan sus sonidos y por tanto, esclarecer su funcionamiento. Esto es interesante ya que cerca de nuestras costas no se encuentran ballenas beluga y realizar un modelado adaptado a los cetáceos que conviven en nuestro hábitat puede tener una mayor repercusión, sobre todo a la hora de discernir entre sonidos producidos por ellos y los creados artificialmente por el hombre.

De hecho, el estudio y caracterización de los ruidos antropogénicos, a la hora de poder clasificar entre ellos y un sonido realizado por algún animal, es una de las líneas futuras más interesantes que provoca esta tesis doctoral. Muchos de los sonidos captados, ya sean en cautividad o en mar abierto, requieren un análisis exhaustivo, no sólo para esclarecer los comportamientos y entresijos de las capacidades de los sistemas fonadores de los mamíferos marinos, sino para algo mucho más interesante y con mucha más transversalidad, la monitorización y clasificación de los ruidos producidos en estos entornos.

Directivas tan relevantes para el sector de las energías marinas renovables como la Directiva Marco de Estrategias Marinas (DMEM) (Marine Strategy Directive 2008/56/EC), en relación al descriptor 11, y la Directiva de Estudios de Impacto Ambiental (EIA) (2011/92/EU), muestran este interés, requiriendo la monitorización y evaluación del ruido submarino generados por estas actividades humanas desarrolladas en el medio marino-marítimo. Futuros objetivos de I+D+i de la Comisión Europea responden a esta temática de investigación, la cual está, además, alineada con una de las prioridades principales indicadas por el Plan de Acción de la Energía Azul (Blue Energy Action Plan): "*there are environmental issues to be faced, including the need for more research and*

better information on environmental impacts"

Además, la DMEM recoge el Anexo I como uno de los descriptores del Buen Estado Ambiental, el descriptor 11 (D11) que se define como: "La introducción de energía, incluido el ruido subacuático, se sitúa en niveles que no afectan de manera adversa al medio marino". La gran expansión y distribución de parques de energías marinas renovables en Europa hace que este D11 tenga especial relevancia en cuanto a la introducción de energía en el medio marino por parte de este sector de energía y el impacto que puede generar en el ecosistema marino, en especial a especies marinas de gran importancia comercial (acuicultura, pesquerías, etc.). Sin embargo, el D11 tiene un carácter especial que lo convierte en una situación de interés, tal y como establece la Decisión de la Comisión Europea (2010/477/UE, de 1 de Septiembre de 2010) sobre criterios y normas metodológicas aplicables al buen estado ambiental de las aguas marinas.

Por tanto, el estudio y caracterización de estos ruidos, dado la problemática actual, unidos al análisis del comportamiento de los cetáceos a partir de sus sonidos, gozará de un interés y será un ámbito que servirá como punto de partida donde comenzar a aplicar todo lo mostrado en la presente tesis doctoral.

Apéndices

Apéndice A

Aplicaciones: Sistema integrado de detección y clasificación de eventos acústicos submarinos (SAMARUC)

A.1. Introducción

A lo largo de la tesis doctoral se han explicado y mostrado las características de la fisionomía del sistema de producción de sonidos de los mamíferos marinos, en concreto de los odontocetos, estudiando los diferentes sonidos que son capaces de producir, introduciendo un modelo adaptado que permita la creación o síntesis de dichos sonidos. El análisis de los parámetros utilizados en el modelado ha permitido obtener con cuales de ellos es viable realizar una clasificación en función de la naturaleza del sonido. El potencial de estos parámetros se ha explotado a la hora de diseñar y desarrollar algoritmos de detección y clasificación de los sonidos o señales subacúaticas de forma automática.

Para que los algoritmos diseñados se ejecuten con un coste computacional bajo y con el objetivo de lograr una independencia tecnológica a la hora de adquirir señales subacúaticas con un control exhaustivo y una calidad óptima, se ha diseñado y construido un dispositivo de monitorización acústica pasiva (PAM) denominado SAMARUC. La posibilidad de incluir los algoritmos programados permite lograr una gran eficacia de los recursos del sistema (capacidad computacional, gestión del consumo de las baterías, etc.) y con ello posibilitar su comercialización. Un prototipo plenamente funcional de SAMARUC, ha sido fabricado y está siendo usado en aplicaciones medioambientales.

El trabajo realizado en esta tesis doctoral se enmarca dentro de las investigaciones realizadas en el Grupo de Tratamiento de Señal (GTS) del iTEAM. **La redacción de este apéndice pretende mostrar la concienciación e importancia a la hora de transferir la investigación descrita en esta tesis doctoral.** En la primera sección se presenta tanto la tecnología diseñada a la hora de incrementar el potencial de los algoritmos diseñados, como las aplicaciones en las que es posible su utilización. A continuación se muestra un resumen del estado del arte y las tecnologías ya existentes en el mercado evaluando las ventajas e inconvenientes de la tecnología diseñada. Además, se

ha realizado un estudio exhaustivo del mercado donde se pretende introducir el producto para la posterior elaboración de un correcto plan de negocio y un planteamiento adecuado de las estrategias de promoción y marketing. Para finalizar el capítulo se presentarán las líneas futuras y conclusiones a lograr.

A.2. Pasos realizados y descripción de la tecnología

En este apartado se describen los primeros pasos que se llevaron a cabo y la evolución que sufrió la tecnología diseñada. La sinergia establecida entre el GTS y el departamento de investigación del Oceanográfico de Valencia fue clave. La necesaria optimización de los recursos disponibles en la institución biológica, junto a la dilatada experiencia del GTS en el procesado de señal, hacen posible la extrapolación de las características y técnicas habituales de un campo tan ampliamente estudiado como es el reconocimiento de la voz humana a un nuevo ámbito, los sonidos subacuáticos, sean tanto a señales de animales como posibles ruidos que intervienen en el mundo marino.

El primer acuerdo de colaboración fue financiado por la Cátedra Telefónica para la realización de algoritmos de detección y clasificación de señales bioacústicas. Este resultado permitió a los biólogos involucrados analizar de forma óptima la inmensa cantidad de archivos de audio submarino grabados en los distintos fondeos que habían realizado tanto en cautividad como en mar abierto. Meses más tarde, y ante varios medios de comunicación se presentó el sistema de detección y clasificación diseñado por el grupo en las instalaciones del Oceanográfico de Valencia consiguiendo estrechar lazos para próximas colaboraciones.

A raíz de dicha colaboración se contacta con la asociación local sin ánimo de lucro para el estudio y conservación del entorno Xaloc. Xaloc, entidad consultora medioambiental del Gobierno Balear en el parque natural de Cabrera y del Gobierno Valenciano en el parque natural de las islas Columbretes, se encarga del control de las especies marinas cerca de dichos parques empleando como herramientas habituales dispositivos PAM. Esta organización muestra interés por el trabajo realizado y sugiere la adaptación de los algoritmos diseñados a entornos no controlados: mares y océanos, donde existe mucha más diversidad de señales acústicas que en cautividad (ej. ruidos provocados por el hombre, vocalizaciones de un mayor abanico de especies de animales).

Para ello se estableció como prioritario adaptar los algoritmos diseñados a la detección y clasificación de los ruidos antropogénicos, es decir, ruidos provocados por el ser humano (voz humana, prospecciones petrolíferas o de cualquier otra índole, motores de barco) o por sismos, los cuales afectan al bienestar de la fauna a proteger para su aplicación tanto en entornos en cautividad (privados) como en mar abierto. Algoritmos que pudieran ser embebidos en dispositivos PAM, aprovechando la capacidad de obtener datos que poseen estos dispositivos.

Los equipos PAM que Xalòc utilizaba procedían de EEUU (EAR, MARU) y tanto los alquileres de los mismos, los gastos de distribución, así como la facilidad de uso del equipo eran altamente mejorables. Siendo así, se planteó la posibilidad de realizar un dispositivo que mejorara dichas cualidades y que además fuese un paso adelante respecto

al resto de la tecnología existente en el mercado, incorporando los algoritmos diseñados embebidos dentro de él.

Gracias a la financiación de un proyecto del programa de Investigación Fundamental no Orientada del Ministerio de Economía y Competitividad, pasados cinco meses se presentó la primera versión del dispositivo. Sus cualidades están por encima del estado del arte de los equipos PAM actuales: un sistema de almacenaje fácil de utilizar, una carcasa capaz de aguantar hasta 1000 metros de profundidad, un hidrófono con un ancho de banda mayor de lo habitual y una electrónica sencilla y versátil, formada básicamente por un microcontrolador ezdsp5535 y su compilador Code Composer Studio en su versión 4 junto con el software de cálculo científico MATLAB y el compilador de éste para la inclusión de los algoritmos diseñados, un interruptor magnético para el fácil encendido de SAMARUC, almacenaje para las baterías, placa base con reguladores/amplificadores y tarjeta de memoria SD/micro. Más específicamente los componentes principales que conforman el dispositivo PAM son los mostrados en la Tabla A.1, donde se han establecido varias configuraciones adaptando el hardware a las condiciones específicas de los algoritmos diseñados, planteando prototipos de carcasas para fondeos de corta y de larga duración y para poca o mucha profundidad, además de la compra de hidrófonos adaptados a la aplicación elegida.

Componentes	**Config. 1**	**Config. 2**	**Config. 3**	**Config. 4**
Carcasa	Profundidad 100m. Corta Duración. 1500€	Profundidad 100m. Larga Duración.	Profundidad 1000m. Corta Duración.	Profundidad 1000m. Larga Duración.
Hidrófono	CR1. Omnidireccional. Profundidad 100m.	CR1. Omnidireccional. Profundidad 100m.	CR1. Omnidireccional. Profundidad 100m.	CR1. Omnidireccional. Profundidad 100m.
Electrónica	Microcontrolador ezdsp5535. Interruptor magnético. Tarjeta microSD			

Tabla A.1: Componentes hardware.

Durante el año 2014 se realizaron las primeras pruebas de estanqueidad en mar abierto; el dispositivo se fondeó en el parque natural de Cabrera en tandas de cuatro en cuatro meses (se han realizado tres fondeos hasta la fecha). Estas primeras pruebas en condiciones reales han permitido no sólo comprobar la viabilidad física del proyecto (profundidad, vacío, flotabilidad) sino además comprobar que se cumplen los objetivos de computación (cantidad de horas de grabación, frecuencia de muestreo, sensibilidad de la señal grabada) pudiendo así comparar la calidad de la señal grabada con los otros dispositivos existentes en el mercado. Se pretende seguir con estos fondeos todavía durante unos meses más, hasta tener un control pleno del sistema en condiciones de funcionamiento reales. En la Fig. A.1 se muestra una foto del SAMARUC.

Se considera fundamental continuar con estos fondeos, al menos dos más, para seguir optimizando algunos de los parámetros; por ejemplo, la disminución de la frecuencia de muestreo supone una bajada en la calidad de la señal (reduce el ancho de banda), pero a cambio permite alargar la duración de la autonomía del dispositivo. Adquirir el control absoluto del comportamiento del equipamiento bajo determinadas condiciones es crucial

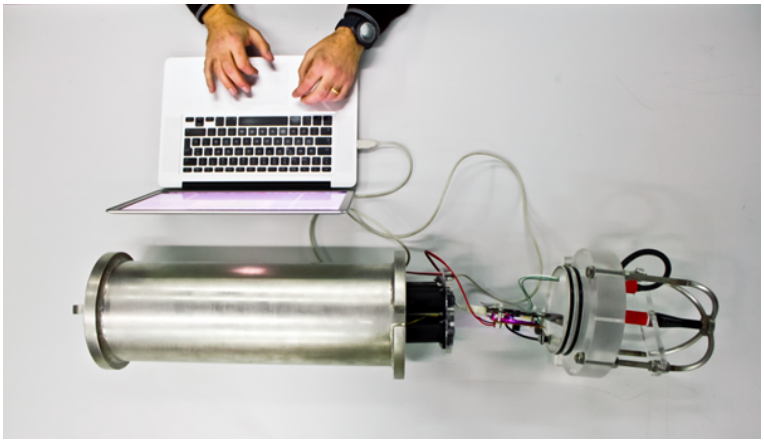


Figura A.1: Detalle del dispositivo de monitorización acústica pasiva SAMARUC mientras se realizan labores de programación de software en su electrónica.

para el futuro disfrute de su versatilidad en muy distintas aplicaciones.

Tal y como muestra la Fig. A.2 este hito supone un punto de inflexión entre el trabajo de investigación desarrollado en la universidad y su transferencia a la industria medioambiental. Básicamente, se ha comprobado que disponemos de un producto que satisface las necesidades del sector medioambiental que hasta ahora se conformaba con realizar un procesado de los datos a posteriori, sin ningún software adaptado y pagando un precio excesivo por dispositivos PAM nada actualizados a las nuevas herramientas disponibles en el mercado. La clave: la combinación entre un software que permite un análisis sencillo, rápido y transparente al usuario, con un dispositivo PAM que permita su introducción al medio acuático y su correcta aplicación. El resultado ha sido un sistema que consideramos podría ser convertido en un producto con un posible interés comercial y con facilidad de expansión a diferentes mercados.

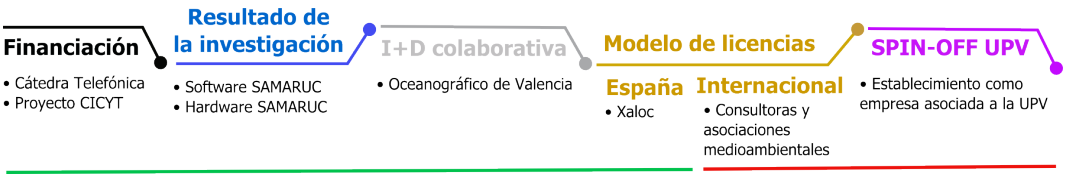


Figura A.2: Cronograma del origen de la tecnología y de las acciones futuras. Las líneas horizontales verde y roja marcan respectivamente los pasos que se han dado y los que quedan por dar.

Durante estos últimos cuatro años, esta línea de investigación se ha consolidado en el GTS lo cual ha permitido seguir avanzando tanto a nivel científico, con el estudio de los mecanismos de producción acústica de mamíferos submarinos (fruto de la cual ha surgido la presente tesis doctoral), como a nivel técnico, con el desarrollo de una nueva tecnología que permite ampliar las fronteras del sector medioambiental marino. El dispositivo puede

ser empleado para el control del entorno marino y del medio ambiente a través de las señales acústicas de animales marinos y ruidos antropogénicos. Más concretamente, a continuación se enumeran las **aplicaciones** para las que se puede utilizar SAMARUC:

I. Detección y control de la presencia de prospecciones sísmicas no autorizadas.

II. Controlar el buen estado ambiental (nivel acústico) del medio marino (que incluye los vertidos de energía acústica) de obligado cumplimiento para los estados de la UE (Directiva 2008/56/CE sobre la Estrategia Marina).

III. Detección de microsismos como los producidos por la inyección de gas.

IV. Gestión portuaria y tráfico marítimo: estima del tráfico en puertos a través de la firma acústica de los diferentes buques.

V. Caracterización de la biodiversidad de una zona marina: identificación de las diferentes especies a través de sus sonidos.

VI. Detección del paso de cetáceos, control de flujos y establecimiento de patrones migratorios.

VII. Estima de la densidad de individuos de una especie a través de sus sonidos (requiere software de terceras partes).

VIII. Control del ruido antropogénico y del nivel acústico submarino en piscifactorías. Estudios sobre la influencia de éste en algunas especies como el atún rojo.

Además, puede ser fácilmente extendido a diferentes aplicaciones a mercados ajenos al sector medioambiental. Entre las posibles aplicaciones se encuentra la detección acústica de fugas: detección de pequeñas fugas en un red de canalización o tanque de agua.

El software desarrollado en la tesis doctoral ha sido adaptado para la detección de eventos acústicos submarinos, tanto para la detección de cetáceos, como para la detección de diferentes tipos de ruidos. Tiene la posibilidad de ser empleado tanto dentro del SAMARUC o en un ordenador personal (versión PC) para aprovechar la necesidad que tienen muchos clientes de analizar datos que ya poseen y así obtener resultados de manera sencilla. Esta versión PC nos ofrece la posibilidad de rentabilizar el software diseñado. En ambos casos su uso resulta sencillo para el cliente final, proporcionando resultados concretos y rápidos sin necesidad de ningún conocimiento técnico. La Fig. A.3 muestra un ejemplo de los resultados que puede proporcionar el sistema.

Como ya se ha comentado, los equipos PAM que Xaloc utilizaba eran mejorables. El nuevo producto diseñado, SAMARUC, permite a esta y otras instituciones la extracción directa desde el dispositivo de los ficheros de datos grabados ya procesados, es decir, detectados y clasificados cada uno de los eventos ocurridos en el intervalo de grabación. Esta tarea no sólo facilita la tarea posterior de identificación de resultados sino que además simplifica la labor de almacenaje de gran cantidad de datos, un problema real en la actualidad con la cantidad de información de la que se dispone.

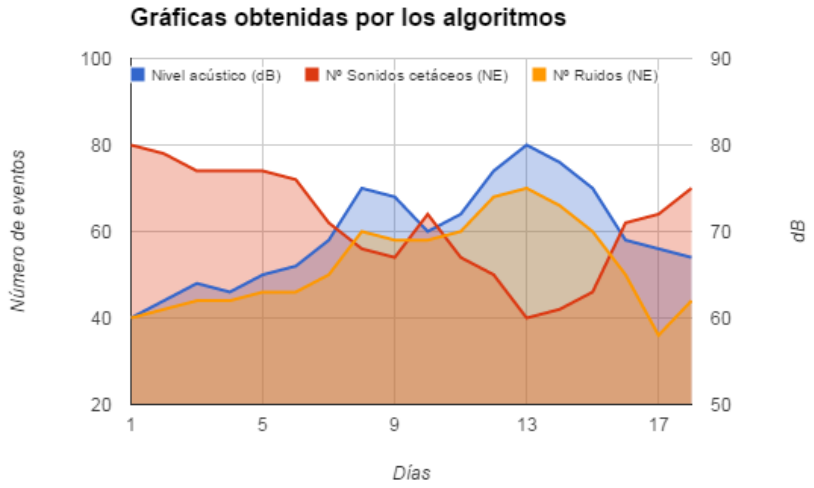


Figura A.3: Ejemplo de las gráficas obtenidas por los algoritmos desarrollados.

La tecnología que aquí se presenta es innovadora respecto a lo existente en el mercado, deja de ser un dispositivo pasivo (que sólo graba) e incorpora algoritmos de procesado embebidos en el dispositivo de almacenaje. Este paso supone un cambio sustancial en los dispositivos de monitorización acústica al poder incorporarse en ellos los avances hechos en el procesado de la señal. No sólo en las técnicas de detección y clasificación sino también los avances en nuevos dispositivos electrónicos cada vez más potentes y versátiles. Este es uno de los principales objetivos presentes desde el primer momento, lograr una tecnología que fuese transparente al usuario final ya que somos conscientes de que nuestros clientes finales son y pueden ser profesionales ajenos a la informática.

A.3. La evolución de SAMARUC

A nivel científico-técnico se pretende continuar con el estudio de las especificaciones técnicas de SAMARUC en su conjunto. Es necesario comprobar todas sus características tanto a nivel de hardware como de software en un despliegue real como son los fondeos que se vienen haciendo en el Parque Natural de Cabrera.

Campaña de medidas en los parques Nacionales de Cabrera y Denia (Cabo de San Antonio). Durante todo el año que dure la prueba de concepto los biólogos e investigadores del Oceanográfico continuarán empleando la unidad SAMARUC de gran profundidad en los parques naturales de Cabrera y en el área marítima del cabo de San Antonio (Denia). Esto incluye la solicitud de permisos, transporte de las unidades de SAMARUC hasta el lugar de destino, alquiler del los barcos y elementos necesarios para realizar el fondeo (boyas de profundidad, eslabones de liberación, etc.). El objetivo de esta tarea es doble: por un lado el obtener fotos, material audiovisual y señales para la web del sistema

SAMARUC y por otro el que los biólogos de estos centros de investigación publiquen más trabajos en los que se dé a conocer el sistema.

El contenido mostrado en este apéndice ha generado la publicación del siguiente artículo en revista:

- Ramón Miralles Ricós, Guillermo Lara Martínez, Alicia Carrión Garcia, Jose Antonio Esteban Simón. "SAMARUC a Programmable system for Passive acoustic monitoring of cetaceans". Waves. Volumen 5. Páginas 69-79. Año 2014.

y la participación en el siguiente congreso:

- Ramón Miralles Ricós, Jose Antonio Esteban Simón, Paula Alonso, Jose Amengual, Beatriz Ramos Lopez y Guillermo Lara Martínez. "On the characterization of Seismic Airgun detonation underwater sounds". OCEANNOISE Vilanova i la Geltrú, Barcelona, España. Año 2015.

Bibliografía

[1] F. Itakura, "Minimum prediction residual principle applied to speech recognition," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 23, pp. 67–72, Feb. 1975.

[2] O. Pierre-Yves, "The production and recognition of emotions in speech: features and algorithms," *International Journal of Human-Computer Studies*, vol. 59, pp. 157–183, July 2003.

[3] M. Kudo and J. Sklansky, "Comparison of algorithms that select features for pattern classifiers," *Pattern Recognition*, vol. 33, pp. 25–41, Jan. 2000.

[4] M. Last, A. Kandel, and O. Maimon, "Information-theoretic algorithm for feature selection," *Pattern Recognition Letters*, vol. 22, pp. 799–811, May 2001.

[5] D. P. Muni, N. R. Pal, and J. Das, "Genetic programming for simultaneous feature selection and classifier design," *IEEE transactions on systems, man, and cybernetics. Part B, Cybernetics: a publication of the IEEE Systems, Man, and Cybernetics Society*, vol. 36, pp. 106–117, Feb. 2006.

[6] S. Nakariyakul and D. P. Casasent, "An improvement on floating search algorithms for feature subset selection," *Pattern Recognition*, vol. 42, pp. 1932–1940, Sept. 2009.

[7] J. Schenk, M. Kaiser, and G. Rigoll, "Selecting Features in On-Line Handwritten Whiteboard Note Recognition: SFS or SFFS?," in *10th International Conference on Document Analysis and Recognition, 2009. ICDAR '09*, pp. 1251–1254, July 2009.

[8] A. Marcano-Cedeño, J. Quintanilla-Domínguez, M. G. Cortina-Januchs, and D. Andina, "Feature selection using Sequential Forward Selection and classification applying Artificial Metaplasticity Neural Network," in *IECON 2010 - 36th Annual Conference on IEEE Industrial Electronics Society*, pp. 2845–2850, Nov. 2010.

[9] D. Heckerman and D. M. Chickering, "Learning Bayesian networks: The combination of knowledge and statistical data," in *Machine Learning*, pp. 20–197, 1995.

[10] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification, pp. 1-654.* John Wiley & Sons, INC, New York: Wiley-Interscience, 2 ed., Oct. 2000.

[11] C. M. Bishop, *Pattern Recognition and Machine Learning, pp. 142-146.* Springer, 1 ed., 2006.

[12] D. Gillespie, D. K. Mellinger, J. Gordon, D. McLaren, P. Redmond, R. McHugh, P. Trinder, X.-Y. Deng, and A. Thode, "PAMGUARD: Semiautomated, open source software for realtime acoustic detection and localization of cetaceans.," *The Journal of the Acoustical Society of America*, vol. 125, pp. 2547–2547, Apr. 2009.

[13] D. K. Mellinger, "Acoustic feature extraction and classification in ishmael," *The Journal of the Acoustical Society of America*, vol. 134, pp. 3986–3986, Nov. 2013.

[14] D. S. Houser, D. A. Helweg, and P. W. Moore, "Classification of dolphin echolocation clicks by energy and frequency distributions," *The Journal of the Acoustical Society of America*, vol. 106, pp. 1579–1585, Sept. 1999.

[15] P. T. Madsen, I. Kerr, and R. Payne, "Echolocation clicks of two free-ranging, oceanic delphinids with different food preferences: false killer whales pseudorca crassidens and risso's dolphins grampus griseus," *The Journal of experimental biology*, vol. 207, pp. 1811–1823, May 2004.

[16] J. A. Thomas, C. F. Moss, and M. Vater, *Echolocation in Bats and Dolphins.* University of Chicago Press, 2004.

[17] V. M. Janik, L. S. Sayigh, and R. S. Wells, "Signature whistle shape conveys identity information to bottlenose dolphins," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 103, pp. 8293–8297, May 2006.

[18] L. S. Sayigh, H. C. Esch, R. S. Wells, and V. M. Janik, "Facts about signature whistles of bottlenose dolphins, tursiops truncatus," *Animal Behaviour*, vol. 74, pp. 1631–1642, Dec. 2007.

[19] P. T. Madsen, M. Lammers, D. Wisniewska, and K. Beedholm, "Nasal sound production in echolocating delphinids (tursiops truncatus and pseudorca crassidens) is dynamic, but unilateral: clicking on the right side and whistling on the left side," *The Journal of experimental biology*, vol. 216, pp. 4091–4102, Nov. 2013.

[20] C. Nikias and A. Petropulu, *Higher-Order Spectra Analysis a Nonlinear Signal Processing Framework, pp. 123-138.* Englewood Cliffs, New Jersey: Prentice Hall, 1993.

[21] R. Miralles, L. Vergara, A. Salazar, and J. Igual, "Blind detection of nonlinearities in multiple-echo ultrasonic signals," *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, vol. 55, pp. 1–11, Aug. 2008.

[22] E. Verteletskaya, K. Sakhnov, and B. Simak, "Pitch detection algorithms and voiced/unvoiced classification for noisy speech," in *Systems, Signals and Image Processing, 2009. IWSSIP 2009. 16th International Conference*, pp. 1–5, Oct. 2009.

[23] J. N. W. Fitch and H. Herzel, "Calls out of chaos: the adadtative significance of nonlinear phenomena in mammalian vocal production," *Animal behaviour*, vol. 55, pp. 407–418, Aug. 2008.

[24] R. S. Payane and S. McVay, "Songs of humpback whales," *Science*, vol. 173, pp. 585–597, Aug. 1971.

[25] T. A. Wilson, G. S. Beavers, M. A. DeCoster, D. K. Holger, and M. D. Regenfuss, "Experiments on the fluid mechanics of whistling," *The Journal of the Acoustical Society of America*, vol. 50, pp. 366–372, July 1971.

[26] C. W. Turl and R. H. Penner, "Target detection: Beluga whale and bottlenose dolphin echolocation abilities compared," *The Journal of the Acoustical Society of America*, vol. 74, pp. S74–S74, Nov. 1983.

[27] W. W. L. Au, D. A. Carder, R. H. Penner, and B. L. Scronce, "Demonstration of adaptation in beluga whale echolocation signals," *The Journal of the Acoustical Society of America*, vol. 77, no. 2, p. 726, 1985.

[28] D. Bernoulli, "Hidrodynamics," in *Hidrodynamics*, 1738.

[29] B. L. Lonsbury-Martin, F. P. Harris, B. B. Stagner, M. D.Hawkins, and G. K. Martin, "Distortion product emissions in humans. I. Basic properties in normally hearing subjects," *The Annals of Otology, Rhinology & Laryngology. Supplement*, vol. 147, pp. 3–14, May 1990.

[30] S. H. Ridgway, D. A. Carder, R. F. Green, A. S. Gaunt, S. L. L. Gaunt, and W. E. Evans, "Electromyographic and pressure events in the nasolaryngeal system of dolphins during sound production," in *Animal Sonar Systems* (R.-G. Busnel and J. F. Fish, eds.), no. 28 in NATO Advanced Study Institutes Series, pp. 239–249, Springer US, Jan. 1980.

[31] R. S. Mackay and H. M. Liaw, "Dolphin vocalization mechanisms," *Science (New York, N.Y.)*, vol. 212, pp. 676–678, May 1981.

[32] M. Amundin and S. H. Andersen, "Bony nares air pressure and nasal plug muscle activity during click production in the harbour porpoise, phocoena phocoena, and the bottlenosed dolphin, tursiops truncatus," *Journal of Experimental Biology*, vol. 105, pp. 275–282, July 1983.

[33] S. H. Ridgway and D. A. Carder, "Nasal pressure and sound production in an echolocating white whale, delphinapterus leucas," in *Animal Sonar* (P. E. Nachtigall and P. W. B. Moore, eds.), no. 156 in NATO ASI Science, pp. 53–60, Springer US, Jan. 1988.

[34] K. Norris, *The evolution of acoustic mechanisms in odontocete cetaceans.* Envolution and Environment, New Haven: Yale University Press, e. t. drake ed., 1968.

[35] T. W. Cranford, M. Amundin, and K. S. Norris, "Functional morphology and homology in the odontocete nasal complex: implications for sound generation," *Journal of morphology*, vol. 228, pp. 223–285, June 1996.

[36] M. Johnson, P. T. Madsen, W. M. X. Zimmer, N. A. de Soto, and P. L. Tyack, "Foraging blainville's beaked whales (mesoplodon densirostris) produce distinct click types matched to different phases of echolocation," *The Journal of experimental biology*, vol. 209, pp. 5038–5050, Dec. 2006.

[37] P. W. Moore, L. A. Dankiewicz, and D. S. Houser, "Beamwidth control and angular target detection in an echolocating bottlenose dolphin (tursiops truncatus)," *The Journal of the Acoustical Society of America*, vol. 124, pp. 3324–3332, Nov. 2008.

[38] K. S. Norris, "Some problems of echolocation in cetaceans," 1964.

[39] K. J. Diercks, R. T. Trochta, C. F. Greenlaw, and W. E. Evans, "Recording and analysis of dolphin echolocation signals," *The Journal of the Acoustical Society of America*, vol. 49, pp. 135–135, Jan. 1971.

[40] K. J. Dormer, "Mechanism of sound production and air recycling in delphinids: Cineradiographic evidence," *The Journal of the Acoustical Society of America*, vol. 65, pp. 229–239, Jan. 1979.

[41] J. G. Mead, "Anatomy of the external nasal passages and facial complex in the delphinidae (mammalia: Cetacea)," 1975.

[42] W. E. Evans and J. H. Prescott, "Observations of the sound production capabilities of the bottlenose porpoise: a study of whistles and clicks.," in *Zoologica New York*, 47, pp. 121–128, 1962.

[43] J. C. Lilly, "Vocal behavior of the bottlenose dolphin," *Proceedings of the American Philosophical Society*, vol. 106, pp. 520–529, Dec. 1962.

[44] J. E. Heyning, *Comparative facial anatomy of beaked whales (Ziphiidae) and a systematic revision among the families of extant Odontoceti.* UCLA, 1986.

[45] P. T. Madsen, F. H. Jensen, D. Carder, and S. Ridgway, "Dolphin whistles: a functional misnomer revealed by heliox breathing," *Biology letters*, vol. 8, pp. 211–213, Apr. 2012.

[46] M. Amundin, "Helium effects on the click frequency spectrum of the harbor porpoise, phocoenaphocoena," *The Journal of the Acoustical Society of America*, vol. 90, pp. 53–59, July 1991.

[47] "The sources of sound in birdsong," in *The Physics of Birdsong*, Biological and Medical Physics, Biomedical Engineering, pp. 47–60, Springer Berlin Heidelberg, Jan. 2005.

[48] "Complex oscillations," in *The Physics of Birdsong*, Biological and Medical Physics, Biomedical Engineering, pp. 79–97, Springer Berlin Heidelberg, Jan. 2005.

[49] "Anatomy of the vocal organ," in *The Physics of Birdsong*, Biological and Medical Physics, Biomedical Engineering, pp. 37–46, Springer Berlin Heidelberg, Jan. 2005.

[50] S. Nowicki, "Vocal tract resonances in oscine bird sound production: evidence from birdsongs in a helium atmosphere," *Nature*, vol. 325, pp. 53–55, Jan. 1987.

[51] M. O. Lammers and M. Castellote, "The beluga whale produces two pulses to form its sonar signal," *Biology Letters*, vol. 5, pp. 297–301, June 2009.

[52] P. T. Madsen, D. Wisniewska, and K. Beedholm, "Single source sound production and dynamic beam formation in echolocating harbour porpoises (phocoena phocoena)," *The Journal of experimental biology*, vol. 213, pp. 3105–3110, Sept. 2010.

[53] W. E. Evans, "Echolocation by marine delphinids and one species of freshwater dolphin," *The Journal of the Acoustical Society of America*, vol. 54, pp. 191–199, July 1973.

[54] J. C. Lilly, *Communication Between Man and Dolphin.* New York, N.Y: Julian Press, May 1987.

[55] R. L. Brill and P. J. Harder, "The effects of attenuating returning echolocation signals at the lower jaw of a dolphin (tursiops truncatus)," *The Journal of the Acoustical Society of America*, vol. 89, pp. 2851–2857, June 1991.

[56] S. O. Murray, E. Mercado, and H. L. Roitblat, "Characterizing the graded structure of false killer whale (pseudorca crassidens) vocalizations," *The Journal of the Acoustical Society of America*, vol. 104, pp. 1679–1688, Sept. 1998.

[57] G. A. Campbell and R. M. Foster, *Fourier integrals for practical applications.* American Telephone and Telegraph Company. Technical publications. Mathematical physics. MonographB-584, New York: Bell telephone laboratories], 1931.

[58] E. M. Haacke and J. L. Patrick, "Reducing motion artifacts in two-dimensional fourier transform imaging," *Magnetic Resonance Imaging*, vol. 4, no. 4, pp. 359–376, 1986.

[59] E. P. Paschalis, O. Jacenko, B. Olsen, R. Mendelsohn, and A. L. Boskey, "Fourier transform infrared microspectroscopic analysis identifies alterations in mineral properties in bones from mice transgenic for type x collagen," *Bone*, vol. 19, pp. 151–156, Aug. 1996.

[60] P. Athanas and A. Abbott, "Real-time image processing on a custom computing platform," *Computer*, vol. 28, pp. 16–25, Feb. 1995.

[61] Q. Wang, Q. Guo, J. Zhou, and Q. Lin, "Nonlinear joint fractional fourier transform correlation for target detection in hyperspectral image," *Optics & Laser Technology*, vol. 44, pp. 1897–1904, Sept. 2012.

[62] L. Muda, B. K. M., and I. Elamvazuthi, "Voice recognition algorithms using mel frequency cepstral coefficient (MFCC) and dynamic time warping (DTW) techniques," *Journal of Computing*, vol. 2, pp. 138–143, Mar. 2010.

[63] R. N. Bracewell, *The Fourier Transform and Its Applications*. McGraw Hill, 2000.

[64] D. G. Childers, D. Skinner, and R. Kemerait, "The cepstrum: A guide to processing," *Proceedings of the IEEE*, vol. 65, pp. 1428–1443, Oct. 1977.

[65] S. Ahmadi and A. Spanias, "Cepstrum-based pitch detection using a new statistical v/UV classification algorithm," *IEEE Transactions on Speech and Audio Processing*, vol. 7, pp. 333–338, May 1999.

[66] A. M. Noll, "Shorttime spectrum and cepstrum techniques for vocalpitch detection," *The Journal of the Acoustical Society of America*, vol. 36, pp. 296–302, Feb. 1964.

[67] X. C. Halkias and D. P. W. Ellis, "Estimating the number of marine mammals using recordings of clicks from one microphone," in *In Proc. ICASSP-06*, 2006.

[68] A. Oppenheim and R. Schafer, "From frequency to quefrency: a history of the cepstrum," *IEEE Signal Processing Magazine*, vol. 21, pp. 95–106, Sept. 2004.

[69] M. J. Owren and C. D. Linker, "Some analysis methods that may be useful to acoustic primatologists," in *Current Topics in Primate Vocal Communication* (E. Zimmermann, J. D. Newman, and U. Jürgens, eds.), pp. 1–27, Springer US, Jan. 1995.

[70] G. Manteuffel, B. Puppe, and P. C. Schön, "Vocalization of farm animals as a measure of welfare," *Applied Animal Behaviour Science*, vol. 88, pp. 163–182, Sept. 2004.

[71] G. Mehraei, F. J. Gallun, M. R. Leek, and J. G. W. Bernstein, "Spectrotemporal modulation sensitivity for hearing-impaired listeners: Dependence on carrier center frequency and the relationship to speech intelligibility," *The Journal of the Acoustical Society of America*, vol. 136, pp. 301–316, July 2014.

[72] H. Ding, B. Qian, Y. Li, and Z. Tang, "A Method Combining LPC-Based Cepstrum and Harmonic Product Spectrum for Pitch Detection," in *2006 International Conference on Intelligent Information Hiding and Multimedia*, pp. 537–540, Dec. 2006.

[73] X. Sun, "Pitch determination and voice quality analysis using Subharmonic-to-Harmonic Ratio," in *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 1, pp. I–333–I–336, May 2002.

[74] T. Riede, H. Herzel, K. Hammerschmidt, L. Brunnberg, and G. Tembrock, "The harmonic-to-noise ratio applied to dog barks," *The Journal of the Acoustical Society of America*, vol. 110, no. 4, p. 2191, 2001.

[75] R. Miralles, A. Carrion, D. Looney, G. Lara, and D. P. Mandic, "Characterization of the complexity in short oscillating time series: An application to seismic airgun detonations," pp. 1–15, May 2015.

[76] G. M. Mindlin and R. Gilmore, "Topological analysis and synthesis of chaotic time series," *Physica D: Nonlinear Phenomena*, vol. 58, pp. 229–242, Sept. 1992.

[77] M. Schroeder, "Vocoders: Analysis and synthesis of speech," *Proceedings of the IEEE*, vol. 54, pp. 720–734, May 1966.

[78] N. Jayant, "Digital coding of speech waveforms: PCM, DPCM, and DM quantizers," *Proceedings of the IEEE*, vol. 62, pp. 611–632, May 1974.

[79] M. R. Schroeder and B. Atal, "Code-excited linear prediction(CELP): High-quality speech at very low bit rates," in *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '85.*, vol. 10, pp. 937–940, Apr. 1985.

[80] E. Parmentier, J.-P. Lagardère, Y. Chancerelle, D. Dufrane, and I. Eeckhaut, "Variations in sound-producing mechanism in the pearlfish carapini (carapidae)," *Journal of Zoology*, vol. 276, pp. 266–275, Nov. 2008.

[81] W. H. Thorpe, "The learning of song patterns by birds, with especial reference to the song of the chaffinch fringilla coelebs," *Ibis*, vol. 100, pp. 535–570, Oct. 1958.

[82] R. W. Warner, "The anatomy of the syrinx in passerine birds," *Journal of Zoology*, vol. 168, pp. 381–393, Nov. 1972.

[83] A. S. Gaunt and S. Nowicki, "Sound production in birds: Acoustics and physiology revisited," in *Animal Acoustic Communication* (D. S. L. Hopp, D. M. J. Owren, and D. C. S. Evans, eds.), pp. 291–321, Springer Berlin Heidelberg, Jan. 1998.

[84] J. D. Skinner and C. T. Chimimba, *The Mammals of the Southern African Subregion*. Cambridge University Press, Nov. 2005.

[85] *Mammalogy: Adaptation, Diversity, Ecology*. JHU Press, Sept. 2007.

[86] D. E. F. Sissom, D. A. Rice, and G. Peters, "How cats purr," *Journal of Zoology*, vol. 223, pp. 67–78, Jan. 1991.

[87] J. E. Remmers and H. Gautier, "Neural and mechanical mechanisms of feline purring," *Respiration Physiology*, vol. 16, pp. 351–361, Dec. 1972.

[88] P. R. Marler and H. Slabbekoorn, *Nature's Music: The Science of Birdsong*. Academic Press, Oct. 2004.

[89] M. Castellote and F. Fossa, "Measuring acoustic activity as a method to evaluate welfare in captive beluga whales (delphinapterus leucas)," *Aquatic Mammals*, vol. 32, pp. 325–333, Sept. 2006.

[90] S. King, "An introduction to statistical parametric speech synthesis," *Sadhana*, vol. 36, pp. 837–852, Oct. 2011.

[91] D. W. Linzey, *Vertebrate Biology*. JHU Press, Dec. 2011.

[92] M. Greibus and L. Telksnys, "Segmentation analysis using synthetic speech signals," *Electronics and Electrical Engineering*, vol. 18, Oct. 2012.

[93] T. W. Cranford, "In search of impulse sound sources in odontocetes," in *Hearing by Whales and Dolphins* (W. W. L. Au, R. R. Fay, and A. N. Popper, eds.), no. 12 in Springer Handbook of Auditory Research, pp. 109–155, Springer New York, Jan. 2000.

[94] J. A. Thomas, C. F. Moss, and M. Vater, *Echolocation in Bats and Dolphins*. University of Chicago Press, 2004.

[95] B. Atal and M. R. Schroeder, "Predictive coding of speech signals and subjective error criteria," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 27, pp. 247–254, June 1979.

[96] B. Atal and J. Remde, "A new model of LPC excitation for producing natural-sounding speech at low bit rates," in *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '82.*, vol. 7, pp. 614–617, May 1982.

[97] J. Xu, L. Xing, A. Perkis, and Y. Jiang, "On the properties of mean opinion scores for quality of experience management," in *2011 IEEE International Symposium on Multimedia (ISM)*, pp. 500–505, Dec. 2011.

[98] L. Malfait, J. Berger, and M. Kastner, "P.563 amp;#8212; the itu-t standard for single-ended speech quality assessment," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, pp. 1924–1934, Nov. 2006.

[99] "Wideband extension to recommendation p.862 for the assessment of wideband telephone networks and speech codecs," *International Telecommunication Union, ITU-T Rec. P.862.2, CH-Genova*, 2005.

[100] J. B. Alonso, J. de Leon, I. Alonso, and M. A. Ferrer, "Automatic detection of pathologies in the voice by HOS based parameters," *EURASIP J. Appl. Signal Process.*, vol. 2001, p. 275284, Dec. 2001.

[101] C. R. Norrenbrock, U. Heute, and F. Hinterleitner, "On the use of vocal-tract approximations for instrumental quality assessment," in *ITG Symposium; Proceedings of Speech Communication; 10*, pp. 1–4, Sept. 2012.

[102] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, pp. 600–612, Apr. 2004.

[103] S. Kandadai, J. Hardin, and C. D. Creusere, "Audio quality assessment using the mean structural similarity measure," in *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 221–224, Mar. 2008.

[104] M. Cooper and J. Foote, "Summarizing popular music via structural similarity analysis," in *Applications of Signal Processing to Audio and Acoustics, 2003 IEEE Workshop on.*, pp. 127–130, Oct. 2003.

[105] A. Perovi, Z. orevi, M. Paskota, A. Takai, and A. Jovanovi, "Automatic recognition of features in spectrograms based on some image analysis methods," *Acta Polytechnica Hungarica*, vol. 10, no. 2, pp. 153–172, 2013.