UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

FACULTAD DE INGENIERIA

SISTEMA DE IDENTIFICACIÓN AUTOMÁTICA DEL LENGUAJE HABLADO EN ARCHIVOS MULTIMEDIA DE VOZ

TESIS

QUE PARA OBTENER EL TÍTULO DE:

INGENIERO EN TELECOMUNICACIONES

PRESENTA:

OLVERA ZAMBRANO, MAURICIO MICHEL

ASESOR: ESCOBAR SALGUERO, LARRY HIPÓLITO

Ciudad Universitaria, CD. MX 2017

Sistema de Identificación Automática del Lenguaje Hablado en Archivos Multimedia de Voz

Mauricio Michel Olvera Zambrano

16 de marzo de 2017

Agradecimientos

Primeramente a *Dios*, por darme la oportunidad de vivir y por estar conmigo en cada paso que doy, por fortalecer mi corazón e iluminar mi mente y por haber puesto en mi camino a aquellas personas que han sido mi soporte y compañía durante toda mi vida.

A *mis padres* por ser el pilar fundamental en todo lo que soy, en toda mi educación, tanto académica, como de la vida, por su incondicional apoyo perfectamente mantenido a través del tiempo.

A la *Universidad Nacional Autónoma de México*, mi alma máter, por abrirme sus puertas para ser mejor persona y un buen profesional, así como por darme la oportunidad de brincar el charco y cumplir mi sueño de estudiar en *Purdue University*, donde viví una de las mejores experiencias de mi vida y que llevaré grabada siempre en mi corazón.

A mi director de tesis *Larry Escobar*, por todas las oportunidades que me ha brindado y por estos dos años de trabajo conjunto, tiempo en el cual se ha tomado el arduo trabajo de transmitirme sus diversos conocimientos, orientaciones y consejos. Su persistencia, paciencia y motivación han sido fundamentales para mi formación como profesional.

Al *Dr. Rodolfo Neri Vela*, a quien admiro inmensamente y quién desde la escuela secundaria me inspiró a estudiar Ingeniería en Telecomunicaciones.

A mis profesores: *Mauricio Ortega*, *Bohumil Psenicka*, *Chih-Chun Wang* y *Mirelle Boutin*, quienes con su gran pasión por la docencia me introdujeron al fascinante mundo del Procesamiento Digital de Señales.

A mis hermanas, *Vyvyan* y *Nuria*, por estar conmigo y apoyarme siempre, las quiero mucho.

A mis sobrinos, *Emiliano* y *Santiago*, para que vean en mí un ejemplo a seguir.

A *Celina*, *Fátima*, *Erick* y *Alfredo*, por todos estos años de amistad y que aunque muchas veces lejos, siempre están cerca de mí.

A mis compañeros y excompañeros del *Laboratorio de Procesamiento Digital de Señales*: *José*, *Alfonso*, *Mickey*, *Nacho*, *Yolo* y especialmente a *Iván*, por ayudarme en la obtención de los modelos computacionales necesarios para la implementación de mi sistema; a *Luis*, por revisar mi

trabajo escrito y corregirlo para mejorarlo; y a *Samuel*, por brindarme su amistad y confiar en mí para emprender proyectos juntos.

A *Mirian Catarino Urcino*, por obsequiarme todo ese amor que me motiva y me alienta a seguir siempre adelante. Este logro también es tuyo. ¡Te amo!

Al proyecto PAPIME PE100616 *Servidor para Prácticas de Procesamiento Digital de Señales en Tiempo Real* que me permitió realizar los procesos computacionales fundamentales para el desarrollo de este trabajo.

Resumen

En este trabajo se presenta una extensa investigación sobre la forma en que el ser humano es capaz de producir y percibir la voz, y cómo es que a través del procesamiento digital de señales, la teoría de la probabilidad y el aprendizaje automático se puede replicar mediante una computadora la habilidad que los humanos tienen para identificar el idioma hablado en un segmento de voz, para lo cual se diseña e implementa un sistema automático de identificación de idiomas capaz de detectar el idioma hablado en un archivo multimedia de voz, dentro del siguiente conjunto de idiomas: inglés, español, francés, alemán, ruso y japonés.

Abstract

In this work I present an extensive research on the way in which human beings are able to produce and perceive speech, and how it is that through digital signal processing, probability theory and selected topics in machine learning, the human ability to identify the language of a spoken utterance can be replicated using a computer. Therefore I design and implement an automatic language identification system capable of identifying which of the following six languages: english, spanish, french, german, russian and japanese, is being spoken within a multimedia voice file.

Contenido

Tablas

Figuras

1

Introducción

La *identificación automática del lenguaje hablado* (LID, por sus siglas en inglés) consiste en reconocer a través de un proceso computacional automático, el idioma que se habla en un segmento digital de voz. Desde 1970 se han realizado investigaciones en esta área y a partir de entonces los sistemas LID han avanzado en complejidad. Entre algunas de las aplicaciones más importantes de dichos sistemas se encuentran los sistemas multilingües de diálogo hablado, implementados en quioscos de información localizados en aeropuertos internacionales y lugares turísticos; los sistemas de traducción multilingüe, que tienen como interfaz un sistema LID cuya entrada de voz puede estar en varios idiomas; los sistemas de indexación, búsqueda y clasificación de archivos de datos de audio y corpora de voz, los cuales se conforman de múltiples idiomas y se encuentran alojados en bases de datos muy grandes; e incluso en compañías telefónicas y centros de llamadas también se suelen utilizar sistemas LID para atender llamadas entrantes de usuarios que hablan lenguas extrajeras y transferirlas a operadores que hablan su mismo idioma.

1.1. Objetivo

Realizar una investigación extensiva acerca de la forma en que el ser humano produce y percibe la voz, con la finalidad de diseñar e implementar a partir de la teoría del procesamiento digital de señales, la teoría de la probabilidad y algunos temas selectos del aprendizaje automático, un sistema capaz de identificar de manera automática el idioma hablado en un segmento digital de voz, dentro del siguiente conjunto de idiomas: inglés, español, francés, alemán, ruso y japonés.

1.2. Método

Los sistemas de reconocimiento del lenguaje hablado son usualmente categorizados en diferentes enfoques, cada uno de los cuales emplea diversas características para la discriminación de idiomas, tales como las acústicas y fonéticas, como pueden ser los componentes espectrales o el inventario fonológico de cada idioma; las fonotácticas, que definen la estructura silábica permisible del idioma; las prosódicas, como la duración, el tono y entonación; y las léxicas, que abarcan características de las palabras y sintaxis.

Debido a que las investigaciones en la identificación automática de idiomas han confirmado que tanto las características acústicas, como las fonotácticas son las referencias del lenguaje más efectivas, este trabajo de tesis pretende la creación de un sistema LID basado exclusivamente en las características acústicas del habla, ya que a través de este enfoque es posible obtener buenos resultados de identificación empleando tiempos cortos de procesamiento.

Los repertorios fonéticos de los idiomas difieren significativamente entre uno y otro, aunque existen fonemas afines a distintos idiomas. Las diferencias entre los repertorios fonéticos implican que cada idioma tiene un conjunto único de fonemas y por consiguiente su propia distribución de características acústicas y fonéticas.

Las características acústicas y fonéticas de la señal de voz son extraídas a través de diversas técnicas, las más comunes se basan en:

- El análisis de los coeficientes de predicción lineal (LPC).
- El análisis de los coeficientes cepstrales de frecuencia Mel (coeficientes MFCC).

Como métodos de modelado y clasificación de las características acústicas se encuentran:

- La cuantización vectorial (VQ).
- Los modelos ocultos de Markov (HMM).
- Las máquinas de soporte vectorial (SVM).
- Los modelos de mezclas Gaussianas (GMM).

En un principio, los coeficientes LPC y la cuantización vectorial fueron las técnicas utilizadas para la extracción y clasificación de las características acústicas del habla en los primeros sistemas LID, y con la evolución de dichos sistemas, otras técnicas de extracción de características

y clasificación fueron adoptadas y también combinadas para realizar el modelado acústico de idiomas.

En este trabajo se emplearán los coeficientes MFCC, así como los coeficientes cepstrales delta desplazados (SDC), los cuales son una extensión de los coeficientes MFCC, para la extracción de características acústicas de las señales de voz, así como los GMMs para la clasificación y modelado de las características acústicas de cada idioma. En la actualidad los coeficientes MFCC, SDC, y los GMMs representan el estado del arte en los sistemas de identificación automática de idiomas y por consiguiente son las técnicas de extracción y modelado más ampliamente utilizadas en los sistemas LID que siguen un enfoque acústico.

1.3. Organización del Trabajo

En el capítulo 2: *Producción y Percepción de Voz*, se presentan los aspectos fisiológicos más importantes acerca de la manera en que el ser humano produce y percibe la voz, con la finalidad de caracterizar el sistema de producción de voz a través de un enfoque fuente-filtro y entender cómo son procesadas por el oído las ondas de presión acústica.

En el capítulo 3: *Caracterización Acústico-Fonética de la Señal de Voz*, se describen las características acústicas y fonéticas más importantes presentes en la señal de voz. También se describen los conceptos más importantes relacionados con su representación en el dominio de la frecuencia, tales como la frecuencia fundamental, los formantes de voz y el espectrograma.

En el capítulo 4: *Procesamiento Digital de Señales de Voz*, se aborda la teoría fundamental de señales y sistemas, en la cual se describen las técnicas matemáticas más importantes que permiten manipular señales digitales en distintos dominios. Particularmente se hace énfasis en el dominio de la frecuencia, donde se presenta de manera extensa un estudio de la teoría de Fourier.

En el capítulo 5: *Modelado del Lenguaje Hablado*, se presenta una breve introducción al aprendizaje automático y se describen los conceptos matemáticos más importantes de la teoría de probabilidad. Además se realiza un estudio detallado de la distribución Gaussiana, a partir de la cual se introducen los modelos de mezclas Gaussianas, útiles para representar características de los fenómenos físicos del mundo real.

En el capítulo 6: *Diseño e Implementación del sistema LID*, se conjunta la información de los capítulos anteriores para realizar el diseño y la implementación de un sistema capaz de identificar de forma automática el idioma hablado en segmentos de voz. Se presenta el diagrama de bloques de un sistema LID cuyo enfoque es puramente acústico, describiendo de manera detallada cada una de las fases que lo conforman.

En el capítulo 7: *Pruebas y Resultados*, se presentan dos configuraciones del sistema LID: un sistema dependiente del género del hablante y otro completamente independiente de éste. Se detalla la forma en que se generan los modelos acústicos de los idiomas y cuáles son sus eficiencias. Además, se exponen los resultados obtenidos por cada una de las configuraciones del sistema LID y se realiza una comparación entre éstos.

Finalmente, en el capítulo 8: *Conclusiones*, se presentan las conclusiones con base en los resultados obtenidos.

2

Producción y Percepción de la Señal de Voz

Debido a la naturalidad con la que todos los días el ser humano utiliza el lenguaje hablado para comunicarse, es común que los procesos de producción y percepción de la voz los realice de manera casi inconsciente, sin embargo, los movimientos articulatorios de los labios, lengua, cuerdas vocales, y otros órganos del sistema vocal, así como los movimientos físicos de los sistemas mecánicos que conforman el sentido del oído, se encuentran entre los más sutiles de todas las acciones realizadas por el ser humano en su día a día. Actualmente, las máquinas con la habilidad de no sólo entender voz, sino también de comunicarse mediante ésta, se componen de tecnologías de reconocimiento y síntesis de voz sofisticadas que han contribuido a que el desarrollo de interfaces hombre-máquina se simplifique y sean cada vez más natural. Sin embargo, para conseguir funciones que sean incluso más cercanas a las realizadas por los seres humanos, se debe conocer más sobre la forma de operación de los mecanismos a través de los cuales la voz es producida y percibida, para así desarrollar nuevas tecnologías que hagan uso eficiente de estas funciones. El entendimiento de los mecanismos en los cuales se basa la voz, ayuda a clarificar cómo es que el cerebro humano procesa la información y cómo la tecnología puede ser capaz de desarrollar dispositivos que imiten este procesamiento.

2.1. Producción de Voz

El proceso de producción de voz comienza desde el momento en el que una persona formula una idea en su mente acerca de lo que quiere transmitir mediante el habla. Luego, la idea generada es trasladada al aparato vocal a través de nervios sensoriales, que mediante la ejecución de una serie de instrucciones neuromusculares por parte del hablante, provocan que las cuerdas vocales vibren cuando sea debido y también que le den forma al tracto vocal, de tal manera que una secuencia apropiada de sonidos de voz sea creada y hablada por el locutor, produciendo como resultado final una onda de presión acústica que es propagada a través del aire [1].

2.1.1. Articuladores

Los *articuladores* son aquellos órganos del cuerpo humano involucrados en la producción de voz, órganos que fundamentalmente desempeñan funciones biológicas primarias que facilitan la respiración y la alimentación, y que como parte del proceso evolutivo del ser humano adquirieron la habilidad de articular sonidos, es decir, adoptaron de forma secundaria funciones lingüísticas que facilitaron la comunicación. Estos órganos son flexibles por naturaleza, por lo que su forma y tamaño cambian, implicando una serie de contracciones musculares que dan como resultado la gran variedad de articulaciones de los sonidos que el ser humano es capaz de producir mediante el habla. Su estudio es llevado a cabo por una rama particular de la Fonética llamada *Fonética Articulatoria*.

Los músculos en el pecho que el ser humano emplea para respirar producen el flujo de aire necesario para generar casi todos los sonidos del habla; por otra parte, los músculos en la laringe producen muchas modificaciones diferentes al flujo de aire que va desde el pecho hasta la boca. Después de pasar por la laringe, el aire viaja a través del tracto vocal, del cual escapa hacia la atmósfera por la boca o por las fosas nasales.

El ser humano posee un conjunto muy grande y complejo de músculos capaces de producir cambios en la forma del tracto vocal, y para entender cómo es que los diferentes sonidos del habla se originan, es necesario familiarizarse con las distintas partes que componen al tracto vocal. La figura 2.1 muestra un diagrama que es utilizado frecuentemente en el estudio de la Fonética. Dicho diagrama representa una vista lateral de la cabeza humana, como si ésta hubiera sido cortada por la mitad y que señala los articuladores principales utilizados en el habla, los cuales son descritos por [2] de la siguiente manera:

- El **paladar suave** permite que el aire pase a través de la nariz y de la boca, aunque durante el habla usualmente es levantado de manera que el aire no pueda escapar por la nariz. En su lado más bajo, este articulador puede ser tocado por la lengua, permitiendo la generación de un tipo particular de consonantes denominadas *consonantes velares*, como la *g* y *k*.

- El **paladar duro** es una superficie lisa y curva que constituye la pared superior de la cavidad bucal. Las consonantes que son articuladas con la lengua en una posición elevada cercana al paladar duro son llamadas *consonantes palatales*. El sonido de la *y* en la palabra *rayo* es un ejemplo.

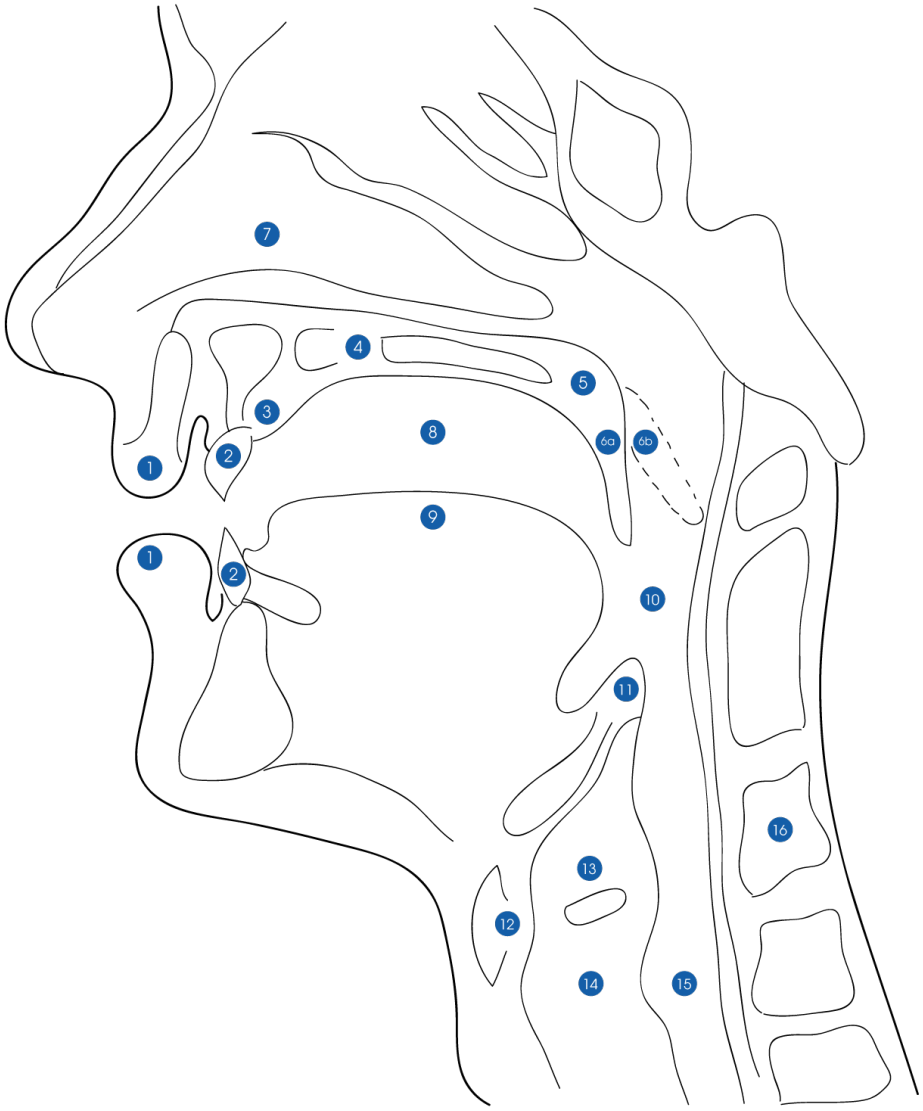


Figura 2.1: El aparato articulatorio humano. (1) Labios, (2) Dientes, (3) Reborde alveolar, (4) Paladar duro, (5) Paladar suave (velo del paladar), (6a) Úvula relajada, (6b) Úvula levantada (7) Pasaje nasal, (8) Boca (pasaje oral), (9) Lengua, (10) Faringe (garganta), (11) Epiglotis, (12) Laringe, (13) Cuerdas vocales y glotis, (14) Tráquea, (15) Esófago, y (16) Espina dorsal. Adaptado de [3].

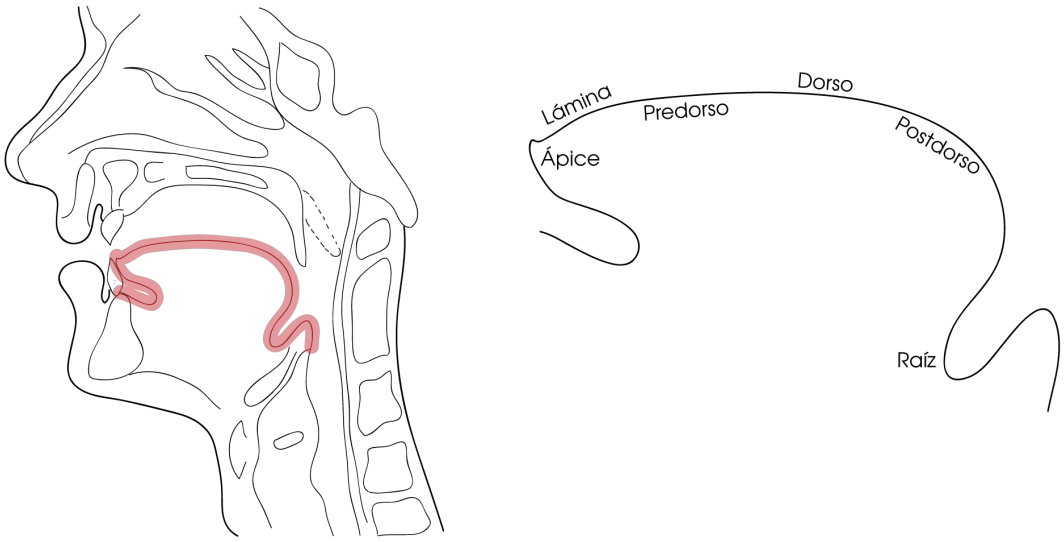


Figura 2.2: Subdivisiones de la lengua. Adaptado de [2].

- El **reborde alveolar** se encuentra entre los dientes frontales superiores y el paladar duro. Los sonidos producidos con la lengua tocando esta parte, tales como la *t*, *d*, o *n*, son denominados *alveolares*.

- La **lengua** es un articulador muy importante que puede moverse hacia muchas partes dentro de la cavidad bucal y también adoptar diferentes formas. Las distintas partes que conforman a este articulador se muestran en la figura 2.2.

- Los **dientes** son órganos articuladores duros cuyas raíces se originan en el reborde alveolar del maxilar. Los dientes laterales superiores se encuentran en contacto con la lengua en la mayoría de los sonidos del habla. Aquellos que son producidos con la lengua tocando los dientes frontales superiores son llamados *dentales*.

- Los **labios** son fundamentales en el proceso del habla. Pueden juntarse uno con el otro para producir sonidos bilabiales, o estar en contacto con los dientes para generar sonidos *labiodentales*.

- La **faringe** es un tubo que comienza justo arriba de la laringe. Mide aproximadamente 7 cm de longitud en las mujeres y alrededor de 8 cm en los hombres, y en su extremo más alto se divide en dos partes, una de ellas es la parte trasera de la boca y la otra es el comienzo del camino que lleva a la cavidad nasal.

Los articuladores descritos anteriormente son los que se utilizan principalmente en la producción de sonidos y se encuentran ubicados por encima de la **laringe**, la cual también puede ser descrita como un articulador, al igual que las **mandíbulas**; especialmente la mandíbula inferior, que tiene un movimiento mucho mayor durante el habla.

Los órganos articuladores pueden clasificarse en *articuladores activos* y *articuladores pasivos*. Los articuladores activos son aquellos que realizan todo o la mayoría del movimiento articulatorio en la producción de los sonidos de voz. Por otra parte, los articuladores pasivos, que en su mayoría se encuentran conectados directamente con el cráneo, son aquellos que realizan poco o ningún movimiento durante el habla, por lo que simplemente se limitan a recibir el contacto de los articuladores activos. En la Tabla 2.1 se muestran listados los órganos articuladores de acuerdo a esta clasificación.

Articuladores Activos	**Articuladores Pasivos**
Ápice de la lengua	Paladar duro
Lámina de la lengua	Labio superior
Cuerpo de la lengua	Reborde alveolar
Raíz de la lengua	Dientes superiores
Labio inferior	Paladar suave (velo del paladar)
Dientes inferiores	Úvula
Laringe	Faringe (pared faríngea)

Tabla 2.1: Clasificación de los órganos articuladores principales [4].

2.1.2. El Tracto Vocal

El *tracto vocal* es el espacio acústico donde los sonidos del habla se propagan. De acuerdo a la definición convencional, dicho espacio es la formación anatómica que conduce a los sonidos vocales de la glotis a los labios. En una definición más amplia, el tracto vocal incluye todos los espacios de aire donde se lleva a cabo una variación de presión acústica durante la producción del habla. En los jóvenes, de acuerdo a [5], la longitud del tracto vocal es de 14 cm en las mujeres y de 16.5 cm en los hombres. Considerando que el tracto vocal sufre un alargamiento a lo largo de la vida del ser humano, en los adultos el valor representativo de la longitud del tracto vocal es 15 cm para las mujeres y 17.5 cm para los hombres [6].

Un modelo realístico de la forma del tracto vocal consiste en un tubo que varía como una función del tiempo y desplazamiento a lo largo del eje de propagación del sonido. No obstante, la formulación de tal modelo para caracterizar la forma variante en el tiempo del tracto vocal puede ser demasiado compleja, por lo que un método para simplificarlo radica en representar al tracto vocal como una concatenación de tubos acústicos sin pérdidas, como se muestra en la figura 2.3.

El modelo del tracto vocal completo consta de una secuencia de tubos con áreas transversales A_k y longitudes l_k. Si un gran número de tubos de longitudes cortas es utilizado en el modelo, entonces la estructura general de tubos concatenados se aproxima a la de un tubo de área transversal variante [7]. El área de la sección transversal de cada tubo y la velocidad del aire determinan la presión del sonido y la velocidad de volumen, magnitudes que a su vez determinan la forma en que se produce la voz.

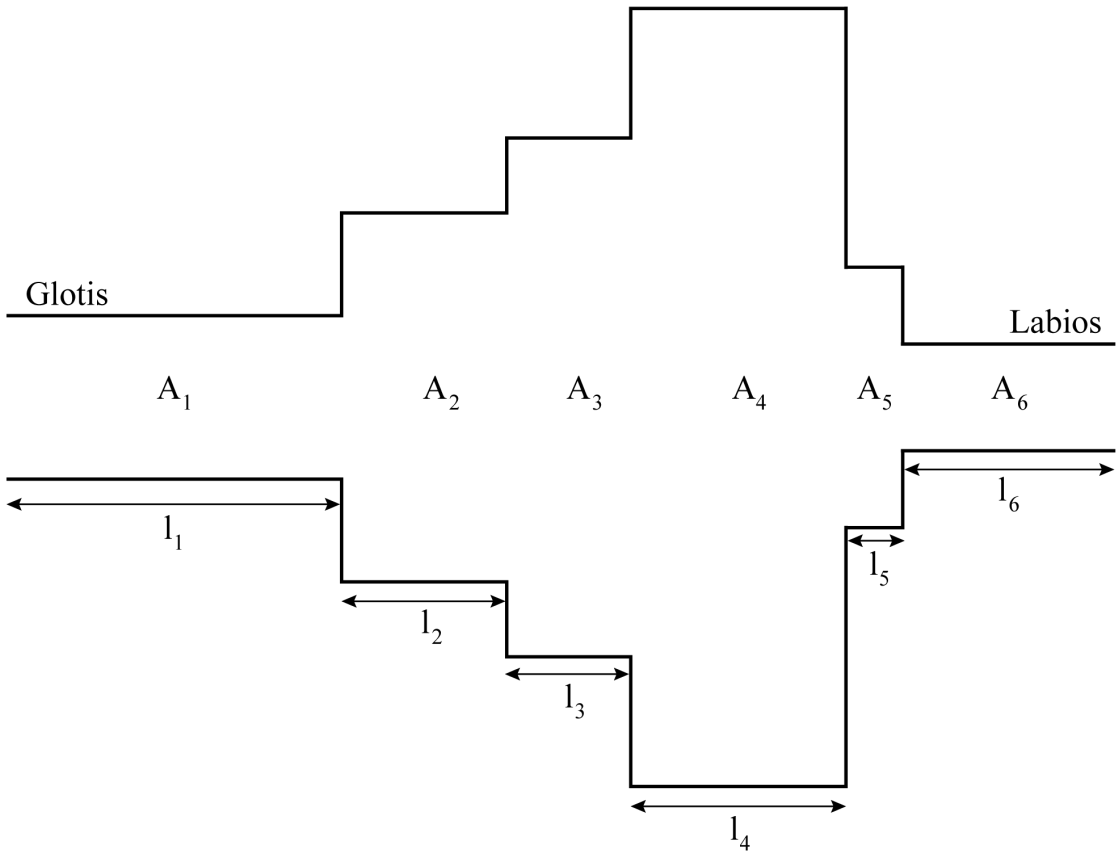


Figura 2.3: Modelo acústico del tracto vocal. Adaptado de [7].

2.1.3. El Mecanismo de Voz

Los sonidos de voz se caracterizan por un gran número de articulaciones diferentes. Pero las articulaciones por sí mismas no producen ningún sonido. Cuando se hace forzar aire a través de la laringe, con la tensión de las cuerdas vocales ajustada de manera que vibren en una oscilación relajada, se produce sonido [8]. No obstante, la vibración de las cuerdas vocales es la misma para todos los sonidos generados de esta forma, siempre produciendo pulsos cuasi-periódicos de aire, los cuales son modificados acústicamente mientras se propagan a través del tracto vocal, cuya configuración depende de la posición de los órganos articuladores. De manera que el habla, con toda su variedad de sonidos, puede modelarse a través de un sistema como el mostrado en la figura 2.4, el cual está constituido por dos elementos principales:

1. Una **fuente** de excitación que determina cómo el aire suministrado por los pulmones se pone en movimiento, generando uno de dos tipos de sonidos:
    - **Sonidos voceados**, en donde las cuerdas vocales se cierran y abren rítmicamente para convertir el aire proveniente de los pulmones en una onda de presión acústica, cuya forma de onda se asemeja a un tren de pulsos. La frecuencia de estos pulsos

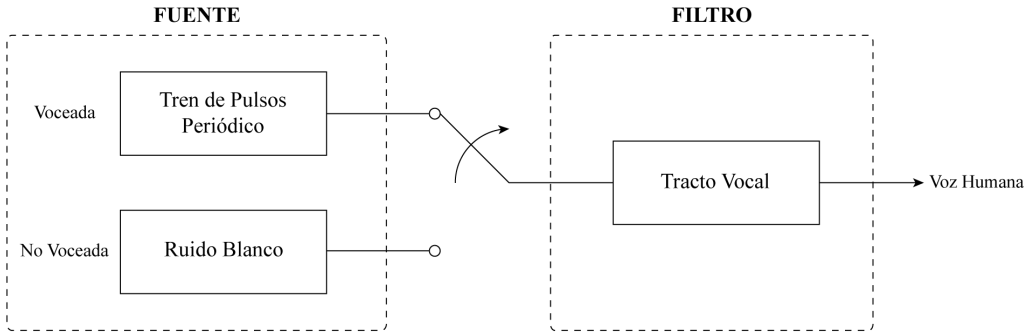


Figura 2.4: Modelo del Sistema de producción de voz humano.

define la tonalidad de un sonido voceado prolongado. En la figura 2.5 (a) se muestra la forma de onda de un sonido voceado.

- **Sonidos no voceados**, en donde, en contraste con los sonidos voceados, las cuerdas vocales no vibran, y los sonidos se producen al constreñir en el tracto vocal el aire proveniente de los pulmones, provocando una turbulencia que genera un sonido relativamente aleatorio, cuya calidad se asemeja más al ruido blanco. Las ondas de presión acústica resultantes oscilan mucho más rápido que las de los sonidos voceados, y además usualmente tienen una amplitud menor que la de éstas. En la figura 2.5 (b) se muestra la forma de onda de un sonido no voceado.

2. Un **filtro** cuya conformación espectral es realizada por el tracto vocal, y que sirve como medio para moldear o refinar el sonido generado por la fuente. La convolución en el dominio del tiempo del filtro con la fuente produce el sonido deseado.

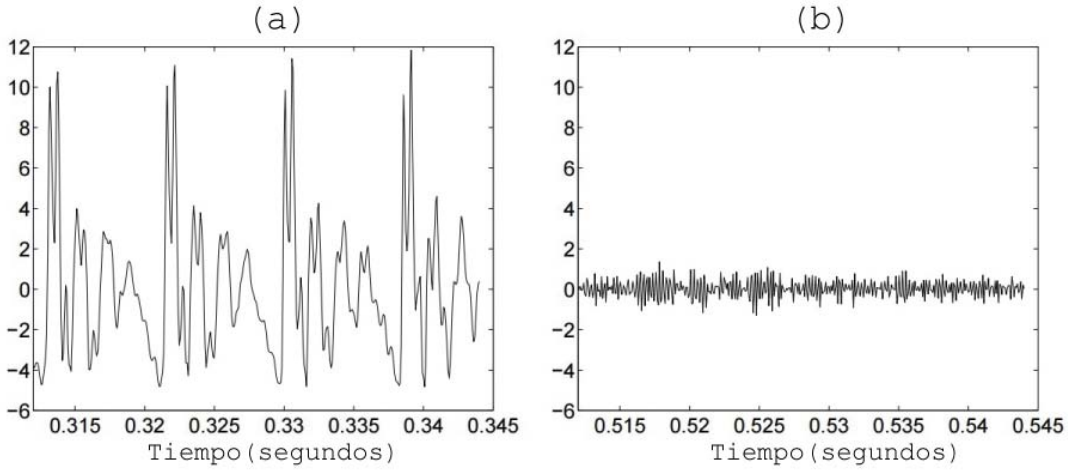


Figura 2.5: Distinción entre sonidos voceados y no voceados. (a) Segmento voz voceada, (b) Segmento de voz no voceada. Adaptado de [9].

La figura 2.6 muestra un diagrama básico de cómo el enfoque *fuente-filtro* que modela el sistema de producción de voz humano, es aplicado a los órganos involucrados en el proceso del habla. Los pulmones, las cuerdas vocales y la tráquea pertenecen a la *fuente*, mientras que las distintas cavidades, el velo del paladar y la lengua forman parte del *filtro*.

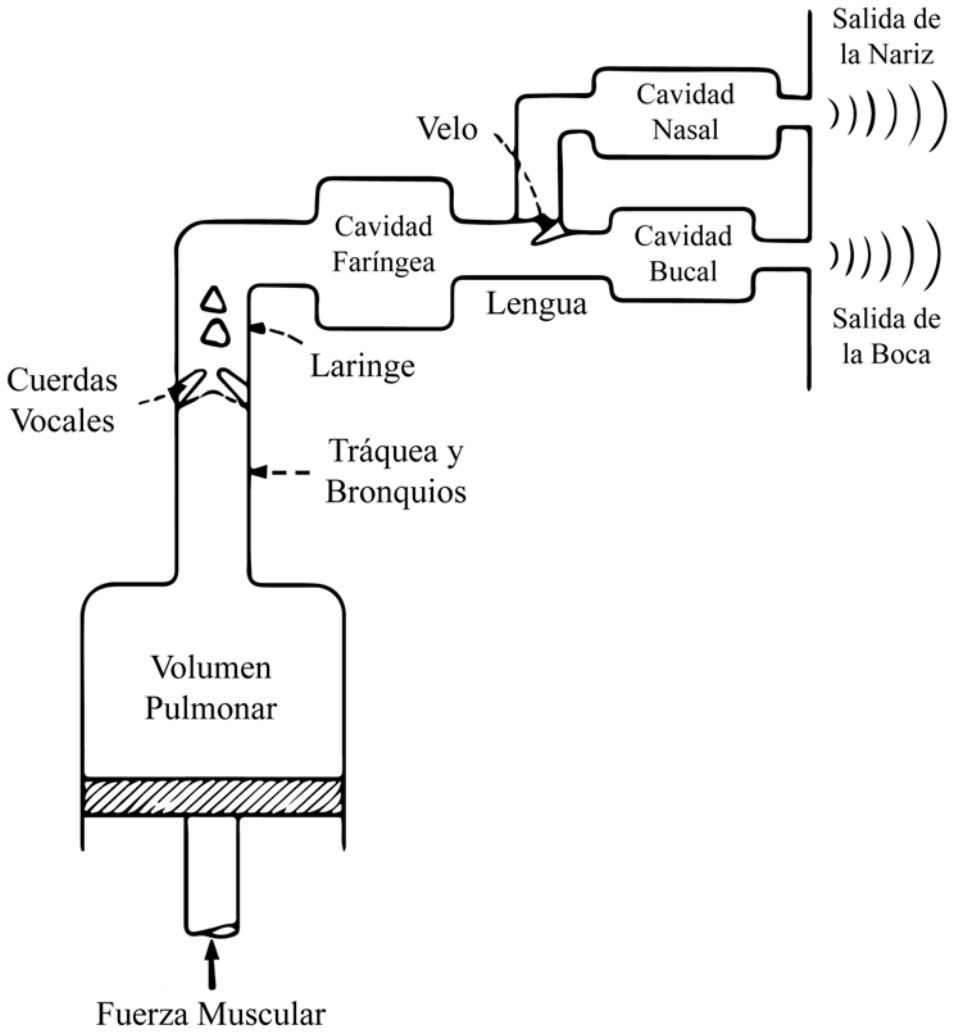


Figura 2.6: Esquema del mecanismo fisiológico de la producción del habla. Adaptado de [1].

2.2. Percepción de Voz

La percepción de voz comienza con la recepción de las ondas de sonido que arriban a los oídos. En ese momento, el oído procesa la onda de presión acústica, primero convirtiéndola en un patrón de vibraciones mecánicas en la membrana basilar, y luego representando este patrón como una serie de pulsos para ser transmitidos a lo largo del nervio auditivo hasta llegar al cerebro [10]. Este proceso es muy complejo e involucra varias etapas distintas en las que la información perceptual es extraída, y en las cuales se refleja la división anatómica principal del oído [11].

2.2.1. Fisiología del Oído

El oído humano, como se muestra en la figura 2.7, consta principalmente de tres secciones para procesar el sonido: el *oído externo*, el *oído medio* y el *oído interno*.

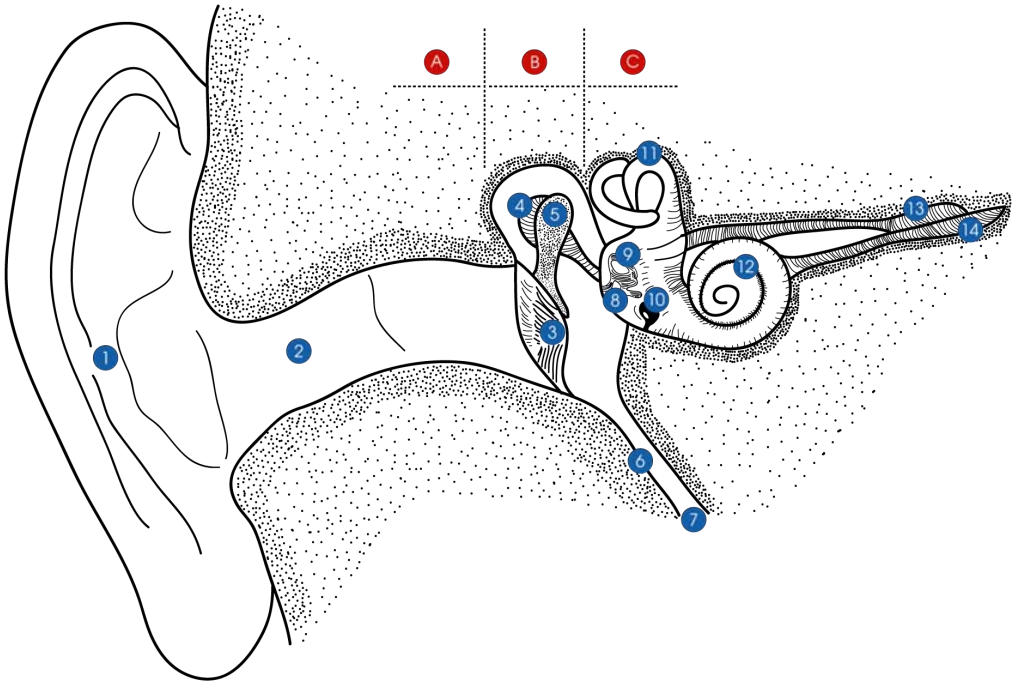


Figura 2.7: Vista esquemática del oído humano. (A) Oído externo, (B) Oído medio, (C) Oído interno, (1) Pabellón auricular, (2) Canal auditivo externo, (3) Tímpano, (4) Yunque, (5) Martillo, (6) Trompa de Eustaquio, (7) Cavidad nasal, (8) Estribo, (9) Ventana oval, (10) Ventana redonda, (11) Aparato vestibular (canales semicirculares), (12) Cóclea, (13) Nervio vestibular, y (14) Nervio coclear. Adaptado de [12].

2.2.1.1. El Oído Externo

El *oído externo* está constituido por dos partes. La parte visible se conoce como *aurícula*, o bien, como *pabellón auricular*, una estructura consistente de varias prominencias redondeadas formadas esencialmente de cartílago. Tiene como funciones principales: enfocar las ondas de sonido hacia el canal auditivo externo del oído y asistir en la detección de las fuentes de sonido. También protege la entrada del canal auditivo, tanto de ataques físicos como de cantidades excesivas de sonido.

La parte no visible del oído externo es el *canal auditivo*, el cual forma un tubo a lo largo del cual viaja el sonido. La longitud del canal es de aproximadamente 2.5 cm y conduce al tímpano. En su interior hay pequeños vellos y glándulas que secretan cerumen, una sustancia que actúa como una barrera que impide que partículas de polvo suspendidas en el aire, insectos y otros elementos pequeños se aproximen al tímpano. El canal actúa como un pequeño amplificador para los sonidos cuyas frecuencias se encuentran entre los 3,000 y 4,000 Hz, de manera que los sonidos débiles a esas frecuencias son más perceptibles. También ayuda a proteger al tímpano de cambios de temperatura y humedad, así como de daños físicos [11].

2.2.1.2. El Oído Medio

El *oído medio* es una cavidad llena de aire de 1.3 cm de longitud y 6 cm^3 de volumen [10]. Comienza en la *membrana timpánica* o *tímpano*. Este órgano de forma aproximadamente circular, yace transversalmente al canal auditivo externo en un ángulo de alrededor de 55°. Se compone de un tejido fibroso con propiedades elásticas importantes que le permiten vibrar cuando es alcanzado por las ondas de sonido. Su forma y tensión hacen que las vibraciones se concentren en una protuberancia cercana a su centro, de donde son transferidas al primero de los huesos del oído medio, el cual está firmemente unido a la membrana.

La función primaria del oído medio es convertir las ondas de sonido en vibraciones mecánicas, que a su vez son transmitidas al oído interno. En el proceso, las vibraciones son amplificadas enormemente (hasta un valor de 30 dB) al momento de alcanzar al oído interno. Esta labor es realizada por el conjunto de huesos más pequeño del cuerpo humano, conocidos como *osículos auditivos*. Estos huesos fueron nombrados de acuerdo a su forma: el *martillo*, el cual está unido al tímpano, el *yunque*, y el *estribo*, el cual encaja en la *ventana oval*, una abertura en la pared ósea que separa al oído medio del oído interno [11].

2.2.1.3. El Oído Interno

El *oído interno* está conformado por un conjunto de pequeñas cavidades interconectadas y varios pasajes dentro del cráneo. Contiene a los *canales semicirculares*, los cuales controlan el sentido fisiológico del equilibrio, y a la *cóclea*, un tubo en forma de espiral de aproximadamente 3.5 cm de longitud, enrollado 2.6 veces, cuya función principal es convertir las vibraciones mecánicas producidas por el oído medio en impulsos eléctricos nerviosos capaces de ser transmitidos al cerebro [10].

La estructura más relevante del oído interno es la cóclea. En la figura 2.8 se muestra una vista esquemática de la sección transversal del *ducto coclear*. La espiral está dividida por la *membrana basilar* en dos compartimentos: la *rampa vestibular* y la *rampa timpánica*, las cuales están llenas

de un líquido viscoso. Las vibraciones entran a este fluido a través de la ventana oval y la rampa vestibular, y son transmitidas por toda la cóclea a lo largo de la membrana basilar, que actúa como un banco de filtros, realizando un análisis espectral no uniforme del sonido [12].

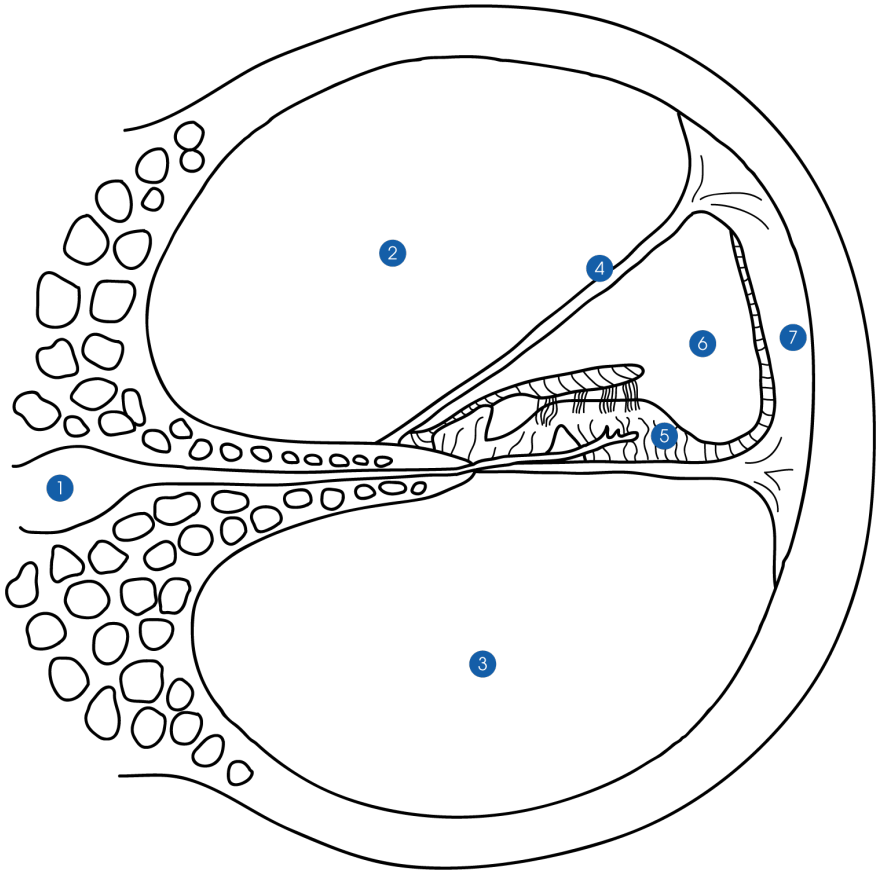


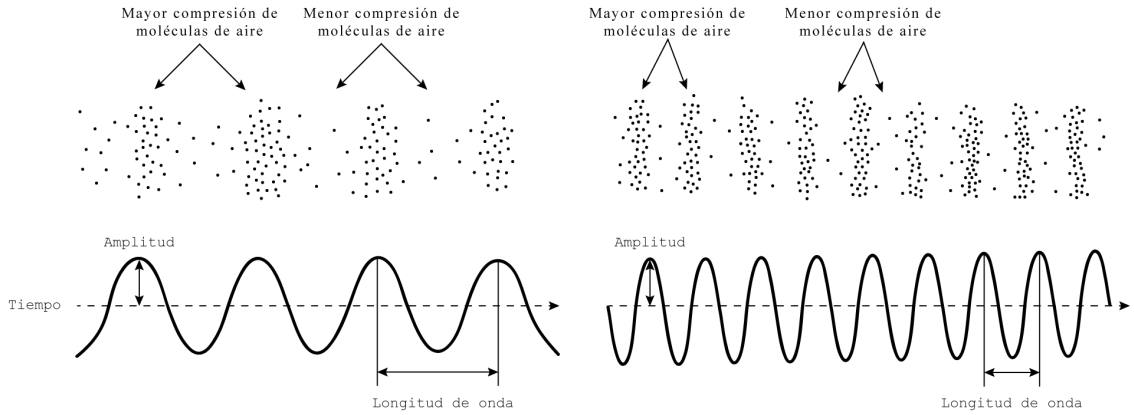
Figura 2.8: Sección transversal del ducto coclear. (1) Fibra de nervio coclear, (2) Rampa vestibular llena de líquido perilinfático, (3) Rampa timpánica llena de líquido perilinfático, (4) Membrana de Reissner, (5) Órgano de Corti, (6) Conducto coclear lleno de líquido endolinfático, y (7) Membrana basilar. Adaptado de [11].

2.2.2. El Sonido

El *sonido* es una variación de presión debida a las vibraciones de las moléculas en un medio elástico. En el estudio de la producción de voz, generalmente se trata la propagación del sonido a través del aire: las partículas de aire son perturbadas por los movimientos de los órganos vocales, especialmente de las cuerdas vocales. Pero en el estudio de la percepción de voz, el aire no es el único medio involucrado, ya que el proceso de la audición humana requiere que las vibraciones de sonido en el aire sean transformadas primero en vibraciones mecánicas (a través del mecanismo óseo del oído medio), luego en cambios hidráulicos (a través del líquido dentro del oído

medio), y finalmente en impulsos nerviosos (a lo largo del nervio auditivo al cerebro) [11].

El sonido es representado gráficamente por una curva que exhibe la forma de una onda en un determinado instante de tiempo. En la figura 2.9 se muestra la representación gráfica de una onda de sonido y su relación con los cambios de presión en el aire. Se trata de una *onda transversal*, la cual no tiene parecido alguno con la *onda longitudinal* real que se propaga por el aire [13]. Sin embargo, este modelo gráfico es utilizado para simplificar la presentación de la información referente a la onda de sonido.



Sonido de longitud de onda larga (sonido de baja frecuencia)

Sonido de longitud de onda corta (sonido de alta frecuencia)

Figura 2.9: Representación gráfica de las ondas de sonido. Adaptado de [14].

2.2.3. Atributos Físicos y Perceptuales del Sonido

La *Psicoacústica* es la disciplina científica que estudia la percepción del sonido. Se encarga de modelar la relación existente entre los atributos perceptuales de un sonido y las propiedades físicas medibles que lo caracterizan [10]. En la Tabla 2.2 se listan las propiedades físicas principales del sonido ligadas a una cualidad perceptual.

Atributo Físico	Atributo Perceptual
Intensidad	Volumen
Frecuencia fundamental	Tonalidad
Forma espectral	Timbre
Tiempo de inicio y fin	Ritmo
Fase	Ubicación

Tabla 2.2: Relación entre los atributos físicos y perceptuales del sonido [10].

2.2.3.1. Intensidad y Volumen

En 1933, los científicos estadounidenses Harvey Fletcher y Wilden Munson de los Laboratorios Bell, reportaron por primera vez en una publicación titulada "Volumen, su definición, medición y cálculo" -*"Loudness, its definition, measurement and calculation"*-, los resultados de un estudio acerca de la forma real en que los seres humanos oyen. La investigación que condujeron demostró que aunque los sonidos con un mayor nivel de intensidad usualmente suenan más fuerte, la sensibilidad del oído varía con la frecuencia y la calidad del sonido. Por ejemplo, los sonidos de baja frecuencia cuya intensidad es moderada o baja, son muy difíciles de ser percibidos por el oído, no así los sonidos de frecuencias más altas. Sin embargo, a medida que la intensidad aumenta, las diferencias entre las diferentes frecuencias se igualan [15]. En la figura 2.10 se muestra la gráfica de las *curvas de igual volumen* reportada por Fletcher y Munson en su publicación. Las curvas indican que la respuesta del mecanismo auditivo humano está en función de la frecuencia y los niveles de presión sonora (SPL).

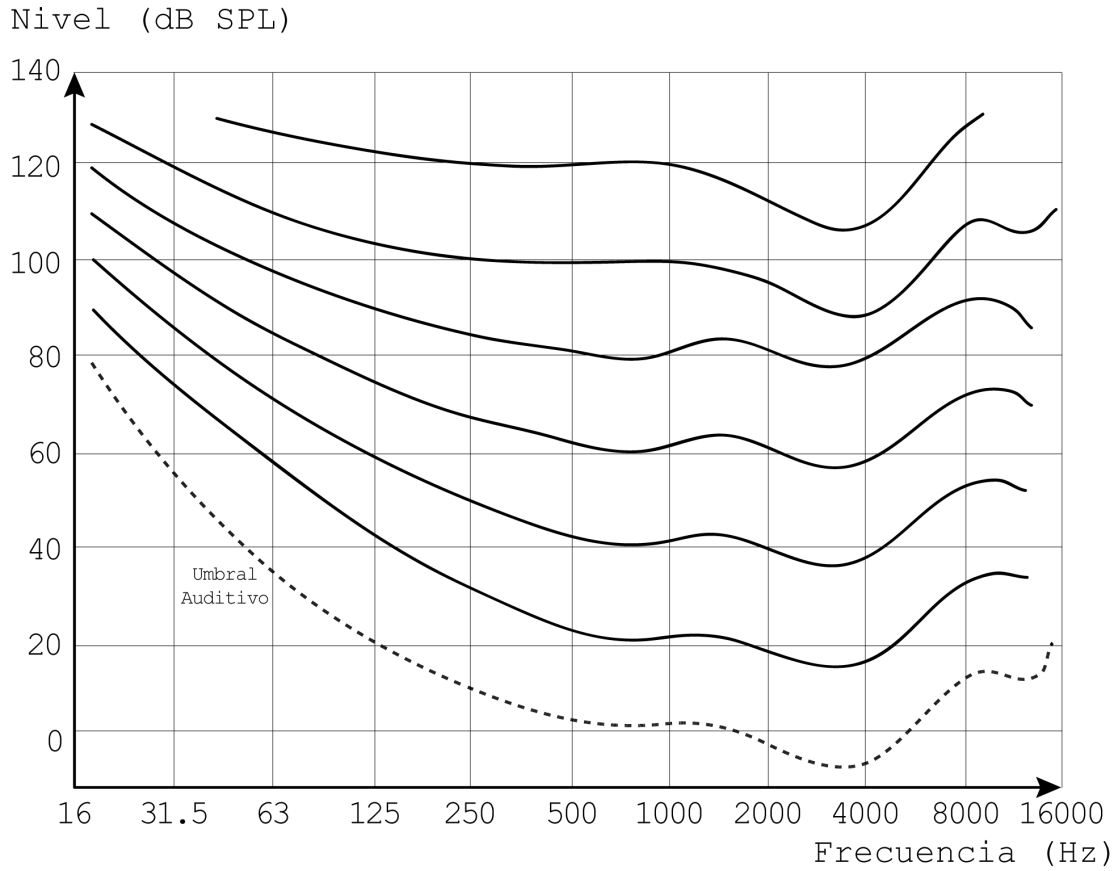


Figura 2.10: Curvas de igual volumen. Adaptado de [6].

2.2.3.2. Tonalidad y Frecuencia

La *tonalidad* (en inglés, *pitch*) es probablemente el atributo perceptual más importante del sonido. Sin esta característica, el habla consistiría en susurros y también sería muy difícil para los humanos la identificación de fuentes de sonido [16]. En [17], el Instituto Nacional Americano de Estándares (*ANSI*, por sus siglas en inglés) define la tonalidad como *"aquel atributo de la sensación auditiva en términos del cual los sonidos pueden ser ordenados en una escala que se extiende de alto a bajo"*. En este sentido, la frecuencia de un *tono puro*, es de decir, un sonido producido por una onda de presión acústica de una sola vibración, se correlaciona con la sensación de tonalidad. En general, entre más alta sea la frecuencia de un sonido, el ser humano percibe más alto la tonalidad del sonido.

2.2.3.3. La Escala Mel

Existen varios métodos para medir la dependencia que hay entre la tonalidad de un tono puro y su frecuencia. Por ejemplo, se puede obtener una relación *tonalidad-frecuencia* mediante una estimación de magnitud. También se puede utilizar un método para *duplicar* o *demediar*, en el cual una persona ajusta la frecuencia de un tono de comparación, hasta que subjetivamente suene el doble o la mitad de alto que la tonalidad de un tono de prueba con una frecuencia establecida por el experimentador [18]. El resultado clásico de este experimento se muestra en la gráfica de la figura 2.11. Se trata de la *escala Mel*, medida en 1937 por los científicos Stevens, Volkmann, y Newman. En el eje horizontal de la gráfica se representa la frecuencia medida en

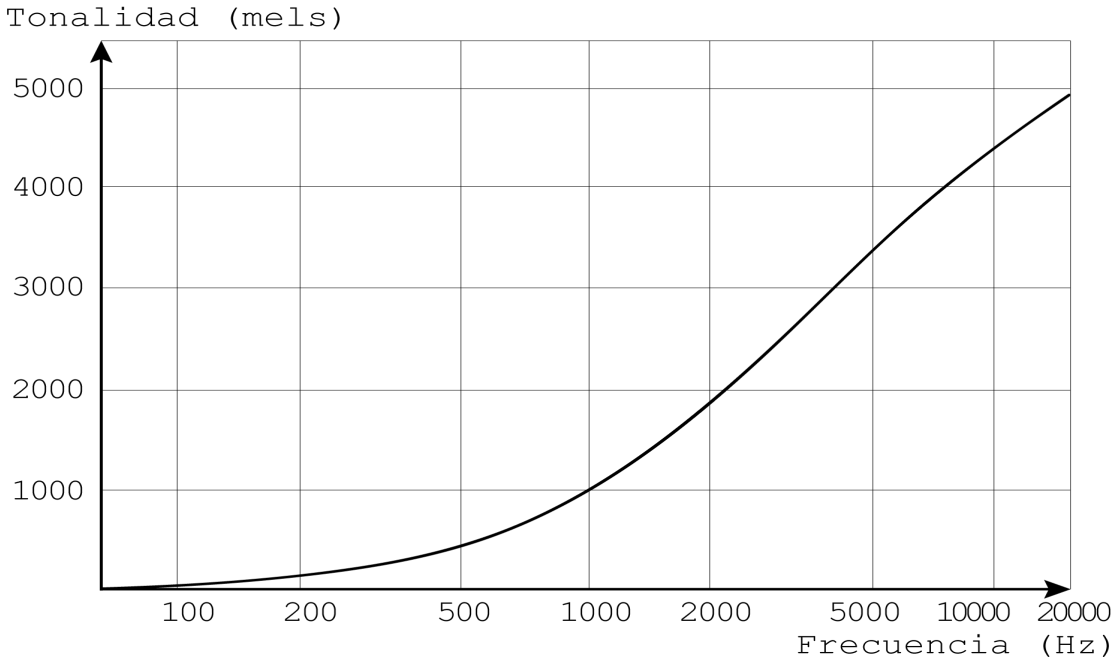


Figura 2.11: Relación de la tonalidad (en mels) y la frecuencia de un tono puro. Adaptado de [19].

Hertz, y en el eje vertical la tonalidad medida en *mels*, unidad derivada de la palabra inglesa *melody*. La escala tiene como referencia arbitraria una tonalidad de 1000 mels asignada a un valor de 1000 Hz. Se puede observar claramente que la tonalidad no es lineal en frecuencia, por lo que la relación entre ambos atributos no es idéntica. Un tono que suena en promedio dos veces más alto recibe un valor de 2000 mels, mientras que un tono que suena solamente la mitad de alto tiene una tonalidad de 500 mels.

2.2.3.4. La Escala Bark

La *escala Bark* es una escala perceptual alternativa y proporcional a la escala Mel. Fue propuesta en 1961 por el científico Eberhard Zwicker, y nombrada así, en honor al científico Heinrich Barkhausen, quien propuso las primeras mediciones subjetivas de la intensidad del sonido [20]. La escala está basada en el análisis espectral no uniforme de sonidos que realiza la membrana basilar, órgano fundamental del oído interno. Este órgano es delgado y flexible en su base y más grueso en el extremo próximo a la ventana oval y al tímpano; como resultado, la base responde a sonidos de alta frecuencia, y la raíz a sonidos de frecuencias bajas. Como se muestra en la figura 2.12, cada punto de la membrana basilar es sensible a un margen específico de frecuencias, por lo que cada punto puede ser considerado como un filtro pasobanda, cuya respuesta en frecuencia se hace cada vez más amplia conforme aumenta la frecuencia.

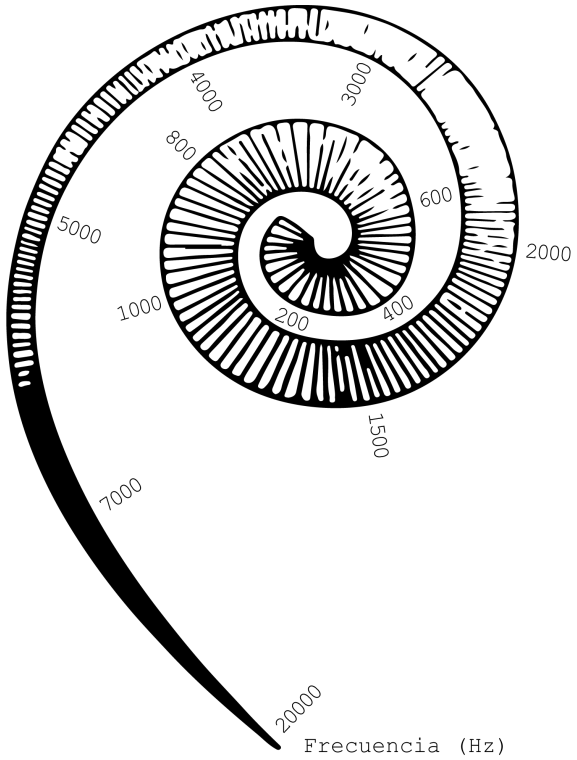


Figura 2.12: Frecuencias representativas a lo largo de la membrana basilar. Adaptado de [21].

Pero, debido a que los puntos en la membrana basilar no pueden vibrar de manera independiente el uno del otro, los filtros se encuentran traslapados significativamente [12]. En la figura 2.13 se representa esquemáticamente un banco de 15 filtros que refleja aproximadamente la sensibilidad de frecuencia del sistema auditivo.

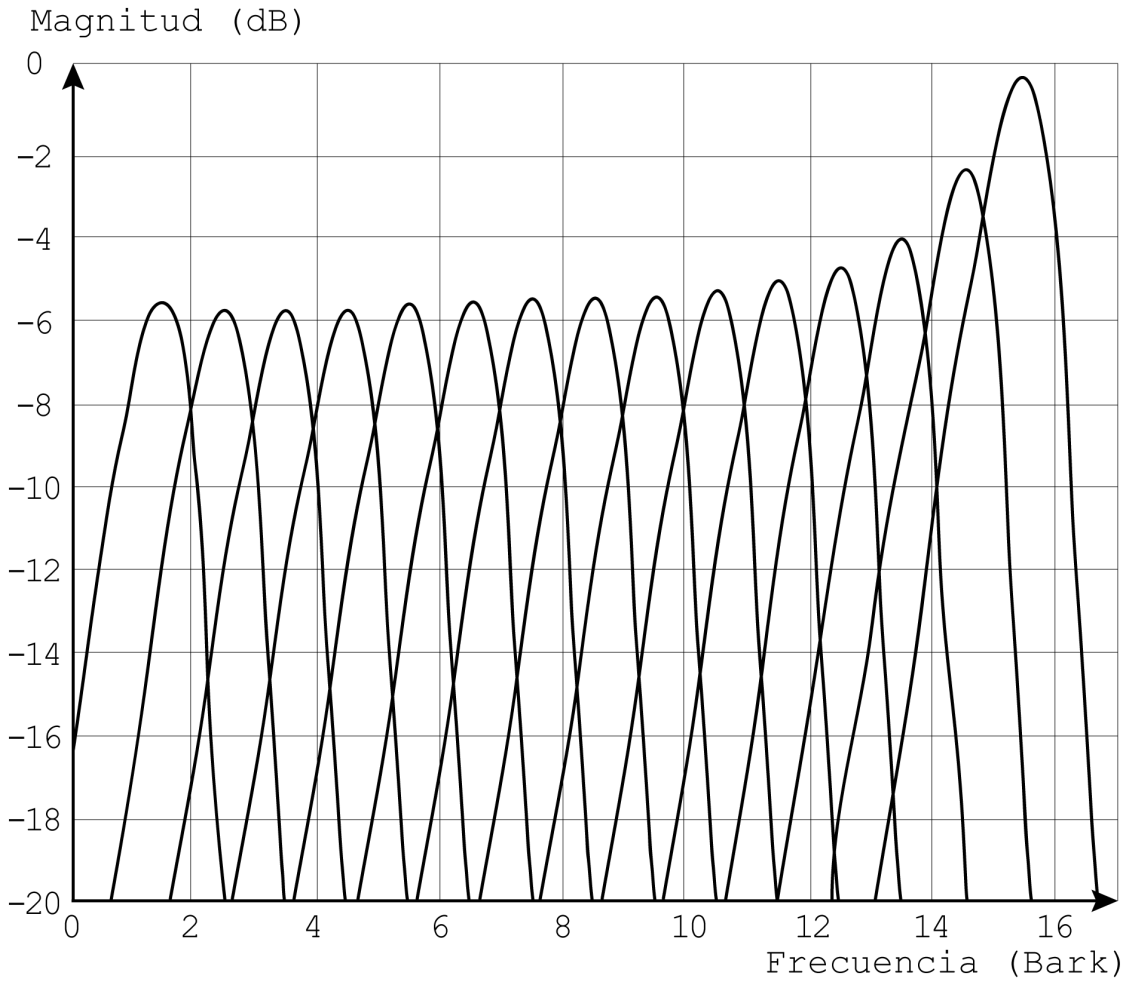


Figura 2.13: Banco de filtros Bark. Adaptado de [22].

La escala Bark, cuyas frecuencias centrales se listan en la Tabla 2.3, varía desde 1 a 24 Barks, correspondientes a las primeras 24 *bandas críticas* del oído, es decir aquellas bandas de frecuencias con las que dos o más tonos puros excitan casi las mismas células auditivas de la membrana basilar [23]. Mediante la escala Bark se muestra que la resolución perceptual del oído es más fina en las frecuencias más bajas y menos sensible a las frecuencias altas del rango audible humano.

Banda Crítica [Bark]	Frecuencia Central [Hz]	Banda Crítica [Bark]	Frecuencia Central [Hz]
1	50	13	1850
2	150	14	2150
3	250	15	2500
4	350	16	2900
5	450	17	3400
6	570	18	4000
7	700	19	4800
8	840	20	5800
9	1000	21	7000
10	1170	22	8500
11	1370	23	10500
12	1600	24	13500

Tabla 2.3: Escala de frecuencia Bark [20].

2.2.3.5. Timbre

El *timbre* es aquel atributo de la sensación auditiva en términos del cual un oyente puede juzgar que dos sonidos presentados de manera similar y con el mismo volumen y tonalidad son diferentes [17]. De forma que si una nota musical de un tono dado se toca satisfactoriamente con exactamente la misma intensidad en dos instrumentos musicales diferentes, una persona podría distinguir claramente la diferencia entre los dos sonidos producidos y referirlos a cada instrumento. En el caso de la voz humana, el timbre es parcialmente responsable de que la voz de cada persona suene diferente.

2.2.4. Enmascaramiento Auditivo

El *enmascaramiento auditivo* es un fenómeno ocasionado por las vibraciones mecánicas que se producen en la membrana basilar. Fundamentalmente ocurre cuando un sonido no puede ser percibido si otro sonido cercano en frecuencia tiene un nivel más alto [10]. El hecho de que un sonido se vuelva inaudible debido al enmascaramiento, se puede cuantificar con respecto al umbral de audición. Como se muestra en la figura 2.14, un tono intenso, denominado *enmascarador*, tiende a elevar el umbral de audición alrededor de su ubicación en el eje de la frecuencia. De manera que todos los componentes espectrales cuyo nivel es más bajo que el del umbral elevado son enmascarados y por consiguiente no podrán ser escuchados. Del mismo modo, cualquier componente espectral cuyo nivel se encuentre por encima del umbral elevado no es enmascarado y por lo tanto será escuchado.

En la figura 2.14 se observa además, que el efecto de enmascaramiento es mucho mayor para las frecuencias por encima de la frecuencia enmascaradora que para las frecuencias que se encuentran por debajo, ya que la caída del umbral desplazado es menos abrupto por encima que por debajo del tono enmascarador [12].

El concepto de enmascaramiento es ampliamente utilizado en señales de audio para lograr

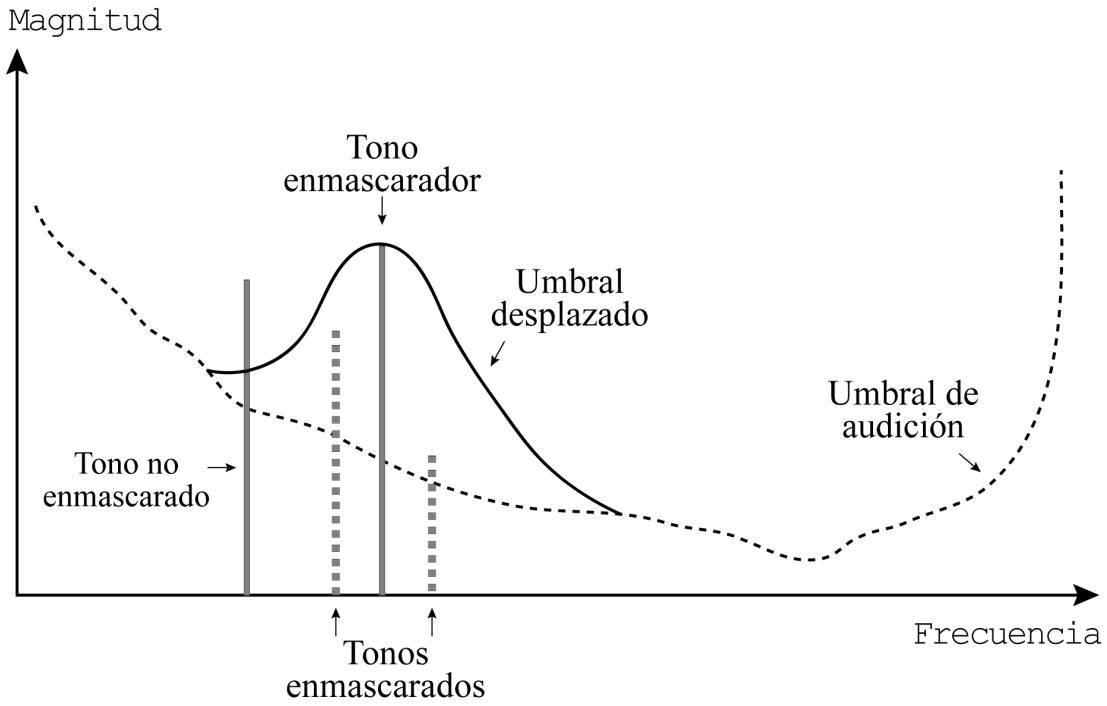


Figura 2.14: Efecto de enmascaramiento auditivo. Adaptado de [12].

representaciones con tasas de datos más bajas, y que al mismo tiempo conserven un alto grado de fidelidad perceptual, lo cual es posible al remover áreas de la señal donde el umbral de audición es elevado por fuertes componentes espectrales en la señal.

2.3. Resumen

En este capítulo se han presentado conceptos clave sobre la forma en la cual el ser humano produce y percibe la voz. En el caso de la producción de voz, se han descrito todos los órganos del cuerpo humano capaces de articular sonidos, así como el medio acústico donde se propagan. Estos elementos en conjunto, permiten caracterizar el sistema de producción de voz a través de un enfoque fuente-filtro. En el caso de la percepción de voz, se ha descrito la división anatómica principal del oído, con la finalidad de entender cómo es que la ondas de presión acústica son procesadas. También se presenta la relación existente entre los atributos físicos y perceptuales del sonido, así como el fenómeno de enmascaramiento auditivo.

3

Caracterización Acústico-Fonética de la Señal de Voz

El habla varía en respuesta a muchas circunstancias, por lo que no es posible tener un conocimiento completo de la estructura de los sonidos de un idioma. Además, los idiomas se encuentran en constante evolución, razón por la cual no puede haber una descripción final de los sonidos de los idiomas. La siguiente generación de hablantes siempre hablarán un poco diferente de sus predecesores, y pueden incluso crear sonidos que nunca habían sido utilizados anteriormente por los humanos. Sin embargo, una descripción y clasificación de los sonidos que el aparato fonador humano es capaz de producir recae en la segmentación de la cadena hablada. La mayoría de la literatura concerniente a la Fonética de los dos últimos siglos asume que un análisis significativo del habla puede ser llevado a cabo dividiendo la señal de voz en pequeños fragmentos que puedan reconocerse como sonidos significativos del habla. Éstos pueden ser descritos mediante términos formales con entidades teóricas tales como fonemas o segmentos. La descripción acústica de estos segmentos, o de manera más general, el estudio de sus características físicas, descompone totalmente la complejidad de los sonidos del habla.

3.1. Representación Acústica de la Voz

La teoría que explica la acústica radiada en términos del mecanismo vocal que los produce se denomina *teoría acústica de la producción del habla*. Con esta teoría es posible entender no solamente qué da lugar a las variaciones radiadas de tiempo/presión de la señal de voz, sino también de dónde provienen sus propiedades de frecuencia fundamental, armónicos y envolvente espectral, cruciales para la señalización de la tonalidad y la calidad de las vocales y consonantes.

3.1.1. Tonos Puros y Ondas Complejas

El análisis acústico de los sonidos del habla se basa fundamentalmente en las funciones matemáticas seno y coseno. En el caso de un tono puro, la onda de presión acústica generada corresponde a una simple función senoidal. Dicha onda posee dos características muy importantes: una frecuencia particular y una amplitud determinada. La *frecuencia* es el número de veces por segundo que una onda se repite, y se mide en ciclos por segundo o también Hertz (Hz). Por otro lado, la *amplitud* es la cantidad máxima de desplazamiento sufrido por las partículas en el medio desde su posición antes de ser alteradas. En la figura 3.1 (a) y 3.1 (b) se muestran la gráficas de un par de tonos puros de 10 y 3 Hz respectivamente y cuya amplitud es igual a 1.

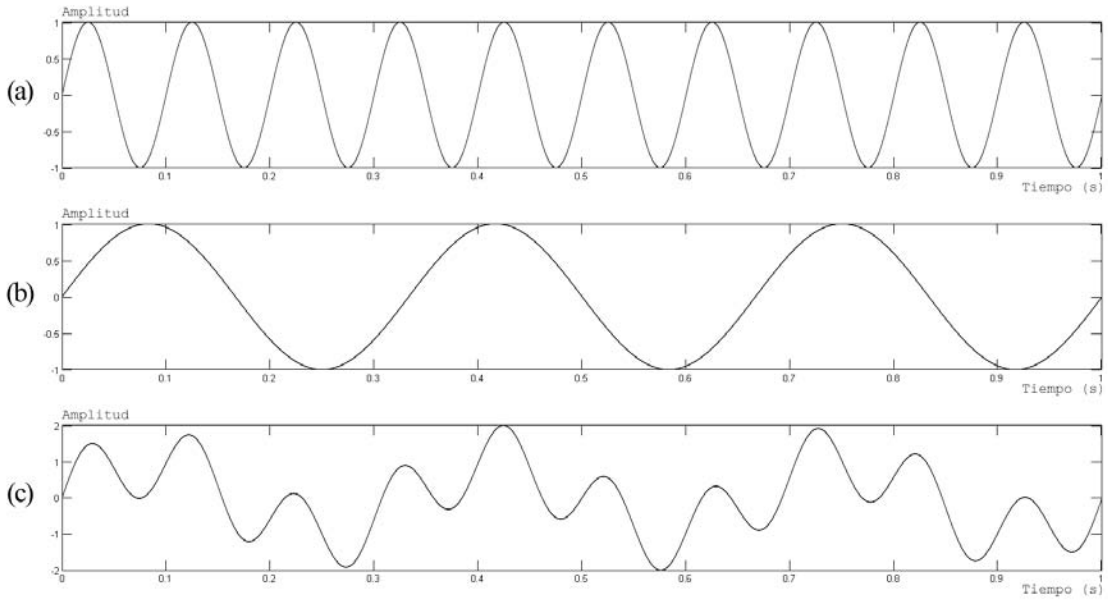


Figura 3.1: Tonos puros y ondas complejas. (a) Tono puro de 10 Hz. (b) Tono puro de 3 Hz. (c) Onda compleja conformada por la combinación de los tonos (a) y (b).

Cuando dos o más tonos puros con distintos valores de frecuencia y amplitud se combinan, se genera como resultado una *onda compuesta* o *compleja* como la que se muestra en la figura 3.1 (c). Los sonidos de voz, al igual que la gran mayoría de los sonidos que el ser humano es capaz de percibir, son siempre conformados por ondas complejas, ya que son el resultado del uso simultáneo de diversas fuentes de vibración que ocurren en el sistema de producción de voz.

Sin embargo, así como la mezcla de funciones senoidales da lugar a este tipo de ondas, también es posible descomponer ondas complejas en una serie de funciones senoidales de diferentes frecuencias, cada una de las cuales representa una componente frecuencial. Al análisis matemático involucrado en combinar tonos puros para producir ondas complejas, y viceversa, se conoce como *Análisis de Fourier*, en honor al matemático francés Jean-Baptiste Joseph Fourier, quién fue el primero en descubrir este concepto.

3.1.2. Frecuencia Fundamental

A la componente frecuencial más baja de una onda compleja se le conoce como *frecuencia fundamental*, generalmente abreviada como F_0, y para el oído humano representa una medida de qué tan alta o baja es la tonalidad de la voz de una persona.

El intervalo de frecuencias que un adulto joven es capaz de oír es extremadamente amplio —aproximadamente de 20 a 20,000 Hz. No es posible oír vibraciones con frecuencias más bajas (*infrasónicas*) o más altas (*supersónicas*) que este intervalo. Sin embargo, las frecuencias localizadas en ambos extremos del intervalo de audición humano tienen muy poco significado para el habla, ya que la zona en la cual se concentra la mayoría de la información relevante de la señal de voz recae en las frecuencias comprendidas entre 100 y 4,000 Hz [11].

La frecuencia fundamental de un adulto típico es de aproximadamente 120 Hz, con un intervalo de variación de 80 a 300 Hz, mientras que una mujer adulta típica posee una frecuencia fundamental de aproximadamente 220 Hz que puede variar hasta los 500 Hz [11], [24]. Por otro lado, los niños y los bebés tienen en promedio frecuencias fundamentales mucho más altas, las cuales incluso pueden llegar a sobrepasar 1,500 Hz [25]. La razón por la cual los hombres adultos exhiben un intervalo de frecuencias fundamentales más bajo que las mujeres y niños de cualquier sexo, se fundamenta en el hecho de que los hombres poseen laringes varios centímetros más largas [26]. La frecuencia fundamental varía como una función de la edad en ambos sexos, y su magnitud es más alta a temprana y avanzada edad. En la figura 3.2 se exhibe la frecuencia fundamental como una función de la edad en hablantes adultos típicos.

Aunque muchos estudios han reportado diferentes intervalos de frecuencia fundamental para cada sexo, algunos de ellos han demostrado que el intervalo de F_0 depende de diversos factores, tales como el idioma de los hablantes, el grupo étnico al que pertenecen, el tipo de textos utilizados para realizar las pruebas, el tipo de discurso empleado, así como el estado emocional de los hablantes.

La tabla 3.1 resume los resultados de varias investigaciones realizadas para obtener el promedio de F_0 para hombres y mujeres, e incluye solamente aquellas investigaciones en las que los hablantes (todos ellos adultos) realizaron las mismas tareas. En todos los reportes, a excepción del realizado por Phil Rose en 1991, la F_0 promedio fue claramente más alta y el intervalo de F_0 en Hz notoriamente más amplio para las mujeres que para los hombres.

Los valores muy altos de la F_0 promedio observada en los hablantes masculinos de Wú, uno de los principales dialectos del chino, son bastante notables. Muestran que el promedio de la F_0 utilizada en el habla pertenece al conjunto de propiedades que pueden ser prescritas por convención social [37]. Sin embargo, este fenómeno no es único, pues un promedio elevado de la F_0 puede ser también observado en uno de los dialectos suecos hablado en la provincia de Småland [38].

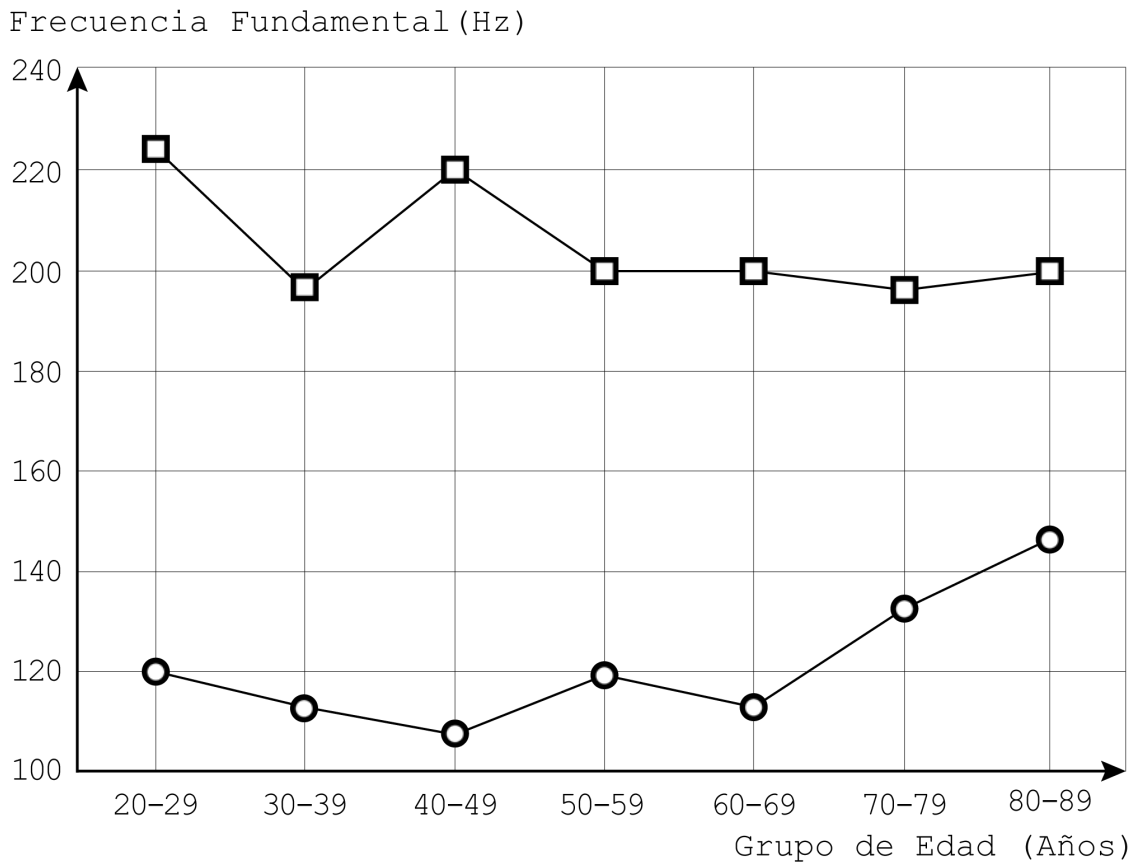


Figura 3.2: La frecuencia fundamental como función de la edad en hablantes adultos típicos. □ = Mujeres, O = Hombres. Datos masculinos de [27]; datos femeninos de [28].

En la mayoría de los idiomas del mundo, los hablantes, en sus conversaciones cotidianas, suelen utilizar la frecuencia fundamental más baja que son capaces de producir, ya que fisiológicamente se requiere un menor esfuerzo para sostener la fonación durante el habla. Por el contrario, la frecuencia fundamental más alta del intervalo de F_0, que solamente puede ser alcanzada con las cuerdas vocales extremadamente tensas, es únicamente aproximada en casos excepcionales, como por ejemplo cuando se grita para pedir ayuda en situaciones de emergencia [39].

3.1.3. Armónicos

Ya que los sonidos de voz involucran más de una onda senoidal, existen otras cantidades de energía que son generadas por la vibración de la cuerdas vocales, todas las cuales están correlacionadas con la onda senoidal básica mediante una simple relación matemática: todas son múltiplos de la frecuencia fundamental. Por consiguiente una F_0 de 200 Hz producirá un conjunto de frecuencias que tendrán un valor de 400 Hz, 600 Hz, 800 Hz, y así sucesivamente. Estas

Investigación	Idioma	Sexo	Edad	F_0 [Hz]
Rappaport (1958), [29]	Alemán	m		129
		f		238
Chevrie-Muller et al. (1967), [30]	Francés	m	20-61	145
		f	19-72	226
Takefuta et al. (1972), [31]	Inglés	m		127
		f		186
Chen (1974), [32]	Mandarín	m	30-50	108
		f	30-50	184
Boë et al. (1975), [33]	Francés	m		118
		f		207
Kitzing (1979), [34]	Sueco	m	21-70	110
		f	21-70	193
Johns-Lewis (1986), [35]	Inglés	m	24-49	101
		f	24-49	182
Pegoraro Krook (1988), [36]	Sueco	m	20-79	113
		f	20-79	188
Rose (1991), [37]	Wú	m	25-62	170
		f	25-62	187

Tabla 3.1: Frecuencia fundamental promedio en hombres y mujeres adultos de acuerdo a nueve investigaciones.

frecuencias múltiplos de F_0 son conocidas como *armónicos*, y son numeradas secuencialmente. Sin embargo, el ser humano normalmente no es capaz de oír los armónicos como tonos separados, ya que entre más altos son, su amplitud es cada vez más baja que la de la frecuencia fundamental. No obstante, la presencia de armónicos añade una gran riqueza no solamente a los sonidos del habla, sino también a los sonidos producidos por instrumentos musicales, y en general a muchos otros tipos de sonidos. La ausencia de armónicos en la voz humana, haría que ésta sonara plana y poco interesante [40].

3.1.4. Espectro de Sonido

Mediante un espectro de sonido es posible representar las componentes frecuenciales de las ondas complejas que conforman a los sonidos. Éste se obtiene al realizar un análisis espectral del cual resulta una gráfica que resume visualmente las diversas frecuencias que componen al sonido, cuyos ejes horizontal y vertical simbolizan la frecuencia y amplitud de las componentes, respectivamente.

En la figura 3.3 se muestra el espectro de una onda compleja como bien puede ser la producida por la cuerda de una guitarra. La altura de cada línea representa la amplitud de la onda senoidal presente, la cual es usualmente mostrada como un nivel relativo de presión sonora

que se mide en Pascales, o también es mostrada en decibeles. Las componentes frecuenciales se encuentran equiespaciadas en caso de tratarse de ondas periódicas o separadas de manera no uniforme en el caso de ondas aperiódicas.

3.1.5. Formantes

Los diversos componentes acústicos de los sonidos del habla, que representan la manera en que el tracto vocal resuena durante la articulación de sonidos, puede también mostrarse a través de un espectro de sonido.

En el espectro, la amplitud de algunas frecuencias es mucho mayor que otras. De hecho, es posible ver varios picos de energía acústica en cada caso, reflejando los puntos principales de resonancia en el tracto vocal. Estos picos son conocidos como *formantes*, y al igual que los armónicos, son numerados secuencialmente, desde el más bajo hasta el más alto: la *primer formante* o *F1*, la *segunda formante* o *F2*, y así sucesivamente. La figura 3.4 muestra el espectro de la vocal /i/, la cual fue articulada por un hombre cuya voz tiene una frecuencia fundamental de 120 Hz. Se observa que la primer formante, *F1*, alcanza su máximo alrededor de 360 Hz, la segunda formante, *F2*, tiene un valor de 2,280 Hz, y la tercer formante, *F3*, tiene una amplitud de aproximadamente 3,000 Hz.

Las formantes son características muy importantes de los sonidos de voz. Todas las vocales y algunas consonantes tienen formantes. En el caso de las vocales, especialmente las dos primeras son fundamentales para que el ser humano sea capaz de diferenciarlas unas de otras, o bien, para reconocer repeticiones de una misma vocal incluso cuando es producida por diferentes hablantes [11].

3.1.6. Espectrograma

Durante el habla, los sonidos de voz cambian, de otra manera no sería posible comunicarse. En este sentido, un espectro que proporciona información frecuencial acerca de un sonido individual, no es suficiente para representar la naturaleza cambiante de los sonidos que se producen, y que es fundamental para comprender un mensaje. Lo que se requiere es una representación

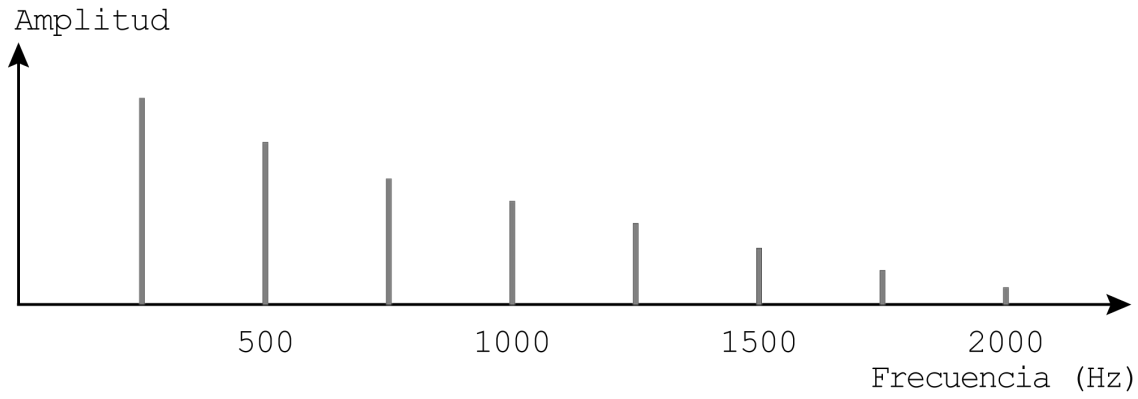


Figura 3.3: Espectro de sonido.

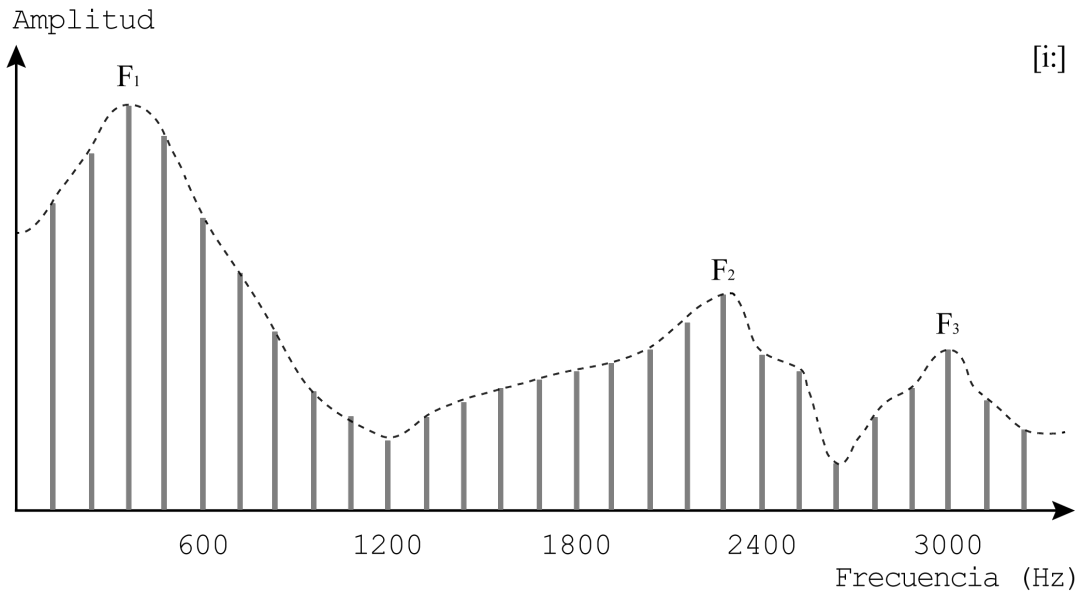


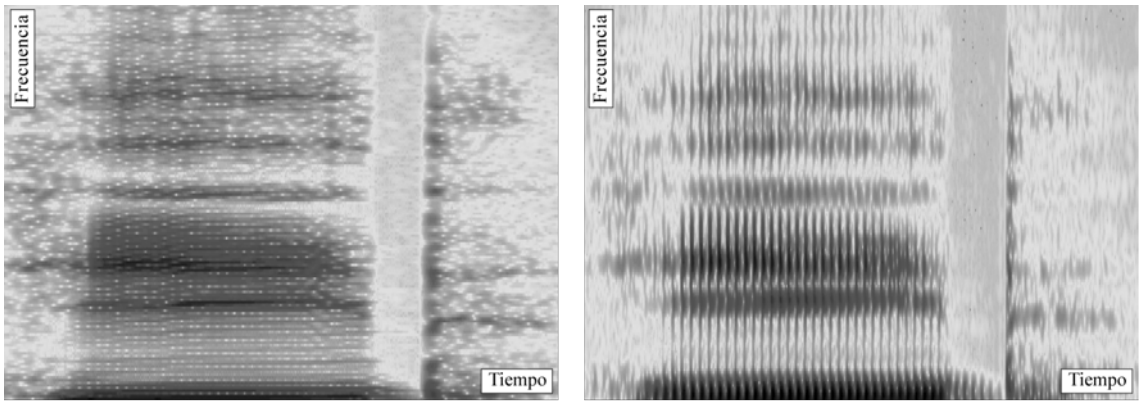
Figura 3.4: Espectro de sonido de la vocal /i/.

visual de la forma en que el espectro de sonido cambia con el tiempo, y esto se logra a través de un *espectrograma* [41]. Un espectrograma es entonces, una gráfica que muestra cómo el contenido frecuencial de una señal cambia con respecto al tiempo. Además, proporciona información más compleja que la que puede ser obtenida mediante la forma de onda de la señal en el dominio del tiempo, tal como la intensidad de las diferentes frecuencias presentes en la señal, cuya representación está dada por marcas de color blanco y negro o también multicolor; entre más intensa sea una frecuencia particular contenida en la señal más oscura es la marca.

Existen dos tipos de espectrograma: de *banda estrecha* y de *banda ancha*. Ambos son comúnmente utilizados para analizar señales de voz y su diferencia radica en la forma en la cual es llevado a cabo el análisis. En la figura 3.5 se muestran los dos tipos de espectrograma para un segmento muy corto de voz. En el caso del espectrograma de banda estrecha se analiza el intervalo de frecuencias de voz en pequeñas bandas (*generalmente 45 Hz*), mientras que en el espectrograma de banda ancha el análisis se realiza empleando bandas de frecuencia mucho más amplias (*generalmente 300 Hz*), lo cual hace que las formantes resalten claramente, por lo que este tipo de espectrograma resulta ser más útil en el procesamiento digital de señales de voz [11].

3.2. Representación Fonética de la Voz

El estudio de los sonidos de voz a menudo se divide en dos grandes disciplinas: la *Fonética* y la *Fonología*. En años recientes, sin embargo, los dos campos han coincidido cada vez más en su alcance [42]. De manera tradicional, la *Fonética* se ocupa de las propiedades físicas medibles de los sonidos de voz, es decir, la forma precisa en que los órganos articuladores son empleados para producir ciertos sonidos, y las características de las ondas de sonido resultantes [43]. En



a) Espectrograma de banda estrecha b) Espectrograma de banda ancha

Figura 3.5: Tipos de espectrograma

este sentido la Fonética responde a los siguientes cuestionamientos:

- Cómo se producen los sonidos de voz.
- Cuántos sonidos diferentes ocupan los idiomas.
- Cómo viaja a través del aire un sonido de voz.
- Cómo el oído humano registra los sonidos de voz.
- Cómo se puede medir el habla.

Por otro lado, la *Fonología* estudia la manera en que los idiomas organizan los sonidos en diferentes patrones, por lo que esta disciplina se encarga de atender las siguientes tareas:

- Cómo los idiomas organizan los sonidos para distinguir diferentes palabras.
- Qué tipos de restricciones o limitantes ponen los idiomas en las secuencias de sonidos.
- Qué tipos de cambios o alteraciones experimentan los sonidos si se presentan secuencias ilícitas.
- Cómo son organizados los sonidos en componentes más grandes, tales como sílabas, palabras o frases.

3.2.1. Alfabeto Fonético Internacional

El *Alfabeto Fonético Internacional* (*AFI*) es un alfabeto publicado por primera vez en 1888, que representa los sonidos de un idioma en forma escrita [45]. Fue creado por la Asociación Fonética Internacional (*API*, por sus siglas en francés), cuyo objetivo es promover el estudio de la Fonética y las diferentes aplicaciones prácticas de esa ciencia. Desde su fundación en 1886 la API se dio a la tarea de desarrollar una serie de símbolos convenientes de usar, pero a su

EL ALFABETO FONETICO INTERNACIONAL (actualizado en 2005)

CONSONANTES (INFRAGLOTALES)

	LABIAL		CORONAL				DORSAL			RADICAL		GLOTAL
	BILABIAL	LABIODENTAL	DENTAL	ALVEOLAR	POSTALVEOLAR	RETROFLEJA	PALATAL	VELAR	UVULAR	FARÍNGEA	EPIGLOTAL	
NASAL	m	ɱ		n		ɳ	ɲ	ŋ	N			
OCLUSIVA	p b	pˠ bˠ		t d		ʈ ɖ	c ɟ	k g	q G		ʡ	ʔ
FRICATIVA	ɸ β	f v	θ ð	s z	ʃ ʒ	ʂ ʐ	ç ʝ	x ɣ	χ ʁ	ħ ʕ	H ʢ	h ɦ
APROXIMANTE		ʋ		ɹ		ɻ	j	ɰ				
VIBRANTE MÚLTIPLE	ʙ			r					ʀ		ʜ	
VIBRANTE SIMPLE		ⱱ		ɾ		ɽ						
FRICATIVA LATERAL				ɬ ɮ			ʎ̝	ʟ̝				
APROXIMANTE LATERAL				l		ɭ	ʎ	ʟ				
VIBR. SIMPLE LATERAL				ɺ								

Las consonantes alineadas a la izquierda son sordas, las alineadas a la derecha sonoras. Las casillas en gris son articulaciones consideradas imposibles.

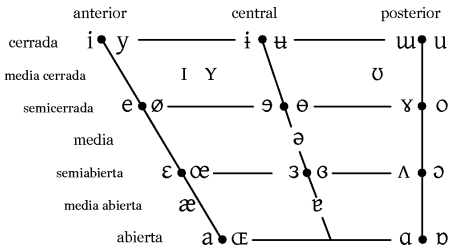
CONSONANTES (SUPRAGLOTALES)

CLIC	IMPLOSIVA	EYECTIVA
ʘ bilabial	ɓ bilabial	’ como en:
| dental	ɗ dental / alveolar	pʼ bilabial
! (post)alveolar	ʄ palatal	tʼ dental / alveolar
ǂ palatoalveolar	ɠ velar	kʼ velar
‖ lateral alveolar	ʛ uvular	sʼ fricativa alveolar

CONSONANTES (COARTICULADAS)

ʍ fricativa labiovelar sorda
w aproximante labiovelar sonora
ɥ aproximante labiopalatal sonora
ɕ fricativa alveopalatal sorda
ʑ fricativa alveopalatal sonora
ʃ͡ʒ ʃ y x simultáneas
k͡p t͡s Africadas y dobles articulaciones pueden representarse con dos símbolos atados con una cuña

VOCALES



SUPRASEGMENTALES

ˈ acento principal ˌ acento secundario
ːˑ larga ˑ semilarga
˘ corta ˇ breve
. rotura silábica ‿ enlace

ENTONACIÓN

| grupo menor (pie)
‖ grupo mayor (entonación)
↗ ascenso global ↘ descenso global

TONO

NIVEL	CONTORNO
e̋ ˥ extra alto	e᷄ ascendente
é ˦ alto	ê descendente
ē ˧ medio	e᷅ ascendente alto
è ˨ bajo	e᷈ ascendente bajo
ȅ ˩ extra bajo	e᷉ descendente alto
NIVELACIÓN	e᷌ descendente bajo
↓ - 1 tono	e᷇ ascendente descendente
↑ + 1 tono	e᷆ descendente ascendente

DIACRÍTICOS En algunos pueden aparecer arriba: n̊. En superíndice: tˢ (tendencia fricativa), bʱ (sonora mate), ʔa (ataque glotal), ᵊ (*schwa* epentético), oᵘ (diptongación)

SILABICIDAD Y TENDENCIA		FONACIÓN		ARTICULACIÓN PRIMARIA		ARTICULACIÓN SECUNDARIA			
ɹ̩ n̩	silábica	n̥ d̥	ensordecida	t̪ d̪	dental	tʷ dʷ	labializada	ɔ̹ x̹	más labializada
e̯ u̯	no silábica	s̬ t̬	sonorizada	t̺ d̺	apical	tʲ dʲ	palatalizada	ɔ̜ x̜ʷ	menos labializada
tʰ dʰ	aspirada	b̤ a̤	sonora mate	t̻ d̻	laminal	tˠ dˠ	velarizada	ẽ z̃	nasalizada
dⁿ	tendencia nasal	b̰ a̰	sonora estridente	u̟ t̟	avanzada	tˤ dˤ	faringizada	ɔˠ ɜˠ	rotacismo
dˡ	tendencia lateral	b̪̰ a̪̰	estridente	i̠ t̠	retraída	ɬ ɮ	velarizada o faríngizada	e̝ o̝	base de la lengua avanzada
d˺	tendencia no audible	t̼ d̼	linguolabial	ë ä	centralizada	ě ŭ	medio centralizada	e̞ o̞	base de la lengua retraída
e̬ β̞	descenso lingual (β es aproximante bilabial)			e̝ ɹ̝	ascenso lingual (ɹ es fricativa alveolar sonora no sibilante)				

Figura 3.6: Alfabeto Fonético Internacional [44].

vez lo suficientemente comprensibles para hacer frente a la extensa variedad de sonidos que constituyen a los diferentes idiomas del mundo; y también para motivar el uso de su notación entre las personas que realizan actividades relativas al lenguaje.

El *AFI* está basado en el alfabeto Romano, el cual no sólo tiene la ventaja de ser ampliamente conocido, sino que además incluye letras y símbolos adicionales de diversas fuentes, y que son necesarios debido a que la cantidad de sonidos en los idiomas es mucho más grande que el número de letras presentes en el alfabeto Romano.

El *AFI* puede ser utilizado para distintos propósitos. Por ejemplo, en un diccionario se emplea como una forma de mostrar la pronunciación de las palabras; o en el análisis de voz es usado para anotar la información acústica de la señal.

La notación empleada por el *AFI* está fundamentada en suposiciones teóricas acerca de la mejor manera en que la voz puede ser analizada. Algunas de éstas incluyen:

- Las características lingüísticamente relevantes del habla.
- La representación parcial del habla como una secuencia de sonidos discretos o *segmentos*.
- La división de los segmentos de voz en dos categorías principales: *consonantes* y *vocales*.
- La descripción fonética de las consonantes y vocales de acuerdo a la manera en que se producen y sus características auditivas.
- Aspectos *suprasegmentales*, tales como el estrés y el tono, cuya representación es independiente de los segmentos.

En la figura 3.6 se muestra un conjunto de tablas que resumen la organización del *AFI* y que reflejan las suposiciones antes mencionadas. No obstante, para presentar el *AFI* es necesario hacer referencia a ejemplos de palabras de algún idioma en particular. Sin embargo, es importante mencionar que debido a que todos los idiomas tienen diferentes acentos y variaciones en la pronunciación de su vocabulario, cuando un sonido del *AFI* es ejemplificado mediante una palabra, ésta representa el sonido que puede ser escuchado sin que necesariamente el mismo sonido ocurra en cada pronunciación [46].

3.2.2. Características Acústicas de las Vocales y Consonantes

Las *vocales* y *consonantes* son probablemente las dos categorías más importantes en la descripción del habla. En Fonética, las vocales se distinguen de las consonantes en términos de cómo son articuladas en el tracto vocal, y de los patrones asociados de energía acústica. Por otro lado, las consonantes son definidas como sonidos producidos por un cierre en el tracto vocal, o por un estrechamiento que es lo suficientemente marcado para que el aire no pueda escapar sin producir fricción audible. Las vocales son sonidos que no tienen tal constricción: el aire escapa de una forma relativamente ininterrumpida a través de la boca o nariz. Por consiguiente, es relativamente fácil "sentir" la articulación de las consonantes; mientras que las vocales, las cuales involucran únicamente movimientos leves de la lengua y labios, son difíciles de localizar en este sentido, pero son más fáciles de distinguir auditivamente [11]. En la tabla 3.2 se muestra la clasificación de las vocales y consonantes.

3.2.2.1. Vocales

Todos los sonidos vocálicos presentan estructuras de formantes bien definidas. La figura 3.7 presenta un esquema que refleja la estructura formántica de las vocales del Español. Los valores de las formantes varían considerablemente según el hablante y en función de diversos factores individuales tales como la edad, sexo o características anatómicas [47]; o también en función de la realización del habla, dependiente de factores contextuales tales como la velocidad o la formalidad del discurso [48].

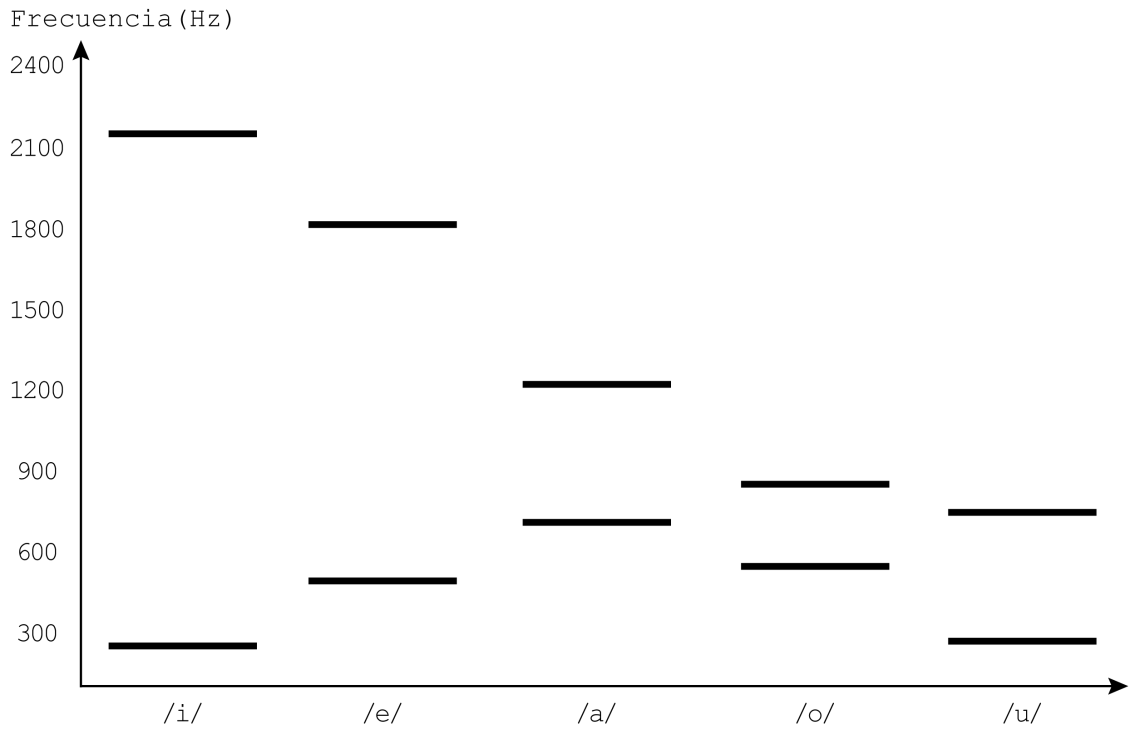


Figura 3.7: Estructura formántica de las vocales [49].

De acuerdo a [50], el popular proyecto *Sounds of Speech* de la Universidad de Iowa en los Estados Unidos de América, las vocales se clasifican de la siguiente manera:

- **Por la posición de la lengua.** Estas vocales son sonidos en los que no se constriñe el tracto vocal severamente sino que se crea una configuración abierta y global determinada principalmente por la posición de la lengua.
  - **Altas.** Las vocales altas se caracterizan por un movimiento de elevación de la lengua hacia el techo de la boca dejando una abertura relativamente estrecha por donde fluye el aire.
  - **Medias.** Las vocales medias se caracterizan por un movimiento de elevación de la lengua hacia el techo de la boca dejando una abertura más amplia que las vocales altas.

  - **Baja.** En la vocal baja el movimiento de elevación de la lengua es muy ligero de tal forma que ésta queda ocupando el hueco de la mandíbula inferior.

- **Diptongos.** Son secuencias de dos vocales que pertenecen a una misma sílaba. La vocal de mayor abertura desempeña el papel de núcleo silábico mientras que la otra pasa a funcionar como una semiconsonante o una semivocal según preceda o siga al núcleo.

  - **Crecientes.** Son diptongos donde la primera vocal es más cerrada, y por consiguiente, menos prominente que la segunda. Es decir, que los diptongos crecientes empiezan con una semiconsonante y terminan con una vocal.
  - **Decrecientes.** Son diptongos donde la segunda vocal es más cerrada, y por consiguiente, menos prominente que la primera. Es decir, que los diptongos decrecientes empiezan con una vocal y terminan con una semivocal.

- **Semiconsonantes.** Son vocales que preceden otra vocal típicamente más abierta, que es la que actúa como el núcleo de la sílaba a la que las dos vocales pertenecen. Las semiconsonantes son, por lo tanto, el elemento inicial de un diptongo.

  - **Anterior.** La semiconsonante anterior resulta de un breve movimiento de abertura en que la lengua se desplaza desde una posición palatal cerrada a la posición de cualquier otra vocal siguiente.
  - **Posterior.** La semiconsonante posterior resulta de un breve movimiento de abertura en que la lengua se desplaza de una posición labiovelar cerrada a la posición de cualquier otra vocal siguiente.

- **Semivocales.** Son vocales que siguen otra vocal típicamente más abierta, que es la que actúa como el núcleo de la sílaba a la que pertenecen las dos vocales. Las semivocales son, por lo tanto, el elemento final de un diptongo.

  - **Anterior.** La semivocal anterior resulta de un breve movimiento de cerrazón en que la lengua pasa de la posición de una vocal precedente a la posición alta y anterior de [i].
  - **Posterior.** La semivocal posterior resulta de un breve movimiento de cerrazón en que la lengua se desplaza de la posición de una vocal precedente a la posición alta y posterior de [u].

3.2.3. Consonantes

Las consonantes son casi siempre sonidos que aparecen al principio o al final de una vocal. Son producidas a partir de gestos de la lengua y labios, cuyo movimiento fluido da lugar a formas específicas. Los gestos son difíciles de describir dado que resulta mucho más fácil asociar una consonante con las posiciones de los órganos vocales que caracterizan el sonido a producir. Estas posiciones pueden ser descritas bastante bien si se consideran los siguientes factores: lo que hacen las cuerdas vocales; la ubicación dentro de la boca donde el sonido es producido, y lo que le ocurre al flujo de aire proveniente de los pulmones [51].

De acuerdo a [50], el popular proyecto *Sounds of Speech* de la Universidad de Iowa en los Estados Unidos de América, las consonantes se clasifican de la siguiente manera:

- **Modo.** Se refiere a la manera como se produce un sonido y cómo el flujo de aire es modificado a su paso por la cavidad bucal.
  - **Oclusivas.** Son consonantes caracterizadas por un bloqueo total del flujo de aire causado por una obstrucción completa que se crea cuando un articulador activo hace contacto con un articulador pasivo.
  - **Fricativas.** Son consonantes que se articulan forzando el aire a través de una hendidura estrecha creada por la fricción entre dos articuladores pero sin que se interrumpa el flujo de aire.
  - **Africadas.** Son consonantes cuya articulación incluye una fase de obstrucción total seguida de una fase de fricción. Durante la fase de obstrucción total el flujo de aire se interrumpe momentáneamente mientras que durante la fase de fricción el aire escapa forzadamente.
  - **Nasales.** Son consonantes cuya articulación requiere una obstrucción total en la cavidad oral acompañada de un descenso velar que permite que el aire fluya a través de la cavidad nasal.
  - **Espirantes.** Son consonantes en las que un articulador activo se aproxima a un articulador pasivo formando así una hendidura amplia por la que el aire escapa sin causar ruido turbulento.
  - **Laterales.** Son consonantes en las que la lengua produce un bloqueo central pero el aire escapa lateralmente porque los lados de la lengua descienden y se contraen para formar así canales laterales por los que el aire fluye continuamente.
  - **Vibrantes.** Son consonantes caracterizadas por un movimiento vibratorio del articulador activo sin que se interrumpa el flujo de aire.
- **Lugar.** Se refiere a la zona del tracto vocal donde un articulador activo actúa sobre un articulador pasivo para modificar la corriente de aire y matizar así el sonido.
  - **Labial.** Se refiere a los sonidos que se articulan por la acción de los labios.
  - **Coronal.** Se refiere a los sonidos que se articulan por la acción de la corona (ápice) de la lengua.
  - **Palatal.** Se refiere a los sonidos que se articulan con el dorso de la lengua elevándose hacia el paladar.
  - **Dorsal.** Se refiere a los sonidos que se articulan por la acción del dorso de la lengua.
  - **Gutural.** Se refiere a los sonidos que se articulan en la faringe o en la laringe.
- **Voz.** Se refiere a la vibración de las cuerdas vocales durante la articulación de un sonido.
  - **Sordas.** Son consonantes en cuya articulación las cuerdas vocales no entran en vibración.
  - **Sonoras.** Son consonantes en cuya articulación las cuerdas vocales entran en vibración.

Vocales	Posición de lengua	Altas	[i] [u]
		Medias	[e] [o]
		Baja	[a]
	Semiconsonantes	Anterior	[j]
		Posterior	[w]
	Semivocales	Anterior	[i̯]
		Posterior	[u̯]
	Diptongos	Crecientes	[ja] [je] [jo] [ju] [wa] [we] [wo] [wi]
		Decrecientes	[ai̯] [au̯] [ei̯] [eu̯] [oi̯] [ou̯]
Consonantes	Modo	Oclusivas	[p] [b] [t̪] [d̪] [k] [g]
		Fricativas	[f] [θ] [s] [s̪] [ʃ] [ʒ] [ʝ] [x] [χ] [h]
		Africadas	[ʧ] [ʤ]
		Nasales	[m] [ɱ] [n̥] [n̪] [n] [ň] [ɲ] [ŋ]
		Espirantes	[β̞] [ð̞] [ɣ]
		Laterales	[l̥] [l̪] [l] [ľ] [ʎ]
		Vibrantes	[ɾ] [r]
	Lugar	Labial	Bilabial: [p] [b] [β̞] [m] Labio-dental: [f̪] [ɱ]
		Coronal	Interdental: [θ] [ð̞] [l̥] [n̥] [t̪] [d̪] [n̪] [l̪] Alveolar: [ɾ] [r] [l] [n] [s] [s̪] Alveopalatal: [ʃ] [ʒ] [ʧ] [ʤ] [ň] [ľ]
		Palatal	[ʝ] [ʎ] [ɲ]
		Dorsal	Velar: [k] [g] [x] [ɣ] [ŋ] Uvular: [χ]
		Gutural	Glotal: [h]
	Voz	Sordas	[p] [t̪] [k] [f] [θ] [s] [s̪] [ʃ] [x] [χ] [h] [ʧ]
		Sonoras	[b] [d̪] [g] [ʒ] [ʤ] [m] [ɱ] [n̥] [n̪] [n] [ň] [ɲ] [ŋ] [β̞] [ð̞] [ʝ] [ɣ] [l] [l̥] [l̪] [ľ] [ʎ] [ɾ] [r]

Tabla 3.2: Clasificación de las vocales y consonantes según [50].

3.3. Resumen

En este capítulo se han presentado las características acústicas y fonéticas más importantes de la señal de voz. Se ha descrito su representación en el dominio de la frecuencia haciendo énfasis en los conceptos de frecuencia fundamental, armónicos y formantes, y su visualización mediante espectros de sonido y espectrogramas. También se ha presentado el Alfabeto Fonético Internacional como una serie de símbolos que representan los sonidos de cualquier lenguaje oral, así como una descripción de las principales caracteríscas de los sonidos de las vocales y consonantes, las dos categorías más importantes referentes al habla.

4

Procesamiento Digital de Señales de Voz

Como consecuencia de los avances tecnológicos en el área de la electrónica y computación, el procesamiento digital de señales ha sufrido un desarrollo extensivo tanto teórico como experimental, que ha impulsado de manera creciente el uso de las computadoras para establecer formas efectivas que permiten la transferencia de información entre seres humanos y máquinas. En este sentido, el intercambio de información a través de la voz, resulta una forma muy conveniente en la interacción hombre-máquina, ya que es el método de comunicación más natural y por consecuencia el más ampliamente utilizado por los humanos [52]. Este hecho ha sido fundamental para el crecimiento de una de las áreas más importantes del procesamiento digital de señales: el procesamiento digital de señales de voz.

4.1. Fundamentos de Procesamiento Digital de Señales

El *Procesamiento Digital de Señales* (*PDS*), como el término lo sugiere, es el procesamiento de señales empleando una computadora. Los fundamentos presentados en esta sección introducen conceptos básicos sobre señales y el análisis de sistemas, así como técnicas generales sobre la manipulación de señales en distintos dominios.

4.1.1. Señales

Las señales se encuentran en prácticamente todas las áreas en las que se desenvuelve el ser humano. Su presencia permite la comunicación no sólo entre personas sino también entre otros seres vivos e incluso máquinas. Sin embargo, definir precisamente lo que es una señal, es una tarea complicada, por lo que de manera general, cualquier cosa que transporta información acerca de la naturaleza, estado o comportamiento de algún fenómeno, se considera una señal. Algunos ejemplos de señales son: la voz humana, los cantos de los pájaros y el aroma de las flores.

Matemáticamente, una señal se representa mediante una función de valor real o complejo de una o varias variables reales. Bajo este contexto, los términos de señal y función pueden emplearse de manera indistinta. Cuando la función depende de una sola variable, se dice que la señal es unidimensional. Una señal de voz, la temperatura ambiental máxima diaria, la precipitación pluvial en un lugar, son todos ejemplos de señales de una dimensión. Por otro lado, cuando la función es dependiente de dos o más variables, se dice que la señal es multidimensional. Una imagen es un ejemplo que representa una señal de dos dimensiones.

4.1.1.1. Clasificación de Señales

Generalmente, las señales pueden clasificarse con base en su naturaleza y características, por lo que su representación y procesamiento depende de su tipo. De manera habitual son clasificadas en dos grandes grupos como *señales continuas* y *señales discretas*.

Una **señal continua** como la que se muestra en la figura 4.1 (a), es una función matemáticamente continua, es decir, que está definida para cada valor de su variable independiente. En una representación en el dominio del tiempo, dicha señal se denota empleando la variable independiente t, que representa al tiempo en segundos, como $x(t)$. En otras representaciones, la variable independiente puede simbolizar cualquier otra cantidad física distinta del tiempo.

Por otra parte, una **señal discreta** como la que se muestra en la figura 4.1 (b), está definida únicamente para valores discretos de su variable independiente. Matemáticamente, las señales discretas se representan como una secuencia de números x, en la cual el enésimo número en la secuencia se denota $x(n)$, donde n es un número entero.

Desde otros puntos de vista las señales también pueden ser clasificadas en:

- **Señales Determinísticas**: las cuales son funciones que están completamente definidas en el tiempo, por lo que su valor no es incierto, ya que puede ser predecible en cualquier instante de tiempo. El patrón de la señal es regular y puede ser caracterizado matemáticamente por una función del tiempo conocida. La figura 4.2 (a) muestra una señal senoidal como ejemplo de una señal determinística.

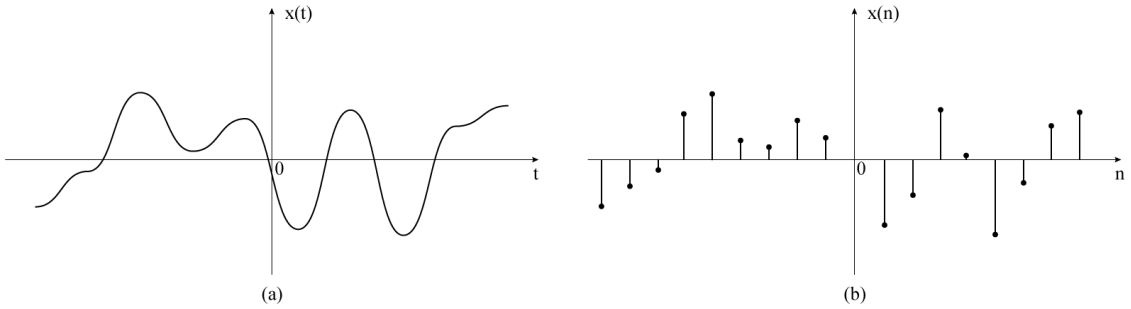


Figura 4.1: (a) Señal continua. (b) Señal discreta.

- **Señales Aleatorias**: las cuales se caracterizan por poseer un patrón irregular y cuya ocurrencia es aleatoria por naturaleza. Además, los valores de la señal en cada instante de tiempo no pueden ser predecibles. Su comportamiento es meramente probabilístico y puede ser analizado mediante procesos estocásticos. En la figura 4.2 (b) se muestra el ejemplo de una señal aleatoria.

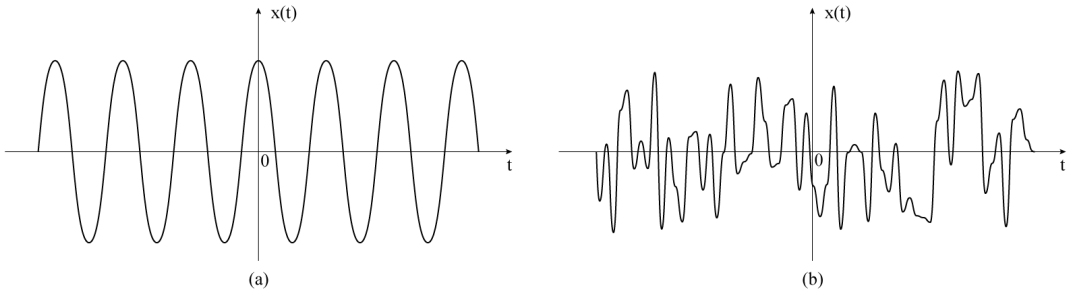


Figura 4.2: (a) Señal determinística. (b) Señal aleatoria.

- **Señales Periódicas**: las cuales poseen un patrón definido que se repite una y otra vez tal y como se muestra en la figura 4.3 (a). Matemáticamente, una señal continua $x(t)$ es periódica si existe un valor positivo de T para el cual

$$x(t) = x(t + T) \tag{4.1}$$

para todos los valores de t. Es decir, que una señal periódica tiene la propiedad de ser modificada por un desplazamiento de tiempo T. En dicho caso, se dice que $x(t)$ es una señal periódica con periodo T. Al valor positivo más pequeño de T que satisface la ecuación 4.1 se le conoce como *periodo fundamental*, T_0, de la señal.

Para una señal discreta, la condición de periodicidad puede ser escrita como

$$x(n) = x(n + N) \tag{4.2}$$

para todos los valores de n. El periodo fundamental, N_0, de la señal es el valor entero positivo más pequeño de N para el cual se satisface la ecuación 4.2. En la figura 4.3 (b) se muestra una señal periódica en tiempo discreto.

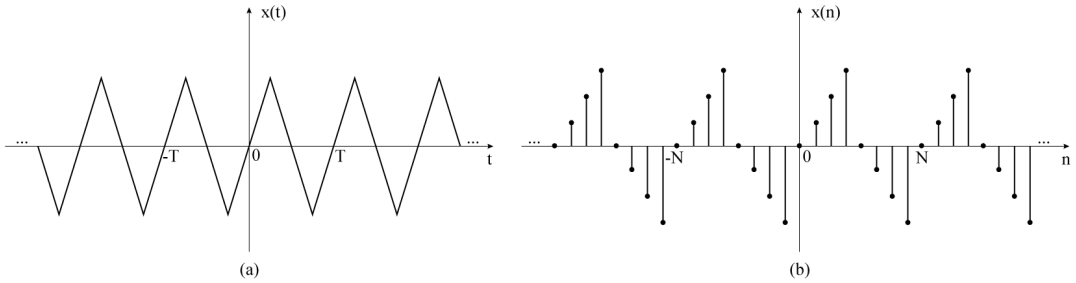


Figura 4.3: (a) Señal continua periódica. (b) Señal discreta periódica.

- **Señales Aperiódicas**: las cuales son funciones cuyos patrones son irregulares y no logran satisfacer las ecuaciones 4.1 o 4.2. La figura 4.4 muestra un par de señales aperiódicas en tiempo continuo y tiempo discreto.

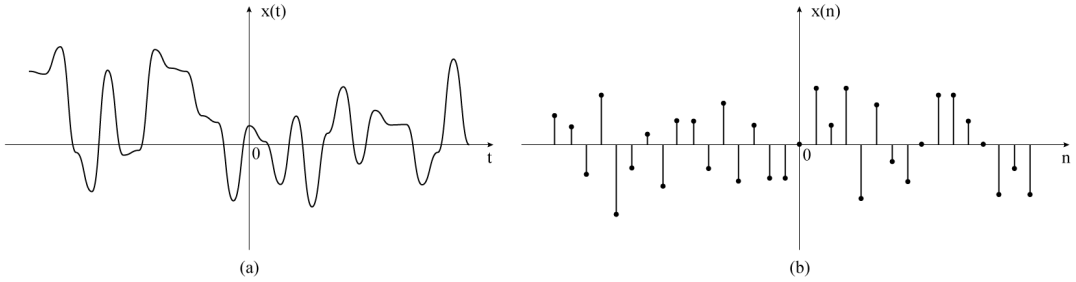


Figura 4.4: (a) Señal continua aperiódica. (b) Señal discreta aperiódica.

- **Señales Par e Impar**. Si una señal presenta simetría en el dominio del tiempo, es decir, que es idéntica a su reflejo respecto del origen, es una señal par. Matemáticamente, una señal par satisface la siguiente relación:

  Para una señal continua,

  $$x(t) = x(-t) \tag{4.3}$$

  Para una señal discreta,

  $$x(n) = x(-n) \tag{4.4}$$

  Por otro lado, si una señal presenta anti-simetría, es decir, que no es idéntica respecto al origen, entonces es una señal impar y satisface la siguiente relación:

  Para una señal continua,

  $$x(t) = -x(-t) \tag{4.5}$$

  Para una señal discreta,

  $$x(n) = -x(-n) \tag{4.6}$$

  La figura 4.5 muestra algunos ejemplos de señales par e impar en tiempo continuo y tiempo discreto.

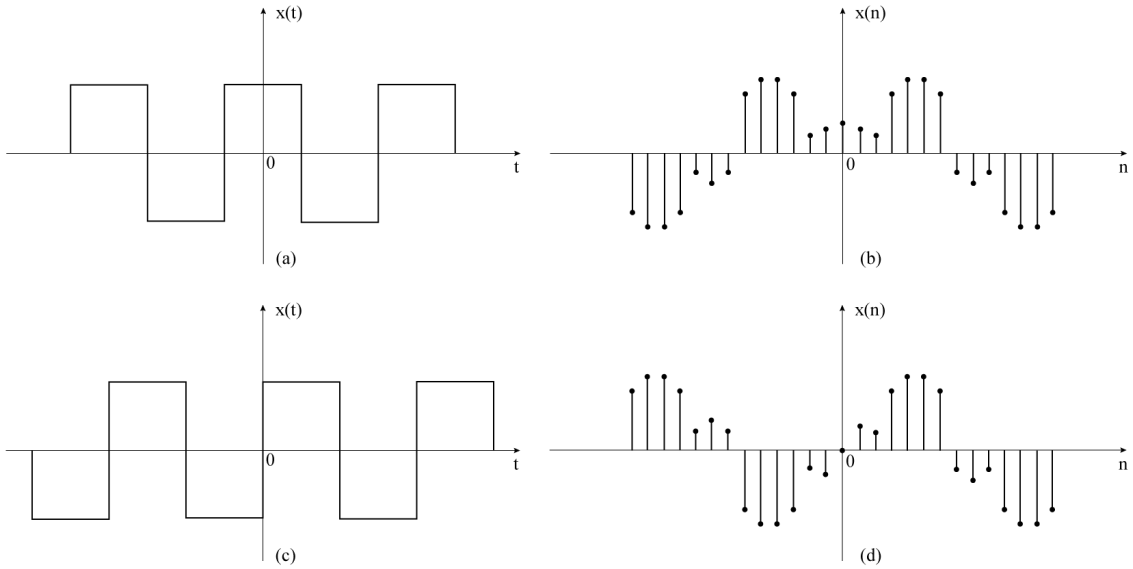


Figura 4.5: Ejemplos de señales par e impar en tiempo continuo y tiempo discreto. (a) Señal continua par. (b) Señal discreta par. (c) Señal continua impar. (d) Señal discreta impar.

Una señal puede ser expresada como la suma de sus componentes par e impar como se ilustra en la figura 4.6. En el caso de señales continuas, matemáticamente esta consideración puede expresarse de la siguiente manera:

$$x(t) = x(t)_{par} + x(t)_{impar} \tag{4.7}$$

donde,

$$x(t)_{par} = \frac{1}{2}[x(t) + x(-t)] \tag{4.8}$$

$$x(t)_{impar} = \frac{1}{2}[x(t) - x(-t)] \tag{4.9}$$

El mismo tipo de descomposición se aplica para señales discretas:

$$x(n) = x(n)_{par} + x(n)_{impar} \tag{4.10}$$

donde,

$$x(n)_{par} = \frac{1}{2}[x(n) + x(-n)] \tag{4.11}$$

$$x(n)_{impar} = \frac{1}{2}[x(n) - x(-n)] \tag{4.12}$$

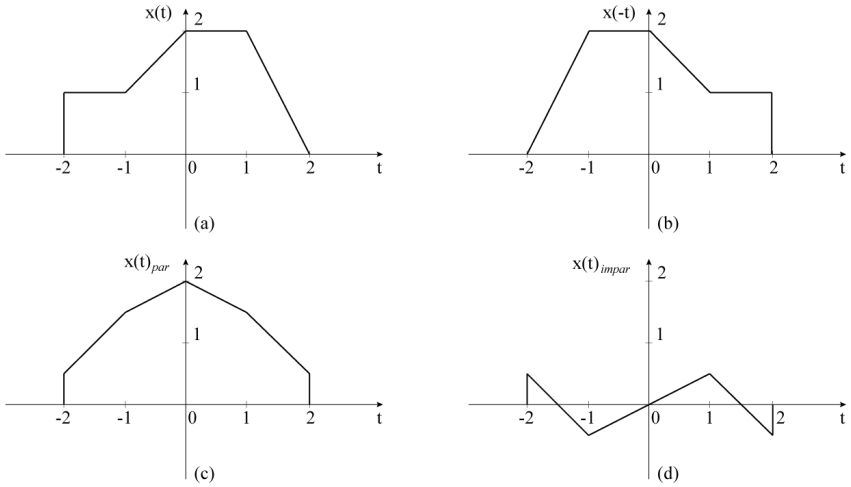


Figura 4.6: Descomposición de una señal continua en sus componentes par e impar. (a) Señal continua. (b) Señal reflejada. (c) Componente par de la señal. (d) Componente impar de la señal.

- **Señales de Energía**: las cuales tienen energía finita y potencia promedio igual a cero, i.e., la señal es de energía si $0 < E < \infty$, y $P = 0$, donde la energía y potencia de una señal están definidas por las siguientes ecuaciones:

  - **Energía**

    Para una señal continua,

    $$E_x = \int_{-\infty}^{\infty} |x(t)|^2 \mathrm{d}t \tag{4.13}$$

    Para una señal discreta,

    $$E_x = \sum_{n=-\infty}^{\infty} |x(n)|^2 \tag{4.14}$$

  - **Potencia**

    Para una señal continua,

    $$P_x = \lim_{T\to\infty} \frac{1}{T} \int_{-\infty}^{\infty} |x(t)|^2 \mathrm{d}t \tag{4.15}$$

    Para una señal discreta,

    $$P_x = \frac{1}{N} \sum_{n=0}^{N-1} |x(n)|^2 \tag{4.16}$$

- **Señales de Potencia**: las cuales tienen potencia promedio finita y energía infinita, i.e., la señal es de potencia si $0 < P < \infty$, y $E = \infty$. Si la señal no satisface cualquiera de las dos condiciones anteriores, entonces no es una señal de potencia ni de energía.

4.1.1.2. Transformaciones de la Variable Independiente

En el análisis de señales, frecuentemente ocurren tres operaciones básicas sobre la variable independiente para facilitar el procesamiento de la señal. Estas operaciones son: *escalamiento en el tiempo*, *reflexión o inversión en el tiempo* y *desplazamiento en el tiempo*.

El **escalamiento en el tiempo** expande o comprime una señal a lo largo del eje del tiempo. Una señal continua $x(t)$ es escalada en el tiempo al multiplicar la variable de tiempo t por una constante positiva b, para producir $x(bt)$. Un factor positivo de b puede expandir ($0 < b < 1$) o comprimir ($b > 1$) la señal en el tiempo.

Por otro lado, una secuencia de tiempo discreto $x(n)$ puede ser comprimida al multiplicar n por un número entero k, para producir la secuencia escalada en el tiempo $x(nk)$, o puede ser expandida al dividir n por un número entero m, para producir la secuencia escalada en el tiempo $x(n/m)$. Debido a que la variable independiente n solamente puede tomar valores enteros, la operación de escalamiento en el tiempo origina la aparición de muestras iguales a cero cuando la señal se expande, o la desaparición de muestras cuando la señal se comprime.

En la figura 4.7 se representa gráficamente la transformación de escalamiento en el tiempo aplicada a un señal en tiempo continuo y tiempo discreto.

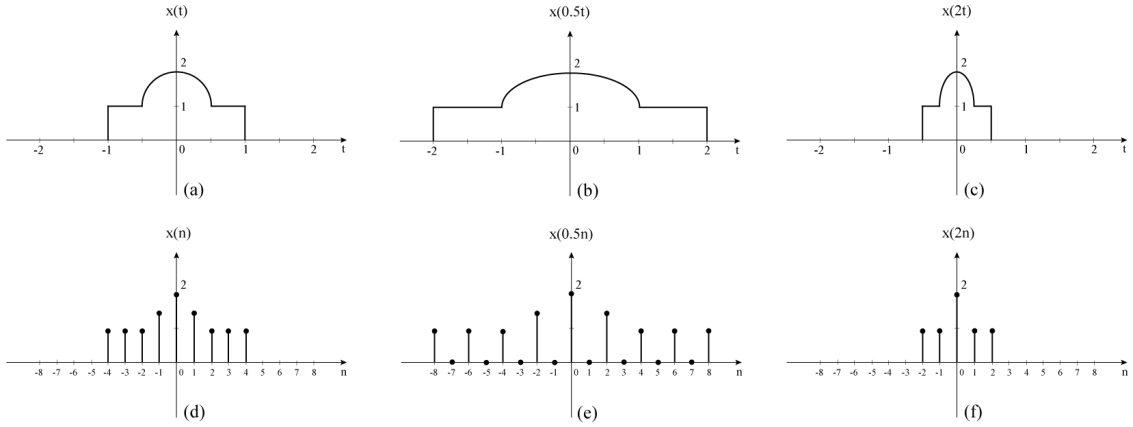


Figura 4.7: Escalamiento en el tiempo. (a) Señal continua $x(t)$. (b) Señal continua $x(t)$ expandida. (c) Señal continua $x(t)$ comprimida. (d) Señal discreta $x(n)$. (e) Señal discreta $x(n)$ expandida. (f) Señal discreta $x(n)$ comprimida.

La **reflexión** o **inversión en el tiempo** refleja la señal alrededor de $t = 0$, o alrededor de $n = 0$, provocando una inversión de la señal en el tiempo.

Una señal continua $x(t)$ es invertida en el tiempo al reemplazar t por $-t$, para producir $x(-t)$. De manera similar, una secuencia discreta $x(n)$ es invertida al reemplazar n por $-n$, para producir la señal invertida $x(-n)$. La figura 4.8 ilustra gráficamente la transformación de reflexión en el tiempo aplicada a una señal continua y a otra en tiempo discreto.

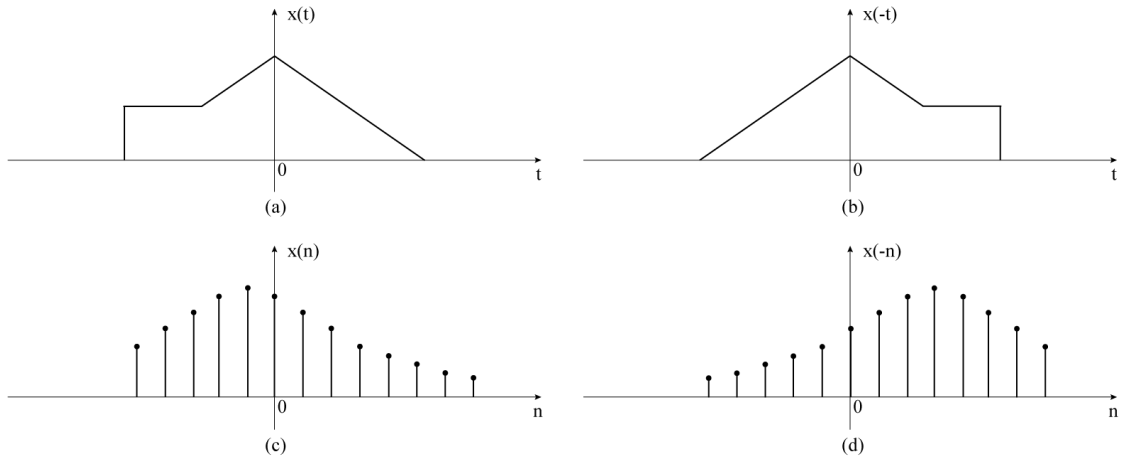


Figura 4.8: Reflexión o inversión en el tiempo. (a) Señal continua $x(t)$. (b) Señal continua reflejada $x(-t)$. (c) Señal discreta $x(n)$. (d) Señal discreta reflejada $x(-n)$.

El **desplazamiento en el tiempo** consiste en desplazar la señal a lo largo del eje del tiempo. En el caso de una señal continua $x(t)$, la variable independiente t es reemplazada por $t-t_0$, para producir $x(t-t_0)$. Cuando $t_0 > 0$, el desplazamiento de la señal es hacia la derecha, produciendo un atraso en la señal, mientras que cuando $t_0 < 0$, el desplazamiento de la señal es hacia la izquierda, produciendo un adelanto en la señal.

De la misma forma, para una señal discreta $x(n)$, la variable independiente n es reemplazada por $n - n_0$, para producir $x(n - n_0)$. Cuando $n_0 > 0$, se tiene un atraso en la señal, mientras que cuando $n_0 < 0$, se tiene un adelanto en la señal.

En la figura 4.9 se representa gráficamente la transformación de desplazamiento en el tiempo aplicada a un par de señales en tiempo continuo y tiempo discreto.

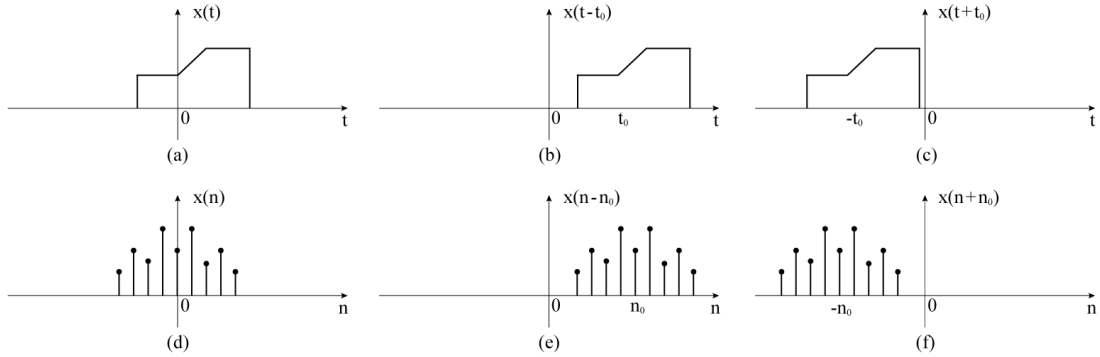


Figura 4.9: Desplazamiento en el tiempo. (a) Señal continua $x(t)$. (b) Señal continua retrasada $x(t - t_0)$. (c) Señal continua adelantada $x(t + t_0)$. (d) Señal discreta $x(n)$. (e) Señal discreta retrasada $x(n - n_0)$. (f) Señal discreta adelantada $x(n + n_0)$.

También es posible aplicar secuencialmente las operaciones de escalamiento, desplazamiento e inversión en el tiempo, a una señal para obtener otra de la forma $x(\alpha t + \beta)$ en el caso de señales continuas, o $x(\alpha n + \beta)$ en el caso de señales discretas. La señal resultante posee la misma forma de $x(t)$ o $x(n)$ independientemente del orden en el cual las operaciones son aplicadas. Sin embargo, la señal será expandida si $|\alpha| < 1$, comprimida si $|\alpha| > 1$, invertida en el tiempo si $\alpha < 0$, o presentará un desplazamiento en el tiempo si β es un número diferente de cero.

4.1.1.3. Funciones Singulares

Las *funciones singulares* son una clasificación muy importante de las señales aperiódicas. Son esenciales ya que con ellas es posible representar una gran variedad de señales. La función *impulso unitario*, es la función singular más básica, y todas las demás funciones singulares, tales como las funciones *escalón unitario*, *rampa unitaria* o *pulso unitario* pueden ser obtenidas al integrarla o diferenciarla repetidamente. A continuación se describen las funciones singulares más importantes.

- **Función Impulso Unitario**
  La función *impulso unitario*, también conocida como función *Delta* o *Delta de Dirac*, se denota como $\delta(t)$ en tiempo continuo, y como $\delta(n)$ en tiempo discreto.

  En tiempo continuo, $\delta(t)$ se define como:

$$\delta(t) = \begin{cases} \infty & t = 0 \\ 0 & t \neq 0 \end{cases} \tag{4.17}$$

  con la condición de que,

$$\int_{-\infty}^{\infty} \delta(t)\mathrm{d}t = 1 \tag{4.18}$$

  Las ecuaciones 4.17 y 4.18 indican que el área de la función impulso unitario es uno y que está confinada a un intervalo infinitesimal en el eje del tiempo y concentrado en $t = 0$.

  La función impulso unitario es muy útil en el análisis de sistemas y señales en tiempo continuo, ya que puede emplearse para:

  1. Muestrear señales.
  2. Descomponer señales en componentes elementales.
  3. Caracterizar la respuesta de una clase de sistemas.

  En tiempo discreto, la función impulso unitario es conocida como función *Delta de Kronecker*, *secuencia impulso unitario* o *muestra unitaria*. Dicha función se define como:

$$\delta(n) = \begin{cases} 1 & n = 0 \\ 0 & n \neq 0 \end{cases} \tag{4.19}$$

  La figura 4.10 muestra la representación gráfica de la función impulso unitario en tiempo continuo y discreto.

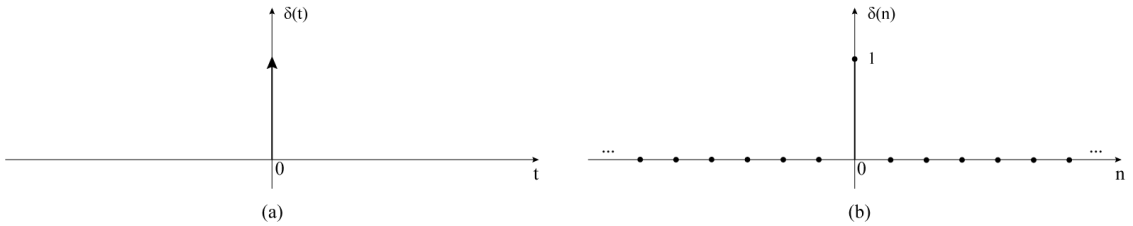


Figura 4.10: (a) Función impulso unitario en tiempo continuo. (b) Función impulso en tiempo discreto o secuencia o muestra unitaria.

- **Función Escalón Unitario**
  La función *escalón unitario*, se denota como $u(t)$ en tiempo continuo, y como $u(n)$ en tiempo discreto.

  En tiempo continuo, la función escalón unitario $u(t)$ se define como:

$$u(t) = \begin{cases} 1 & t \geq 0 \\ 0 & t < 0 \end{cases} \tag{4.20}$$

  La relación de la función escalón unitario con la función impulso unitario está dada por las siguientes ecuaciones:

$$\frac{\mathrm{d}u(t)}{\mathrm{d}t} = \delta(t) \tag{4.21}$$

  o bien,

$$\frac{\mathrm{d}}{\mathrm{d}t}[u(t - t_0)] = \delta(t - t_0) \tag{4.22}$$

$$\int_{-\infty}^{\infty} \delta(t)\mathrm{d}t = u(t) \tag{4.23}$$

  o de manera general,

$$u(t - t_0) = \int_{-\infty}^{t} \delta(\tau - t_0)\mathrm{d}\tau = \begin{cases} 1 & t > t_0 \\ 0 & t < t_0 \end{cases} \tag{4.24}$$

  En tiempo discreto, la función escalón unitario se conoce como *secuencia escalón unitario* o *escalón unitario discreto*. Se denota $u(n)$ y se define como:

$$u(n) = \begin{cases} 1 & n \geq 0 \\ 0 & n < 0 \end{cases} \tag{4.25}$$

La secuencia escalón unitario $u(n)$ puede ser construida a partir de una secuencia o muestra unitaria como se indica a continuación:

$$u(n) = \sum_{k=-\infty}^{n} \delta(k) \tag{4.26}$$

o bien,

$$u(n) = \sum_{k=0}^{\infty} \delta(n-k) \tag{4.27}$$

Sin embargo, la secuencia o muestra unitaria también puede ser construida a partir de la secuencia escalón unitario como:

$$\delta(n) = u(n) - u(n-1) \tag{4.28}$$

La figura 4.11 muestra la representación gráfica de la función escalón unitario en tiempo continuo y tiempo discreto.

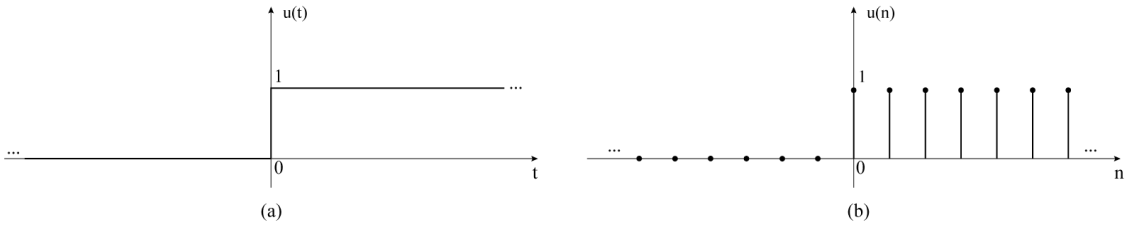


Figura 4.11: (a) Función escalón unitario en tiempo continuo. (b) Función escalón unitario en tiempo discreto.

- **Función Rampa Unitaria**
  La función *rampa unitaria*, se denota como $r(t)$ en tiempo continuo y como $r(n)$ en tiempo discreto.

  En tiempo continuo, la función rampa, $r(t)$, puede ser construida al integrar la función impulso unitario dos veces o al integrar la función escalón unitario una vez, esto es,

$$r(t) = \int_{-\infty}^{t}\int_{-\infty}^{\alpha} \delta(\tau)\mathrm{d}\tau\mathrm{d}\alpha = \int_{-\infty}^{t} u(\alpha)\mathrm{d}\alpha = \int_{0}^{t} \mathrm{d}\alpha \tag{4.29}$$

  es decir,

$$r(t) = \begin{cases} 0 & t < 0 \\ t & t > 0 \end{cases} \tag{4.30}$$

En tiempo discreto, la función rampa se conoce como *secuencia rampa* y se define como:

$$r(n) = \begin{cases} 0 & n < 0 \\ n & n > 0 \end{cases} \tag{4.31}$$

La figura 4.12 muestra la representación gráfica de la función rampa en tiempo continuo y tiempo discreto.

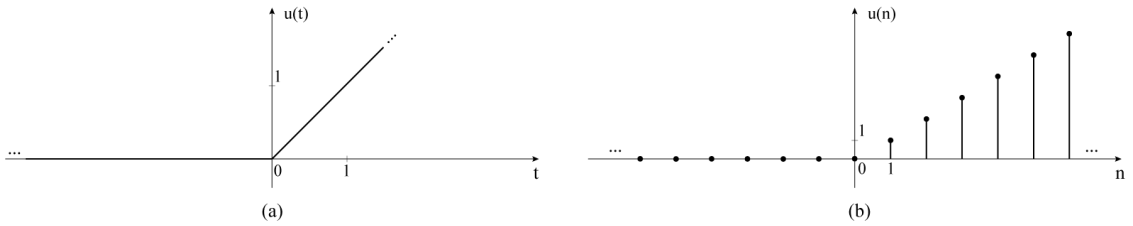


Figura 4.12: (a) Función rampa unitaria en tiempo continuo. (b) Función rampa unitaria en tiempo discreto o secuencia rampa.

- **Función Pulso Unitario**
  La función *pulso unitario*, también conocida como *ventana* o *pulso rectangular*, se denota $\Pi(t)$ en tiempo continuo y se define como:

$$\Pi(t) = \begin{cases} 1 & |t| \leq \frac{1}{2} \\ 0 & |t| > \frac{1}{2} \end{cases} \tag{4.32}$$

En términos de la función escalón unitario $u(t)$, la función pulso unitario se define como:

$$\Pi(t) = u(t + \tfrac{1}{2}) - u(t - \tfrac{1}{2}) \tag{4.33}$$

La figura 4.13 muestra la representación gráfica de la función pulso unitario.

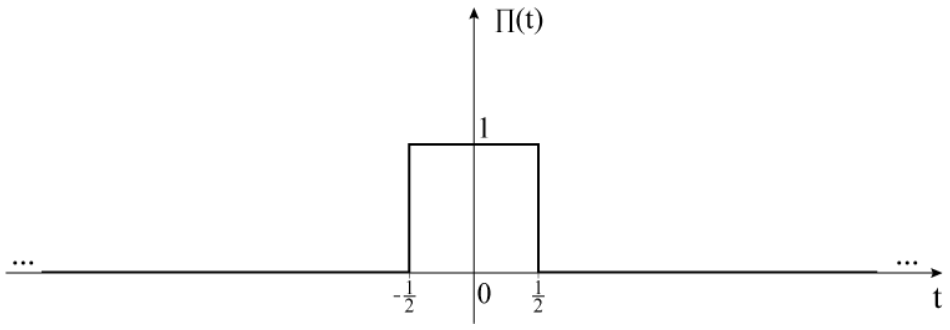


Figura 4.13: Función pulso unitario

4.1.1.4. Propiedades de la Función Impulso

En tiempo continuo algunas de las propiedades más importantes son:

- **Propiedad de Escalamiento**
$$\delta(at) = \frac{\delta(t)}{|a|} \tag{4.34}$$

- **Propiedad de Multiplicación**
La multiplicación de una función $f(t)$ por la función impulso en el instante $t = t_0$ es equivalente a la función $f(t)$ valuada en el instante $t = t_0$, es decir,
$$f(t)\delta(t - t_0) = f(t_0)\delta(t - t_0) \tag{4.35}$$

- **Propiedad de Muestreo**
Para cualquier función $f(t)$ continua en $t = t_0$,
$$\int_{-\infty}^{\infty} f(t)\delta(t - t_0)\mathrm{d}t = f(t_o) \tag{4.36}$$
La propiedad de muestreo de la función delta de Dirac, permite medir el valor de $f(t)$ en el instante $t = t_0$.

- **Propiedad de Convolución**
La convolución de una función $f(t)$ con una función impulso $\delta(t)$, da como resultado la misma señal $f(t)$, es decir,
$$\int_{-\infty}^{\infty} f(\tau)\delta(t - \tau)\mathrm{d}\tau = f(t) \tag{4.37}$$
De igual forma, la convolución de una función $f(t)$ con una función impulso desplazada un valor t_0, es equivalente a desplazar la función $f(t)$ el mismo valor t_0, es decir,
$$f(t) * \delta(t - t_0) = f(t - t_0) \tag{4.38}$$

En tiempo discreto, algunas de las propiedades más importantes de la función impulso son:

- **Propiedad de Multiplicación**
Si una función discreta $x(n)$ es multiplicada por una secuencia impulso unitario desplazada $\delta(n - k)$, el resultado será una secuencia que vale cero en todos los puntos excepto en $x(k)$, es decir,
$$x(n)\delta(n - k) = x(k)\delta(n - k) \tag{4.39}$$

- **Propiedad de Convolución**
La convolución de una secuencia discreta $x(n)$ con una secuencia impulso unitario $\delta(n)$, produce la misma función $x(n)$, es decir,
$$x(n) = x(n) * \delta(n) \tag{4.40}$$

o bien,

$$x(n) = \sum_{k=-\infty}^{\infty} x(k)\delta(n-k) \tag{4.41}$$

La ecuación 4.40 establece que cualquier señal discreta $x(n)$ puede ser representada mediante la sumatoria de un número infinito de impulsos unitarios ponderados y desplazados, donde el peso o amplitud de cada impulso $\delta(n-k)$ es el valor de $x(k)$.

- **Propiedad de Muestreo**
  La multiplicación de una secuencia discreta $x(n)$, con una secuencia impulso unitario $\delta(n)$ desplazada un valor k, permite aislar el valor de la secuencia discreta $x(n)$ valuada en el instante $n = k$, es decir,

$$x(k) = \sum_{n=-\infty}^{\infty} x(k)\delta(n-k) \tag{4.42}$$

4.1.2. Sistemas Discretos

En el contexto de procesamiento digital de señales, un *sistema* puede definirse como una descripción cuantitativa de un proceso físico capaz de transformar una señal. En otras palabras, y de forma más precisa, un sistema es una especie de "*caja negra*" (visto como una abstracción matemática) que transforma de manera determinística, señales de entrada en señales de salida [53].

Un sistema discreto puede definirse matemáticamente mediante una función de transformación, denotada T, que opera sobre una señal de entrada $x(n)$, para producir una señal de salida $y(n)$, es decir,

$$y(n) = \mathrm{T}\{x(n)\} \tag{4.43}$$

La ecuación 4.43 resume la relación existente entre sistemas y señales para el caso de una variable independiente, donde $y(n)$ es la respuesta del sistema a la señal de excitación $x(n)$. En la figura 4.14 se ilustra gráficamente esta relación.

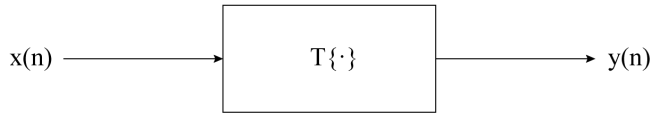


Figura 4.14: Representación gráfica de un sistema discreto.

4.1.2.1. Clasificación de Sistemas

Los sistemas discretos pueden clasificarse de la siguiente manera:

- **Sistemas estáticos o sin memoria**
  Un sistema es *estático* si su salida en un momento depende únicamente del valor de la entrada en ese momento en particular, es decir, la respuesta del sistema no depende de

valores pasados o futuros de la señal de entrada. Es por ello que los sistemas estáticos también son conocidos como sistemas *sin memoria*.

Por ejemplo, la función $y(n) = x(n)$ representa un sistema estático o sin memoria, ya que el sistema pasa la entrada a la salida de forma directa, en este caso sin efectuar algún procesamiento.

- **Sistemas dinámicos o con memoria**
  Un sistema es *dinámico* o *con memoria* si su salida en un momento específico depende del valor de la entrada en ese momento en particular, pero además de valores anteriores de la señal.

  Por ejemplo, la función $y(n) = x(n) + x(n - N)$ describe a un sistema dinámico con memoria de duración N.

- **Sistemas causales y no causales**
  Un sistema es *causal* si su salida en un momento específico no depende de valores futuros de la señal de entrada, sino de únicamente del valor presente o valores anteriores de la entrada. Si la respuesta del sistema es dependiente de valores futuros de la señal de entrada, entonces el sistema es *no causal*. Normalmente, este tipo de sistemas no es realizable.

  Por ejemplo,

  - La función $y(n) = x(n - 1)$ describe a un sistema *causal* porque $y(n)$ depende únicamente de $x(n - 1)$, valor anterior de $x(n)$.
  - La función $y(n) = x(n) + x(n + 1)$ describe a un sistema *no causal*, ya que $y(n)$ depende de $x(n + 1)$, valor futuro de $x(n)$.

- **Sistemas lineales y no lineales**
  Un sistema es *lineal* si cumple con el principio de *superposición*, es decir, que satisface las siguientes propiedades:

  - ***Aditividad***: donde la salida del sistema excitado de manera simultánea por dos señales independientes, es igual a la suma algebraica de sus salidas para cada una de las entradas aplicadas individualmente. Matemáticamente, esta propiedad es descrita como:
    $$\mathrm{T}\{x_1(n) + x_2(n)\} = \mathrm{T}\{x_1(n)\} + \mathrm{T}\{x_2(n)\} \tag{4.44}$$
  - ***Homogeneidad***: donde la salida del sistema a una sola entrada independiente es proporcional a la entrada. Matemáticamente, esta propiedad es descrita como:
    $$\mathrm{T}\{ax(n)\} = a\mathrm{T}\{x(n)\} \tag{4.45}$$

  Las dos propiedades anteriores que definen a un sistema lineal pueden ser combinadas en la siguiente ecuación que describe el principio de superposición aplicado a dos señales,

  $$\mathrm{T}\{ax_1(n) + bx_2(n)\} = a\mathrm{T}\{x_1(n)\} + b\mathrm{T}\{x_2(n)\} \tag{4.46}$$

De forma general, el principio de superposición es descrito matemáticamente como:

$$\mathrm{T}\left\{\sum_{i=1}^{N} a_i x_i(n)\right\} = \sum_{i=1}^{N} a_i \mathrm{T}\{x_i(n)\} \tag{4.47}$$

La ecuación 4.47 establece que la salida o respuesta de un sistema lineal que ha sido excitado por una suma ponderada de N señales arbitrarias, es igual a la misma suma ponderada de las respuestas individuales del sistema para cada una de las N señales, lo cual implica que,

$$y(n) = \mathrm{T}\left\{\sum_{i=1}^{N} a_i x_i(n)\right\} = \sum_{i=1}^{N} a_i x_i(n) \tag{4.48}$$

En la figura 4.15 se ilustra gráficamente el principio de superposición.

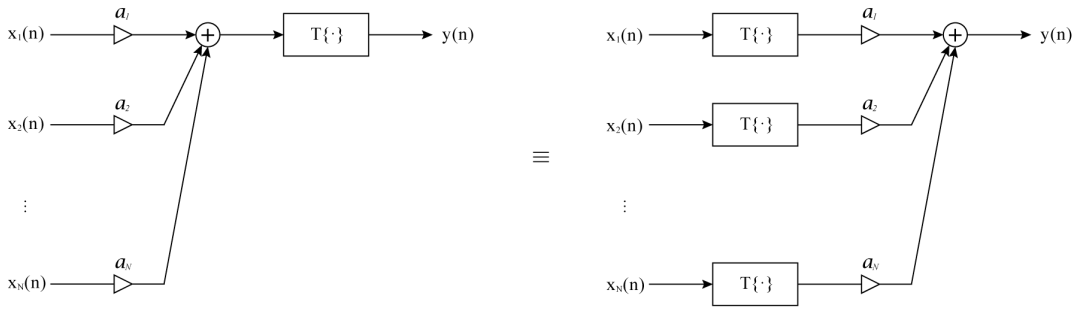


Figura 4.15: Representación gráfica del principio de superposición. Adaptado de [54].

- **Sistemas variantes e invariantes en el tiempo**
  Un sistema es *invariante* en el tiempo si un desplazamiento en el tiempo ocurrido en la señal de entrada, produce el mismo desplazamiento en el tiempo en la salida. Analíticamente, si se tiene que:

$$x(n) \longrightarrow y(n), \tag{4.49}$$

  el sistema será invariante en el tiempo si cumple con:

$$x(n - n_0) \longrightarrow y(n - n_0), \tag{4.50}$$

  es decir, la respuesta del sistema no es dependiente del tiempo en que se aplica la entrada. En tanto que para un sistema *variante* en el tiempo, la salida es dependiente del tiempo en que la entrada es aplicada.

- **Sistemas estables e inestables**
  Un sistema es *estable* si una señal de entrada $x(n)$ acotada en amplitud, produce una señal de salida $y(n)$ también acotada en amplitud, es decir,

$$|x(n)| \leq A_x < \infty, \qquad |y(n)| \leq A_y < \infty, \qquad \forall n \tag{4.51}$$

  Este tipo de sistemas son denominados sistemas *BIBO*, por las siglas en inglés *Bounded Input Bounded Output*. Si para cualquier entrada $x(n)$ acotada en amplitud la salida no es acotada, es decir, es infinita en amplitud, entonces el sistema es *inestable*.

4.1.2.2. Sistemas Discretos Lineales e Invariantes en el Tiempo

Dentro de la clasificación general de los sistemas discretos, los esquemas de clasificación basados en las propiedades de linealidad e invarianza en el tiempo son los más comunes e importantes. Debido a que una gran variedad de procesos físicos poseen estas propiedades, los sistemas más simples pero a la vez más frecuentemente utilizados para modelar dichos procesos, son los *sistemas discretos lineales e invariantes en el tiempo* o *sistemas discretos LIT*, ya que son sistemas que combinan ambas propiedades.

4.1.2.3. Respuesta al Impulso y la Suma de Convolución

Un hecho importante respecto al comportamiento de los sistemas discretos LIT, es que la respuesta del sistema ante cualquier entrada está determinada completamente por su respuesta a una señal de entrada en particular: la secuencia o muestra unitaria en el instante $n = 0$ [55].

La *respuesta al impulso*, como se denomina a la respuesta del sistema ante una muestra unitaria de entrada, se denota por $h(n)$, es decir,

$$h(n) = \mathrm{T}\{\delta(n)\} \tag{4.52}$$

Para un sistema invariante en el tiempo, la ecuación 4.52 se puede expresar como:

$$h(n-k) = \mathrm{T}\{\delta(n-k)\} \tag{4.53}$$

De la relación existente entre una señal de entrada y otra de salida a partir de una función de transformación expresada por la ecuación 4.43 y la representación de cualquier entrada arbitraria como una suma de impulsos ponderados establecida en la ecuación 4.41, se tiene que,

$$y(n) = \mathrm{T}\left\{\sum_{k=-\infty}^{\infty} x(k)\delta(n-k)\right\} \tag{4.54}$$

Aplicando a la ecuación 4.54 el principio de superposición establecido en la ecuación 4.46,

$$y(n) = \sum_{k=-\infty}^{\infty} x(k)\mathrm{T}\{\delta(n-k)\} \tag{4.55}$$

y sustituyendo la ecuación 4.53 en la ecuación 4.55, se obtiene,

$$y(n) = \sum_{k=-\infty}^{\infty} x(k)h(n-k) \tag{4.56}$$

Lo cual significa que un sistema discreto LIT puede ser caracterizado completamente por su respuesta al impulso $h(n)$, en el sentido de que, dado $h(n)$, es posible emplear la ecuación 4.56 para conocer la salida $y(n)$ ante cualquier entrada arbitraria $x(n)$ [56].

La ecuación 4.56 se conoce como *suma de convolución*, la cual también puede representarse mediante la siguiente notación:

$$y(n) = x(n) * h(n) \tag{4.57}$$

donde $*$ representa al operador *convolución*, de manera que $y(n)$ es la convolución de la señal de entrada $x(n)$, con la respuesta al impulso $h(n)$ del sistema.

4.1.2.4. Propiedades de la Convolución e Interconexión de Sistemas Discretos LIT

Figura 4.16: Sistema discreto en términos de su respuesta al impulso

La relación existente entre la respuesta del sistema $y(n)$ y las secuencias discretas $x(n)$ y $h(n)$, se representa gráficamente en la figura 4.16. Dicha representación resulta fundamental para comprender las propiedades de la convolución en función de la interconexión de sistemas discretos LIT. Algunas de las propiedades más importantes se describen a continuación.

- **Conmutatividad**
  La propiedad *conmutativa* establece que la señal de entrada $x(n)$ y la respuesta al impulso $h(n)$ pueden intercambiar la función que desempeñan dentro de un sistema, de manera que es posible que $h(n)$ represente una señal de excitación y x(n) la respuesta al impulso del sistema. Gráficamente, esta propiedad se ilustra en la figura 4.17, y matemáticamente se describe como:

$$x(n) * h(n) = h(n) * x(n) \tag{4.58}$$

Figura 4.17: Interpretación de la propiedad conmutativa de la convolución.

- **Asociatividad**
  La propiedad *asociativa* establece que la salida $y_1(n)$ de un sistema con respuesta al impulso $h_1(n)$ y entrada $x(n)$, que se convierte en la entrada de un segundo sistema con respuesta al impulso $h_2(n)$, es igual a la salida $y(n)$ de un sistema único con respuesta al impulso

$$h(n) = h_1(n) * h_2(n) \tag{4.59}$$

  cuando se aplica a la entrada la misma señal de excitación $x(n)$. La figura 4.18. muestra gráficamente esta propiedad y matemáticamente se se describe como:

$$[x(n) * h_1(n)] * h_2(n) = x(n) * [h_1(n) * h_2(n)] \tag{4.60}$$

Figura 4.18: Interpretación de la propiedad asociativa de la convolución.

- **Distributividad**
  La propiedad *distributiva* de la convolución ilustrada gráficamente en la figura 4.19, establece que,

$$x(n) * [h_1(n) + h_2(n)] = x(n) * h_1(n) + x(n) * h_2(n) \tag{4.61}$$

  es decir, si dos sistemas discretos LIT con respuesta al impulso $h_1(n)$ y $h_2(n)$, respectivamente, son excitados por una misma señal de entrada $x(n)$, la suma de las respuestas individuales de los sistemas es igual a la respuesta de únicamente un sistema con respuesta al impulso

$$h(n) = h_1(n) + h_2(n) \tag{4.62}$$

Figura 4.19: Interpretación de la propiedad distributiva de la convolución.

4.1.3. Análisis de Fourier

La representación de Fourier de señales continuas y discretas es muy importante en el procesamiento de señales, ya que proporciona un método para trasladar señales a otro dominio en el cual puedan ser manipuladas [57]. Este otro dominio corresponde al *dominio de la frecuencia*, cuya representación y métodos resultantes, históricamente emergieron de la transformación de señales en el dominio del tiempo en otras formas útiles. Dos de las más notables son la *Serie de Fourier* y la *Transformada de Fourier*, desarrolladas por el matemático francés Jean-Baptiste Joseph Fourier.

4.1.3.1. Respuesta en Frecuencia

Al igual que la respuesta al impulso, la *respuesta en frecuencia* es muy útil en la caracterización de los sistemas discretos LIT. La respuesta en frecuencia define la forma en que cambia en amplitud (compleja) una exponencial compleja cuando pasa por un sistema.

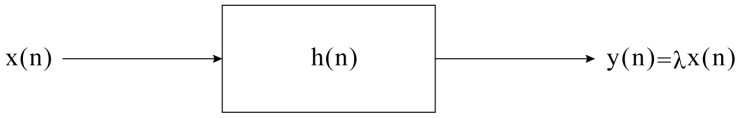


Figura 4.20: Eigenfunción de un sistema discreto LIT.

Las *eigenfunciones* de un sistema invariante en el tiempo son secuencias que al excitar un sistema, pasan directamente hacia la salida de éste con solamente un cambio complejo en amplitud. Es decir, si la entrada al sistema es una señal $x(n)$, la salida es $y(n) = \lambda x(n)$, donde λ, el

eigenvalor, generalmente depende de la entrada $x(n)$. Este concepto se ilustra gráficamente en la figura 4.20. Señales de la forma:

$$x(n) = e^{j\omega n} \qquad -\infty < n < \infty \tag{4.63}$$

donde ω es una constante, son eigenfunciones de sistemas discretos lineales e invariantes en el tiempo. Esto puede demostrarse a partir de la suma de convolución:

$$\begin{aligned} y(n) = h(n) * x(n) &= \sum_{k=-\infty}^{\infty} h(k)x(n-k) \\ &= \sum_{k=-\infty}^{\infty} h(k)e^{j\omega(n-k)} = e^{jn\omega} \sum_{k=-\infty}^{\infty} h(k)e^{-j\omega k} \\ &= H(e^{j\omega})e^{j\omega n} \end{aligned}$$

Por consiguiente, el eigenvalor, el cual se denota $H(e^{j\omega})$, es:

$$H(e^{j\omega}) = \sum_{k=-\infty}^{\infty} h(k)e^{-j\omega k} \tag{4.64}$$

donde $H(e^{j\omega})$ es, en general, de valores complejos y depende de la frecuencia ω de la exponencial compleja. Por lo tanto, puede ser escrita en términos de su parte *real* e *imaginaria*, es decir,

$$H(e^{j\omega}) = H_R(e^{j\omega}) + jH_I(e^{j\omega}) \tag{4.65}$$

o en términos de su *magnitud* y *fase*,

$$H(e^{j\omega}) = |H(e^{j\omega})|e^{j\phi_h(\omega)} \tag{4.66}$$

donde

$$|H(e^{j\omega})|^2 = H(e^{j\omega})H^*(e^{j\omega}) = H_R^2(e^{j\omega}) + H_I^2(e^{j\omega}) \tag{4.67}$$

y

$$\phi_h(\omega) = \tan^{-1}\frac{H_I(e^{j\omega})}{H_R(e^{j\omega})} \tag{4.68}$$

4.1.3.2. Filtros Selectivos en Frecuencia

En muchas aplicaciones del procesamiento digital de señales, algunas veces es necesario realizar transformaciones a los componentes frecuenciales de una señal. Este proceso se denomina *filtrado* y es llevado a cabo por *filtros digitales*. De manera que un filtro digital describe un sistema lineal invariante en el tiempo capaz de amplificar, atenuar o suprimir frecuencias específicas de una señal.

Los filtros pueden ser caracterizados en términos de sus propiedades, tales como linearidad, invarianza en el tiempo, causalidad, estabilidad, etc., y también pueden ser clasificados en términos de la forma de su respuesta en frecuencia.

Los filtros cuya respuesta en frecuencia es la unidad a lo largo de un cierto intervalo de frecuencias y es cero en las frecuencias restantes, se denominan *filtros selectivos en frecuencia* [56]. Específicamente, a los intervalos donde la magnitud de la respuesta en frecuencia es constante se les conoce como *bandas de paso*, y a aquellos en los cuales es nula se les conoce como *bandas suprimidas*. Las frecuencias que marcan los límites de las bandas de paso y bandas suprimidas se denominan *frecuencias de corte* [57]. La figura 4.21 muestra la respuesta en frecuencia de varios tipos de filtros ideales.

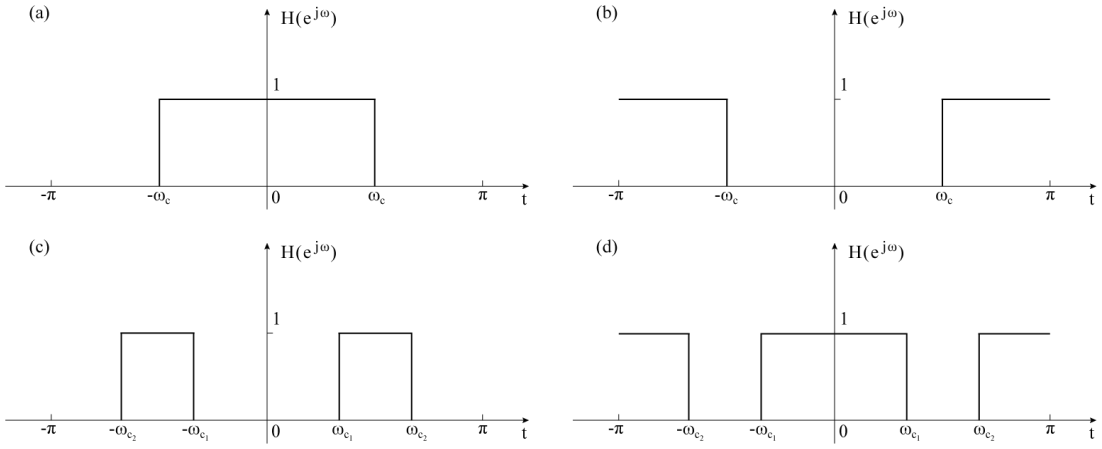


Figura 4.21: Filtros selectivos en frecuencia. (a) Filtro pasobajas. (b) Filtro pasoaltas. (c) Filtro pasobanda. (d) Filtro supresor de banda.

4.1.3.3. Transformada de Fourier en Tiempo Discreto

La respuesta en frecuencia de un sistema discreto LIT puede determinarse mediante una suma de multiplicaciones de la respuesta al impulso $h(n)$ con una exponencial compleja $e^{-jn\omega}$, evaluada para cada valor de n, donde $-\infty < n < \infty$. En este sentido, la *transformada de Fourier en tiempo discreto* o *DTFT* de una secuencia $x(n)$, se define como:

$$X(e^{j\omega}) = \sum_{n=-\infty}^{\infty} x(n)e^{-j\omega n} \tag{4.69}$$

donde $x(n)$ puede representarse mediante una integral de Fourier de la forma,

$$x(n) = \frac{1}{2\pi}\int_{-\pi}^{\pi} X(e^{j\omega})e^{j\omega n}d\omega \tag{4.70}$$

La ecuación 4.70 representa la *transformada de Fourier en tiempo discreto inversa* o *IDTFT*, que puede ser vista como una descomposición de la secuencia $x(n)$ en una combinación lineal de exponenciales complejas cuyas frecuencias se encuentran en un intervalo de longitud 2π.

En el tiempo continuo, las contrapartes naturales de las ecuaciones 4.69 y 4.70 son:

$$X(\omega) = \int_{-\infty}^{\infty} x(t)e^{-j\omega t}\mathrm{d}t \tag{4.71}$$

$$x(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} X(\omega)e^{j\omega t}\mathrm{d}\omega \tag{4.72}$$

Las ecuaciones 4.71 y 4.72 definen a la *transformada de Fourier* y *transformada de Fourier inversa* en el tiempo continuo, respectivamente.

Sin embargo, para que la DTFT de una secuencia exista, la sumatoria en la ecuación 4.69 debe converger, es decir, se requiere que $x(n)$ sea absolutamente sumable:

$$\sum_{n=-\infty}^{\infty} |x(n)| = S < \infty \tag{4.73}$$

De manera que la respuesta en frecuencia de un sistema discreto LIT, $H(e^{j\omega})$, es la DTFT de la respuesta al impulso $h(n)$, la cual puede representarse mediante una integral de Fourier como:

$$h(n) = \frac{1}{2\pi}\int_{-\pi}^{\pi} H(e^{j\omega})e^{j\omega n}\mathrm{d}\omega \tag{4.74}$$

donde $H(e^{j\omega})$, de acuerdo a la ecuación 4.69 se calcula como:

$$H(e^{j\omega}) = \sum_{n=-\infty}^{\infty} h(n)e^{-jn\omega} \tag{4.75}$$

Dadas las secuencias discretas $x(n)$ y $y(n)$, y sus respectivas transformadas de Fourier $X(e^{j\omega})$ y $Y(e^{j\omega})$, algunas de las propiedades más importantes de la DTFT se listan en la tabla 4.1.

4.1.3.4. Serie Discreta de Fourier

La ecuación 4.2 establece que una secuencia periódica $x(n)$ con periodo N, se define como:

$$x(n+N) = x(n), \qquad \forall n \tag{4.76}$$

La representación de una señal periódica $x(n)$ en una *serie discreta de Fourier* o *DFS* con periodo fundamental N y con una frecuencia fundamental digital $\omega_0 = \frac{2\pi}{N}$, está dada por:

$$x(n) = \sum_{k=0}^{N-1} X(k)e^{j\frac{2\pi kn}{N}} \qquad n = 0, 1, ...(N-1) \tag{4.77}$$

donde $X(k)$ representa los coeficientes de la serie de Fourier. La expresión matemática para $X(k)$ puede obtenerse al multiplicar ambos lados de la ecuación 4.77 por $e^{-j\frac{2\pi ln}{N}}$, donde l es un entero y evaluando para $n = 0, ..., (N-1)$, se obtiene,

$$\sum_{n=0}^{N-1} x(n)e^{-j\frac{2\pi ln}{N}} = \sum_{n=0}^{N-1}\left\{\sum_{k=0}^{N-1} X(k)e^{j\frac{2\pi kn}{N}} \cdot e^{-j\frac{2\pi ln}{N}}\right\} \tag{4.78}$$

Propiedad	Secuencia	DTFT
Linealidad	$ax(n) + by(n)$	$aX(e^{j\omega}) + bY(e^{j\omega})$
Desplazamiento	$x(n - n_0)$	$e^{-jn_0\omega}X(e^{j\omega})$
Reflexión	$x(-n)$	$X(e^{-j\omega})$
Modulación	$e^{jn\omega_0}x(n)$	$X(e^{j(\omega-\omega_0)})$
Convolución	$x(n) * y(n)$	$X(e^{j\omega})Y(e^{j\omega})$
Conjugación	$x^*(n)$	$X^*(e^{-j\omega})$
Derivación	$nx(n)$	$j\frac{\mathrm{d}X(e^{j\omega})}{\mathrm{d}\omega}$
Multiplicación	$x(n)y(n)$	$\frac{1}{2\pi}\int_{-\pi}^{\pi} X(e^{j\theta})Y(e^{j(\omega-\theta)})\mathrm{d}\theta$

Tabla 4.1: Propiedades de la Transformada de Fourier en Tiempo Discreto [57].

Intercambiando el orden de las sumatorias,

$$\sum_{n=0}^{N-1} x(n)e^{-j\frac{2\pi ln}{N}} = \sum_{k=0}^{N-1} X(k)\left\{\sum_{n=0}^{N-1} e^{j\frac{2\pi(k-l)n}{N}}\right\} \tag{4.79}$$

donde,

$$\begin{aligned} e^{j\frac{2\pi(k-l)n}{N}} &= \cos((k-l)\tfrac{2\pi}{N}n) + j\sin((k-l)\tfrac{2\pi}{N}n) \\ &= 1 + 0 \\ &= 1 \qquad k - l = 0 \pm N \pm ... \end{aligned}$$

Por consiguiente,

$$\sum_{n=0}^{N-1} x(n)e^{\frac{j2\pi(k-l)n}{N}} = \begin{cases} N & k - l = 0 \pm N \pm 2N \pm ... \\ 0 & otro\,caso \end{cases} \tag{4.80}$$

Con el desarrollo anterior, el lado derecho de la ecuación 4.78 puede reducirse a

$$\sum_{n=0}^{N-1} x(n)e^{-j\frac{2\pi ln}{N}} = NX(l) \qquad l = 0, 1, 2, ..., (N-1) \tag{4.81}$$

Intercambiando el índice l por k, se obtiene

$$X(k) = \frac{1}{N}\sum_{n=0}^{N-1} x(n)e^{-j\frac{2\pi kn}{N}} \qquad k = 0, 1, 2, ..., (N-1) \tag{4.82}$$

La ecuación 4.82 representa los coeficientes de la DFS, donde cada uno de los términos de la sumatoria es periódico con periodo N.

4.1.3.5. Transformada Discreta de Fourier

Mediante la ecuación 4.82 es posible que una secuencia de duración finita $0 \leq n \leq (N-1)$ y longitud N en el dominio del tiempo, sea trasladada al dominio de la frecuencia como una secuencia $X(k)$ de la misma longitud N, donde $k = 0, 1, ...(N-1)$. La transformación se realiza al dominio de la frecuencia dado que $X(k)$, para cualquier valor de k, representa el coeficiente de Fourier para la exponencial compleja discreta con frecuencia igual al k-ésimo harmónico de la frecuencia fundamental $\frac{2\pi}{N}$. Con base en lo anterior, se define la *Transformada Discreta de Fourier* o *DFT*, como:

$$X(k) = \sum_{n=0}^{N-1} x(n) e^{-j\frac{2\pi kn}{N}} \qquad k = 0, 1, 2, ..., (N-1) \tag{4.83}$$

Dado que la DFT es una transformación reversible, la *transformada discreta de Fourier inversa* o *IDFT* se define como:

$$x(n) = \frac{1}{N} \sum_{k=0}^{N-1} X(k) e^{j\frac{2\pi kn}{N}} \tag{4.84}$$

Por conveniencia en notación, la DFT y la IDFT son comúnmente escritas en términos de la cantidad compleja W_N, definida por

$$W_N = e^{-j\frac{2\pi}{N}} \tag{4.85}$$

De forma que las ecuaciones que definen la DFT y la IDFT pueden ser expresadas como:

$$X(k) = \sum_{n=0}^{N-1} x(n) W_N^{kn} \tag{4.86}$$

$$x(n) = \frac{1}{N} \sum_{k=0}^{N-1} X(k) W_N^{-kn} \tag{4.87}$$

La comparación de la definición de la DTFT en la ecuación 4.69 con la definición de la DFT en la ecuación 4.83, resulta en que los coeficientes de la DFT son muestras de la DTFT, es decir,

$$X(k) = \sum_{n=0}^{N-1} x(n) e^{-j\frac{2\pi kn}{N}} = \sum_{n=-\infty}^{\infty} x(n) e^{-j\frac{2\pi kn}{N}} = X(e^{j\omega})\big|_{\omega=\frac{2\pi k}{N}} \tag{4.88}$$

En la tabla 4.2 se muestran algunas propiedades importantes de la DFT.

4.1.3.6. Transformada Rápida de Fourier

La *transformada rápida de Fourier* o *FFT*, es una versión rápida de la DFT. La FFT utiliza algunos algoritmos ingeniosos para realizar lo mismo que la DFT, pero en un tiempo mucho menor. Por consiguiente, la FFT reduce la complejidad computacional de la DFT del orden $O(N^2)$ al orden $O(N \log_2 N)$.

Propiedad	Secuencia	DFT
Linealidad	$ax(n) + by(n)$	$aX(k) + bY(k)$
Periodicidad	$x(n + mN) = x(n)$	$X(k + mN) = X(k)$
Reflexión	$x(-n)$	$X(-k)$
Desplazamiento	$x(n - n_0)$	$X(k)e^{-j\frac{2\pi n_0 k}{N}}$
Modulación	$x(n)e^{j\frac{2\pi k_0 n}{N}}$	$X(k - k_0)$
Multiplicación	$x(n)y(n)$	$\frac{1}{N}[X(k) \otimes Y(k)]$

Tabla 4.2: Propiedades de la Transformada Discreta de Fourier [58].

La historia de la FFT comienza en 1805, cuando Carl Friedrich Gauss intentó determinar la órbita de ciertos asteroides. De este modo, desarrolló la transformada discreta de Fourier, incluso antes de que Joseph Fourier publicara sus resultados en 1822. Para calcular la DFT, Gauss inventó un algortimo que es equivalente al de Cooley y Tukey publicado en 1965, tiempo en el que la milicia estadounidense estaba interesada en un método para detectar pruebas soviéticas nucleares. Entre 1805 y 1965, varios científicos desarrollaron métodos eficientes para calcular la DFT, pero ninguno de ellos fue tan general como el método desarrollado por Gauss o Cooley y Tukey. La tabla 4.3 resume los principales descubrimientos de métodos eficientes para el cálculo de la DFT.

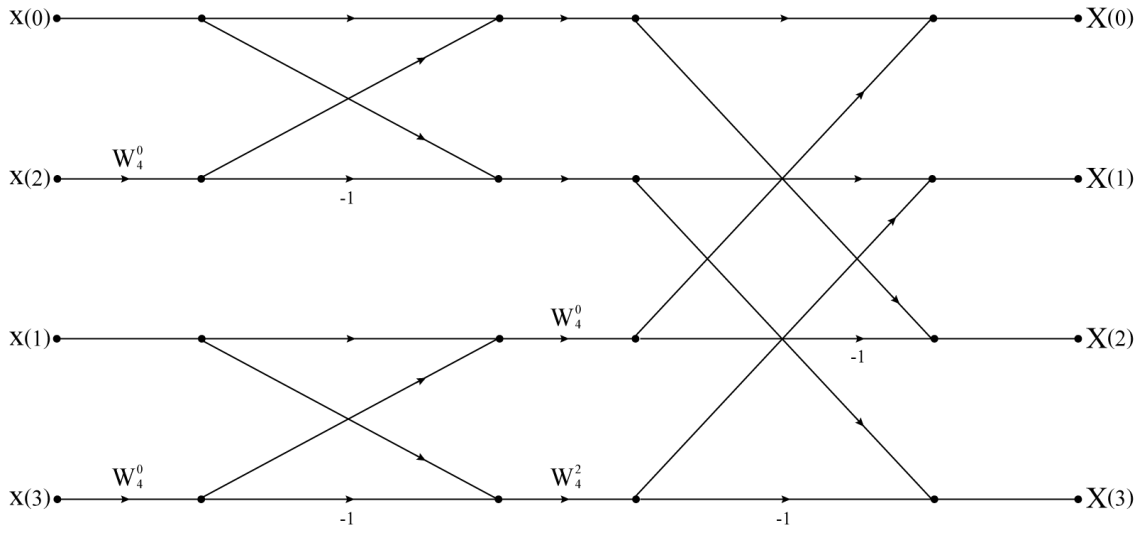


Figura 4.22: Diagrama de mariposa de la FFT de cuatro puntos.

Investigador(es)	Fecha	Aplicación
C. F. Gauss [59]	1805	Interpolación de órbitas de cuerpos celestes
F. Carlini [60]	1828	Análisis armónico de presión barométrica
A. Smith [61]	1846	Corrección de desviaciones de brújulas en barcos
J. D. Everett [62]	1860	Modelado de desviaciones de temperatura bajo tierra
C. Runge [63]	1903	Análisis armónico de funciones
K. Stumpff [64]	1939	Análisis armónico de funciones
Danielson y Lanczos [65]	1942	Difracción de rayos X en cristales
L. H. Thomas [66]	1948	Análisis armónico de funciones
I. J. Good [67]	1958	Análisis armónico de funciones
Cooley y Tukey [68]	1965	Análisis armónico de funciones
S. Winograd [69]	1976	Uso de la teoría de la complejidad para análisis armónico

Tabla 4.3: Principales descubrimientos de métodos eficientes para el cálculo de la DFT [70].

Existen diversas variantes para el cálculo rápido de la DFT. La eficiencia de los algoritmos consiste en realizar cálculos más pequeños de la DFT, explotando las propiedades de simetría y periodicidad de la exponencial compleja $W_N^{kn} = e^{-j(2\pi/N)kn}$. A los algoritmos que descomponen la secuencia discreta $x(n)$ en subsecuencias más pequeñas, son conocidos como algoritmos de *decimación en el tiempo*. El algoritmo de la FFT de *Cooley* y *Tukey* está basado en este enfoque. La figura 4.22 muestra el diagrama del cálculo de una DFT de cuatro puntos ($N = 4$) empleando un algoritmo de decimación en el tiempo. El cálculo se realiza en dos etapas; primero se calculan dos DFTs de dos puntos, y luego una de cuatro puntos. El cálculo básico realizado en cada etapa consiste en tomar dos números a y b, multiplicar b por W_N^r, y luego sumar y restar el producto de a para formar un nuevo par de números A y B. A este cálculo se le conoce como *mariposa*, ya que el diagrama de flujo se asemeja a una mariposa.

También es posible derivar algoritmos que primero descomponen la secuencia de salida $X(k)$ en subsecuencias sucesivas más pequeñas. Estos algoritmos se denominan algoritmos de *deci-*

mación en frecuencia. El algoritmo de *Sande* y *Tukey* es un ejemplo del uso de este enfoque. Otra clase de FFTs subdivide el conjunto inicial de datos de longitud N, en pequeñas potencias de 2, por ejemplo si $N = 2$, se obtienen FFTs de *base 4*, o si $N = 8$, se obtienen FFTs de *base 8*. Estas pequeñas transformaciones se llevan a cabo por secciones de código altamente optimizado que toman ventaja de las simetrías existentes en ese particular N. Por ejemplo, si $N = 4$, los senos y cosenos trigonométricos que ingresan al algoritmo son todos ± 1 o 0, de manera que muchas multiplicaciones son eliminadas, dejando en gran parte sumas y restas.

Existen también algoritmos de la FFT para conjuntos de datos de longitud N que no son potencia de 2, los cuales consisten en subdividir la secuencia inicial en subsecuencias sucesivas más pequeñas, pero no mediante factores de 2, sino por cualquier factor primo que divida a N. Entre más grande es el factor primo de N, más bajo es el rendimiento del método. Si la longitud de la secuencia discreta es un número primo, entonces no es posible realizar una subdivisión, y la transformada de Fourier se lleva a cabo empleando N^2 operaciones.

Los algoritmos de *Winograd* son otros algoritmos para el cálculo rápido de la DFT. Son análogos a las FFTs de base-4 y base-8. Winograd derivó códigos altamente óptimos para tomar transformadas discretas de Fourier de una longitud pequeña, e.g., para $N = 2, 3, 4, 5, 7, 8, 11, 13, 16$. El algoritmo también emplea una forma ingeniosa para combinar los subfactores, pues involucra un método para el reordenamiento de los datos antes y después del procesamiento, permitiendo una reducción significativa en el número de multiplicaciones en el algoritmo [71].

4.1.3.7. Transformada Coseno Discreta

La parte real de la transformada de Fourier de una señal real es la transformada de Fourier de la componente par de la señal. Es decir, que si la señal es una función par, entonces su transformada de Fourier es real. Este hecho hace posible poder construir a partir de una secuencia discreta, una función par simétrica cuya DFT sea real. A esta transformación se le conoce como *transformada coseno discreta* o *DCT*.

La DCT es comúnmente utilizada en codificación de imágenes, compresión de datos, algoritmos adaptables y otras aplicaciones del procesamiento digital de señales, debido a sus propiedades como la capacidad de compactación de la energía, su aproximación a la estadísticamente óptima *transformada Karhunen-Loéve* (*KLT*) para decorrelacionar una señal, y a las ventajas computacionales para el cálculo rápido de los coeficientes. La DCT descompone una señal en un grupo de señales cosenoidales ortogonales, denominadas funciones base. La transformada toma un conjunto de muestras y entrega a la salida una serie de coeficientes [72].

Existen ocho tipos de transformadas coseno. Los tipos DCT I-IV se derivan de DFTs pares reales de orden par y los tipos DCT V-VIII son derivados de DFTs pares reales de orden impar. En la práctica solamente los tipos de DCT I-V son utilizados.

Las definiciones de las cuatro variantes principales de la DCT tienen diferencias pequeñas. En estas ecuaciones de transformación, los elementos de una secuencia discreta $x(0), ..., x(N-1)$ de longitud N, son transformados en secuencias $X(0), ..., X(N-1)$. En la literatura se definen varias alternativas de escalamiento para las transformadas, sin embargo, en la práctica el requerimiento es que la transformada y su transformada inversa tenga juntas una ganancia unitaria. A continuación se definen los cuatro tipos principales de la DCT.

- **DCT-I**
  Esta variante de la DCT es raramente utilizada. Los coeficientes de transformación $X(i)$, donde $i = 0, 1, 2, ..., N-1$, se calculan de acuerdo a la siguiente ecuación:

$$X(i) = \frac{1}{2}(x(0) + (-1)^i x(N-1)) + \sum_{k=1}^{N-2} x(k) \cos\left(\frac{ik\pi}{N-1}\right) \tag{4.89}$$

- **DCT-II**
  En esta variante, los coeficientes de transformación $X(i)$, se calculan de acuerdo a la siguiente ecuación:

$$X(i) = \sum_{k=0}^{N-1} x(k) \cos\left(\frac{i(k+\frac{1}{2})\pi}{N}\right) \tag{4.90}$$

- **DCT-III**
  Esta variante se relaciona con la DCT-II y se define por medio de la siguiente ecuación:

$$X(i) = \frac{1}{2}x(0) \sum_{k=1}^{N-1} x(k) \cos\left(\frac{i(k+\frac{1}{2})\pi}{N}\right) \tag{4.91}$$

  En muchas aplicaciones de codificación, el par de transformadas más empleado involucra a la DCT-II y DCT-III, ya que una es la transformada inversa de la otra.

- **DCT-IV**
  Esta variante de la DCT, se define como:

$$X(i) = \sum_{k=0}^{N-1} x(k) \cos\left(\frac{(i+\frac{1}{2})(k+\frac{1}{2})\pi}{N}\right) \tag{4.92}$$

4.1.4. Muestreo y Reconstrucción de Señales Continuas

El muestreo es una de las operaciones fundamentales en el procesamiento digital de señales, ya que proporciona un mecanismo para convertir señales continuas en señales discretas. Esta conversión se lleva a cabo tomando de la señal continua un número suficiente de muestras adquiridas cada cierto intervalo de tiempo, con las cuales la señal puede ser representada en su totalidad. Este hecho se establece en el *teorema del muestreo*, que de manera más específica, determina la frecuencia mínima con la cual una señal continua debe ser muestreada para convertirla en una señal discreta sin que exista pérdida de información y que a su vez, de acuerdo a esta condición, la señal original pueda ser reconstruida perfectamente a partir de sus muestras [73]. Una manera conveniente para representar el muestreo de una señal continua y entender los efectos producidos en el domino de la frecuencia es a través de los operadores $comb$ y rep [74], cuya definición se establece en las secciones siguientes.

4.1.4.1. Operador $comb$

El operador $comb$ aplicado a una función $f(t)$, consiste en multiplicar dicha función con un tren de impulsos periódico $p_T(t)$, es decir,

$$comb_T[f(t)] = f(t)p_T(t) \tag{4.93}$$

donde $p_T(t)$ se define matemáticamente como:

$$p_T(t) = \sum_{n=-\infty}^{\infty} \delta(t - nT) \tag{4.94}$$

o bien, mediante su representación en una serie de Fourier compleja como:

$$p_T(t) = \frac{1}{T} \sum_{n=-\infty}^{\infty} e^{jn\frac{2\pi}{T}t} \tag{4.95}$$

Sustituyendo la ecuación 4.94 en la ecuación 4.93,

$$comb_T[f(t)] = f(t) \sum_{n=-\infty}^{\infty} \delta(t - nT) \tag{4.96}$$

La ecuación 4.96 se puede reescribir como:

$$comb_T[f(t)] = \sum_{n=-\infty}^{\infty} f(t)\delta(t - nT) \tag{4.97}$$

Empleando la propiedad de multiplicación de la función impulso establecida en la ecuación 4.35, el operador $comb_T$ aplicado a una función $f(t)$, también puede definirse como:

$$comb_T[f(t)] = \sum_{n=-\infty}^{\infty} f(nT)\delta(t - nT) \tag{4.98}$$

4.1.4.2. Operador rep

El operador rep denota un proceso mediante el cual una función es replicada periódicamente. Aplicado a una función $f(t)$, consiste en convolucionar dicha función con un tren de impulsos periódico $p_T(t)$, es decir,

$$rep_T[f(t)] = f(t) * p_T(t) \tag{4.99}$$

Sustituyendo la ecuación 4.94 en la ecuación 4.99,

$$rep_T[f(t)] = f(t) * \sum_{n=-\infty}^{\infty} \delta(t - nT) \tag{4.100}$$

La ecuación 4.100 se puede reescribir como:

$$rep_T[f(t)] = \sum_{n=-\infty}^{\infty} f(t) * \delta(t - nT) \tag{4.101}$$

Empleando la propiedad de convolución de la función impulso establecida en la ecuación 4.38, el operador rep_T aplicado a una función $f(t)$, también puede definirse como:

$$rep_T[f(t)] = \sum_{n=-\infty}^{\infty} f(t - nT) \tag{4.102}$$

4.1.4.3. Teorema del Muestreo

El muestreo de una señal continua limitada en banda, es decir, de una señal cuya transformada de Fourier es exactamente cero fuera de una banda finita de frecuencias ($X(w) = 0\, para\, |\omega| \geq \omega_M$), puede llevarse a cabo aplicando el operador $comb$ a la señal que se desea muestrear. El resultado de esta operación, y cuyo desarrollo se muestra gráficamente en la figura 4.23 (a)-(c), es una señal muestreada $x_s(t)$, que consiste en un tren de impulsos cuya amplitud de cada impulso es igual a las muestras de la señal original $x(t)$, en intervalos equiespaciados un valor T, es decir,

$$x_s(t) = comb_T[x(t)] = \sum_{n=-\infty}^{\infty} x(nT)\delta(t - nT) \tag{4.103}$$

Analíticamente, los efectos en el dominio de la frecuencia del muestreo en el dominio del tiempo, se pueden conocer al aplicar la transformada de Fourier a la ecuación 4.103, esto es,

asumiendo que

$$\mathscr{F}\{x(t)\} \longrightarrow X(\omega) \tag{4.104}$$

entonces,

$$\mathscr{F}\{comb_T[x(t)]\} = \mathscr{F}\{x(t)p_T(t)\} \tag{4.105}$$

De acuerdo al teorema de la convolución en la frecuencia, el cual afirma que

$$\mathscr{F}\{f_1(t)f_2(t)\} = \frac{1}{2\pi}F_1(\omega) * F_2(\omega) \tag{4.106}$$

y la transformada de Fourier del tren de impulsos $p_T(t)$, dada por

$$\mathscr{F}\{p_T(t)\} = \omega_o \sum_{n=-\infty}^{\infty} \delta(\omega - n\omega_0), \qquad \omega_0 = \frac{2\pi}{T} \tag{4.107}$$

se tiene que la ecuación 4.105 puede ser expresada como:

$$\mathscr{F}\{comb_T[x(t)]\} = \frac{1}{2\pi}\left[X(\omega) * \omega_o \sum_{n=-\infty}^{\infty} \delta(\omega - n\omega_0)\right] \tag{4.108}$$

Sustituyendo ω_0 y reacomodando los términos del lado derecho de la ecuación 4.108, se tiene que,

$$\mathscr{F}\{comb_T[x(t)]\} = \frac{1}{T}\sum_{n=-\infty}^{\infty} X(\omega) * \delta(\omega - n\omega_0) \tag{4.109}$$

Finalmente, empleando la propiedad de convolución de la función impulso establecida en la ecuación 4.38, se obtiene la transformada de Fourier de $comb_T[x(t)]$:

$$\mathscr{F}\{comb_T[x(t)]\} = \frac{1}{T}\sum_{n=-\infty}^{\infty} X(\omega - n\omega_0) \tag{4.110}$$

Comparando el lado derecho de la ecuación 4.110 con la definición del operador rep_T aplicado a una función $f(t)$, dada en la ecuación 4.102, se observa que,

$$\mathscr{F}\{comb_T[x(t)]\} = \frac{1}{T} rep_{\omega_0}[X(\omega)] \tag{4.111}$$

por lo cual,

$$x_s(t) \xrightarrow{\mathscr{F}} X_s(\omega) \qquad X_s(\omega) = \frac{1}{T} rep_{\omega_0}[X(\omega)] \tag{4.112}$$

La ecuación 4.112 muestra que la transformada de Fourier de $x_s(t)$, es una función que replica periódicamente el espectro de la señal original $x(t)$. La figura 4.23 (d)-(f) ilustra gráficamente este proceso.

De acuerdo a la figura 4.23 (f), el espectro de una señal muestreada idealmente es una repetición periódica escalada en amplitud del espectro original. Ya que la señal $x(t)$ es una señal limitada en banda y ha sido muestreada de manera que la frecuencia de muestreo es al menos dos veces mayor que la frecuencia máxima contenida en la señal, las repeticiones periódicas de $X(\omega)$ no se traslapan. Cuando el muestreo no cumple con esta condición, se produce un fenómeno conocido como *aliasing*, el cual produce traslapes en el espectro de la señal muestreada tal como se muestra en la figura 4.24, ocasionando inevitablemente la pérdida de información. Esta idea se establece en el *teorema del muestreo*, o *teorema de Nyquist-Shannon*, en el cual también se enuncia que si la condición anterior matemáticamente expresada como:

$$\omega_s > 2\omega_M, \qquad \omega_s = \frac{2\pi}{T} \tag{4.113}$$

se cumple, entonces que la señal puede ser reconstruida perfectamente a partir de sus muestras. Este procedimiento se lleva a cabo empleando un filtro pasobajas con ganancia T y frecuencia de corte igual a $\frac{\omega_s}{2}$ para recuperar $X(\omega)$ de $X_s(\omega)$, es decir,

$$X(\omega) = \begin{cases} TX_s(\omega), & |\omega| \leq \frac{\omega_s}{2} \\ 0, & otro\,caso \end{cases} \tag{4.114}$$

En la figura 4.25 (a)-(d) se muestra el proceso de reconstrucción de una señal continua a partir de sus muestras utilizando un filtro ideal pasobajas. Sin embargo, la condición de Nyquist para reconstruir perfectamente una señal continua a partir de sus muestras es solamente una condición suficiente más no necesaria, pues existen señales que a pesar de no cumplir con este teorema pueden ser reconstruidas en su totalidad. La figura 4.25 (e)-(h) muestra un ejemplo de reconstrucción perfecta cuando no se cumple la condición de Nyquist.

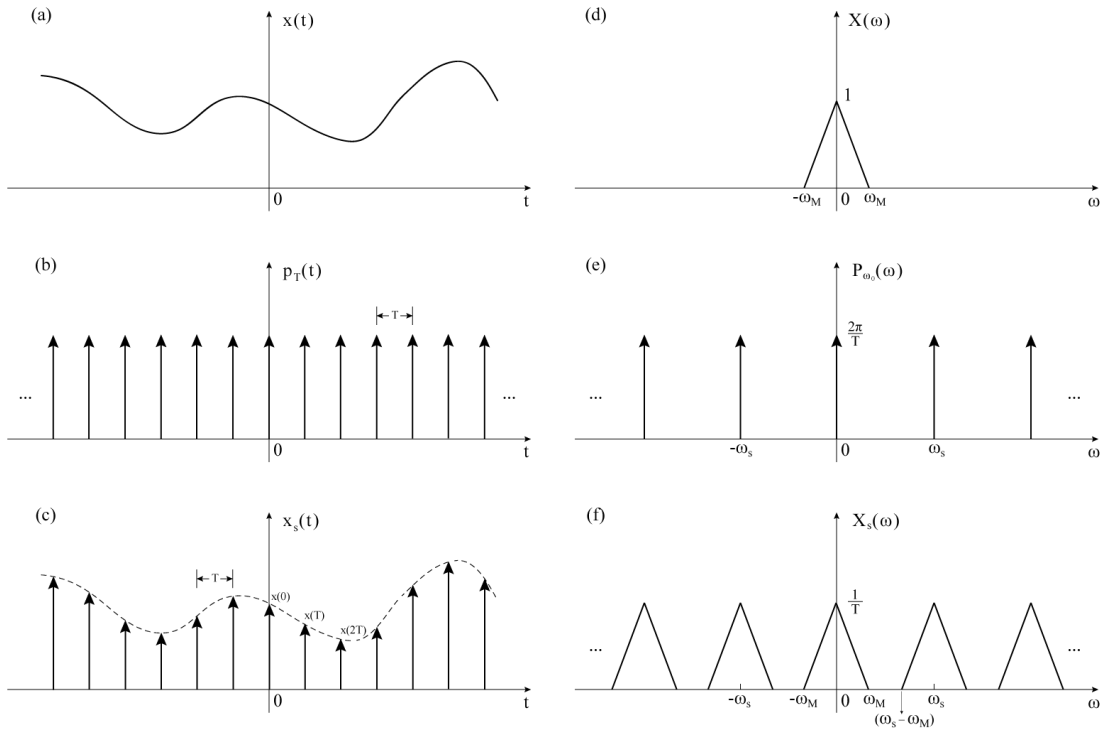


Figura 4.23: Muestreo de una señal continua. (a) Señal continua. (b) Tren periódico de impulsos. (c) Señal muestreada. (d) Espectro de una señal continua. (e) Espectro de un tren periódico de impulsos. (f) Espectro de una señal muestreada.

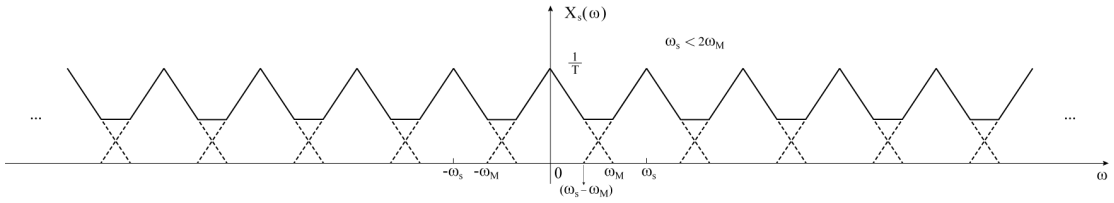


Figura 4.24: Efecto de aliasing en el dominio de la frecuencia debido a un submuestreo.

4.2. Análisis de la Señal de Voz en el Dominio del Tiempo

El análisis de la señal de voz en el dominio del tiempo permite una interpretación física sencilla y simplicidad en el cálculo de parámetros relevantes del habla [75]. Entre las características más importantes que se encuentran con facilidad en el análisis temporal están las estadísticas de la forma de onda de la señal, la frecuencia fundamental, la energía, así como la tasa de cruces por cero y la autocorrelación, las cuales proporcionan detalles espectrales sin emplear necesariamente métodos formales de análisis espectral [76].

En la mayoría de los esquemas de procesamiento de voz, existe una suposición fundamental que considera que las propiedades de la señal de voz cambian de forma relativamente lenta

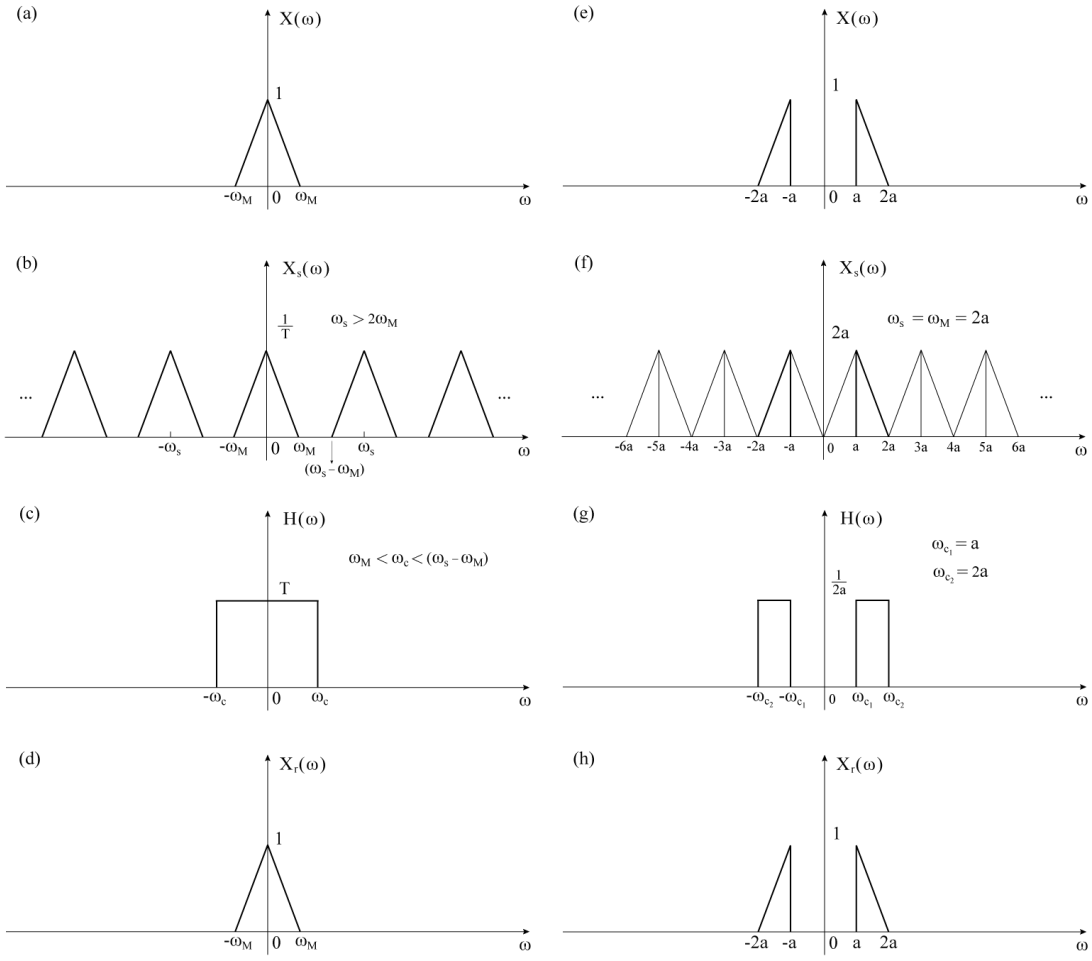


Figura 4.25: Reconstrucción ideal de una señal continua. (a) Espectro $X(\omega)$ de una señal continua $x(t)$. (b) Espectro $X_s(\omega)$ de una señal muestreada $x_s(t)$. (c) Espectro $H(\omega)$ de un filtro pasobajas ideal. (d) Espectro $X_r(\omega)$ de una señal continua recuperada a partir de sus muestras. (e) Espectro $X(\omega)$ de una señal pasobanda. (f) Espectro $X_s(\omega)$ de una señal pasobanda muestreada a una frecuencia menor a la establecida por la condición de Nyquist. (g) Espectro $H(\omega)$ de un filtro pasobanda. (h) Espectro $X_r(\omega)$ de la señal pasobanda recuperada a partir de sus muestras.

con respecto al tiempo. De manera que cuando la señal de voz es analizada sobre periodos de tiempo entre 5 y 100 ms, sus características se pueden consideran estacionarias [1], es decir, que sus parámetros estadísticos tales como la media, varianza y potencia de las componentes espectrales, entre otros, se mantienen constantes. Esta suposición conduce a una variedad de métodos de procesamiento en *tiempo corto* en los cuales segmentos muy pequeños de la señal de voz son aislados y procesados como si fueran segmentos cortos de un sonido prolongado con propiedades fijas [77]. Del procesamiento de la señal resulta una secuencia distinta que sirve como una nueva representación de la señal de voz.

4.2.1. Entramado y Ventaneo de la Señal de Voz

Debido a la naturaleza no estacionaria de la señal de voz, su análisis en el dominio del tiempo se lleva a cabo tomando pequeñas porciones de la señal a la vez. Cada segmento de voz seleccionado se conoce como *trama de análisis*, por lo que la señal de voz es procesada trama por trama, comúnmente en intervalos superpuestos, hasta que toda la región de voz es cubierta por al menos una de las tramas. La figura 4.26 muestra la segmentación en tramas superpuestas de una señal de voz.

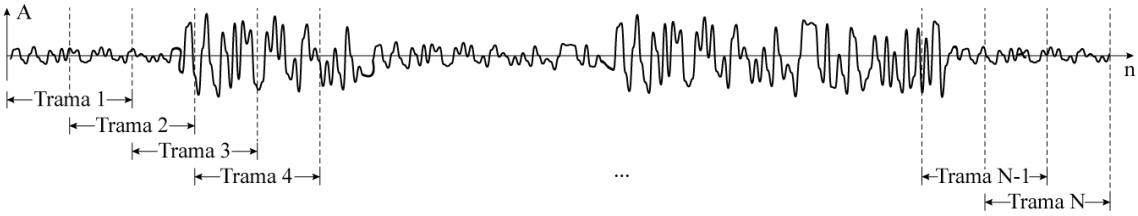


Figura 4.26: Entramado de una señal de voz. Las tramas se superponen 50 %.

De forma más específica, a la técnica empleada para segmentar una señal en tramas de un número finito de muestras se le conoce como *ventaneo*, que consiste en la multiplicación de una señal con una función denominada *ventana* cuya amplitud es cero excepto en la región de interés. El uso de ventanas es importante, ya que es necesario considerar cómo tratar los bordes de las tramas para reducir los componentes espectrales generados por el proceso de segmentación.

4.2.1.1. Tipos de Ventanas

Existen varios tipos de ventanas, cada una de las cuales tiene sus propias características y su uso es adecuado para distintas aplicaciones. Algunas de las ventanas más empleadas en el procesamiento digital de señales se ilustran en la figura 4.27 y de acuerdo a [78], se definen matemáticamente como:

- **Ventana rectangular**
  Se define como:
$$w(n) = \begin{cases} 1, & |n| \leq \frac{N-1}{2} \\ 0, & otro\,caso \end{cases} \tag{4.115}$$

- **Ventana triangular**
  Se define como:
$$w(n) = \begin{cases} 1 - \frac{|n|}{N/2}, & |n| \leq \frac{N-1}{2} \\ 0, & otro\,caso \end{cases} \tag{4.116}$$

- **Ventana de Hanning**
  Se define como:
$$w(n) = \begin{cases} 0.5\left[1 + \cos\left(\frac{2\pi}{N}n\right)\right], & |n| \leq \frac{N-1}{2} \\ 0, & otro\,caso \end{cases} \tag{4.117}$$

- **Ventana de Hamming**

  Se define como:

$$w(n) = \begin{cases} 0.54 + 0.46\cos\left(\frac{2\pi}{N}n\right), & |n| \leq \frac{N-1}{2} \\ 0, & otro\,caso \end{cases} \tag{4.118}$$

- **Ventana de Blackman**
  Se define como:

$$w(n) = \begin{cases} 0.42 + 0.5\cos\left(\frac{2\pi n}{N}n\right) + 0.08\cos\left(\frac{2\pi}{N}2n\right), & |n| \leq \frac{N-1}{2} \\ 0, & otro\,caso \end{cases} \tag{4.119}$$

- **Ventana de Kaiser-Bessel**
  Se define como:

$$w(n) = \begin{cases} \frac{I_0\left(\beta\sqrt{1-(\frac{n}{N/2})^2}\right)}{I_{0(\beta)}}, & |n| \leq \frac{N-1}{2} \\ 0, & otro\,caso \end{cases} \tag{4.120}$$

4.2.2. Energía

De acuerdo con la ecuación 4.14, la energía de una señal discreta se define como:

$$E_x = \sum_{n=-\infty}^{\infty} |x(n)|^2$$

Sin embargo, la ecuación anterior para el cálculo de la energía tiene poca utilidad si la señal discreta es variante en el tiempo. Por consiguiente, su uso sobre una señal de voz carece totalmente de significado, pues ofrece poca información acerca de las propiedades de la señal de voz que son dependientes del tiempo.

Para conocer las variaciones en el tiempo de la energía en una señal de voz, es necesario tener una representación diferente de la señal. Es por ello que la mayoría de las técnicas de análisis de la señal de voz en el dominio del tiempo se representan matemáticamente mediante la siguiente forma:

$$Q_n = \sum_{m=-\infty}^{\infty} T\{x(m)\}w(n-m) \tag{4.121}$$

en donde una transformación $T\{\cdot\}$ que puede ser lineal o no lineal, aplicada a una señal de voz, denotada por $x(n)$, es convolucionada con una ventana $w(n)$, usualmente de longitud finita. En la figura 4.28 se muestra el diagrama de bloques de este proceso.

Si la transformación $T\{\cdot\}$ en la ecuación 4.121 realiza la operación de elevar al cuadrado, es decir, que aplicada a una señal $x(n)$, $T\{x(n)\} = x^2(n)$, entonces Q_n corresponde a la *energía* en tiempo corto. Debido a que la señal es elevada al cuadrado, los valores presentes en la señal cuya amplitud es alta, son enfatizados con el cálculo de la energía, lo cual ayuda a reflejar la variación

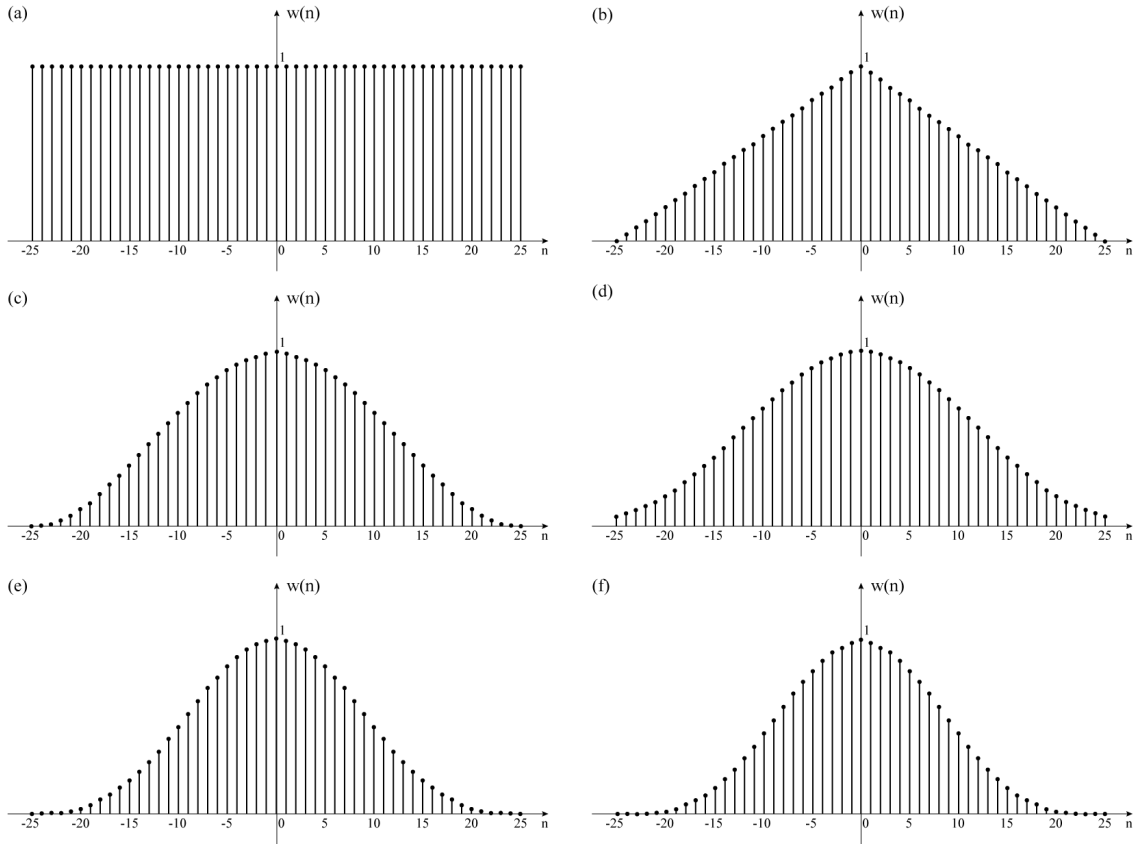


Figura 4.27: Diferentes tipos de ventanas ($N = 25$). (a) Rectangular. (b) Triangular. (c) Hanning. (d) Hamming. (e). Blackman (f). Kaiser-Bessel. Adaptado de [78].

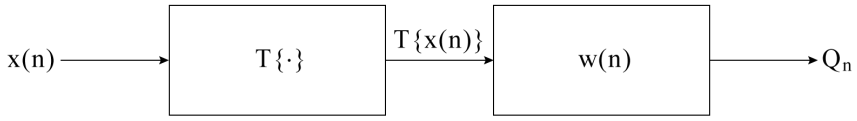


Figura 4.28: Análisis de la señal de voz en el dominio del tiempo.

en amplitud de sonidos voceados y no voceados. Matemáticamente, la energía en tiempo corto se define como:

$$E_n = \sum_{m=-\infty}^{\infty} [x(m)w(n-m)]^2 \tag{4.122}$$

A diferencia de la ecuación 4.14, la ecuación 4.122 es un vector y no un solo valor.

4.2.3. Magnitud

Una de las desventajas de la función de energía definida en la ecuación 4.122, es que es muy sensible a grandes niveles de energía, por lo que muestra a muestra las variaciones grandes en la

señal $x(n)$ son enfatizadas. Una forma simple de aligerar el problema es que la transformación $T\{\cdot\}$ de la ecuación 4.121, en lugar de elevar al cuadrado la función sobre la cual opera, calcule su magnitud absoluta, es decir, que $T\{x(n)\} = |x(n)|$. Por consiguiente, la función de *magnitud* puede definirse como:

$$M_n = \sum_{m=-\infty}^{\infty} |x(m)|w(n-m) \tag{4.123}$$

4.2.4. Cruces por Cero

Los *cruces por cero* indican el número de veces que una señal atraviesa el nivel cero en cualquiera de los dos sentidos [79]. Un cruce por cero se determina cuando dos muestras sucesivas de una señal tienen signos algebraicos diferentes. Matemáticamente, la tasa de cruces por cero se define como:

$$Z_n = \frac{1}{2} \sum_{m=-\infty}^{\infty} |\text{sgn}[x(m)] - \text{sgn}[x(m-1)]|w(n-m) \tag{4.124}$$

donde,

$$\text{sgn}[x(n)] = \begin{cases} 1, & x(n) \geq 0 \\ -1, & x(n) \leq 0 \end{cases} \tag{4.125}$$

El cálculo de los cruces por cero es simplemente una medida del contenido frecuencial de una señal. El modelo de producción de voz sugiere que la energía de un sonido voceado se concentra por debajo de los 3 kHz, mientras que en los sonidos no voceados, la mayoría de la energía se encuentra en las frecuencias altas. Debido a que las altas frecuencias implican tasas de cruces por cero altas, y bajas frecuencias implican tasas de cruces por cero bajas, existe una fuerte correlación entre la tasa de cruces por cero y la distribución de la energía, con la frecuencia. Con base en lo anterior, una generalización razonable, sin embargo imprecisa, es que si la tasa de cruces por cero es alta, entonces la señal de voz es no voceada, mientras que si la tasa de cruces por cero es baja, la señal de voz es voceada [77].

4.2.5. Función de Autocorrelación

Dadas dos señales de energía finita $x(n)$ y $y(n)$, la *correlación* entre $x(n)$ y $y(n)$, denotada por la función r_{xy}, se define como:

$$r_{xy}(\ell) = \sum_{n=-\infty}^{\infty} x(n)y(n-\ell) \tag{4.126}$$

donde el índice ℓ es un parámetro de desplazamiento y el subíndice xy indica que la secuencia $x(n)$ permanece fija mientras la secuencia $y(n)$ se desplaza ℓ unidades. La correlación es útil para medir el grado en que dos señales son similares. La técnica basada en el cálculo de la correlación entre una señal y una versión retardada de la misma, es decir, el caso especial donde

$y(n) = x(n)$ es llamada *autocorrelación*, que analíticamente se define como:

$$r_{xx}(\ell) = \sum_{n=-\infty}^{\infty} x(n)x(n-\ell) \tag{4.127}$$

En el procesamiento digital de señales de voz, la función de autocorrelación puede ser empleada para encontrar la frecuencia fundamental o la tonalidad (pitch) de una señal, sin embargo, se requiere el uso de su forma analítica en tiempo corto, la cual se define como:

$$R_n(\ell) = \sum_{m=-\infty}^{\infty} x(m)w(n-m)x(n-\ell)w(n-m+\ell) \tag{4.128}$$

4.3. Análisis de la Señal de Voz en el Dominio de la Frecuencia

Al igual que las propiedades temporales de la señal de voz se mantienen fijas en intervalos cortos de tiempo, las propiedades espectrales de la señal de voz también pueden asumirse invariables bajo el mismo análisis en tiempo corto. La caracterización de las propiedades espectrales de las señales de voz se logra mediante la representación variante en el tiempo de la transformada de Fourier.

4.3.1. Transformada de Fourier en Tiempo Discreto de Tiempo Corto

Las propiedades espectrales de la señal de voz pueden ser analizadas mediante la *transformada de Fourier en tiempo discreto de tiempo corto* o *STDTFT* definida matemáticamente por:

$$X_n(e^{j\omega}) = \sum_{m=-\infty}^{\infty} x(m)w(n-m)e^{-j\omega m} \tag{4.129}$$

donde $w(n-m)$ es una ventana que determina la porción de la señal de entrada que se enfatiza específicamente en un índice de tiempo n. La STDTFT es una representación bidimensional de la señal unidimensional $x(n)$, ya que es una función de dos variables: el índice de tiempo n, el cual es discreto, y la frecuencia ω, la cual es continua.

Mediante un cambio de variables, la ecuación 4.129 puede definirse de manera alternativa como:

$$X_n(e^{j\omega}) = \sum_{m=-\infty}^{\infty} x(n-m)w(m)e^{-j\omega(n-m)} \tag{4.130}$$

$$= e^{-j\omega n} \sum_{m=-\infty}^{\infty} x(n-m)w(m)e^{j\omega m} \tag{4.131}$$

Si se define,

$$\tilde{X}_n(e^{j\omega}) = \sum_{m=-\infty}^{\infty} x(n-m)w(m)e^{j\omega m} \tag{4.132}$$

entonces $X_n(e^{j\omega})$ puede expresarse como:

$$X_n(e^j\omega) = e^{-j\omega n}\tilde{X}_n(e^{j\omega}) \tag{4.133}$$

Si en la ecuación 4.133 se asume que el índice de tiempo n permanece fijo, entonces $X_n(e^{j\omega})$ es simplemente la transformada de Fourier de la secuencia $w(n-m)x(m), -\infty < m < \infty$. Por consiguiente, $X_n(e^{j\omega})$ tiene las mismas propiedades de la DTFT. Sin embargo, si se considera que $X_n(e^{j\omega})$ es una función del índice de tiempo n con la frecuencia ω fija, las ecuaciones 4.129 y 4.132 se definen en forma de una convolución.

4.3.2. Espectrograma

La STDTFT es una colección de DTFTs que difieren por la posición de la ventana que segmenta la señal de voz. Esto puede ser visualizado en una imagen conocida como *espectrograma*. Como se mencionó en el capítulo anterior, un espectrograma muestra la forma en que las caracerísticas espectrales de la señal evolucionan con el tiempo. Un espectrograma es creado al colocar las DTFTs verticalmente en una imagen, asignando una columna diferente a cada segmento de tiempo. Como convención, normalmente la frecuencia aumenta de abajo hacia arriba, y el tiempo de izquierda a derecha. El valor del pixel en cada punto de la imagen es proporcional a la magnitud (o magnitud al cuadrado) del espectro en una cierta frecuencia en algún punto particular del tiempo.

Para señales cuasi-periódicas como la señal de voz, los espectrogramas son divididos en dos categorías de acuerdo a la longitud de la ventana que segmenta la señal. Los espectrogramas de *banda ancha* utilizan una ventana con una longitud comparable a la de un solo periodo, lo cual resulta en una alta resolución en el dominio del tiempo, pero una baja resolución en el dominio de la frecuencia. Este tipo de espectrogramas se caracteriza por tener bandas verticales, que corresponden a las regiones de alta y baja energía dentro de un solo periodo de la señal. Por otro lado, en los espectrogramas de *banda estrecha*, la ventana es lo suficientemente grande para capturar varios periodos de la señal. Aquí la resolución en el tiempo es menor para dar una mayor resolución al contenido espectral. En este tipo de espectrogramas los armónicos de la frecuencia fundamental pueden observarse como bandas horizontales. La figura 4.29 muestra los dos tipos de espectrograma de un segmento de voz.

En el cálculo de la STDTFT se emplean ventanas separadas un determinado número de muestras, el cuál puede especificarse en términos de la superposición o traslape existente entre ventanas sucesivas. El criterio para decidir la cantidad de traslape incluye la longitud de la ventana, la resolución deseada en el tiempo, y la tasa a la cual las características de la señal cambian con el tiempo.

4.3.3. Bancos de Filtros

El método de análisis espectral mediante *bancos de filtros* permite no solamente analizar los diferentes rangos de frecuencias presentes en una señal de voz, sino también separar los dos elementos del sistema que modela al habla: la fuente y el filtro. Esta separación es posible, ya que cada uno de los componentes de la señal de voz posee funciones lingüísticas diferentes e independientes. Mientras que la fuente se encarga de controlar la tonalidad, el filtro controla la

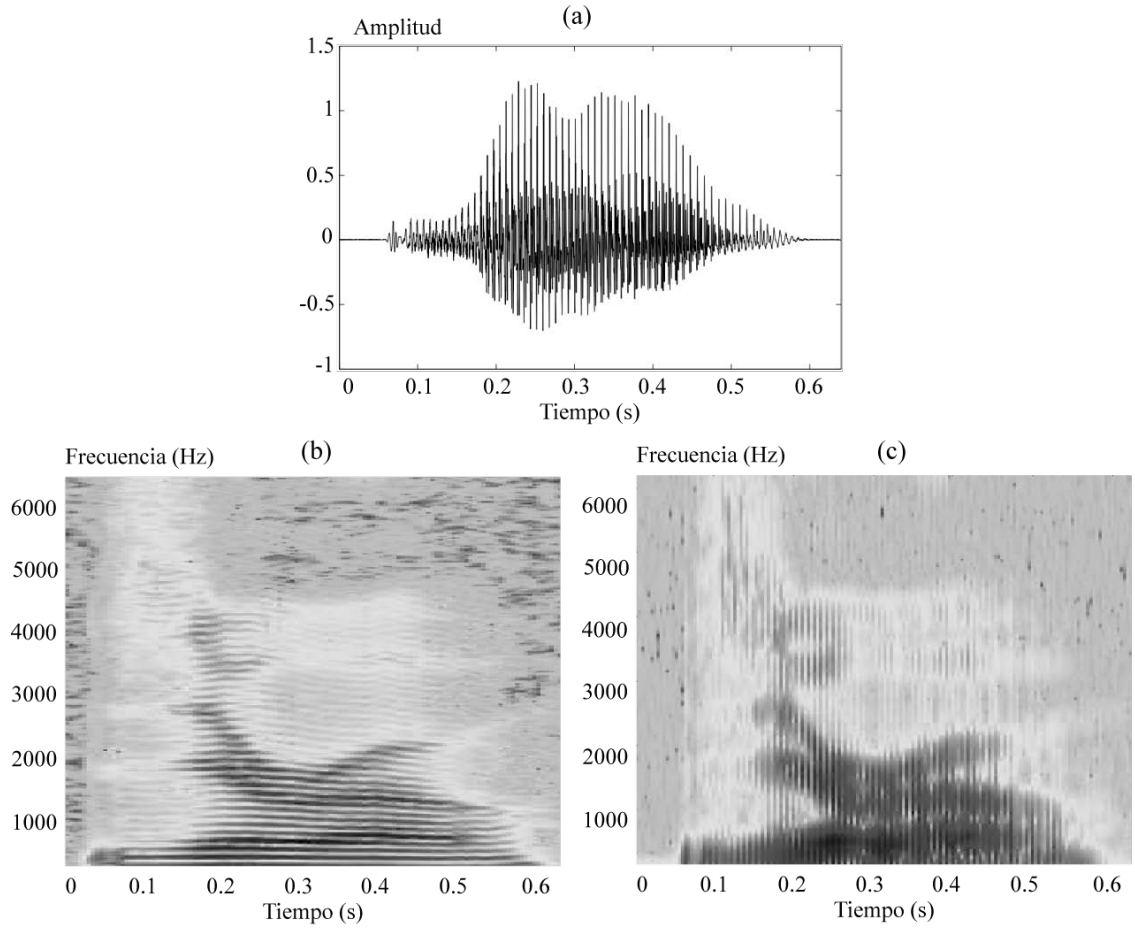


Figura 4.29: Espectrogramas de un segmento de voz. (a) Forma de onda de la palabra inglesa "zero". (b) Espectrograma de banda estrecha con ventana de 41 ms. (c) Espectrograma de banda ancha con ventana de 5 ms.

envolvente espectral y la posición de los formantes de voz, los cuales determinan qué fonemas son producidos.

El análisis mediante banco de filtros es llevado a cabo al crear en primera instancia una serie de *contenedores*, cada uno de los cuales se encuentra centrado en una frecuencia en particular. El análisis mediante bancos de filtros en el dominio del tiempo puede realizarse creando un filtro paso-banda, que permita que las frecuencias cercanas o ubicadas en la frecuencia central de cada contenedor pasen, pero que atenúe las demás frecuencias, reduciendo su magnitud a cero. Después del filtrado, la cantidad de energía se calcula y se asigna a ese contenedor. Una vez que esto se ha realizado para cada contenedor, se obtiene una representación de cómo varía la energía de acuerdo a la frecuencia. El efecto de la operación de filtrado es difuminar el efecto de los armónicos individuales, de manera que la representación final es en gran parte la del tracto vocal, independientemente de la fuente. De forma alternativa, se puede realizar el análisis mediante banco de filtros en el dominio de la frecuencia, es decir, en el espectro. En este caso, se

toma la magnitud del espectro y se multiplica por una función ventana centrada en cada una de las frecuencias centrales de los contenedores. Esto tiene el efecto de fijar toda la energía fuera de la ventana a cero, permitiendo medir únicamente la energía dentro de la ventana. Lo anterior se ilustra en la figura 4.30 (a). En algunas ocasiones, los contenedores no se encuentran espaciados de manera lineal a lo largo del eje de la frecuencia, sino que se definen logarítmicamente o de acuerdo a una escala perceptual, lo cual se ilustra en la figura 4.30 (b).

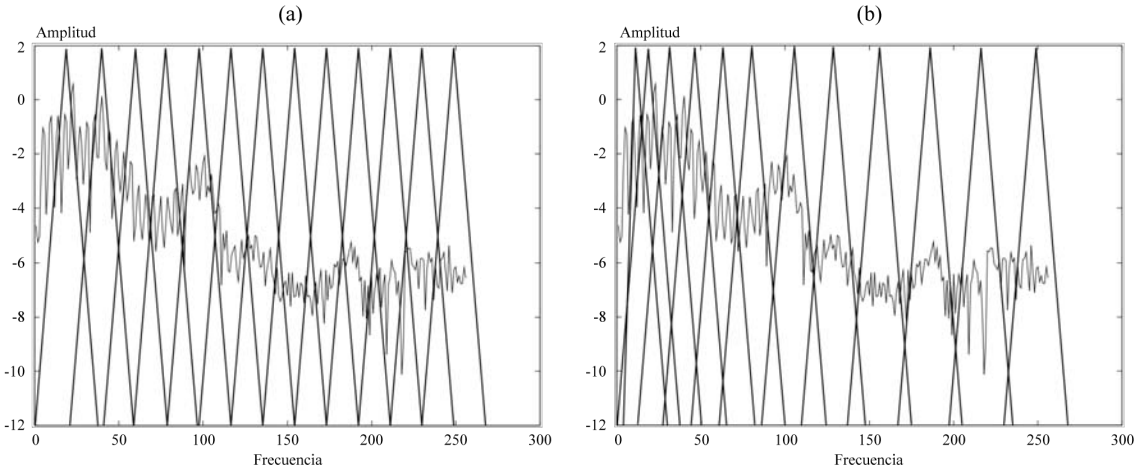


Figura 4.30: Análisis mediante bancos de filtros. (a) Filtros equiespaciados. (b) Filtros espaciados logarítmicamente [80].

4.3.4. Análisis Cepstral

Existen varias definiciones y acepciones del *cepstro*, sin embargo, una de las definiciones más comunes establece que el cepstro es la DFT inversa de la magnitud logarítmica de la DFT de una señal, es decir,

$$c(n) = \mathscr{F}^{-1}\{\log|\mathscr{F}\{x(n)\}|\} \tag{4.134}$$

donde $\mathscr{F}$ representa la DFT y $\mathscr{F}^{-1}$ la DFT inversa. Para una trama de la señal de voz, el cepstro se define como:

$$c(n) = \sum_{n=0}^{N-1} \log\left(\left|\sum_{n=0}^{N-1} x(n)e^{-j\frac{2\pi}{N}kn}\right|\right)e^{j\frac{2\pi}{N}kn} \tag{4.135}$$

El cepstro, al igual que el análisis mediante banco de filtros, permite separar los componentes del sistema que modela al habla.

La magnitud del espectro de una señal periódica contiene armónicos a intervalos igualmente espaciados. En la figura 4.31 (a) se ilustra este hecho. Debido a los efectos producidos por la ventana de truncamiento, los armónicos no se visualizan como funciones delta en el espectro, sino que aparecen más redondeados.

En la mayoría de las señales, la amplitud de los armónicos disminuye rápidamente a medida que la frecuencia incrementa. Este efecto puede reducirse al comprimir el espectro con respecto a la amplitud, lo cual puede conseguirse fácilmente calculando el logarítmo del espectro. Esta

operación produce que las amplitudes relativas de los armónicos estén mucho más cerca que en un espectro de magnitud absoluta. La figura 4.31 (b) muestra este efecto.

Ahora bien, si el espectro logarítmico se trata como una forma de onda, es decir como una señal, entonces esta nueva representación puede describirse como una especie de señal cuasi-periódica con algún tipo de modulación en amplitud, ya que algunos periodos tienen una amplitud mayor que otros. Debido a que la tasa a la cual los periodos de la señal cambian es mayor a la tasa de cambio de la amplitud, es posible separar estas características si se calcula la DFT de la señal, obteniéndose una nueva representación, conocida como cepstro. La figura 4.31 c) muestra esta representación. El cepstro muestra la periodicidad de la señal como una espiga que se asemeja a una función delta, pero debido a que los periodos no son sinusoidales, se presentan armónicos múltiplos de la frecuencia fundamental.

El efecto que causa la variación de amplitud también se encuentra presente en el cepstro, sin embargo, ya que ésta es mucho más lenta que la variación de los periodos, se localiza en la parte inferior del cepstro. Ya que la variación de amplitud representa la envolvente espectral y las espigas los armónicos, las operaciones realizadas para calcular el cepstro producen una representación de la señal en la que estos dos componentes se localizan en posiciones diferentes, permitiendo separarlos fácilmente mediante un filtro.

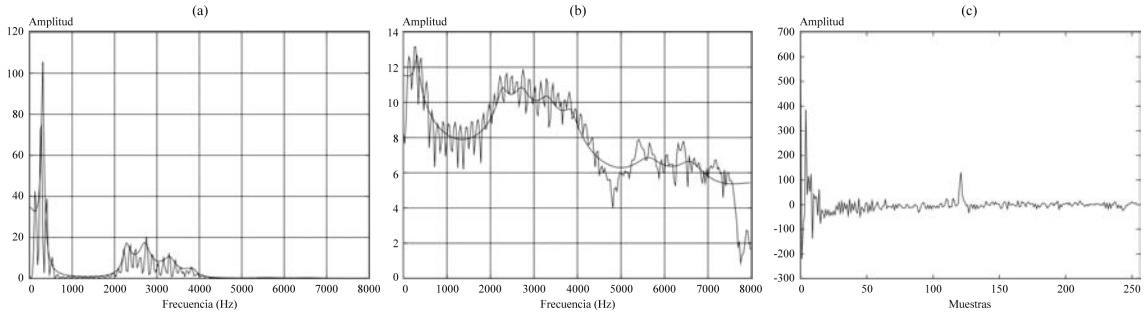


Figura 4.31: Cálculo del cepstro. (a) Espectro de magnitud y envolvente. (b) Espectro logarítmico y envolvente espectral. (c) Cepstro [80].

Si se calcula la DFT de una señal y luego se le aplica la DFT inversa, se obtiene por supuesto la señal original. En este sentido, el cálculo del cepstro difiere en dos aspectos importantes. Primero, solamente se hace uso de la magnitud del espectro, por lo que la información contenida en la fase es descartada. Segundo, al aplicar la DFT inversa a la magnitud del espectro, se obtiene un resultado diferente al que produciría la DFT inversa de un espectro normal. La operación logaritmo escala los armónicos, enfatizando su periodicidad, también asegura que el cepstro sea la suma de la fuente y el filtro, en vez de su convolución.

El cepstro es útil porque divide la señal en una envolvente espectral, dada por los primeros coeficientes, y una fuente dada por la espiga. El análisis subsecuente comúnmente toma únicamente una de estas dos partes. Por ejemplo, si el interés se encuentra en el tracto vocal, se emplean los coeficientes de la parte inferior del cepstro, mientras que si el interés radica en la tonalidad (pitch) y en el comportamiento de la glotis, entonces se mantiene la espiga.

En la mayoría de los casos en los que la envolvente espectral es requerida, resulta fundamental mantener la parte inferior del cepstro, sin convertirla de vuelta al dominio de la frecuencia.

Los coeficientes bajos del cepstro forman una representación muy compacta de la envolvente y poseen la deseable propiedad de modelado estadístico de ser independientes, de manera que solamente sus medias y varianzas son requeridas para proporcionar una distribución estadística precisa. Por esta razón, el cepstro es utilizado en la mayoría de los sistemas de reconocimiento de voz para representar las características acústicas de la señal de voz. Por otro lado, la parte superior del cepstro, es utilizada como la base de los algoritmos de *detección de pitch* [80].

4.4. Resumen

En este capítulo se han presentado los conceptos básicos más importantes involucrados en el procesamiento digital de señales de voz. Primero se han descrito los fundamentos del análisis de sistemas y señales, así como las técnicas que permiten manipular señales en diferentes dominios. Particularmente se ha hecho énfasis en el dominio de la frecuencia, donde se han discutido diversas transformaciones, tales como la transformada de Fourier en tiempo discreto (DTFT) para señales aperiódicas; la serie discreta de Fourier (DFS) para representar secuencias discretas periódicas; la transformada discreta de Fourier (DFT) que ofrece una representación espectral para secuencias finitas periódicas y que a su vez sus coeficientes representan muestras de la DTFT; la transformada coseno discreta (DCT), similar a la DFT, pero la cual realiza operaciones sobre valores reales; así como la transformada rápida de Fourier (FFT), algoritmo que permite reducir enormemente el número de operaciones realizadas por la DFT. También se ha presentado el teorema del muestreo, el cual establece de manera suficiente, más no necesaria, que cualquier señal limitada en banda puede ser reconstruida perfectamente a partir de sus muestras si es muestreada a una frecuencia dos veces mayor que su ancho banda ($\omega_s > 2\omega_M$). Además de los fundamentos anteriores, se han discutido varias técnicas de análisis de la señal de voz en el dominio del tiempo que permiten obtener características espectrales sin emplear métodos formales de análisis espectral. Esto es llevado a cabo analizando la señal de voz sobre periodos de tiempo lo suficientemente cortos para considerar sus características estacionarias. En cuanto al análisis de la señal de voz en el dominio de la frecuencia, se ha descrito la transformada de Fourier en tiempo discreto de tiempo corto (STDTFT) para representar en el dominio de la frecuencia las variaciones de las características espectrales de la señal de voz, así como un par de técnicas que permiten separar los componentes del sistema que modela al habla en una fuente y un filtro, como lo son el análisis mediante bancos de filtros y el análisis cepstral, definido como la transformada inversa de Fourier de la magnitud del espectro logarítmico de una señal.

5

Modelado del Lenguaje Hablado

La teoría de la probabilidad es una de las ramas de las matemáticas con mayor aplicación en el campo de la ingeniería. La probabilidad, con todos sus modelos y técnicas, es una herramienta poderosa para el manejo de datos y problemas de incertidumbre. En el procesamiento digital de señales de voz, la probabilidad juega un papel fundamental en la creación de modelos capaces de representar características esenciales del lenguaje hablado. Específicamente, los modelos probabilísticos del habla ayudan a estimar la distribución de varios fenómenos relacionados con el lenguaje natural con el propósito de incrementar el rendimiento de los sistemas que procesan voz. Desde el primer modelo significativo propuesto en 1980 para esta labor, se han realizado muchos intentos para mejorar el estado del arte. Irónicamente, algunas de las técnicas más exitosas utilizadas para modelar las características acústicas de la señal de voz emplean muy poco conocimiento de lo que la voz, el habla o el lenguaje representan.

5.1. Modelado Acústico

5.1.1. Coeficientes Cepstrales de Frecuencia Mel

Los *coeficientes cepstrales de frecuencia Mel* o coeficientes *MFCC* por sus siglas en inglés, modelan la distribución de la energía espectral en una forma perceptualmente significativa. Se derivan del cepstro de una señal analizada en tiempo corto, el cual emplea una escala de frecuencia no lineal que se aproxima al comportamiento del sistema auditivo humano. Estos coeficientes son ampliamente utilizados en muchas aplicaciones del procesamiento digital de voz. Fueron introducidos por los investigadores Steven B. Davis y Paul Mermelstein en su publicación titulada *"Comparación de representaciones paramétricas para reconocimiento de palabras monosilábicas en frases habladas continuamente"* en 1980, y debido a su efectividad y robustez bajo diversas condiciones han sido estado del arte a partir de entonces [81]. Antes de la introducción de estos coeficientes, los *coeficientes de predicción lineal* (*LPCs*) y los *coeficientes lineales de predicción cepstral* (*LPCC*) fueron utilizados en diversas aplicaciones del procesamiento digital de voz.

Cada paso en el proceso de creación de los MFCCs toma en cuenta consideraciones tanto perceptuales como computacionales. A continuación se describe el cálculo de los MFCCs de acuerdo a [10].

Dada la DFT de una señal, definida por la ecuación 4.83 como:

$$X(k) = \sum_{n=0}^{N-1} x(n) e^{-j\frac{2\pi kn}{N}}, \qquad k = 0, 1, 2, ..., (N-1)$$

Se define un banco de M filtros triangulares, $m = 1, 2, 3, ..., M$, donde m es el índice cada filtro dado por:

$$H_m(k) = \begin{cases} 0 & k < f(m-1) \\ \dfrac{2(k - f(m-1))}{(f(m+1) - f(m-1))(f(m) - f(m-1))} & f(m-1) \leq k \leq f(m) \\ \dfrac{2(f(m+1) - k)}{(f(m+1) - f(m-1))(f(m+1) - f(m))} & f(m) < k \leq f(m+1) \\ 0 & k > f(m+1) \end{cases} \tag{5.1}$$

Estos filtros calculan el espectro promedio alrededor de cada frecuencia central con anchos de banda cada vez mayores. La figura 5.1 ilustra gráficamente un banco de filtros compuesto por 6 filtros triangulares. De manera alternativa, los filtros pueden calcularse como:

$$H_m^{'}(k) = \begin{cases} 0 & k < f(m-1) \\ \dfrac{k - f(m-1)}{f(m) - f(m-1)} & f(m-1) \leq k \leq f(m) \\ \dfrac{f(m+1) - k}{f(m+1) - f(m)} & f(m) < k \leq f(m+1) \\ 0 & k > f(m+1) \end{cases} \tag{5.2}$$

donde,

$$\sum_{m=0}^{M-1} H_m^{'}(k) = 1 \tag{5.3}$$

Sean f_l y f_h las frecuencias inferior y superior, respectivamente, del banco de filtros, F_s la frecuencia de muestreo, M el número total de filtros, y N el tamaño de la FFT. Los valores de los límites $f(m)$ están espaciados uniformemente de acuerdo a la escala de Mel, es decir,

$$f(m) = \left(\frac{N}{F_s}\right) B^{-1}\left(B(f_l) + m\frac{B(f_h) - B(f_l)}{M+1}\right) \tag{5.4}$$

donde B representa la escala de Mel descrita en la sección 2.2.3.3 y se define matemáticamente como:

$$B(f) = 1127 \ln\left(1 + \frac{f}{700}\right) \tag{5.5}$$

y su inversa B^{-1} como:

$$B^{-1}(b) = 700(e^{b/1127} - 1) \tag{5.6}$$

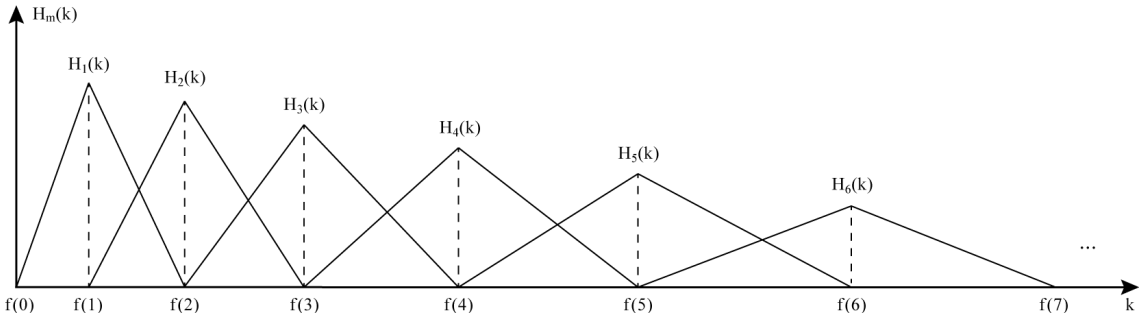


Figura 5.1: Ejemplo de un banco de filtros en el cálculo de los coeficientes MFCC [10].

El cálculo de la energía logarítmica a la salida de cada filtro se define como:

$$s(m) = \ln\left(\sum_{k=0}^{N-1} |X_a(k)|^2 H_m(k)\right), \qquad 0 \leq m < M \tag{5.7}$$

Por consiguiente, el cepstro de frecuencia Mel es la DCT de las M salidas del banco de filtros, es decir,

$$c(n) = \sum_{m=0}^{M-1} S(m) \cos\left(\frac{n(m+\frac{1}{2})\pi}{M}\right), \qquad 0 \leq n < M \tag{5.8}$$

donde el valor de M varía para diferentes implementaciones del banco de filtros, generalmente desde 24 a 40. En aplicaciones del procesamiento digital de voz, usualmente los primeros 12 o 13 coeficientes cepstrales se conservan y los demás son descartados. Es importante mencionar que a pesar de que la definición del cepstro establecida en la sección anterior emplea la DFT inversa en lugar de la DCT, ya que la secuencia $S(m)$ es par, entonces es posible emplear la transformada coseno, particularmente una DCT-II.

5.1.2. Coeficientes Delta

Si bien las características acústicas presentes en la señal de voz son eficazmente representadas a través de los coeficientes MFCC, algunos de los rasgos dinámicos del habla tales como los cambios en la forma de los órganos articuladores, pueden también ser perfectamente representados por una serie de coeficientes conocidos como *deltas* (Δ). Los *coeficientes delta* se estiman calculando las derivadas con respecto del tiempo de los vectores de características acústicas (coeficientes MFCC). A partir de los coeficientes MFCC, los coeficientes delta se pueden calcular empleando la siguiente fórmula de regresión:

$$\Delta(i,t) = \frac{\sum_{n=1}^{N} n(c(i,t+n) - c(i,t-n))}{2\sum_{n=1}^{N} n^2} \tag{5.9}$$

donde $\Delta(i,t)$ corresponde al i-esimo coeficiente delta calculado en el instante t, en términos de los correspondientes coeficientes estáticos $c(i,t+N)$ y $c(i,t-N)$ del vector de características acústicas.

5.1.3. Coeficientes Cepstrales Delta Desplazados

Además de los coeficientes delta para representar rasgos dinámicos del habla, existe otro tipo de coeficientes que permiten captar características importantes presentes en la variaciones de la señal de voz. El uso de los *coeficientes cepstrales delta desplazados* (SDC, por sus siglas en inglés), se llevó a cabo por primera vez en una investigación de B. Bielefeld en 1994 [82]. El cálculo de los coeficientes SDC, al igual que los coeficientes delta, pueden ser obtenidos a partir de los vectores de características acústicas. La figura 5.2 ilustra el proceso de obtención de dichos coeficientes.

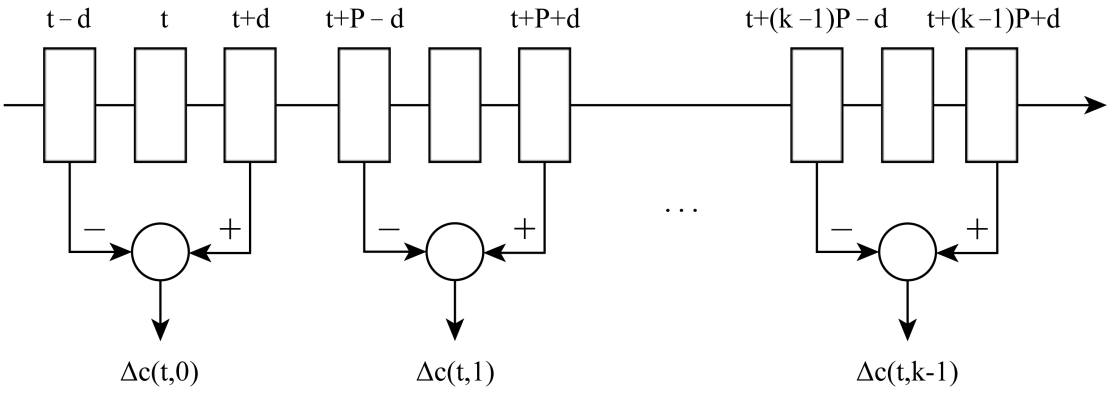


Figura 5.2: Cálculo de los coeficientes SDC en el instante t con parámetros $N-d-P-k$. Adaptado de [83].

Los coeficientes SDC se especifican fundamentalmente por un conjunto de 4 parámetros ($N-d-P-k$) [83], los cuales se describen a continuación:

- $\boldsymbol{N}$: es el número de coeficientes cepstrales calculados en cada trama.
- $\boldsymbol{d}$: representa el adelanto o atraso para el cálculo de los coeficientes delta.

- $\boldsymbol{P}$: es el desplazamiento entre bloques consecutivos.
- $\boldsymbol{k}$: es el número de bloques cuyos coeficientes delta son concatenados para formar el vector final.

En el instante t, el vector de características SDC se define como:

$$SDC(t) = \begin{bmatrix} \Delta c(t,0) \\ \Delta c(t,1) \\ \vdots \\ \Delta c(t,k-1) \end{bmatrix} \tag{5.10}$$

donde $\Delta(c,i) = c(t+iP+d) - c(t+iP-d)$ y representa la i-ésima componente del vector de características calculado en el instante de tiempo t. Cada vector SDC en el instante t emplea kP tramas consecutivas de coeficientes cepstrales.

A partir de los coeficientes MFCCs, los coeficientes delta se pueden calcular empleando la siguiente fórmula de regresión:

$$\Delta c(t+iP) = \frac{\sum_{d=-D}^{D} d \cdot c(t+iP+d)}{\sum_{d=-D}^{D} d^2} \tag{5.11}$$

5.2. Aprendizaje Automático

El *aprendizaje automático* o también conocido en inglés como *machine learning*, puede considerarse como un conjunto de herramientas y métodos que intentan inferir patrones y extraer conocimiento a partir de observaciones hechas del mundo físico [84]. Oficialmente, el término *machine learning* fue empleado por primera vez en 1959 por Arthur Samuel, quien trabajaba para IBM y definió el aprendizaje automático como *"el campo de estudio que permite otorgar a las computadoras la habilidad de aprender sin ser explícitamente programadas"* [85]. Tiempo después Tom Mitchell, jefe del Departamento de Aprendizaje Automático de la Universidad Carnegie Mellon, proporcionó una definición lo suficientemente amplia para incluir a la mayoría de las tareas que convencionalmente se emplean en este campo. En su libro *"Machine Learning"* publicado en 1997, definió formalmente que *"un programa se dice que aprende de la experiencia E con respecto a alguna clase de tareas T y alguna medida de rendimiento P, si su rendimiento en T, medida por P, mejora con la experiencia E."* [86].

De acuerdo a las definiciones anteriores, mediante el aprendizaje automático es posible programar computadoras de manera que puedan *aprender* a realizar una tarea a partir de un análisis de los datos de entrada a su disposición. El análisis de datos es llevado a cabo fundamentalmente por dos esquemas de aprendizaje: el *aprendizaje automático supervisado* y el *aprendizaje automático no supervisado*. Sin embargo, existen además varios enfoques que combinan estos dos esquemas dando lugar a otras alternativas para analizar los datos.

5.2.1. Aprendizaje Supervisado

En este esquema de aprendizaje automático, el manejo de los datos se establece de manera explícita, por lo que se busca encontrar la estructura de los datos a partir de una colección de patrones y su caracterización expresada generalmente en forma de *etiquetas*. Las etiquetas pueden proceder de un conjunto finito de valores, donde los distintos valores son denominados *clases*. Las clases son usualmente representadas mediante números enteros. Por lo tanto, cada dato x_k posee una cierta etiqueta ω_k, donde los valores de ω_k proceden de un pequeño conjunto de números enteros $\omega_k \in \{1, 2, 3, ..., c\}$, donde c representa el número de clases. El objetivo de este enfoque de aprendizaje es *clasificar* los datos mediante la construcción de un clasificador definido por una función Φ que genere una etiqueta de clase a la salida, $\Phi(x_k) = \omega_k$. La clasificación es de alguna manera similar al agrupamiento realizado por los métodos de aprendizaje no supervisados, sin embargo, se requiere saber antes de tiempo cómo son definidas las clases. La figura 5.3 muestra varios ejemplos de clasificadores supervisados.

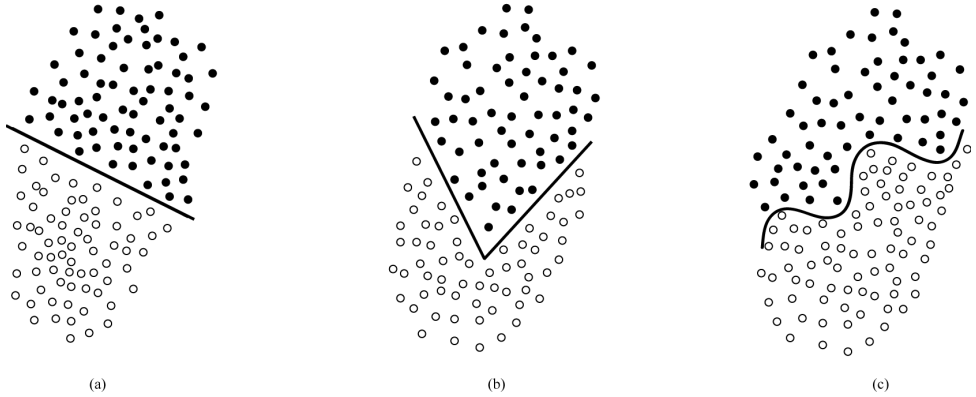


Figura 5.3: Aprendizaje automático supervisado: clasificación de datos en dos dimensiones. (a) Clasificador lineal. (b) Clasificador lineal en partes. (c) Clasificador no lineal. En los tres casos se presentan datos de dos clases representados por puntos blancos y negros.

5.2.2. Aprendizaje No Supervisado

Este enfoque del aprendizaje automático involucra un proceso que se encarga de descubrir o revelar la estructura de los datos sin supervisión alguna, es decir, que en ningún momento se establece explícitamente el manejo de los datos. Algunas de las tareas más comunes del aprendizaje automático no supervisado de acuerdo a [87] son:

- *Clustering* o *agrupamiento*, cuyo objetivo es separar los datos en grupos.
- *Detección de novedad*, que se encarga de identificar las características de los datos que son diferentes a la mayoría.
- *Reducción de dimensionalidad*, la cual pretende representar a los datos con una dimensionalidad menor manteniendo sus características importantes.

Sin embargo, de entre las tareas mencionadas anteriormente, la más relevante en las aplicaciones del aprendizaje automático no supervisado es la orientada al *clustering* o *agrupamiento* de los datos, la cual establece formalmente que dado un conjunto de datos de N dimensiones $X = \{x_1, x_2, x_3, ..., x_N\}$, donde cada x_k se caracteriza por una serie de atributos, se desea determinar la estructura de X, i.e., identificar y describir los *clusters* o *grupos* presentes en el conjunto de datos. La figura 5.4 muestra varios ejemplos de la agrupación de datos en dos dimensiones.

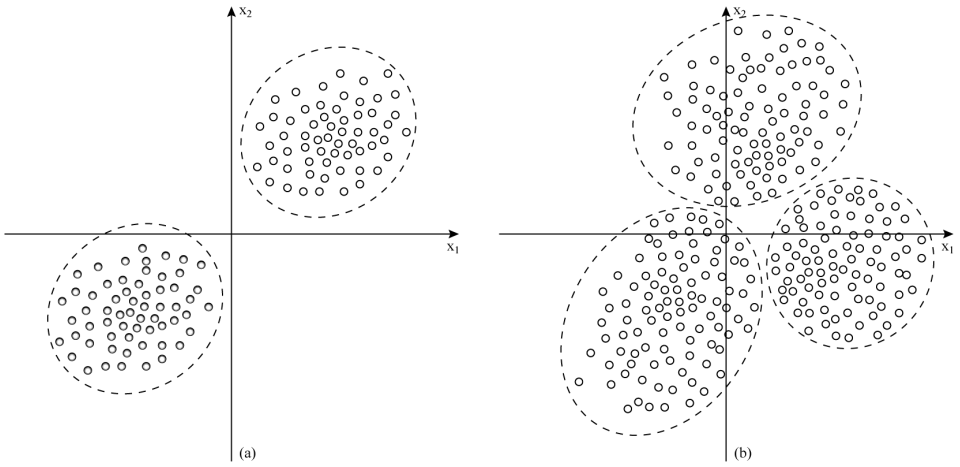


Figura 5.4: Aprendizaje automático no supervisado: agrupamiento de datos. (a) Agrupamiento en dos *clusters* o *grupos*. (b) Agrupamiento en tres *clusters* o *grupos*.

La practicidad de esta tarea es enorme, ya que en esencia, los *clusters* o *grupos* forman una abstracción del conjunto de datos. De manera que en lugar de tratar con muchos puntos de datos, solamente es necesario enfocarse en los pocos grupos identificados, lo cual es evidentemente mucho más conveniente. Sin embargo, los grupos de datos no tienen un carácter numérico; en lugar se perciben como *nubes* de datos y después se opera sobre tales estructuras. En algunas ocasiones, cada grupo posee una semántica bien definida que captura algunas de las partes dominantes y distinguibles de los datos [88].

5.3. Teoría de la Probabilidad

Uno de los conceptos claves en el aprendizaje automático es el de la incertidumbre. La *teoría de la probabilidad* proporciona un marco consistente para la cuantificación y manipulación de la incertidumbre y representa uno de los fundamentos centrales del reconocimiento de patrones.

5.3.1. Conceptos Básicos

5.3.1.1. Experimentos Aleatorios

En teoría de la probabilidad, un *experimento* es cualquier procedimiento capaz de generar resultados observables. Un experimento cuyo resultado puede ser predicho con certeza y que al repetirse bajo las mismas condiciones presenta el mismo resultado, es llamado *experimento*

determinístico. Sin embargo, aquél que no es predecible y que al repetirse bajo las mismas condiciones puede presentar diferentes resultados, es llamado *experimento aleatorio*. Si bien todos los posibles resultados de un experimento pueden conocerse con anticipación, en el caso de los experimentos aleatorios, el resultado de una ejecución en particular no puede ser predicho debido a una serie de causas desconocidas o a la naturaleza intrínseca del fenómeno.

5.3.1.2. Espacio Muestral

El *espacio muestral* de un experimento aleatorio es el conjunto de todos los posibles resultados de un experimento, y es denotado generalmente por la letra griega Ω.

5.3.1.3. Eventos

Un *evento* representa cualquier subconjunto del espacio muestral y comúnmente se denota por las primeras letras del alfabeto en mayúsculas. Cuando un evento consta de únicamente un elemento del espacio muestral se le conoce como *evento simple*. Cuando consta de más de un elemento del espacio muestral se le denomina *evento compuesto*.

5.3.1.4. Espacios de Probabilidad

En la teoría de la probabilidad, comúnmente se hace referencia a la probabilidad de ocurrencia de un evento de naturaleza incierta. Formalmente, un *espacio de probabilidad* se define por la terna $(\Omega, \mathcal{F}, P)$, en donde

- Ω es el espacio de todos los posibles resultados o *espacio muestral*,
- $\mathcal{F} \subseteq 2^{\Omega}$ es el *espacio de eventos*,
- P es la *medida de probabilidad* (o *distribución de probabilidad*) que asigna a un evento $E \in \mathcal{F}$ un valor real entre 0 y 1.

5.3.1.5. Probabilidad Axiomática

Sea el espacio muestral Ω y A un evento asociado con un experimento aleatorio. Entonces la medida de probabilidad P del evento A, denotado por $P(A)$, se define como un número real que satisface los siguientes axiomas:

1. $P(A)$ es un número no negativo. Para todo $A \in \mathcal{F}$, $P(A) \geq 0$.
2. $P(\Omega) = 1$.
3. Si A y B son eventos mutuamente excluyentes en el espacio muestral Ω, entonces la probabilidad de la unión de los eventos es igual a la suma de sus probabilidades, es decir,

$$P(A \cup B) = P(A) + P(B)$$

El término *mutuamente excluyente* empleado en el último axioma puede ser explicado de la siguiente manera: se dice que un conjunto de eventos es mutuamente excluyente si la ocurrencia de cada uno de ellos excluye la ocurrencia de los otros. Dos eventos A y B son mutuamente excluyentes si A ocurre y B no ocurre y viceversa. En otras palabras, A y B no pueden ocurrir simultáneamente, i.e., $P(A \cap B) = 0$.

5.3.1.6. Teoremas Elementales de la Probabilidad

En el desarrollo de la teoría de la probabilidad, todos los resultados son derivados directa o indirectamente utilizando los axiomas de probabilidad, como es el caso de los siguientes teoremas elementales :

Teorema 1 *La probabilidad del conjunto vacío es cero, i.e.,* $P(\varnothing) = 0$.

Teorema 2 *Si* A *es un evento cualquiera, entonces* $\bar{A}$ *también lo es y su probabilidad es:*

$$P(\bar{A}) = 1 - P(A)$$

donde que $\bar{A}$ *es el evento complementario de* A.

Teorema 3 *Si* A *y* B *son dos eventos cualesquiera, entonces:*

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

Teorema 4 *Si* A, B *y* C *son tres eventos cualesquiera, entonces:*

$$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) - P(A \cap B \cap C)$$

Teorema 5 *Si* $A \subset B$ *entonces* $P(A) \leq P(B)$.

5.3.1.7. Probabilidad Condicional

La *probabilidad condicional* de un evento A, asumiendo que el evento B ya ocurrió, se denota mediante $P(A|B)$, y siempre y cuando $P(A) \neq 0$, se define como:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \tag{5.12}$$

y se puede reescribir como:

$$P(A \cap B) = P(B)P(A|B) \tag{5.13}$$

La ecuación 5.13 se conoce como el *teorema de la multiplicación* de la probabilidad.

Las siguientes propiedades se deducen de la definición de probabilidad condicional:

- Si $A \subset B$, entonces $P(B|A) = P(B)$, ya que $A \cap B = A$.
- Si $B \subset A$, entonces $P(B|A) \geq P(B)$, ya que $A \cap B = B$, y $\frac{P(B)}{P(A)} \geq P(B)$, ya que $P(A) \leq P(\Omega) = 1$.

- Si A y B son dos eventos mutuamente excluyentes, entonces $P(B|A) = 0$, ya que $P(A \cap B) = 0$.
- Si $P(A) > P(B)$, entonces $P(A|B) > P(B|A)$.
- Si $A_1 \subset A_2$, entonces $P(A_1|B) \leq P(A_2|B)$.

5.3.1.8. Independencia

Un conjunto de eventos es independiente si la ocurrencia de cualesquiera de ellos no depende de la ocurrencia o no ocurrencia de los otros.

Cuando dos eventos A y B son independientes, entonces $P(B|A) = P(B)$. Si los eventos A y B son independientes, el teorema de la multiplicación definido por la ecuación 5.13 se reescribe como:

$$P(A \cap B) = P(A)P(B) \tag{5.14}$$

De manera general, el teorema de la multiplicación puede extenderse a cualquier número de eventos independientes, por lo que si $A_1, A_2, ..., A_n$ son n eventos independientes, entonces:

$$P(A_1 \cap A_2 \cap ...A_n) = P(A_1)P(A_2)...P(A_n) \tag{5.15}$$

Cuando la condición anterior se satisface, se dice que los eventos $A_1, A_2, ..., A_n$ son *totalmente independientes*. Algunos teoremas importantes sobre los eventos independientes son:

Teorema 6 *Si los eventos A y B son independientes, entonces los eventos $\bar{A}$ y B, o de forma similar, A y $\bar{B}$, también son independientes.*

Teorema 7 *Si los eventos A y B son independientes, entonces $\bar{A}$ y $\bar{B}$ también lo son.*

5.3.1.9. Probabilidad Total

Si los eventos $B_1, B_2, ..., B_n$ son un conjunto de eventos exhaustivo y mutuamente excluyente del espacio muestral Ω, y A es otro evento asociado u ocasionado por B_i, entonces

$$P(A) = \sum_{i=1}^{n} P(B_i)P(A|B_i) \tag{5.16}$$

5.3.1.10. Teorema de Bayes

Si los eventos $B_1, B_2, ..., B_n$ son un conjunto de eventos exhaustivo y mutuamente excluyente del espacio muestral Ω asociado con un experimento aleatorio y A es otro evento asociado u ocasionado por B_i, entonces

$$P(B_i|A) = \frac{P(B_i)P(A|B_i)}{\sum_{i=1}^{n} P(B_i)P(A|B_i)}, \qquad i = 1, 2, ..., n \tag{5.17}$$

5.3.2. Variables Aleatorias

Las *variables aleatorias* juegan un papel muy importante en la teoría de la probabilidad. El hecho más importante acerca de las variables aleatorias es que no son variables. En realidad son funciones que asignan números reales a cada posible resultado de un experimento aleatorio. Usualmente se denotan empleando las últimas letras del alfabeto en mayúsculas.

Las variables aleatorias pueden ser de dos tipos: *discretas* y *continuas*. Las variables aleatorias *discretas* son aquellas que solamente pueden tomar elementos de un conjunto finito o de un conjunto infinito numerable de valores.

5.3.2.1. Función de Probabilidad

Sea X una variable aleatoria discreta cuyos posibles valores pueden asumirse como $x_1, x_2, ..., x_n, ...$, se denota su *función de probabilidad* o *función masa de probabilidad* por $f_X(x)$ y se define como la probabilidad de que la variable aleatoria X tome algún valor x, es decir,

$$f_X(x) = P(X = x) \tag{5.18}$$

donde el *intervalo* de x, es decir, el conjunto de posibles valores que la variable aleatoria X puede asumir, se denota por $\mathrm{Val}(X)$.

La función de probabilidad tiene las siguientes propiedades:

- $0 \leq f_X(x) \leq 1$
- $\sum_{x \in \mathrm{Val}(X)} f_X(x) = 1$
- $P(a \leq X \leq b) = \sum_a^b f_X(x)$

5.3.2.2. Función de Densidad de Probabilidad

La distribución de una variable aleatoria continua X, se conoce como *función de densidad de probabilidad*, la cual es una función no negativa e integrable tal que,

$$\int_{\mathrm{Val}(X)} f_X(x)\mathrm{d}x = 1 \tag{5.19}$$

La probabilidad de una variable aleatoria continua X distribuida de acuerdo a una función de densidad de probabilidad se define como:

$$P(a \leq X \leq b) = \int_a^b f_X(x)\mathrm{d}x \tag{5.20}$$

La ecuación 5.20 indica que la probabilidad de que la variable aleatoria X tome un valor dentro del intervalo $[a, b]$ puede calcularse como el área bajo la función $f_X(x)$ en dicho intervalo. Además, implica que la probabilidad de una variable aleatoria continuamente distribuida que toma cualquier valor único es cero, es decir, que $P(X = a) = 0$.

La función de densidad de probabilidad tiene las siguientes propiedades:

- $f_X(x) \geq 0, \forall x$
- $\int_{-\infty}^{\infty} f_X(x)\mathrm{d}x = 1$

5.3.2.3. Función de Distribución Acumulativa

La *función de distribución* o *función de distribución acumulativa*, muestra el comportamiento de una variable aleatoria. Si X una variable aleatoria, entonces su función de distribución se denota por $F_X(x)$ y se define como una función que asocia a cada valor real la probabilidad de que la variable aleatoria asuma valores menores o igual que ésta, es decir,

$$F_X(x) = P(X \leq x) \tag{5.21}$$

Si la variable aleatoria X es discreta:

$$F_X(x) = \sum_{-\infty}^{x} f_X(x) \tag{5.22}$$

Si la variable aleatoria X es continua:

$$F_X(x) = \int_{-\infty}^{x} f_X(x)\mathrm{d}x \tag{5.23}$$

La función de distribución acumulativa tiene las siguientes propiedades:

- $0 \leq F_X(x) \leq 1$
- Para el mayor valor en el intervalo de la variable aleatoria X, $F_{X(x)} = 1$, es decir,

$$\lim_{x\to\infty} F_X(x) = 1$$

- Para un valor menor al primer valor en el intervalo de la variable aleatoria X, $F_X(x) = 0$, es decir,

$$\lim_{x\to-\infty} F_X(x) = 0$$

- La función de distribución es no decreciente, es decir, si $a \leq b$, entonces $F_X(a) \leq F_X(b)$
- La probabilidad de que una variable aleatoria se encuentre en el intervalo $(a, b]$ se define como:

$$P(a < x \leq b) = F_X(b) - F_X(a)$$

  o bien, de manera general:

  - Para una variable aleatoria discreta:

$$P(a \leq x \leq b) = F_X(b) - F_X(a) + f_X(a) \tag{5.24}$$

  - Para una variable aleatoria continua:

$$P(a \leq x \leq b) = F_X(b) - F_X(a) \tag{5.25}$$

5.3.2.4. Valor Esperado

Una de las operaciones más comunes sobre una variable aleatoria es el cálculo de su *valor esperado*, también conocido como *media*, *esperanza* o *primer momento*. El valor esperado de una variable aleatoria con distribución de probabilidad $f_X(x)$, se denota mediante $\mathbb{E}(X)$ y se define como:

- Para variables aleatorias discretas:

$$\mathbb{E}(X) = \sum_{x \in \mathrm{Val}(x)} x f_X(x) \tag{5.26}$$

- Para variables aleatorias continuas:

$$\mathbb{E}(X) = \int_{-\infty}^{\infty} x f_X(x) \mathrm{d}x \tag{5.27}$$

También es posible calcular el valor esperado de una función de una variable aleatoria, es decir, si X es una variable aleatoria y g es una función tal que $g(X)$ es una variable con valor esperado finito, entonces,

- Para variables aleatorias discretas:

$$\mathbb{E}[g(X)] = \sum_{x \in Val(x)} g(x) f_X(x) \tag{5.28}$$

- Para variables aleatorias continuas:

$$\mathbb{E}[g(X)] = \int_{-\infty}^{\infty} g(x) f_X(x) \mathrm{d}x \tag{5.29}$$

El valor esperado de una variable aleatoria tiene las siguiente propiedades:

- El valor esperado de una constante c es la misma constante, es decir,

$$\mathbb{E}(c) = c \tag{5.30}$$

- El valor esperado de una variable aleatoria multiplicada por una constante, es igual al valor esperado de la variable aleatoria multiplicada por la constante, es decir,

$$\mathbb{E}(cX) = c\mathbb{E}(X) \tag{5.31}$$

- El valor esperado de una variable aleatoria más una constante es igual al valor esperado de la variable aleatoria más la constante, es decir,

$$\mathbb{E}(X + b) = a\mathbb{E}(x) + c \tag{5.32}$$

- Si $X \geq 0$, entonces $\mathbb{E}(X) \geq 0$

- El valor esperado de una suma de variables aleatorias es igual a la suma de los valores esperados de las variables aleatorias. En el caso de dos variables aleatorias X e Y se tiene que:

$$\mathbb{E}(X+Y) = \mathbb{E}(X) + \mathbb{E}(Y) \tag{5.33}$$

- Si X e Y son dos variables aleatorias independientes, entonces,

$$\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y) \tag{5.34}$$

5.3.2.5. Varianza

La *varianza* de una distribución es una medida de la dispersión de una distribución, también se conoce como *segundo momento* o *momento de segundo orden* y se denota por $\mathrm{Var}(X)$ o bien, mediante σ^2. La varianza se define como:

$$\mathrm{Var}(X) = \mathbb{E}[(X - \mathbb{E}(X))^2] \tag{5.35}$$

Sin embargo, empleando las propiedades del valor esperado se puede derivar una expresión alternativa para la varianza,

$$\begin{aligned} \mathbb{E}[(X - \mathbb{E}(X))^2] &= \mathbb{E}[X^2 - 2\mathbb{E}(X)X + \mathbb{E}(X)^2] \\ &= \mathbb{E}(X^2) - 2\mathbb{E}(X)\mathbb{E}(X) + \mathbb{E}(X)^2 \\ &= \mathbb{E}(X^2) - \mathbb{E}(X)^2 \end{aligned} \tag{5.36}$$

La varianza tiene las siguientes propiedades:

- $\mathrm{Var}(X) \geq 0$
- $\mathrm{Var}(c) = 0$
- $\mathrm{Var}(cX) = c^2\mathrm{Var}(X)$
- $\mathrm{Var}(X + c) = \mathrm{Var}(X)$
- En general, $\mathrm{Var}(X + Y) \neq \mathrm{Var}(X) + \mathrm{Var}(Y)$

5.3.3. Modelos Probabilísticos Comunes

En algunas ocasiones los problemas con incertidumbre coinciden con la forma en que se define una variable aleatoria. Este hecho permite formular modelos para resolver el problema a partir de la función de probabilidad o función de densidad de la variable aleatoria. Los modelos probabilísticos se clasifican en *discretos* y *continuos*. A continuación se presentan las distribuciones más comunes.

5.3.3.1. Modelos Probabilísticos Discretos

- **Distribución Discreta Uniforme**
  La *distribución discreta uniforme* es una de las más simples de todas las distribuciones discretas de probabilidad. Es aquella en la cual la variable aleatoria asume cada uno de sus valores la misma probabilidad. La distribución discreta uniforme se denota $X \sim Uniforme(k)$ y se define como:

$$f_X(x) = \frac{1}{k}, \qquad x = x_1, x_2, ..., x_k \tag{5.37}$$

- **Distribución de Bernoulli**
  La *distribución de Bernoulli* es una de las distribuciones más sencillas para modelar un experimento. Una variable aleatoria distribuida de acuerdo a la distribución de Bernoulli solamente puede tomar dos posibles valores, $\{0, 1\}$, los cuales pueden denominarse *éxito* o *fracaso*, con probabilidades p y $q = 1 - p$, respectivamente. La distribución de Bernoulli se denota $X \sim Bernoulli(p)$ y se define como:

$$f_X(x) = p^x(1-p)^{1-x}, \qquad x = \{0, 1\} \tag{5.38}$$

- **Distribución Binomial**
  La *distribución binomial* representa el número de éxitos y fracasos en n ensayos independientes de Bernoulli para algún valor dado de n. La distribución binomial se denota $X \sim Binomial(n, p)$ y se define como:

$$f_X(x) = \binom{n}{x} p^x(1-p)^{n-x}, \qquad x = 0, 1, ..., n \tag{5.39}$$

- **Distribución Geométrica**
  La *distribución geométrica* representa el número de ensayos de Bernoulli que se requieren para observar por primera vez un éxito. La distribución geometrica se denota $X \sim$*Geométrica(p)* y se define como:

$$f_X(x) = q^{x-1}p, \qquad x = 0, 1, ..., n \tag{5.40}$$

- **Distribución de Pascal**
  La *distribución de Pascal* es una generalización de la distribución geométrica y representa el número de ensayos de Bernoulli que se requieren para observar el r-ésimo éxito, si en cada uno los ensayos se tiene una probabilidad de éxito p. La distribución de Pascal se denota $X \sim Pascal(r, p)$ y se define como:

$$f_X(x) = \binom{x-1}{r-1} p^r q^{x-r}, \qquad x = 0, 1, ..., n \tag{5.41}$$

- **Distribución Hipergeométrica**
  La *distribución hipergeométrica* representa el número de éxitos en n ensayos de Bernoulli

extraídos de una población de tamaño N, con r elementos que tienen la característica de interés. La probabilidad en cada ensayo no se mantiene constante, sino que se ve modificada en función de las extracciones anteriores. La distribución hipergeométrica se denota $X \sim$*Hipergeométrica(r,n,N)* y se define como:

$$f_X(x) = \frac{\binom{r}{x}\binom{N-r}{n-x}}{\binom{N}{n}}, \qquad x = 0, 1, ..., n; x \leq r, n - x \leq N - r \tag{5.42}$$

- **Distribución de Poisson**
  La *distribución de Poisson* es una de las distribuciones discretas más útiles. Se emplea cuando se desea calcular la probabilidad de ocurrencias de un evento en un intervalo continuo. Particularmente, esta distribución modela el número de llegadas de eventos por unidad de tiempo. La distribución de Poisson se denota $X \sim Poisson(\lambda)$ y se define como:

$$f_X(x) = \frac{\lambda^x}{x!}e^{-\lambda}, \qquad x = 0, 1, ... \tag{5.43}$$

5.3.3.2. Modelos Probabilísticos Continuos

- **Distribución Continua Uniforme**
  La *distribución continua uniforme* es la más simple de todas las distribuciones continuas de probabilidad. Se denota $X \sim Uniforme(a, b)$ y se define como:

$$f_X(x) = \frac{1}{b-a}, \qquad a \leq x \leq b \tag{5.44}$$

- **Distribución Exponencial**
  La *distribución exponencial* se emplea comúnmente para representar el tiempo de funcionamiento o de espera. Representa el intervalo (generalmente de tiempo) que transcurre entre eventos que se contabilizan por medio de la distribución de Poisson. La distribución exponencial se denota $X \sim Exponencial(\lambda)$ y se define como:

$$f_X(x) = \lambda e^{-\lambda x}, \qquad x \geq 0 \tag{5.45}$$

- **Distribución Normal**
  La *distribución normal* es también conocida como *distribución Gaussiana*. Ya que muchos problemas reales tienen un comportamiento que se puede aproximar a la distribución normal, es una de las distribuciones más empleadas en la práctica. Se denota $X \sim N(\mu, \sigma^2)$ y se define como:

$$f_X(x) = \frac{1}{(2\pi\sigma^2)^{1/2}}e^{-\frac{1}{2\sigma^2}(x-\mu)^2} \tag{5.46}$$

- **Distribución Normal Estándar**
  La *distribución normal estándar* es un caso particular de la distribución normal, la cual tiene como parámetros $\mu = 0$ y $\sigma^2 = 1$. Se denota $Z \sim N(0, 1)$. El procedimiento para obtener una variable aleatoria con distribución normal estándar a partir de una variable

aleatoria con distribución normal y parámetros cualesquiera, se lleva a cabo mediante un corrimiento y un escalamiento, lo que lleva a la expresión:

$$Z = \frac{X - \mu_X}{\sigma_X}, \qquad X \sim N(\mu_X, \sigma_X^2) \tag{5.47}$$

5.3.4. Variables Aleatorias Conjuntas

En algunas ocasiones es necesario estudiar dos o más características de un experimento. En muchos problemas cuyo estudio se realiza a través de modelos aleatorios, pueden existir múltiples variables, y debido a su interrelación se deben estudiar modelos que describan el comportamiento probabilístico conjunto de dichas variables.

Las *variables aleatorias conjuntas* son variables aleatorias definidas sobre un mismo espacio muestral.

5.3.4.1. Función de Distribución Conjunta y Marginal

Una forma de trabajar con dos variables aleatorias X e Y, es considerar a cada una por separado. De esta manera solamente se necesitan sus funciones de distribución $F_X(x)$ y $F_Y(y)$, respectivamente. Sin embargo, cuando se requiere conocer más acerca de los valores que las variables aleatorias X e Y asumen simultáneamente durante un experimento aleatorio, es necesaria una estructura más compleja conocida como **función de distribución conjunta**. De manera general, esta función de distribución proporciona el comportamiento probabilístico acumulado de una serie de variables aleatorias. En el caso particular donde se tienen dos variables aleatorias se define como:

$$F_{XY}(x, y) = P(X \leq x, Y \leq y) \tag{5.48}$$

Además,

- Si X e Y son dos variables aleatorias conjuntas discretas, entonces:

$$F_{XY}(x, y) = \sum_{u=-\infty}^{x} \sum_{v=-\infty}^{y} f_{XY}(u, v) \tag{5.49}$$

- Si X e Y son dos variables aleatorias conjuntas continuas, entonces:

$$F_{XY}(x, y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f_{XY}(u, v) \mathrm{d}v \mathrm{d}u \tag{5.50}$$

Algunas de las propiedades más importantes de la función de distribución conjunta $F_{XY}(x, y)$ son:

- $0 \leq F_{XY}(x, y) \leq 1$
- $\lim_{x,y \to \infty} F_{XY}(x, y) = 1$
- $\lim_{x,y \to -\infty} F_{XY}(x, y) = 0$

La función de distribución conjunta $F_{XY}(x, y)$ no solamente aporta información relativa a la relación entre las variables aleatorias X e Y, sino que además muestra características de X e Y por separado a través de las **distribuciones marginales**, las cuales se definen como:

$$F_X(x) = \lim_{y\to\infty} F_{XY}(x, y) \tag{5.51}$$

$$F_Y(y) = \lim_{x\to\infty} F_{XY}(x, y) \tag{5.52}$$

5.3.4.2. Función de Probabilidad Conjunta y Marginal

Si $X_1, X_2, ..., X_n$ son variables aleatorias conjuntas discretas, su **función de probabilidad conjunta** se define como:

$$f_{X_1X_2...X_n}(x_1, x_2, ..., x_n) = P(X_1 = x_1, X_2 = x_2, ..., X_n = x_n) \tag{5.53}$$

En particular si se tienen dos variables aleatorias X e Y:

$$f_{XY}(x, y) = P(X = x, Y = y) \tag{5.54}$$

Las propiedades de la función de probabilidad conjunta son:

- $0 \leq f_{XY}(x, y) \leq 1, \qquad \forall x, y$
- $\sum_{x\in \text{Val}(X)} \sum_{y\in \text{Val}(X)} f_{XY}(x, y) = 1$
- $P(x_0 \leq X \leq x_1, y_0 \leq Y \leq y_1) = \sum_{x=x_0}^{x_1} \sum_{y=y_0}^{y_1} f_{XY}(x, y)$

La función de probabilidad conjunta se relaciona con la función de probabilidad para cada variable aleatoria de forma separada mediante la **función de probabilidad marginal**, la cual se define para las variables X e Y, respectivamente como:

$$f_X(x) = \sum_{y\in \text{Val}(y)} f_{XY}(x, y) \tag{5.55}$$

$$f_Y(y) = \sum_{x\in \text{Val}(x)} f_{XY}(x, y) \tag{5.56}$$

5.3.4.3. Función de Densidad Conjunta y Marginal

Si $X_1, X_2, ..., X_n$ son variables aleatorias conjuntas continuas, su **función de densidad conjunta** se define como:

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} f_{X_1X_2...X_n}(x_1, x_2, ..., x_n)\mathrm{d}x_1\mathrm{d}x_2...\mathrm{d}x_n = P((X_1, X_2, ..., X_n) \in \mathbb{R}) \tag{5.57}$$

La función de densidad conjunta tiene las siguentes propiedades:

- $f_{X_1X_2...X_n}(x_1, x_2, ..., x_n) \geq 0 \qquad \forall x_1, x_2, ..., x_n$

- $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} f_{X_1X_2...X_n}(x_1,x_2,...,x_n)\mathrm{d}x_1\mathrm{d}x_2...\mathrm{d}x_n = 1$

Para el caso específico donde se tienen dos variables aleatorias, la función de densidad conjunta se define como:

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{XY}(x,y)\mathrm{d}x\mathrm{d}y = P((X,Y)\in A) \tag{5.58}$$

donde A es cualquier región en el plano xy. La función de densidad conjunta se relaciona con la función de densidad para cada variable aleatoria de forma separada mediante la **función de densidad marginal**, la cual se define para las variables X y Y, respectivamente como:

$$f_X(x) = \int_{-\infty}^{\infty} f_{XY}(x,y)\mathrm{d}y \tag{5.59}$$

$$f_Y(y) = \int_{-\infty}^{\infty} f_{XY}(x,y)\mathrm{d}x \tag{5.60}$$

$$\tag{5.61}$$

5.3.4.4. Distribuciones Condicionales

Las *distribuciones condicionales* son empleadas para conocer cuál es la distribución de probabilidad de una variable aleatoria Y, cuando se sabe que la variable aleatoria X puede tomar un cierto valor x.

Para variables aleatorias discretas, la **función de probabilidad condicional** de X dado Y, asumiendo que $f_X(x) \neq 0$, se define como:

$$f_{Y|X}(y|x) = \frac{f_{XY}(x,y)}{f_X(x)} \tag{5.62}$$

En el caso continuo, la situación es técnicamente un poco más complicada ya que la probabilidad de que una variable aleatoria continua X tome un valor específico x es igual a cero. Sin embargo, ignorando este hecho, se define análogamente al caso discreto, la **función de densidad condicional** de Y dado X siempre y cuando $f_X(x) \neq 0$ igual que la ecuación 5.62.

5.3.4.5. Regla de Bayes

Una fórmula que comúnmente se presenta al intentar derivar una expresión para obtener la probabilidad condicional de una variable dada otra es la regla de Bayes.

Para las variables aleatorias discretas X e Y se establece que,

$$P_{Y|X}(y|x) = \frac{P_{XY}(x,y)}{P_X(x)} = \frac{P_{X|Y}(x|y)P_Y(y)}{\sum_{y'\in \mathrm{Val}(Y)} P_{X|Y}(x|y')P_Y(y')} \tag{5.63}$$

Para las variables aleatorias continuas X y Y se establece que,

$$f_{Y|X}(y|x) = \frac{f_{XY}(x,y)}{f_X(x)} = \frac{f_{X|Y}(x|y)f_Y(y)}{\int_{-\infty}^{\infty} f_{X|Y}(x|y')f_Y(y')\mathrm{d}y'} \tag{5.64}$$

5.3.4.6. Independencia

Dos variables aleatorias X e Y son *independientes* si $F_{XY}(x, y) = F_X(x)F_Y(y)$ para todos los valores de x e y. De forma equivalente, si X y Y son variables aleatorias conjuntas, éstas son independientes si y sólo si la función de densidad conjunta es igual al producto de las marginales, es decir,

$$f_{XY}(x, y) = f_X(x)f_Y(y) \tag{5.65}$$

5.3.4.7. Valor Esperado

Si X e Y son variables aleatorias conjuntas con función de probabilidad o densidad conjunta $f_{XY}(x, y)$ y si $g(X, Y)$ es una función de dichas variables aleatorias, entonces el valor esperado de $g(X, Y)$ se define como:

- Para variables aleatorias discretas:

$$\mathbb{E}[g(X, Y)] = \sum_{x\in \mathrm{Val}(X)} \sum_{y\in \mathrm{Val}(Y)} g(x, y)f_{XY}(x, y) \tag{5.66}$$

- Para variables aleatorias continuas:

$$\mathbb{E}[g(X, Y)] = \int_{-\infty}^{\infty} g(x, y)f_{XY}(x, y)\mathrm{d}x\mathrm{d}y \tag{5.67}$$

5.3.4.8. Covarianza

El concepto de valor esperado puede utilizarse para definir la *covarianza*, la cual permite estudiar la relación existente entre dos variables aleatorias. La covarianza de dos variables aleatorias X e Y se define como:

$$\mathrm{Cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}(X))(Y - \mathbb{E}(Y))] \tag{5.68}$$

Empleando las propiedades del valor esperado se puede derivar una expresión alternativa para la covarianza,

$$\begin{aligned}\mathrm{Cov}(X, Y) &= \mathbb{E}[(X - \mathbb{E}(X))(Y - \mathbb{E}(Y))] \\ &= \mathbb{E}[XY - X\mathbb{E}(Y) - Y\mathbb{E}(X) + \mathbb{E}(X)\mathbb{E}(Y)] \\ &= \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y) - \mathbb{E}(Y)\mathbb{E}(X) + \mathbb{E}(X)\mathbb{E}(Y) \\ &= \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)\end{aligned} \tag{5.69}$$

Algunas propiedades relacionadas con la covarianza se listan a continuación:

- $\mathbb{E}[f(X, Y) + g(X, Y)] = \mathbb{E}[f(X, Y)] + \mathbb{E}[g(X, Y)]$.
- $\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\mathrm{Cov}(X, Y)$.
- Si X e Y son dos variables aleatorias independientes, entonces $\mathrm{Cov}(X, Y) = 0$.
- Si X e Y son dos variables aleatorias independientes, entonces $\mathbb{E}[f(X)g(Y)] = \mathbb{E}[f(X)]\mathbb{E}[g(Y)]$.

5.3.5. Vectores Aleatorios

Cuando se tienen múltiples variables aleatorias, algunas veces es conveniente expresarlas en un vector $X = [X_1 X_2 ... X_n]^{\mathrm{T}}$. Este vector se conoce como **vector aleatorio** y es únicamente una notación alternativa para tratar con n variables aleatorias, de manera que las nociones de probabilidad discutidas anteriormente también son válidas para los vectores aleatorios.

5.3.5.1. Valor Esperado

El valor esperado de un vector aleatorio es un vector cuyos elementos son los valores esperados de las variables aleatorias individuales que conforman al vector, es decir, si

$$g(x) = \begin{bmatrix} g_1(x) \\ g_2(x) \\ \vdots \\ g_n(x) \end{bmatrix} \tag{5.70}$$

entonces,

$$\mathbb{E}[g(x)] = \begin{bmatrix} \mathbb{E}[g_1(X)] \\ \mathbb{E}[g_2(X)] \\ \vdots \\ \mathbb{E}[g_n(X)] \end{bmatrix} \tag{5.71}$$

5.3.5.2. Matriz de Covarianza

Para un vector aleatorio X, su matriz de covarianza Σ es la matriz cuadrada de $n \times n$, cuyos valores están dados por $\sum_{ij} = \mathrm{Cov}(X_i, X_j)$. La matriz de covarianza se define como:

$$\Sigma = \begin{bmatrix} \mathrm{Cov}(X_1, X_1) & \cdots & \mathrm{Cov}(X_1, X_n) \\ \vdots & \ddots & \cdots \\ \mathrm{Cov}(X_n, X_1) & \cdots & \mathrm{Cov}(X_n, X_n) \end{bmatrix} \tag{5.72}$$

La ecuación 5.72 puede ser reescrita en términos del valor esperado como:

$$\begin{aligned} \Sigma &= \begin{bmatrix} \mathbb{E}(X_1^2) - \mathbb{E}(X_1)\mathbb{E}(X_1) & \cdots & \mathbb{E}(X_1 X_n) - \mathbb{E}(X_1)\mathbb{E}(X_n) \\ \vdots & \cdots & \cdots \\ \mathbb{E}(X_n X_1) - \mathbb{E}(X_n)\mathbb{E}(X_1) & \cdots & \mathbb{E}(X_n^2) - \mathbb{E}(X_n)\mathbb{E}(X_n) \end{bmatrix} \\ &= \begin{bmatrix} \mathbb{E}(X_1^2) & \cdots & \mathbb{E}(X_1 X_n) \\ \vdots & \cdots & \vdots \\ \mathbb{E}(X_n X_1) & \cdots & \mathbb{E}(X_n^2) \end{bmatrix} - \begin{bmatrix} \mathbb{E}(X_1)\mathbb{E}(X_1) & \cdots & \mathbb{E}(X_1)\mathbb{E}(X_n) \\ \vdots & \cdots & \vdots \\ \mathbb{E}(X_n)\mathbb{E}(X_1) & \cdots & \mathbb{E}(X_n)\mathbb{E}(X_n) \end{bmatrix} \\ &= \mathbb{E}(XX^{\mathrm{T}}) - \mathbb{E}(X)\mathbb{E}(X)^{\mathrm{T}} = ... = \mathbb{E}[(X - \mathbb{E}(X))(X - \mathbb{E}(X))^{\mathrm{T}}] \end{aligned} \tag{5.73}$$

5.4. Modelado Estadístico

5.4.1. La Distribución Gaussiana

La *distribución Gaussiana* o *distribución normal* fue inicialmente asociada con errores de medición. Su descubrimiento se llevó a cabo en la segunda parte del siglo XVII, cuando el físico y astrónomo italiano Galileo Galilei se dio cuenta de que los errores en sus observaciones astronómicas no eran del todo aleatorios, ya que no solamente los errores pequeños superaban en número a los errores más grandes, sino que además los errores tendían a estar simétricamente distribuidos alrededor de un valor central. En otro contexto matemático, a principios del siglo XVIII el estadístico francés Abraham de Moivre, quien era comúnmente frecuentado para dar consultorías estadísticas a jugadores de apuestas, mostró que ciertas distribuciones binomiales podían ser aproximadas por una misma curva general a medida que el número de eventos incrementaba. Sin embargo, fue hasta la primera década del siglo XIX que los matemáticos Adrien-Marie Legendre y Carl Friedrich Gauss elaboraron la ecuación matemática precisa para esta curva, misma que había sido descubierta por Laplace en 1778 cuando derivó el *teorema del límite central* [89].

A pesar de la contribución de muchos matemáticos a la aparición del concepto de la distribución normal, fue el nombre de Gauss el más fuertemente vinculado al descubrimiento, debido en gran parte a que Gauss la asoció con el método de mínimos cuadrados en su primera publicación sobre este tema en el año de 1809 [90].

5.4.1.1. Distribución Gaussiana Univariante

La *distribución Gaussiana*, también conocida como *distribución normal*, es un modelo ampliamente utilizado para la distribución de variables continuas. Esta distribución proporciona una manera fácil para estimar la incertidumbre en muchos fenómenos que acontecen en el mundo. En el caso de una sola variable x, la distribución Gaussiana puede ser escrita de la siguiente forma:

$$\mathcal{N}(x|\mu,\sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}}\exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\} \tag{5.74}$$

donde la ecuación 5.74 expresa un término exponencial multiplicado por un escalar y satisface los dos siguientes requirimientos para cualquier densidad de probabilidad válida:

$$\mathcal{N}(x|\mu,\sigma^2) > 0 \tag{5.75}$$

$$\int_{-\infty}^{\infty}\mathcal{N}(x|\mu,\sigma^2)\mathrm{d}x = 1 \tag{5.76}$$

Lo que hace a la distribución Gaussiana muy útil e importante son principalmente dos factores. *Primero*, solamente dos parámetros son necesarios para especificar la distribución: la *media*, denotada por μ, y la *varianza*, denotada por σ^2. Estos parámetros además de capturar la esencia de la distribución, son fáciles de calcular e interpretar. La raíz cuadrada de la varianza, denotada por σ, es llamada *desviación estándar*, y el recíproco de la varianza, denotado por $\beta = 1/\sigma^2$, representa la *precisión*.

Con el uso del parámetro de precisión β, la distribución Gaussiana puede reescribirse como:

$$\mathcal{N}(x|\mu,\sigma^2) = \sqrt{\frac{\beta}{2\pi}}\exp\left\{-\frac{\beta}{2}(x-\mu)^2\right\} \tag{5.77}$$

Segundo, matemáticamente la distribución Gaussiana tiene algunas propiedades importantes. Por ejemplo, el producto de varias distribuciones Gaussianas forma otra distribución Gaussiana, de manera que no es necesario preocuparse por encontrar otras formas de distribuciones cuando se realizan operaciones con el modelo Gaussiano.

Por último y más teóricamente, el *teorema del límite central* indica que el valor esperado de cualquier variable aleatoria converge a la distribución Gaussiana, lo cual implica que esta distribución es una elección adecuada para modelar problemas de incertidumbre. Debido a los beneficios tanto teóricos como prácticos, la distribución Gaussiana es ampliamente utilizada.

Cuando la distribución Gaussiana tiene media igual a cero y varianza unitaria se conoce como *distribución normal estándar*. La figura 5.5 muestra la gráfica de esta distribución, la cual es simétrica alrededor de la media. También se puede observar que el valor de $\mathcal{N}(x|\mu,\sigma^2)$ se vuelve muy pequeño a medida que el valor de x se aleja de la media, lo cual es debido al signo negativo dentro de la función exponencial de la ecuación 5.74.

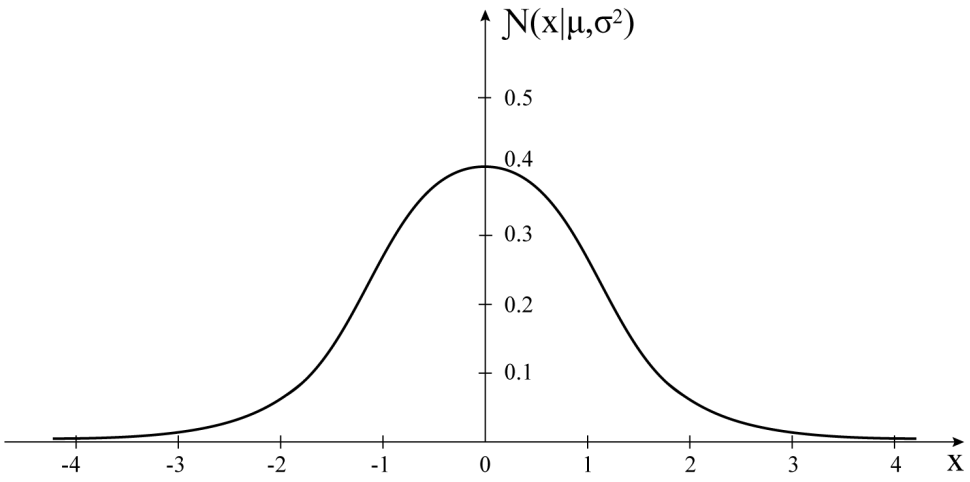


Figura 5.5: Distribución Gaussiana con media igual a cero y varianza unitaria.

De forma general, en la ecuación de la distribución Gaussiana el valor de la media determina el centro de la distribución, o bien, la ubicación del máximo. Críticamente, la forma de la distribución no se ve afectada, pues un cambio en el valor de la media solamente produce un desplazamiento en la distribución. La figura 5.6 (a) y 5.6 (b) muestran un par de distribuciones Gaussianas desplazadas debido a un cambio en el valor de la media con respecto a una distribución normal estándar. La varianza, por su parte, es el parámetro que modifica la dispersión de la distribución. Si la varianza incrementa, entonces la curva Gaussiana se extiende hacia los extremos en comparación con la curva de la distribución normal estándar, por lo que también el valor del máximo decrece de manera que la integral de la distribución siga siendo uno, lo cual cumple las propiedades de una función de densidad de probabilidad. De manera opuesta, si el valor de

la varianza disminuye, entonces la curva se estrecha y el valor del máximo aumenta de forma que la integral de la distribución permanezca siendo uno. Las figuras 5.6 (c) y 5.6 (d) muestran las modificaciones ocurridas a la curva de la distribución Gaussiana cuando el parámetro de la varianza cambia.

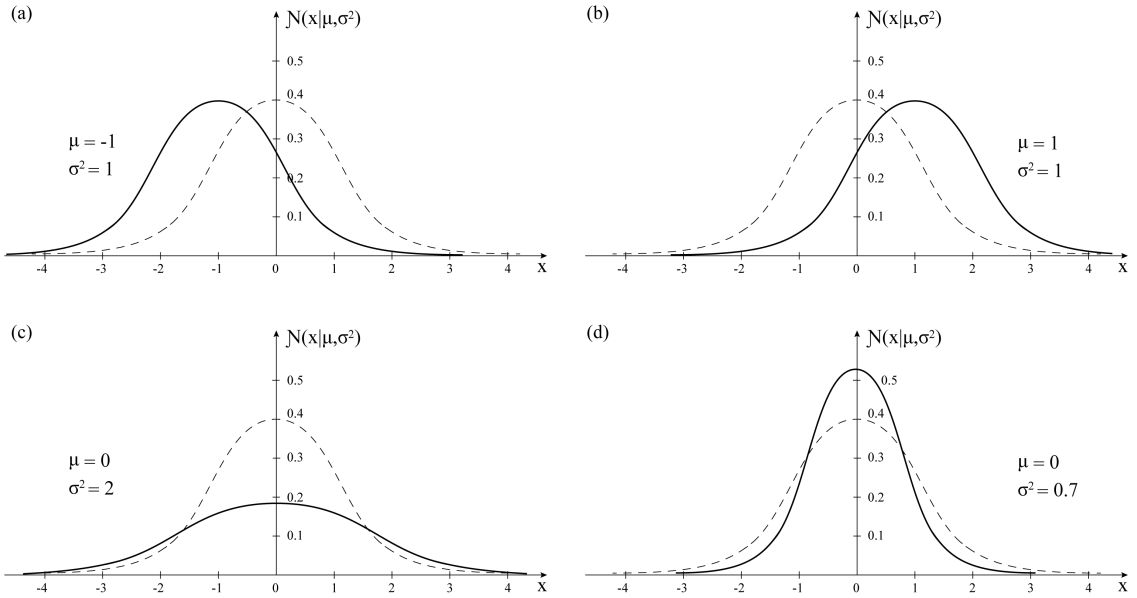


Figura 5.6: Distintas distribuciones Gaussianas. (a) Distribución Gaussiana con $\mu = -1$ y $\sigma^2 = 1$. (b) Distribución Gaussiana con $\mu = 1$ y $\sigma^2 = 1$. (c) Distribución Gaussiana con $\mu = 0$ y $\sigma^2 = 2$. (d) Distribución Gaussiana con $\mu = 0$ y $\sigma^2 = 0.7$

5.4.1.2. Momentos

El valor esperado de x está dado por:

$$\mathbb{E}(x) = \int_{-\infty}^{\infty} \mathcal{N}(x|\mu, \sigma^2) x \mathrm{d}x = \mu \tag{5.78}$$

ya que el parámetro μ representa el valor promedio de x bajo la distribución, es referido como la media. El máximo de una distribución es conocido como su *modo*. En la distribución Gaussiana, el modo coincide con la media. Similarmente, para el momento de segundo orden o varianza,

$$\mathbb{E}(x^2) = \int_{-\infty}^{\infty} \mathcal{N}(x|\mu, \sigma^2) x^2 \mathrm{d}x = \mu^2 + \sigma^2 \tag{5.79}$$

De la ecuación 5.78 y la ecuación 5.79, resulta que la varianza de x en función del valor esperado está dada por:

$$\mathrm{Var}(x) = \mathbb{E}(x^2) - \mathbb{E}(x)^2 = \sigma^2 \tag{5.80}$$

5.4.1.3. Teorema del Límite Central

El *teorema del límite central* establece que, bajo ciertas condiciones, la distribución de probabilidad de la suma de un conjunto de variables aleatorias independientes e idénticamente distribuidas (i.i.d) se aproxima a una distribución Gaussiana a medida que el número de términos en la suma aumenta. Lo anterior se puede ilustrar al considerar N variables $x_1, x_2, ..., x_N$, cada una de las cuales tiene una distribución uniforme en el intervalo $[0, 1]$, y que las distribución de la media es $(x_1 + x_2 + ... + x_N)/N$. Para un número suficientemente grande de N, la distribución tiende a ser una distribución gaussiana.

5.4.1.4. Estimación de Máxima Verosimilitud

La *verosimilitud* es la probabilidad de una observación dados los parámetros de un modelo. Vista como una función de μ y σ^2 se denota como:

$$p(\{x_i\}|\mu, \sigma^2) \tag{5.81}$$

donde el subíndice i indica una observación particular x_i entre múltiples observaciones de x.

Un criterio común para determinar los parámetros en una distribución de probabilidad empleando un conjunto de datos observados, es encontrar los valores de los parámetros que maximizan la función de verosimilitud. En el caso particular de la distribución Gaussiana, se debe determinar el valor los de los parámetros desconocidos μ y σ^2 que maximizan la función de verosimilitud. Matemáticamente lo anterior se puede escribir como:

$$\hat{\mu}, \hat{\sigma}^2 = \arg \max_{\mu,\sigma^2} p(\{x_i\}|\mu, \sigma^2) \tag{5.82}$$

donde $\hat{\mu}$ y $\hat{\sigma}^2$ representan los valores estimados de μ y σ^2 respectivamente.

La función de verosimilitud que se desea maximizar es la probabilidad conjunta de todos los datos observados, los cuales pueden ser intratables si cada instancia de x depende de las demás observaciones. Sin embargo, si se asume que las observaciones x_i son independientes entre sí e idénticamente distribuidas, entonces la probabilidad conjunta puede ser expresada simplemente como el producto de las probabilidades marginales para cada evento por separado, es decir,

$$p(\{x_i\}|\mu, \sigma^2) = \prod_{i=1}^{N} p(x_i|\mu, \sigma^2) \tag{5.83}$$

De acuerdo a la ecuación 5.83, el cálculo de los valores estimados de μ y σ en la ecuación 5.82 se puede reescribir como:

$$\hat{\mu}, \hat{\sigma}^2 = \arg \max_{\mu,\sigma^2} \prod_{i=1}^{N} p(x_i|\mu, \sigma^2) \tag{5.84}$$

Una ventaja importante de la distribución Gaussiana es que a partir de la notación anterior, existe una solución analítica con la cual es posible calcular la estimación de máxima verosimilitud de μ y σ^2.

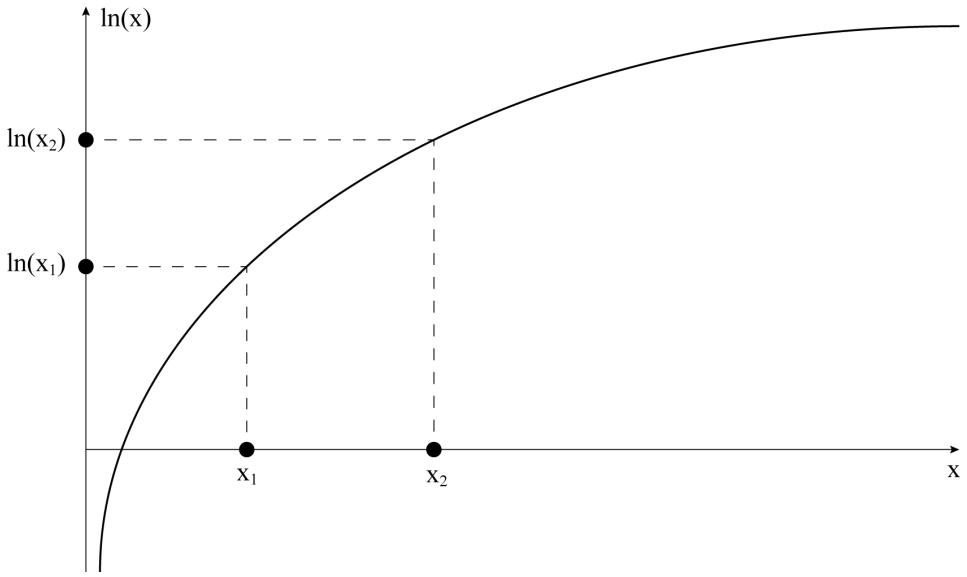


Figura 5.7: Función logarítmica. En una función monótona creciente se tiene la siguiente propiedad: $x_1 \leq x_2 \leftrightarrow \ln(x_1) \leq \ln(x_2)$.

En la práctica, es más conveniente maximizar el logaritmo de la función de verosimilitud, por lo que para el cálculo de las estimaciones es necesario aplicar algunas propiedades de la función logarítmica. En la figura 5.7 se muestra una función de este tipo. Ya que el logaritmo es una función monótona creciente de su argumento, donde si algún valor x^* es el máximo del dominio, entonces el logaritmo de ese valor ($\log x^*$) es también el máximo de la función. En otras palabras, la maximización del logaritmo de una función es equivalente a la maximización de la función en sí. Utilizando esta propiedad en lugar de maximizar la verosimilitud, es posible encontrar los parámetros que maximizan la verosimilitud logarítmica. Matemáticamente, lo anterior se puede expresar como:

$$\arg \max_{\mu,\sigma^2} \prod_{i=1}^{N} p(x_i|\mu,\sigma^2) = \arg \max_{\mu,\sigma^2} \ln \left\{ \prod_{i=1}^{N} p(x_i|\mu,\sigma^2) \right\} \tag{5.85}$$

Tomar el logaritmo no solamente simplifica el análisis matemático subsecuente, sino que además ayuda numéricamente, puesto que el producto de un número grande de probabilidades pequeñas puede llevar a una computadora a experimentar fácilmente errores de precisión numérica, hecho que se resuelve al calcular la suma de las probabilidades logarítmicas. Con el uso de esta propiedad el lado derecho de la ecuación 5.85 es equivalente a la siguiente expresión:

$$\arg \max_{\mu,\sigma^2} \ln \left\{ \prod_{i=1}^{N} p(x_i|\mu,\sigma^2) \right\} = \arg \max_{\mu,\sigma^2} \sum_{i=1}^{N} \ln p(x_i|\mu,\sigma^2) \tag{5.86}$$

Ya que se trata de la distribución Gaussiana,

$$p(x_i|\mu,\sigma^2) = \mathcal{N}(x_i|\mu,\sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{ -\frac{1}{2\sigma^2}(x_i-\mu)^2 \right\} \tag{5.87}$$

Sustituyendo la ecuación 5.87 en el lado derecho de la ecuación 5.86,

$$\arg \max_{\mu,\sigma^2} \sum_{i=1}^{N} \ln p(x_i|\mu,\sigma^2) = \arg \max_{\mu,\sigma^2} \sum_{i=1}^{N} \ln \left\{ \frac{1}{(2\pi\sigma^2)^{1/2}} \exp \left\{ -\frac{1}{2\sigma^2}(x_i-\mu)^2 \right\} \right\} \tag{5.88}$$

Utilizando las propiedades de los logaritmos, el lado derecho de la ecuación 5.88 puede reescribirse como:

$$\arg \max_{\mu,\sigma^2} \sum_{i=1}^{N} \left\{ -\frac{1}{2\sigma^2}(x_i-\mu)^2 - \ln\sqrt{2\pi}\sigma \right\} \tag{5.89}$$

o bien, aplicando nuevamente la propiedad del producto del logaritmo como:

$$\arg \max_{\mu,\sigma^2} \sum_{i=1}^{N} \left\{ -\frac{1}{2\sigma^2}(x_i-\mu)^2 - \ln\sigma - \ln\sqrt{2\pi} \right\} \tag{5.90}$$

Sin embargo, es posible simplificar aún más la ecuación 5.90. Ya que el término $\ln\sqrt{2\pi}$ no varía con los parámetros de interés, es posible ignorarlo sin afectar la solución, por lo que la estimación de los parámetros μ y σ puede expresarse como:

$$\hat{\mu}, \hat{\sigma}^2 = \arg \max_{\mu,\sigma^2} \sum_{i=1}^{N} \left\{ -\frac{1}{2\sigma^2}(x_i-\mu)^2 - \ln\sigma \right\} \tag{5.91}$$

La ecuación 5.91 puede convertirse en un problema de minimización al cambiar $\max$ por $\min$ y tomar el negativo de todos los términos, es decir,

$$\hat{\mu}, \hat{\sigma}^2 = \arg \min_{\mu,\sigma^2} \sum_{i=1}^{N} \left\{ \frac{1}{2\sigma^2}(x_i-\mu)^2 + \ln\sigma \right\} \tag{5.92}$$

Aunque las ecuaciones 5.91 y 5.92 son equivalentes, reescribir el problema como un problema de minimización es la forma estándar en la estimación de máxima verosimilitud. Finalmente, los valores de μ y σ que maximizan la función de máxima verosimilitud se obtienen al tomar derivadas con respecto a la variable deseada y resolviendo la ecuación obtenida.

Sea

$$J(\mu,\sigma) = \sum_{i=1}^{N} \left\{ \frac{1}{2\sigma^2}(x_i-\mu)^2 + \ln\sigma \right\} \tag{5.93}$$

Entonces,

$$\hat{\mu}, \hat{\sigma}^2 = \arg \min_{\mu,\sigma^2} J(\mu,\sigma) \tag{5.94}$$

$J(\mu,\sigma)$ es un símbolo común que representa una función que se quiere minimizar. Si se aplica la condición de optimalidad para optimización convexa, la derivada parcial de primer orden de

J con respecto a μ debe ser cero. Matemáticamente:

$$\frac{\partial J}{\partial \mu} = \frac{\partial}{\partial \mu} \sum_{i=1}^{N} \left\{ \frac{1}{2\sigma^2}(x_i - \mu)^2 + \ln \sigma \right\} \tag{5.95}$$

$$= \sum_{i=1}^{N} \left\{ \frac{\partial}{\partial \mu} \frac{1}{2\sigma^2}(x_i - \mu)^2 \right\} \tag{5.96}$$

$$= \frac{1}{\sigma^2} \sum_{i=1}^{N} (x_i - \mu) = 0 \tag{5.97}$$

De la ecuación 5.97 es posible obtener la estimación de máxima verosimilitud de μ, la cual se denota como $\hat{\mu}$ y se define como:

$$\hat{\mu} = \frac{1}{N} \sum_{i=1}^{N} x_i \tag{5.98}$$

De forma similar, se puede aplicar la misma condición de optimalidad para calcular la estimación de σ^2. En este caso, se puede emplear el valor de $\hat{\mu}$ en lugar de μ como parámetro. Entonces,

$$\frac{\partial J}{\partial \sigma} = \frac{\partial}{\partial \sigma} \sum_{i=1}^{N} \left\{ \frac{1}{2\sigma^2}(x_i - \hat{\mu})^2 + \ln \sigma \right\} \tag{5.99}$$

$$= \left(\frac{\partial}{\partial \sigma} \frac{1}{2\sigma^2} \right) \left(\sum_{i=1}^{N} (x_i - \hat{\mu})^2 \right) - \frac{N}{\sigma} \tag{5.100}$$

$$= \frac{1}{\sigma} \left(N - \frac{1}{\sigma^2} \sum_{i=1}^{N} (x_i - \hat{\mu})^2 \right) = 0 \tag{5.101}$$

De la ecuación 5.101, la estimación de σ^2, denotada por $\hat{\sigma}^2$ se define como:

$$\hat{\sigma}^2 = \frac{1}{N} \sum_{i=1}^{N} (x_i - \hat{\mu})^2 \tag{5.102}$$

La solución de máxima verosimilitud está dada por las ecuaciones 5.98 y 5.102. El parámetro $\hat{\mu}$ representa la *media muestral*, i.e., la media de los valores observados $\{x_i\}$, y el parámetro $\hat{\sigma}^2$ representa la *varianza muestral* medida con respecto a la media muestral $\hat{\mu}$.

5.4.2. La Distribución Gaussiana Multivariante

La *distribución Gaussiana multivariante* o *normal multivariante* es ampliamente utilizada en funciones de densidad de probabilidad conjuntas para variables continuas. Por su capacidad de utilizar múltiples variables, esta distribución aprovecha más características para modelar fenómenos. Matemáticamente, la distribución Gaussiana multivariante se expresa como una exponencial multiplicada por un vector escalar. En D dimensiones se define como:

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \Sigma) = \frac{1}{(2\pi)^{D/2}|\Sigma|^{1/2}} \exp \left\{ -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^{\mathrm{T}} \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right\} \tag{5.103}$$

donde $\mathbf{x}$ es el vector de variables cuya probabilidad intenta ser cuantificada. En contraste con la distribución Gaussiana univariante, la media $\boldsymbol{\mu}$ es ahora un vector D-dimensional, Σ es una matriz cuadrada de $D \times D$ dimensiones que representa la matriz de covarianza, y $|\Sigma|$ denota el determinante de Σ.

En la matriz de covarianza Σ, existen dos componentes clave: los términos presentes en la diagonal y los términos fuera de la diagonal. En el caso particular donde $D = 2$, se tiene que la matriz de covarianza es igual a:

$$\Sigma = \begin{bmatrix} \sigma_{x_1}^2 & \sigma_{x_1}\sigma_{x_2} \\ \sigma_{x_2}\sigma_{x_1} & \sigma_{x_2}^2 \end{bmatrix} \tag{5.104}$$

donde $\sigma_{x_1}\sigma_{x_2} = \sigma_{x_2}\sigma_{x_1}$. Los términos diagonales son varianzas independientes de cada variable, x_1 y x_2. Los términos fuera de la diagonal $\sigma_{x_1}\sigma_{x_2}$ y $\sigma_{x_2}\sigma_{x_1}$ representan la correlación entre las dos variables. De manera general, un componente de correlación representa qué tanto una variable está relacionada con otra.

Para conocer el comportamiento de la distribución Gaussiana multivariante es conveniente analizarla para el caso donde $D = 2$ y a partir de un caso especial en el cual la distribución tiene media igual a cero, varianza unitaria y términos de correlación iguales a cero, es decir, donde los valores de los parámetros de la distribución son:

$$\mathbf{x} = \begin{bmatrix} x & y \end{bmatrix}^{\mathrm{T}}$$
$$\boldsymbol{\mu} = \begin{bmatrix} 0 & 0 \end{bmatrix}^{\mathrm{T}}$$
$$\Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

La función de densidad de probabilidad de la ecuación 5.103 dados los parámetros anteriores se simplifica a la siguiente expresión:

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \Sigma) = \frac{1}{2\pi} \exp\left\{-\frac{x^2 + y^2}{2}\right\} \tag{5.105}$$

En la figura 5.8 (a), se muestra la gráfica de la ecuación 5.105. La representación tridimensional de una distribución Gaussiana de dos dimensiones se asemeja a la forma de una montaña con un solo pico. Si la superficie del pico se cortara por la mitad, la sección transversal tendría exactamente la forma de una distribución Gaussiana de una dimensión. Sin embargo, en algunas ocasiones es más útil dibujar la superficie de la distribución en dos dimensiones, tal y como se muestra en la figura 5.8 (b). La representación en dos dimensiones es la vista superior de una representación en tres dimensiones. Los contornos dibujados conectan los valores de $\mathbf{x}$ con el mismo valor de probabilidad.

En el caso esférico donde la covarianza es una matriz diagonal en la que los elementos de la diagonal tienen valores iguales, los contornos aparecen como círculos. El círculo más interior es donde se ubica el pico o máximo de la distribución, mientras que los círculos exteriores representan las regiones menos probables de la distribución.

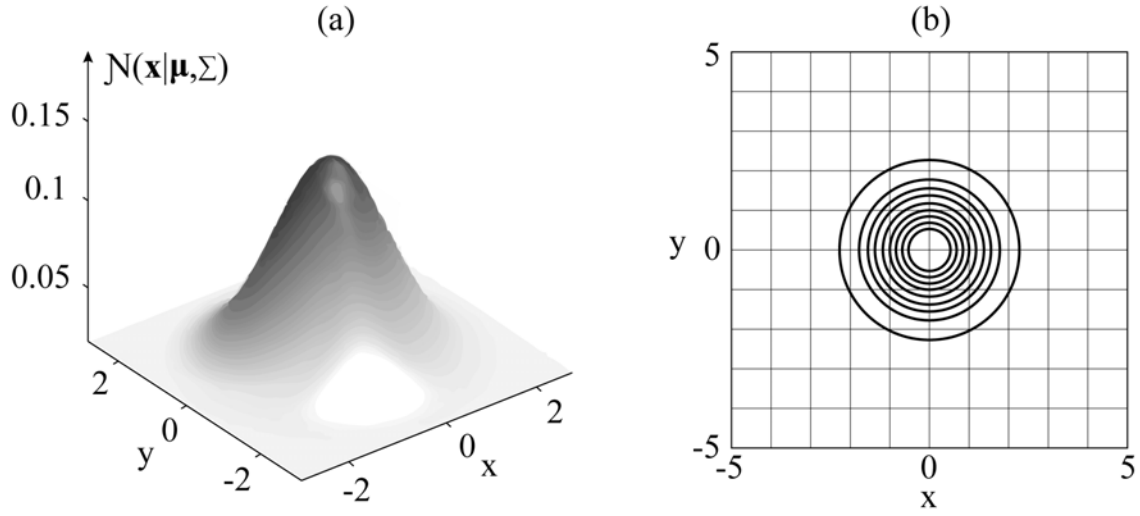


Figura 5.8: Distribución Gaussiana multivariante. (a) Representación en 3D. (b) Representación en 2D.

Al igual que en la distribución Gaussiana univariante, cuando el valor de la media cambia la distribución simplemente es desplazada. La figura 5.9 muestra un par de distribuciones Gaussianas con diferentes valores de media.

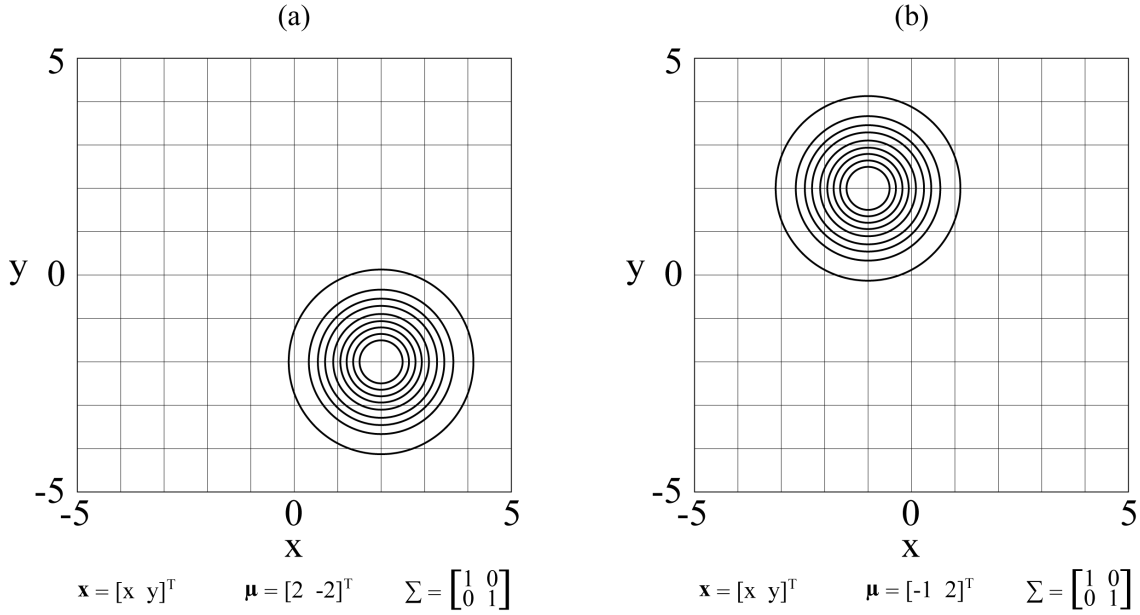


Figura 5.9: Distribución Gaussiana multivariante desplazada. Si $\boldsymbol{\mu}$ tiene elementos cuyo valor es diferente de cero, entonces la distribución puede ser desplazada a lo largo del plano bidimensional xy conservando su forma.

Nuevamente, similar al caso univariante, a medida que el valor de los términos de varianza aumentan, la distribución se extiende a la vez que disminuye el valor del máximo de la distribución. Si por el contrario, el valor de los términos de varianza disminuyen, entonces la distribución se estrecha aumentando el valor del máximo. En la figura 5.10 se ilustra gráficamente este hecho.

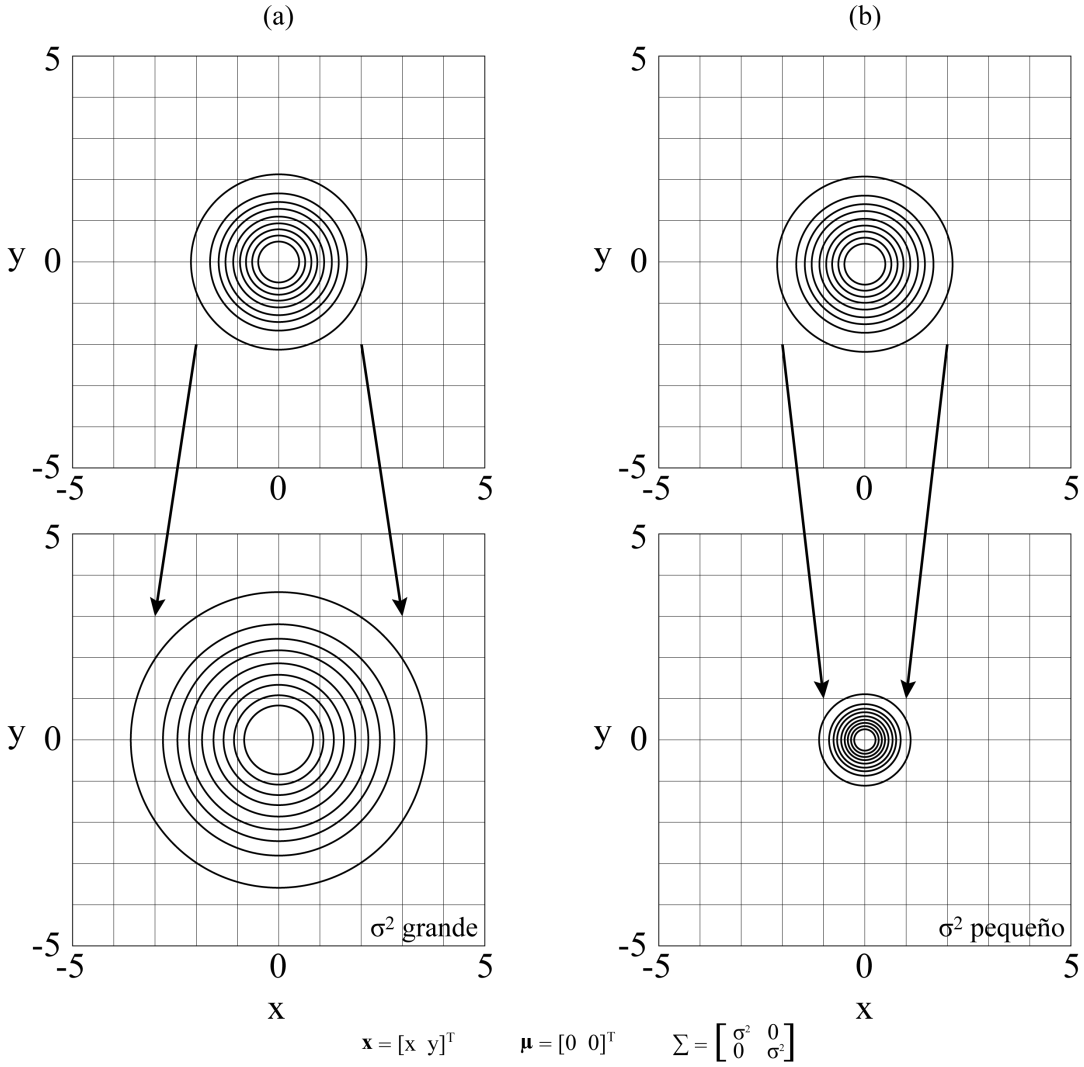


Figura 5.10: Distribución Gaussiana multivariante con diferentes valores de Σ con respecto a una distribución Gaussiana multivariante de varianza unitaria y términos de correlación iguales a cero. (a) Los contornos de la distribución se extienden hacia afuera cuando el valor de las varianzas es grande. (b) Los contornos de la distribución se estrechan cuando el valor de las varianzas es pequeño.

Sin embargo, la matriz de covarianza de la distribución Gaussiana multivariante tiene algunas propiedades que no se ven en el caso univariante. Ya que Σ es una matriz que incluye

términos de correlación en los elementos fuera de la diagonal, si estos términos tienen un valor distinto de cero, entonces la forma de la distribución aparece sesgada, lo cual no ocurre en el caso de una sola variable. La figura 5.11 muestra un par de distribuciones Gaussianas multivariantes sesgadas.

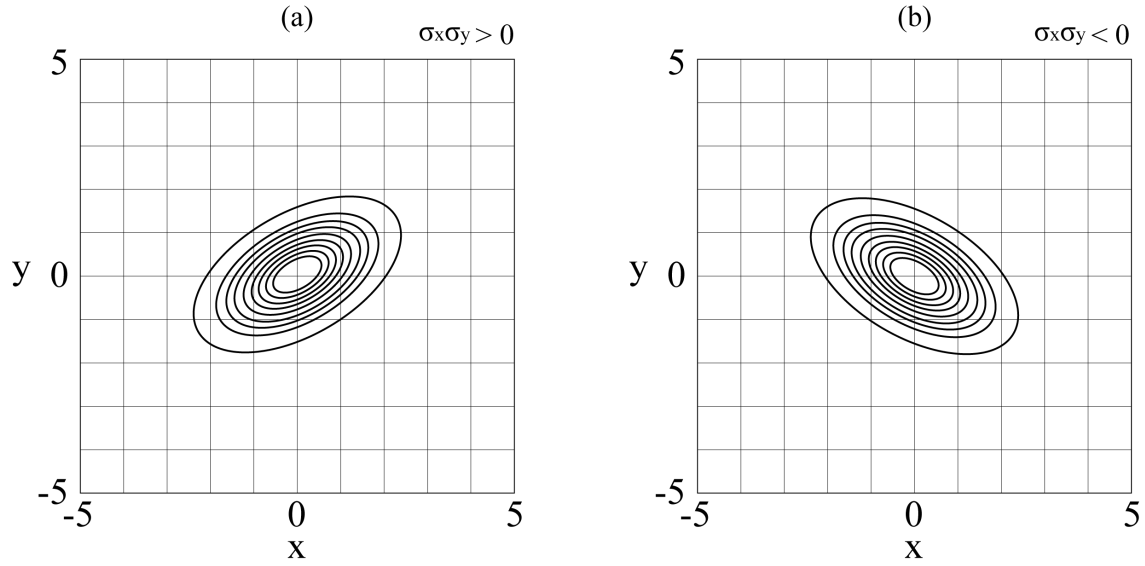


Figura 5.11: Distribución Gaussiana multivariante sesgada. (a) Distribución con términos de correlación mayores a cero. (b) Distribución con términos de correlación menores a cero.

Existen además otras dos propiedades importantes acerca de la matriz de covarianza. La primera es que la matriz de covarianza debe permanecer simétrica y definida positiva, es decir, que los elementos de Σ son simétricos alrededor de la diagonal y los eigenvalores de Σ deben ser positivos. La segunda es que cuando la matriz de covarianza tiene términos de correlación cuyo valor es distinto de cero, es posible encontrar una transformación de coordenadas que permita que la forma parezca simétrica. En este sentido, se puede descomponer la matriz de covarianza para revelar la base de la transformación empleando algoritmos para encontrar eigenvalores.

5.4.2.1. Estimación de Máxima Verosimilitud

Al igual que en el caso univariante, es necesario calcular una estimación de los parámetros del modelo a partir de datos observados. En la distribución Gaussiana multivariante, el interés radica en obtener la media y la matriz de covarianza que maximizan la función de verosimilitud dado un conjunto de observaciones. En este caso, la verosimilitud como función de $\boldsymbol{\mu}$ y Σ se denota como:

$$p(\{\mathbf{x}_i\}|\boldsymbol{\mu}, \Sigma) \tag{5.106}$$

y los parámetros μ y Σ pueden ser estimados como:

$$\hat{\boldsymbol{\mu}}, \hat{\Sigma} = \arg \max_{\boldsymbol{\mu}, \Sigma} p(\{\mathbf{x}_i\}|\boldsymbol{\mu}, \Sigma) \tag{5.107}$$

Asumiendo que los datos x_i son independientes e idénticamente distribuidos, la probabilidad conjunta se expresa como el producto de probabilidades marginales individuales, es decir,

$$\hat{\boldsymbol{\mu}}, \hat{\Sigma} = \arg \max_{\boldsymbol{\mu},\Sigma} \prod_{i=1}^{N} p(\mathbf{x}_i|\boldsymbol{\mu}, \Sigma) \tag{5.108}$$

Del mismo modo que en la distribución Gaussiana de una sola variable, es posible derivar una solución analítica para calcular la estimación de máxima verosimilitud de $\boldsymbol{\mu}$ y Σ. En lugar de maximizar la verosimilitud, se pueden encontrar los parámetros que maximizan la verosimilitud logarítmica, por lo cual,

$$\arg \max_{\mu,\Sigma} \prod_{i=1}^{N} p(\mathbf{x}_i|\boldsymbol{\mu}, \Sigma) = \arg \max_{\mu,\Sigma} \ln \left\{ \prod_{i=1}^{N} p(\mathbf{x}_i|\boldsymbol{\mu}, \Sigma) \right\} \tag{5.109}$$

Sin embargo, ya que el logaritmo de un producto es igual a la suma de los logaritmo de los factores, el lado derecho de la ecuación 5.109 es equivalente a la siguiente expresión:

$$\arg \max_{\mu,\Sigma} \ln \left\{ \prod_{i=1}^{N} p(\mathbf{x}_i|\boldsymbol{\mu}, \Sigma) \right\} = \arg \max_{\mu,\Sigma} \sum_{i=1}^{N} \ln p(\mathbf{x}_i|\boldsymbol{\mu}, \Sigma) \tag{5.110}$$

Como se trata de la distribución Gaussiana Multivariante,

$$p(\mathbf{x}_i|\boldsymbol{\mu}, \Sigma) = \mathcal{N}(\mathbf{x}_i|\boldsymbol{\mu}, \Sigma) = \frac{1}{(2\pi)^{D/2}|\Sigma|^{1/2}} \exp \left\{ -\frac{1}{2}(\mathbf{x}_i - \boldsymbol{\mu})^{\mathrm{T}} \Sigma^{-1} (\mathbf{x}_i - \boldsymbol{\mu}) \right\} \tag{5.111}$$

Sustituyendo la ecuación 5.111 en el lado derecho de la ecuación 5.110, se tiene que:

$$\arg \max_{\mu,\Sigma} \sum_{i=1}^{N} \ln \left\{ \frac{1}{(2\pi)^{D/2}|\Sigma|^{1/2}} \exp \left\{ -\frac{1}{2}(\mathbf{x}_i - \boldsymbol{\mu})^{\mathrm{T}} \Sigma^{-1} (\mathbf{x}_i - \boldsymbol{\mu}) \right\} \right\} \tag{5.112}$$

Utilizando las propiedades de los logaritmos, la ecuación 5.112 se puede reescribir como:

$$\arg \max_{\mu,\Sigma} \sum_{i=1}^{N} \left\{ -\frac{1}{2}(\mathbf{x}_i - \boldsymbol{\mu})^{\mathrm{T}} \Sigma^{-1} (\mathbf{x}_i - \boldsymbol{\mu}) - \frac{1}{2} \ln |\Sigma| - \frac{D}{2} \ln(2\pi) \right\} \tag{5.113}$$

En la ecuación 5.113, el término $\frac{D}{2} \ln(2\pi)$ no varía con los parámetros de interés y puede ser ignorado sin afectar la solución. Por consiguiente, la estimación de los parámetros $\boldsymbol{\mu}$ y Σ se expresa como:

$$\hat{\boldsymbol{\mu}}, \hat{\Sigma} = \arg \max_{\mu,\Sigma} \sum_{i=1}^{N} \left\{ -\frac{1}{2}(\mathbf{x}_i - \boldsymbol{\mu})^{\mathrm{T}} \Sigma^{-1} (\mathbf{x}_i - \boldsymbol{\mu}) - \frac{1}{2} \ln |\Sigma| \right\} \tag{5.114}$$

La ecuación 5.114 puede convertirse en un problema de minimización al cambiar $\max$ por $\min$ y tomar el negativo de los términos, es decir,

$$\hat{\boldsymbol{\mu}}, \hat{\Sigma} = \arg \min_{\mu,\Sigma} \sum_{i=1}^{N} \left\{ \frac{1}{2}(\mathbf{x}_i - \boldsymbol{\mu})^{\mathrm{T}} \Sigma^{-1} (\mathbf{x}_i - \boldsymbol{\mu}) + \frac{1}{2} \ln |\Sigma| \right\} \tag{5.115}$$

Finalmente, los valores de $\boldsymbol{\mu}$ y Σ que maximan la función de máxima verosimilitud se obtienen al tomar derivadas con respecto a la variable deseada y resolviendo la ecuación obtenida.

Sea

$$J(\boldsymbol{\mu}, \Sigma) = \sum_{i=1}^{N} \left\{ \frac{1}{2}(\mathbf{x}_i - \boldsymbol{\mu})^{\mathrm{T}} \Sigma^{-1} (\mathbf{x}_i - \boldsymbol{\mu}) + \frac{1}{2} \ln |\Sigma| \right\} \tag{5.116}$$

Entonces,

$$\hat{\boldsymbol{\mu}}, \hat{\Sigma} = \arg \min_{\mu, \Sigma} J(\boldsymbol{\mu}, \Sigma) \tag{5.117}$$

Si se aplica la condición de optimalidad para optimización convexa, la derivada de primer orden de J con respecto a $\boldsymbol{\mu}$ debe ser cero. El desarrollo matemático se presenta a continuación:

$$\frac{\partial J}{\partial \boldsymbol{\mu}} = \frac{\partial}{\partial \boldsymbol{\mu}} \sum_{i=1}^{N} \left\{ \frac{1}{2}(\mathbf{x}_i - \boldsymbol{\mu})^{\mathrm{T}} \Sigma^{-1} (\mathbf{x}_i - \boldsymbol{\mu}) + \frac{1}{2} |\Sigma| \right\} \tag{5.118}$$

$$= \frac{\partial}{\partial \boldsymbol{\mu}} \sum_{i=1}^{N} \left\{ \frac{1}{2} \boldsymbol{\mu}^{\mathrm{T}} \Sigma^{-1} \boldsymbol{\mu} - \boldsymbol{\mu}^{\mathrm{T}} \Sigma^{-1} \mathbf{x}_i \right\} \tag{5.119}$$

$$= \Sigma^{-1} \sum_{i=1}^{N} \{ \boldsymbol{\mu} - \mathbf{x}_i \} = 0 \tag{5.120}$$

De la ecuación 5.120 se obtiene la estimación de máxima verosimilitud de $\boldsymbol{\mu}$, la cual se denota por $\hat{\boldsymbol{\mu}}$ y se define como:

$$\hat{\boldsymbol{\mu}} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{x}_i \tag{5.121}$$

Del mismo modo, se puede aplicar la condición de optimalidad para calcular la estimación de Σ. Por consiguiente, la derivada de primer orden de J con respecto a Σ debe ser cero. En este caso, el valor de $\hat{\boldsymbol{\mu}}$ es utilizado en lugar de $\boldsymbol{\mu}$ como parámetro. Entonces,

$$\frac{\partial J}{\partial \Sigma} = \frac{\partial}{\partial \Sigma} \sum_{i=1}^{N} \left\{ \frac{1}{2}(\mathbf{x}_i - \hat{\boldsymbol{\mu}})^{\mathrm{T}} \Sigma^{-1} (\mathbf{x}_i - \hat{\boldsymbol{\mu}}) + \frac{1}{2} \ln |\Sigma| \right\} \tag{5.122}$$

$$= \frac{1}{2} \sum_{i=1}^{N} \left\{ -\Sigma^{-1} (\mathbf{x}_i - \hat{\boldsymbol{\mu}})(\mathbf{x}_i - \hat{\boldsymbol{\mu}})^{\mathrm{T}} \Sigma^{-1} + \Sigma^{-1} \right\} \tag{5.123}$$

$$= \frac{1}{2} \Sigma^{-1} \left[- \left\{ \sum_{i=1}^{N} (\mathbf{x}_i - \hat{\boldsymbol{\mu}})(\mathbf{x}_i - \hat{\boldsymbol{\mu}})^{\mathrm{T}} \right\} \Sigma^{-1} + N \cdot \mathrm{I} \right] = 0 \tag{5.124}$$

De la ecuación 5.124 se obtiene la estimación de Σ, la cual se denota por $\hat{\Sigma}$ y se define como:

$$\hat{\Sigma} = \frac{1}{N} \sum_{i=1}^{N} (\mathbf{x}_i - \hat{\boldsymbol{\mu}})(\mathbf{x}_i - \hat{\boldsymbol{\mu}})^{\mathrm{T}} \tag{5.125}$$

Las ecuaciones 5.121 y 5.125 representan la solución de máxima verosimilitud. El parámetro $\hat{\boldsymbol{\mu}}$ representa el vector de medias muestrales y Σ la matriz de covarianza muestral.

5.4.3. Modelo de Mezclas Gaussianas

Aunque la distribución Gaussiana es muy importante y relativamente fácil de manejar, existen algunas limitaciones al utilizar Gaussianas individuales. En el mundo real muchos de los datos adquiridos al analizar fenómenos físicos presentan distribuciones de probabilidad complejas que no pueden ser modeladas adecuadamente por una sola distribución Gaussiana si tienen múltiples modos o si carecen de simetría. Por consiguiente, se necesita un modelo lo suficientemente expresivo para poder ajustarse a cualquier tipo de distribución.

Una forma de abordar el problema anterior, es mediante el uso de un *modelo de mezclas Gaussianas* o *GMM* por sus siglas en inglés, el cual es comúnmente utilizado para modelar datos y para clasificación estadística. Los modelos de mezclas gaussianas (GMMs) son ampliamente conocidos por su habilidad para representar distribuciones arbitrariamente complejas que presentan múltiples modos [91].

En la figura 5.12 se muestran algunos ejemplos de GMMs, donde la curva resultante de la mezcla de múltiples Gaussianas puede tener diversas formas que no permiten definirla matemáticamente mediante una función simple. Sin embargo, si los elementos que conforman la mezcla son elegidos correctamente, cualquier distribución inusual puede ser representada.

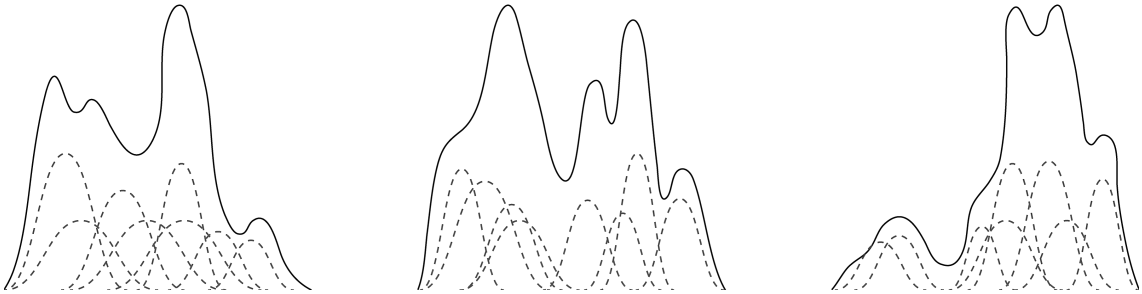


Figura 5.12: Mezclas Gaussianas

En términos generales, un modelo de mezclas Gaussianas es la superposición de varias distribuciones Gaussianas. Dicha superposición se lleva a cabo al tomar combinaciones lineales de distribuciones Gaussianas más básicas, y puede formularse como un modelo probabilístico conocido como *distribución de mezcla* [92]. La combinación lineal de distribuciones Gaussianas puede dar lugar a funciones de densidad muy complejas. Empleando un número lo suficientemente grande de Gaussianas, y ajustando sus medias y covarianzas, así como los coeficientes en la combinación lineal, casi cualquier densidad continua puede ser aproximada con una precisión arbitraria. Matemáticamente, una *mezcla Gaussiana* compuesta por la superposición de K distribuciones Gaussianas se define como:

$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \Sigma_k) \tag{5.126}$$

donde cada densidad Gaussiana $\mathcal{N}(x|\mu_k, \Sigma_k)$ representa a un *componente* de la mezcla Gaussiana con su propia media μ_k y covarianza Σ_k. Los parámetros π_k son los *coeficientes de mezcla* y representan el peso de cada distribución de la mezcla. Si se integran ambos lados de la ecuación

5.126 con respecto a x y se tiene en cuenta que tanto $p(x)$ como los componentes Gaussianos individuales están normalizados, se obtiene que:

$$\sum_{k=1}^{K} \pi_k = 1 \tag{5.127}$$

Además, ya que $p(x) \geq 0$ y $\mathcal{N}(x|\mu_k, \Sigma_k) \geq 0$, entonces $\pi \geq 0$ para todo k. Combinando la condición anterior con la condición de la ecuación 5.127 se obtiene que los coeficientes de mezcla solamente pueden tomar valores en el intervalo $[0, 1]$, es decir,

$$0 \leq \pi_k \leq 1 \tag{5.128}$$

por lo que satisfacen los requerimientos para ser probabilidades.

Una de las propiedades más importantes de la distribución de mezclas Gaussianas, es su propiedad multimodal ($K > 1$ en la ecuación 5.126), en contraste con la propiedad unimodal de la distribución Gaussiana donde $K = 1$. Este hecho hace posible que una mezcla de distribuciones Gaussianas describa adecuadamente a muchos tipos de datos provenientes de fenómenos físicos que exhiben multimodalidad, propiedad que resulta ser representada deficientemente por una sola distribución Gaussiana [91]. La figura 5.13 (b) muestra un GMM de dos componentes.

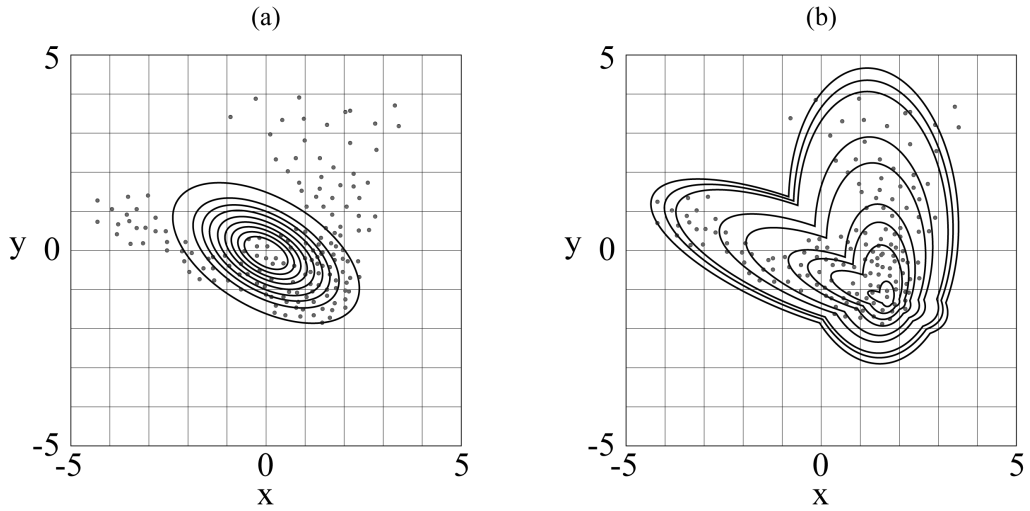


Figura 5.13: Distribución Gaussiana y Modelo de Mezclas Gaussianas de dos dimensiones. (a) Distribución Gaussiana Multivariante bidimensional. (b) Modelo de Mezclas Gaussianas bidimensional (superposición de dos distribuciones Gaussianas).

En contraste con la distribución Gaussiana, un GMM tiene más parámetros, i.e., el número de medias y matrices de covarianza que debe ser especificado, el cual incrementa a medida que el número de mezclas aumenta. Además se tienen los coeficientes de mezcla y el número de componentes Gaussianos que en sí mismo también es un parámetro. Sin embargo, tener más parámetros tiene algunas desventajas. En primer lugar, no existe una solución de máxima verosimilitud para la estimación de los parámetros del modelo, por lo que es necesario utilizar

métodos iterativos de optimización numérica. En segundo lugar, hay más posibilidades de que exista un problema de *sobreajuste* en el modelo.

5.4.3.1. Estimación de Parámetros

Mientras que una sola distribución Gaussiana tiene solamente dos parámetros, un modelo de mezclas Gaussianas tiene múltiples vectores de medias y matrices de covarianza, además de coeficientes que representan el peso de cada distribución, y el número de componentes K en la mezcla. Considerando que los pesos son uniformes ($\pi_k = 1/K$), la estimación de los parámetros radica únicamente en calcular los valores de $\boldsymbol{\mu}$ y Σ que maximizan la función de máxima verosimilitud. El procedimiento matemático para obtener las estimación de estos parámetros es similar al que es llevado a cabo para conocer las estimaciones de máxima verosimilitud de la distribución Gaussiana multivariante dado por las ecuaciones 5.106-5.110, sin embargo, al sustituir el modelo de probabilidad específico de un GMM, resulta que los estimadores son calculados como:

$$\hat{\boldsymbol{\mu}}, \hat{\Sigma} = \arg \max_{\boldsymbol{\mu},\Sigma} \sum_{i=1}^{N} \ln \left\{ \frac{1}{K} \sum_{k=1}^{K} \mathcal{N}(\mathbf{x}_i | \boldsymbol{\mu}_k, \Sigma_k) \right\} \tag{5.129}$$

Sin embargo, la ecuación 5.129 no puede ser simplificada analíticamente debido a la sumatoria de Gaussianas que aparece dentro del logaritmo, pues éste no opera directamente sobre las Gaussianas. Esto implica que los parámetros deben ser estimados mediante cálculos iterativos, aunque cualquier solución encontrada podría no ser una solución óptima global.

5.4.3.2. Algoritmo EM (Expectation-Maximization)

El algoritmo *EM* (*Expectation-Maximization*, por sus siglas en inglés) es un método iterativo de optmización numérica comúnmente utilizado para obtener, bajo ciertas condiciones, la estimación de máxima verosimilitud de los parámetros de un GMM. Para el cálculo se requieren dos cosas: una estimación inicial de $\boldsymbol{\mu}$ y Σ, y una *variable oculta* denotada por z. En este tipo de problemas denominados de *optimización no convexa*, es necesaria una estimación inicial para obtener los parámetros deseados, pues existen muchas soluciones subóptimas llamadas *mínimos locales*, donde la estimación inicial afecta a la solución encontrada.

La variable oculta z esencialmente indica la probabilidad de que el i-ésimo punto sea generado por el k-ésimo componente de la mezcla Gaussiana y se define como:

$$z_k^i = \frac{\mathcal{N}(\mathbf{x}_i | \boldsymbol{\mu}_k, \Sigma_k)}{\sum_{k=1}^{K} \mathcal{N}(\mathbf{x}_i | \boldsymbol{\mu}_k, \Sigma_k)} \tag{5.130}$$

Sea el caso unidimensional donde $K = 2$. En la figura 5.14 se muestran dos distribuciones Gaussianas $g1$ y $g2$, el punto x_i y los valores de x_i valuados en ambas distribuciones denotados como p_1 y p_2. De acuerdo a la ecuación 5.130 los valores de z_1^i y z_2^i se calculan respectivamente como:

$$z_1^i = \frac{p_1}{p_1 + p_2}, \qquad z_2^i = \frac{p_2}{p_1 + p_2}$$

En este caso particular, p_1 es mayor que p_2, de manera que es más probable que x_i sea generado por g_1 que por g_2.

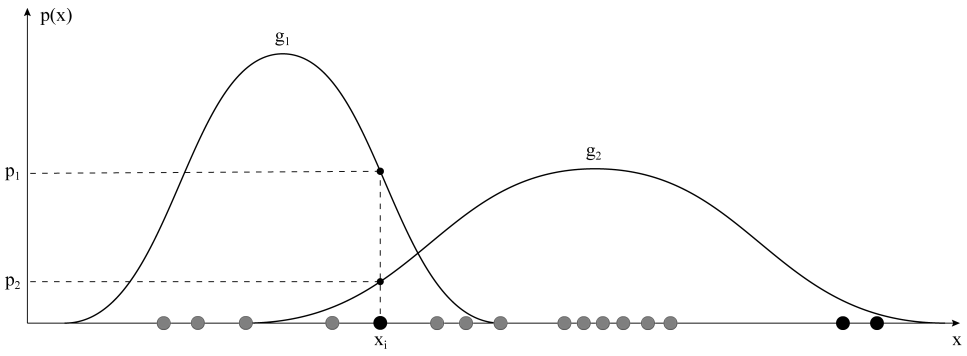


Figura 5.14: Probabilidad de un punto de pertenecer a una distribución. Si p_1 es mayor que p_2, es más probable que el punto x_i haya sido generado por g_1 que por g_2.

Dados los valores de la variable oculta para todos los datos y todos los elementos Gaussianos, es posible redefinir el vector de medias y la matriz de covarianza, ambos ponderados por z como:

$$\hat{\boldsymbol{\mu}}_k = \frac{1}{z_k} \sum_{i=1}^{N} z_k^i \mathbf{x}_i \tag{5.131}$$

$$\hat{\Sigma}_k = \sum_{i=1}^{N} z_k^i (\mathbf{x}_i - \hat{\boldsymbol{\mu}}_k)(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_k)^{\mathrm{T}} \tag{5.132}$$

$$z_k = \sum_{i=1}^{N} z_k^i \tag{5.133}$$

Si la probabilidad de que un punto pertenezca a la k-ésima distribución Gaussiana es pequeña, entonces los puntos contribuyen menos en el cálculo de los parámetros de esa componente Gaussiana en particular.

Los parámetros $\hat{\boldsymbol{\mu}}$ y $\hat{\Sigma}_k$ y las variables ocultas z_k se calculan iterativamente hasta que los valores convergen. Específicamente, los pasos son los siguientes:

1. Se inicializan las medias $\boldsymbol{\mu}$, covarianzas Σ_k y se calcula el valor inicial de la verosimilitud logarítmica (ecuación 5.129).

2. **Paso E (Expectation)**. Se actualiza el valor de las variables ocultas z_k^i con los valores iniciales de $\boldsymbol{\mu}$ y Σ.

3. **Paso M (Maximization)**. Se recalculan los parámetros $\hat{\boldsymbol{\mu}}$ y $\hat{\Sigma}_k$ empleando los valores en turno de las variables ocultas z_k^i.

4. Se evalúa la verosimilitud logarítmica (ecuación 5.129) y se verifica si al igual que alguno de los parámetros, converge. Si el criterio de convergencia no se satisface, entonces se regresa al paso 2.

5.5. Resumen

En este capítulo se ha presentado una breve introducción al aprendizaje automático, cuyo objetivo es desarrollar métodos que puedan detectar automáticamente patrones en un conjunto de datos, y luego utilizarlos para predecir información de interés. Una de las formas más eficientes mediante las cuales una computadora es capaz de aprender es empleando las herramientas de la teoría de la probabilidad, la cual ha sido el pilar de la estadística e ingeniería por siglos. Por esta razón, se han presentado los conceptos básicos más importantes de la teoría de la probabilidad, que en su conjunto pueden aplicarse a cualquier problema de incertidumbre. En el aprendizaje automático, la incertidumbre se presenta de muchas maneras, en cualquier caso es necesario encontrar ya sea la mejor predicción, decisión, o el mejor modelo dado un conjunto de datos. Con fundamento en lo anterior, en este capítulo se discutió ampliamente la distribución Gaussiana, un modelo probabilístico utilizado para la distribución de variables continuas y cuya importancia radica en proporcionar fácilmente una estimación de la incertidumbre de muchos fenómenos físicos. La teoría presentada sobre la distribución Gaussiana, sirvió para introducir el modelo de mezclas Gaussianas o GMM, el cual es de igual manera, un modelo probabilístico que permite representar características de fenómenos físicos. La ventaja más importante que representa este modelo, es que puede ser visto como un método de aprendizaje automático no supervisado que puede ser aplicado en el procesamiento digital de señales de voz debido a la facilidad con la cual se ajusta a los datos de una amplia gama de características del habla basándose en su función de verosimilitud. La estimación de los parámetros del modelo se lleva a cabo mediante una estimación de máxima verosimilitud (ML), la cual es derivada, a diferencia de la distribución Gaussiana, por una técnica de optimización iterativa mediante el uso del algoritmo EM (Expectation-Maximization).

6

Diseño e Implementación del Sistema LID

La *identificación automática del lenguaje hablado* (*LID* por sus siglas en inglés), técnicamente consiste en el proceso automático que realiza una computadora para analizar muestras digitales de voz procedentes de un locutor desconocido e identificar el idioma hablado en dichas muestras. La identificación automática de idiomas plantea muchos problemas de investigación desafiantes, los cuales contribuyen el desarrollo de las tecnologías del lenguaje. En el área de la identificación automática del lenguaje hablado comúnmente se requiere saber qué niveles de información contribuyen de manera más significativa para obtener una identificación de idiomas precisa; cuáles son los mejores parámetros acústicos para representar información específica de cada idioma; qué algoritmos, métodos y enfoques son los más prometedores para implementar un sistema LID; qué recursos específicos del lenguaje son necesarios para modelar las características del habla, y cuál es la correlación entre la duración de la señal de voz a analizar y la eficacia con la cual un sistema LID identifica correctamente un idioma. Algunas de las problemáticas anteriores se derivan naturalmente de la forma en la cual los seres humanos realizan la tarea de identificación de idiomas.

6.1. Sistema de Identificación Automática de Idiomas

Los seres humanos están capacitados para identificar de manera relativamente inmediata los idiomas que no solamente es capaz de hablar, sino también aquéllos que le resultan familiares. Esta facultad de los humanos se ha buscado replicar mediante máquinas por más de medio siglo y en las últimas décadas los avances en la identificación automática del lenguaje hablado han sido notables.

La identificación automática de idiomas ha sido una área de investigación activa por alrededor de 30 años. Los primeros trabajos de investigación en esta área fueron llevados a cabo por R. Leonard y G. Doddington en 1974, quienes adoptaron un enfoque de filtro acústico para la identificación de idiomas [93], y por A. House y E. Neuburg, quienes realizaron la primera contribución a los sistemas LID empleando información específica del lenguaje mediante limitaciones fonotácticas [94]. Diferentes fuentes de información son conocidas por contribuir a la identificación de idiomas, entre las cuales las más importantes son las acústicas, fonéticas, fonémicas y fonotácticas, pero también las características prosódicas, así como las léxicas y morfológicas. Sin embargo, no todas estas características representan la misma importancia para los sistemas LID, ni tampoco tienen la misma facilidad para ser representadas por modelos computacionales. Los enfoques acústico-fonéticos y fonotácticos se benefician por la investigación realizada durante decadas, en principio gracias a la labor de los lingüistas al describir cómo los idiomas se componen de sistemas fonéticos y fonémicos, y más recientemente por científicos cuyo desarrollo en el procesamiento digital de voz ha permitido la creación de modelos computacionales para el reconocimiento automático del habla. Estos factores permitieron que el modelado de características acústico-fonéticas y características fonotácticas se convirtiera en uno de los enfoques más populares para la identificación automática del lenguaje hablado.

Debido al progreso en el desarrollo de los sistemas LID, existe un interés creciente para resolver problemas sutiles concernientes a la identificación automática de idiomas. Los problemas principales involucran la identificación de dialectos y acentos, ya que en un contexto multilingüe las personas no necesariamente se comunican en su lengua materna. En este sentido, la definición del lenguaje hablado es mucho más compleja dado que la señal de voz incluye información acerca del idioma hablado y la lengua materna del hablante. No obstante, este hecho cuestiona si los sistemas LID deberían identificar solamente el idioma hablado en una muestra digital de voz, o si deberían también ser capaces de aportar alguna información adicional sobre el acento del locutor con respecto a su idioma principal.

6.1.1. Características de los Idiomas

La cantidad de idiomas que se hablan en el mundo es muy grande. En [11], el lingüista británico David Crystal estima que existen alrededor de 6,000 idiomas. Sin embargo, otras fuentes [95], [96], [97], estiman que existen entre 4,000 y 8,000 idiomas distintos. El número exacto de idiomas no es conocido y por tanto, muy difícil de establecer. Una de las razones principales por las cuales es complicado estimar el número exacto de idiomas, es el hecho de que el mundo, desde un punto de vista lingüístico, no ha sido estudiado en su totalidad, permitiendo que incluso hoy nuevos idiomas sigan siendo descubiertos. Además, es necesario llegar a un acuerdo para decidir qué es lo que se debería tomar en cuenta; por ejemplo, no es claro si se debe incluir en

el conteo solamente idiomas vivos o lenguas muertas también. Esta distinción es en sí misma problemática pues tampoco se sabe con exactitud cuándo un idioma se considera extinto. En [98], los lingüistas V. Fromkin, R. Rodman y N. Hyams, definen a un idioma extinto cuando éste ya no tiene hablantes nativos, es decir, hablantes que adquirieron el idioma a temprana edad, muy probablemente durante la infancia. Sin embargo, algunos idiomas únicamente tienen un reducido número de hablantes vivos y se enfrentan al hecho de desaparecer una vez que esos hablantes mueran, mientras que otros idiomas, tales como el Latín o el Griego Antiguo, aunque extintos en el sentido de que ya no tienen hablantes nativos, permanecen vivos ya sea a través de la educación o el legado histórico literario. Otra razón por la cual el número total de idiomas es estimado, es debido a la dificultad de trazar una distinción clara entre lenguas genuinamente diferentes y dialectos de la misma lengua. un *dialecto* es comúnmente definido como una variante regional de un idioma que involucra modificaciones en niveles léxicos y gramáticos, contrario a un *acento*, que a pesar de representar también una variante regional, solamente tiene modificaciones en la pronunciación del idioma.

Estas definiciones convencionales son, sin embargo, simplificadas. Primero, muchos idiomas no emplean un sistema de escritura y por consiguiente no tienen una tradición literaria, no obstante son idiomas vivos, algunos de ellos con una rica tradición oral. Segundo, las distinción entre idioma y dialecto no es una división clara, y la clasificación de una variedad particular como un idioma o un dialecto es comúnmente arbitraria y motivada más por consideraciones sociopolíticas que por consideraciones lingüísticas.

Sin embargo, sea cual sea el número exacto de idiomas en el mundo, la gran mayoría de la población mundial habla un conjunto muy limitado de idiomas. De acuerdo a [99], el 95 % de todos los hablantes hace uso solamente del 5 % de los idiomas del mundo. Otro hecho importante es que aproximadamente solamente del 5 al 10 % del total de los idiomas del mundo tienen un sistema de escritura correspondiente [100], lo cual significa que la gran mayoría de idiomas existentes representan un sistema de comunicación verbal. La tabla 6.1 muestra los idiomas más hablados en el mundo de acuerdo al número de hablantes como primera y segunda lengua.

Por otro lado, los inventarios fonéticos de los idiomas, así como sus correspondientes realizaciones fonéticas reflejan características específicas de cada idioma. Los inventarios fonéticos generalmente varían de 20 a 60 símbolos. El idioma Alemán, por ejemplo, posee el doble de síbolos fonéticos que el Español. El tamaño de los inventarios puede incrementar significativamente cuando existe explícitamente información tonal como en el Mandarín o información debida a la duplicación de fonemas como en el caso del Italiano. La distribución de las consonantes y de las vocales, así como la coocurrencia de consonantes y vocales son altamente específicas de cada idioma, y las limitantes fonotácticas son conocidas por ser muy importantes para identificar idiomas. Incluso medidas acústicas como el espectro de la señal de voz exhibe diferencias entre idiomas, lo cual podría estar relacionado con la configuración supralaríngea [102].

6.1.2. Sistemas de Identificación Automática

La identificación automática de idiomas mediante una computadora se puede abordar a través de diferentes enfoques de clasificación. Tradicionalmente, la identificación automática de idiomas ha sido abordada como una tarea de *identificación de conjunto cerrado*, la cual consiste en identificar una entrada de voz como la correspondiente a un idioma perteneciente a

Posición	Idioma	Número de Hablantes
1	Chino Mandarín	1,091 millones
2	Inglés	942 millones
3	Español	518 millones
4	Hindi	380 millones
5	Árabe	352 millones
6	Francés	229 millones
7	Portugués	209 millones
8	Bengalí	208 millones
9	Ruso	201 millones
10	Indonesio	198 millones
11	Urdu	162 millones
12	Alemán	130 millones
13	Japonés	128 millones
14	Panyabí Occidental	90 millones
15	Javanés	84 millones

Tabla 6.1: Idiomas más hablados en el mundo de acuerdo al número de hablantes como primera y segunda lengua [101].

una colección de N número de idiomas conocidos a priori, por lo que se requiere una colección de N modelos dependientes del idioma. Mientras que esta condición es ciertamente de interés científico, este enfoque presenta limitaciones para la mayoría de aplicaciones en la vida real. La representación esquemática de un sistema LID que sigue este enfoque se muestra en la figura 6.1 (a). Sin embargo, la identificación de idiomas también puede visualizarse como una tarea de *detección o verificación de conjunto abierto*. En este caso, el objetivo del sistema es tomar una decisión respecto a si la señal de voz de entrada pertenece o no a un idioma L específico. La señal de voz de entrada puede proceder de un conjunto abierto de idiomas, i.e., que no necesariamente pertenece a alguno de los idiomas modelados. Esta tarea corresponde a un filtrado selectivo de idiomas. Por consiguiente, un sistema de detección de idiomas puede tener como entrada una señal de voz en cualquier idioma e idealmente a la salida es posible obtener un *No* para todos los idiomas excepto para un idioma L específico. En este tipo de sistemas, además de tener un modelo L para el N número de idiomas específicos, existe generalmente un modelo $\bar{L}$ complementario, también denominado como *modelo de referencia universal* (*UBM* por sus

siglas en inglés). Este enfoque, representado en la figura 6.1 (b), resulta más apropiado para las necesidades de las aplicaciones en el mundo real, e.g., el filtrado de idiomas en flujos de audio multilingües. Un enfoque más general y que combina características de los dos enfoques anteriores, corresponde a una tarea de *identificación de conjunto abierto* o *detección multi-objetivo* [103]. Esta tarea consiste en rechazar una entrada de voz si proviene de un idioma desconocido, o de otra manera, en identificar a qué idioma del conjunto de idiomas conocidos pertenece. La identificación de conjunto abierto puede implementarse como un sistema de identificación de conjunto cerrado seguido de un sistema de detección de conjunto abierto o como una serie de K sistemas de detección en paralelo [103].

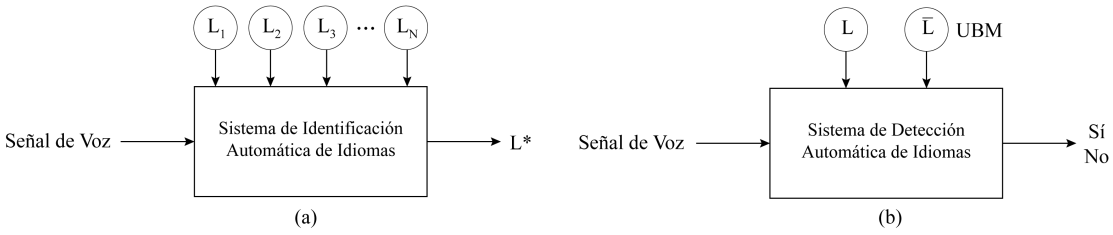


Figura 6.1: Representación esquemática de un sistema LID. (a) Sistema de identificación de idiomas (conjunto cerrado). (b) Sistema de detección de idiomas (conjunto abierto).

6.1.3. Formulación de un Sistema LID

El problema de la identificación automática de idiomas puede ser abordado con la ayuda de una formulación matemática, la cual hace posible descomponer el problema en problemas más simples. Sea X una variable que denota la evidencia acústica (voz) sobre la cual el sistema LID basa su decisión. Sin pérdida de generalidad, se puede hacer la suposición de que X es una secuencia de símbolos provenientes de un alfabeto finito $\mathcal{X}$. Mediante un enfoque estadístico como el presentado en [104], el problema de la identificación automática de idiomas puede enunciarse como:

$$L^* = \arg \max_{L \in \mathcal{L}} P(L|X) \tag{6.1}$$

donde L^* representa el idioma identificado, X es la secuencia de símbolos de la muestra de voz, $\mathcal{L}$ es el conjunto de idiomas potenciales, y $P(L|X)$ es la probabilidad del idioma L dado X. De acuerdo al teorema de Bayes, la ecuación 6.1 puede reformularse como:

$$L^* = \arg \max_{L} P(X|L)P(L) \tag{6.2}$$

donde $P(X|L)$ representa la probabilidad de que la secuencia de símbolos X sea observada cuando el idioma L sea hablado y $P(L)$ es la probabilidad a priori del idioma L. Suponiendo que los diferentes idiomas a identificar son equiprobables, la ecuación 6.2 puede simplificarse como:

$$L^* = \arg \max_{L} P(X|L) \tag{6.3}$$

6.1.4. Modelado del Lenguaje

Los sistemas LID más robustos realizan una extracción de características acústicas muy efectiva y poseen módulos de decisión para la identificación de idiomas. En los sistemas LID de vanguardia [105], [106], [107], [108], el operador $\arg\max_L$ es implementado como un selector del máximo con la ayuda de clasificadores lineales y no lineales. En la figura 6.2 se muestra un diagrama de bloques de los principales componentes de un sistema LID. El *front-end* o *interfaz acústica* tiene como objetivo extraer los vectores de características acústicas más significativas de la señal de voz. El módulo de decisión conocido también como *clasificador back-end*, si se optimiza adecuadamente, contribuye a ganancias de precisión significativas [109].

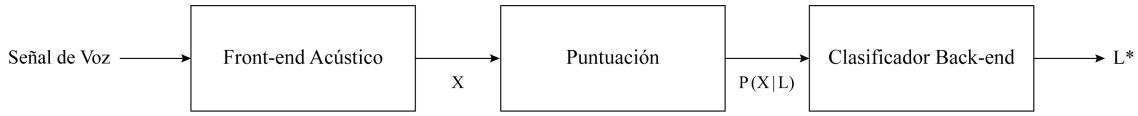


Figura 6.2: Diagrama de bloques de un sistema LID.

Cuando están disponibles únicamente datos de audio para la implementación de un sistema LID, la identificación de idiomas depende exclusivamente de propiedades acústicas. Este enfoque acústico ha sido explorado con resultados exitosos empleando GMMs [110], [111], [112], [113]. Sin embargo, para obtener modelos del lenguaje más complejos se requieren recursos adicionales tales como: etiquetado fonético y segmentación, transcripciones, diccionarios de pronunciación, información morfológica, anotaciones prosódicas, etc. De manera que los diferentes tipos de información específica del lenguaje pueden ser modelados desde niveles puramente acústicos hasta niveles lingüísticos más sofisticados, incluyendo fonemas, información acerca de sistemas vocálicos, fonotáctica, morfología y prosodia [114], [115]. Además, algunos recursos multilingües de datos de audio pueden estar acompañados por transcripciones ortográficas o segmentaciones fonéticas, que no solamente permiten el entrenamiento de los modelos acústicos específicos del lenguaje, sino que también estiman las limitantes fonotácticas de cada idioma. Algunas de las técnicas estándares del reconocimiento del habla pueden ser aplicadas al problema de la identificación automática de idiomas. Emplear niveles léxicos altamente informativos puede resultar en sistemas sumamente complejos, los cuales son potencialmente equivalentes a sistemas multilingües de transcripción. Aunque este enfoque ciertamente garantiza resultados de identificación más confiables, es a expensas de altos costos de desarrollo y operación.

6.1.5. Fases de un Sistema LID

La mayoría de los sistemas LID están conformados por dos fases: una fase de *entrenamiento* y una fase de *reconocimiento* o *identificación*. La fase de entrenamiento consiste en presentar al sistema una serie de ejemplos de señales de voz de una variedad de idiomas para luego ser convertidos en un flujo de vectores de características acústicas. Estos vectores son obtenidos a partir de la segmentación de la señal de voz, e.g. 25 ms, durante los cuales la señal de voz es considerada estacionaria. Los vectores acústicos contienen información espectral o cepstral acerca de la señal de voz. El algoritmo de entrenamiento analiza una secuencia de tales vectores y produce uno o más modelos para cada idioma. La intención de estos modelos es representar las carac-

terísticas fundamentales dependientes del lenguaje de los segmentos de voz de entrenamiento para ser utilizadas durante la siguiente fase del proceso de identificación.

Durante la fase de reconocimiento, se realiza la extracción de características acústicas de un segmento de voz de prueba. Después, la información de los vectores acústicos resultantes son comparados contra cada uno de los modelos generados previamente en el entrenamiento del sistema, y se calcula la probabilidad de que el idioma del extracto de voz coincida con el idioma de los segmentos de voz utilizados para entrenar los modelos. El modelo del idioma más probable de ser correcto es entonces seleccionado [116]. La figura 6.3 muestra el diagrama de bloques de un sistema LID con fases de entrenamiento y reconocimiento.

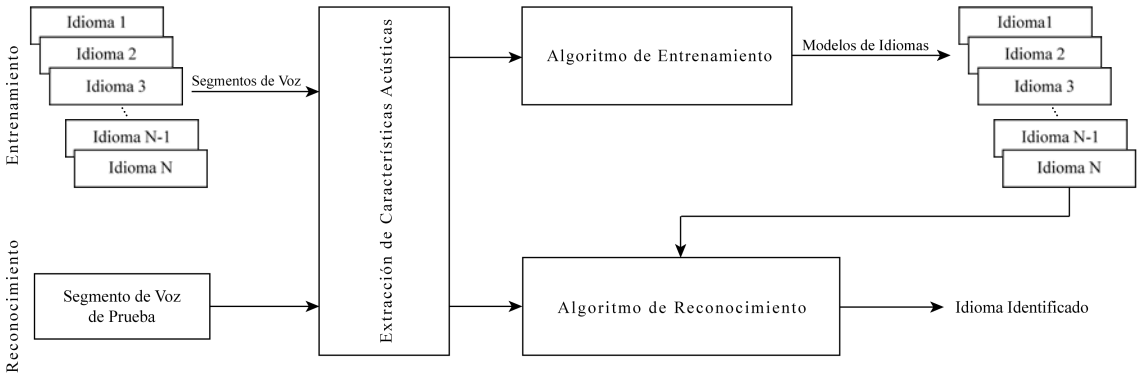


Figura 6.3: Fases de un sistema LID

El sistema LID implementado en este trabajo corresponde a un sistema de conjunto cerrado, por lo que el idioma de todos los fragmentos de voz evaluados por el sistema pertenece exclusivamente a uno de los idiomas modelados en la etapa de entrenamiento del sistema.

6.2. Etapa de Entrenamiento

La fase de entrenamiento tiene como objetivo elaborar modelos de lenguaje λ_N que contengan características propias de cada uno de los idiomas L_N a reconocer durante la tarea de identificación automática. En la figura 6.4 se presenta el diagrama de bloques general del sistema LID con énfasis en los bloques involucrados en la fase de entrenamiento. El primer bloque de esta fase está integrado por una corpora de voz que consta de archivos de audio que contienen voces en cada uno de los idiomas que el sistema tiene como objetivo identificar. Los datos de audio que componen cada corpus de voz son procesados por un bloque de extracción de características, cuya salida genera un flujo de vectores de datos que representan las características acústicas más importantes de las señales de voz presentes en los archivos de audio. Después, los vectores de características acústicas son procesados por un bloque que mediante un algoritmo de entrenamiento, genera un modelo de mezclas Gaussianas para cada uno de los idiomas. A continuación se presenta de manera detallada cada uno de los bloques de la fase de entrenamiento.

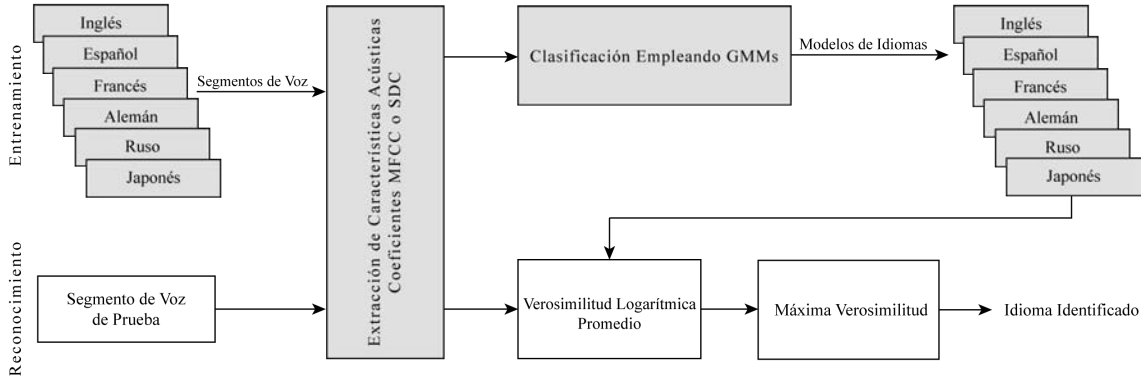


Figura 6.4: Diagrama de bloques del sistema LID con énfasis en los bloques de la fase de entrenamiento.

6.2.1. Corpora de Voz

Un *corpus* de voz es una colección de archivos de audio y sus transcripciones de texto. Al conjunto de varios corpus de voz se le denomina *corpora* de voz. Una corpora de voz es utilizada para construir soluciones para aplicaciones basadas en la voz, una de las más comunes es la creación de modelos acústicos para propósitos de reconocimiento automático del habla.

El enfoque tradicional para construir un corpus de voz consiste en escribir una serie de textos que satisfagan algunos criterios deseados para que los hablantes en turno, mediante el uso de su voz, generen los datos de audio. El proceso de crear contenido para los hablantes, generalmente elegido a partir de un repositorio de muestras textuales que satisface ciertas condiciones, es uno de los pasos más importantes para construir un corpus de voz.

En este trabajo, debido a que el sistema LID a implementar sigue un enfoque puramente acústico, no es necesaria la creación de una corpora de voz con las transcripciones de texto de los archivos de voz, por lo que ésta únicamente se compone de archivos de audio.

La corpora de voz representa una colección de 20 corpus de voz, dos de los cuales constituyen un conjunto de voces masculinas y femeninas, y los dieciocho restantes, están compuestos por una serie de segmentos de voces en seis idiomas distintos. Los idiomas elegidos para la creación de la corpora de voz son: inglés, español, francés, alemán, ruso y japonés.

La base de datos para la creación de cada uno de los corpus de voz constituye una colección de grabaciones de audio de 360 hablantes (180 hombres y 180 mujeres) de diferentes edades y acentos. Las grabaciones de audio fueron extraídas de diferentes fuentes de audio y video en Internet, tales como podcasts de noticias, deportes y entretenimiento, así como audiolibros y conversaciones en películas y entrevistas.

Ya que un sistema LID consiste en dos fases, parte de la información de cada corpus de voz fue utilizada en la fase de entrenamiento y la información restante fue empleada para probar la eficiencia del sistema. La tabla 6.2 muestra la estructura general de la corpora de voz empleada en este trabajo.

# Corpus	Descripción	Contenido	Duración
1	Voces de Hombre	5 hablantes de inglés 5 hablantes de español 5 hablantes de francés 5 hablantes de alemán 5 hablantes de ruso 5 hablantes de japonés	60 min
2	Voces de Mujer		
3	Inglés Voz Hombre	30 hablantes	50 minutos
4	Inglés Voz Mujer		
5	Español Voz Hombre		
6	Español Voz Mujer		
7	Francés Voz Hombre		
8	Francés Voz Mujer		
9	Alemán Voz Hombre		
10	Alemán Voz Mujer		
11	Ruso Voz Hombre		
12	Ruso Voz Mujer		
13	Japonés Voz Hombre		
14	Japonés Voz Mujer		
15	Inglés Mixto	15 hablantes hombres 15 hablantes mujeres	50 minutos
16	Español Mixto		
17	Francés Mixto		
18	Alemán mixto		
19	Ruso Mixto		
20	Japonés Mixto		

Tabla 6.2: Estructura de la corpora de voz. El número total de hablantes es de 360.

6.2.2. Extracción de Características

El proceso que se realiza en el bloque de extracción de características tiene como objetivo transformar la señal de voz de entrada en una secuencia de vectores de características acústicas,

cada uno de los cuales representa la información más importante de la señal en una pequeña ventana de tiempo. La representación de las características acústicas de la señal de voz es llevada a cabo en este trabajo mediante dos tipos de coeficientes: los *coeficientes cepstrales de frecuencia Mel* o *coeficientes MFCC*, los cuales han demostrado ser el mejor enfoque para el diseño de sistemas de identificación automática de idiomas debido a su robustez y eficiencia bajo diversas condiciones; y los *coeficientes cepstrales delta desplazados* o *coeficientes SDC*, los cuales representan una extensión de los coeficientes MFCC y que permiten un alcance temporal mayor para capturar información dinámica de los sonidos de la voz. La figura 6.5 muestra el diagrama de bloques para la obtención de los coeficientes MFCC.

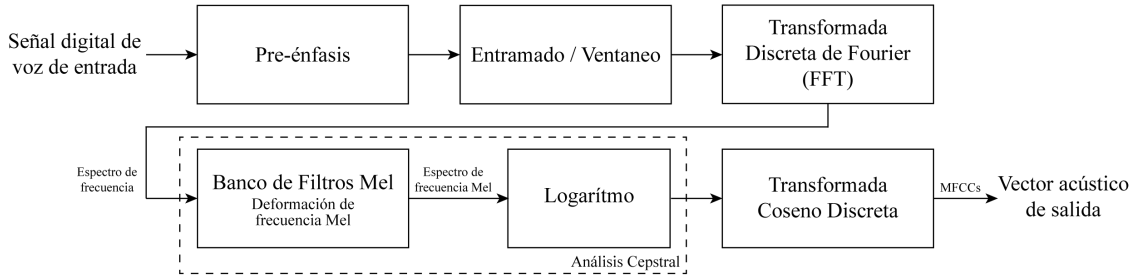


Figura 6.5: Diagrama de bloques para la obtención de coeficientes MFCC.

6.2.2.1. Preénfasis

El primer paso en la extracción de características acústicas consiste en aplicar un filtro a las señales de voz contenidas en los archivos de audio, con la finalidad de compensar la caída de energía espectral que ocurre en las frecuencias altas de la señal de voz causada naturalmente por el pulso glotal y la radiación de los labios. Ya que la intensidad de los sonidos de voz no es lineal con respecto a la frecuencia, ésta disminuye aproximadamente 6 dB por octava. El propósito del *preénfasis* es aumentar la magnitud de las frecuencias más altas, de manera que el espectro tenga un intervalo dinámico similar en toda la banda de frecuencias. Con el fin de incrementar la magnitud de las frecuencias más altas con respecto a la magnitud de las frecuencias más bajas, el preénfasis es llevado a cabo mediante un filtro de respuesta finita al impulso (FIR) paso altas de primer orden, cuya ecuación en diferencias se define como:

$$y(n) = x(n) - \alpha x(n-1) \tag{6.4}$$

donde el parámetro α puede tomar valores comprendidos entre 0 y 1. En esta implementación $\alpha = 0.95$, ya que es el valor más comúnmente utilizado en sistemas de reconocimiento automático de voz, por lo que la ecuación 6.4 se reescribe como:

$$y(n) = x(n) - 0.95x(n-1) \tag{6.5}$$

6.2.2.2. Entramado y Ventaneo

El siguiente paso es particionar la señal en pequeñas tramas de N muestras para mantener constantes las propiedades estadísticas de la señal de voz. Esto permite caracterizar eficiente-

mente los fonemas y subfonemas presentes en la señal. En esta implementación las tramas tienen una duración de 25 ms, cuyo tiempo de desplazamiento es de 15 ms, lo cual significa que las tramas se encuentran superpuestas un $60\,\%$. A la frecuencia de muestreo de 8 KHz el número de muestras contenidas en cada trama es de 200 y el desplazamiento o traslape es igual a 160 muestras. Esto quiere decir que la primera trama de 200 muestras comienza en la primera muestra de la señal y la segunda trama de 200 muestras empieza en la muestra 160 y así sucesivamente con las tramas restantes hasta alcanzar el fin de las muestras totales de la señal. Si el número de muestras presentes en el archivo de voz no se divide en un número par de tramas, se añade un número de muestras cuyo valor es cero al final de la señal para lograr tal división. Después, cada trama es multiplicada por una ventana de Hamming para reducir las discontinuidades abruptas en los bordes y evitar componentes espectrales no deseados.

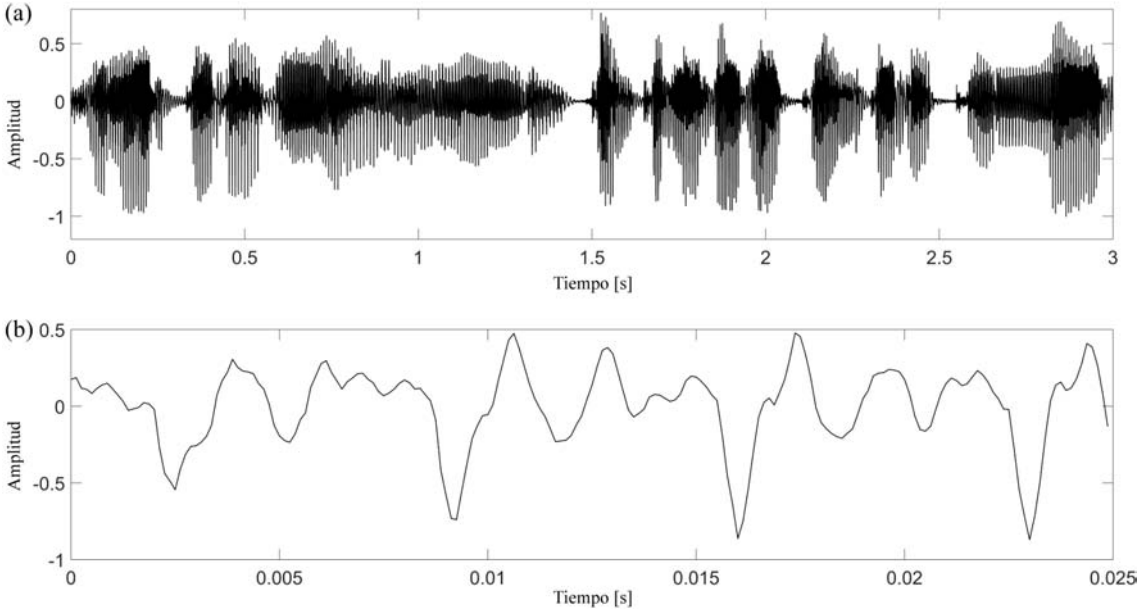


Figura 6.6: Segmentos digitales de voz. (a) Señal de voz de duración igual a 3 s. (b) Trama de voz de duración igual a 25 ms extraída de (a). La forma de onda de la señal extraída se asemeja a una señal cuasi-estacionaria.

6.2.2.3. Cálculo de la DFT

Una vez segmentada la señal de voz en el dominio del tiempo, es necesario transformar cada trama de 200 muestras al dominio de la frecuencia para poder extraer la información espectral importante que está presente en la señal. La herramienta que permite realizar dicha tarea es la DFT, cuyo cálculo se realiza en esta implementación mediante una FFT de 256 puntos y de la cual resultan 256 componentes espectrales. Sin embargo, ya que la transformación es simétrica alrededor de 0 Hz, solamente se conservan 129 puntos.

En la figura 6.6 se muestra la señal de voz contenida en un archivo de audio de 3 segundos y un segmento de 25 ms tomado de dicha señal. Por otra parte, la figura 6.7 muestra una trama

de voz en el dominio del tiempo y su representación en el dominio de la frecuencia, así como las modificaciones que sufre en ambos dominios después del proceso de preénfasis y ventaneo.

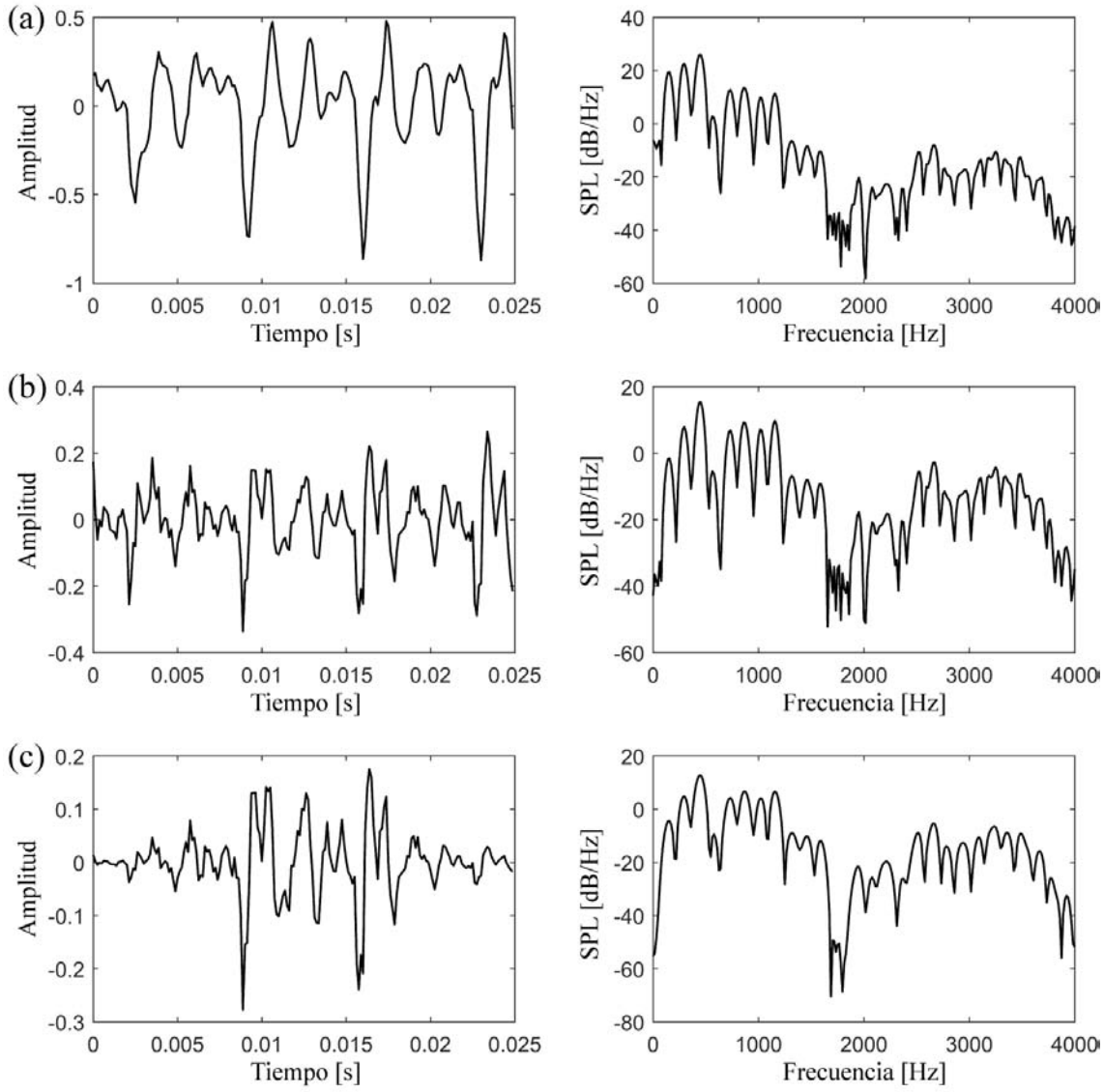


Figura 6.7: Preprocesamiento de una trama de voz. (a) Trama de voz de 25 ms y su espectro calculado mediante una FFT. (b) Trama de voz después del proceso de pre-énfasis. Puede observarse tanto un cambio en la forma de onda en el dominio del tiempo, como una elevación en la magnitud de las frecuencias altas de la señal de voz. (c) Trama de voz después del proceso de ventaneo. Puede observarse un efecto de suavizado en los bordes de la señal y ligeros cambios en el espectro.

6.2.2.4. Cálculo del Banco de Filtros

El espectro obtenido previamente indica la cantidad de energía contenida en cada banda de frecuencias. Sin embargo, ya que la sensibilidad del oído humano varía con la frecuencia y calidad del sonido, es necesario realizar una modificación en el espectro de acuerdo a una escala perceptual. Lo anterior se realiza calculando un banco de filtros espaciados de acuerdo a la escala Mel. El banco de filtros en esta implementación consta de 23 filtros triangulares, cada uno de los cuales es un vector de 129 elementos de acuerdo a las especificaciones de la FFT en el paso anterior. Cada vector tiene en su mayoría elementos cuyo valor es cero, excepto en ciertas regiones del espectro. Para calcular las energías del banco de filtros, cada filtro se multiplica con el espectro y luego se suman los coeficientes. De esta operación resultan 23 números que indican cuánta energía existe en cada uno de los filtros.

El primer paso para obtener los filtros consiste en escoger una frecuencia inferior y una frecuencia superior. Ya que las señales de voz están muestreadas con una frecuencia de muestreo de 8 kHz, la frecuencia superior está limitada a 4 kHz.

Con el uso de la ecuación 5.5 se convierten las frecuencias inferior y superior a Mels. En esta implementación la frecuencia inferior es 32 Hz equivalente a 50.37 Mels, y la frecuencia superior es 4 kHz equivalente a 2,146 Mels. Ya que el banco de filtros consta de 23 filtros, se necesitan 25 puntos, lo cual significa que se requieren adicionalmente de 23 puntos espaciados linealmente entre 50.37 y 2,146 Mels. El valor de dichos puntos se lista en la tabla 6.3. Luego, empleando la ecuación 5.6 los puntos calculados anteriormente se convierten a Hertz. Estos valores se listan en la tabla 6.4. Las frecuencias centrales de los filtros pueden calcularse a partir de la ecuación 5.4, las cuales se listan en la tabla 6.5.

$B(1) = 50.37\,\text{Mels}$	$B(6) = 487\,\text{Mels}$	$B(11) = 923.5\,\text{Mels}$	$B(16) = 1,360.1\,\text{Mels}$	$B(21) = 1,796.7\,\text{Mels}$
$B(2) = 137.7\,\text{Mels}$	$B(7) = 574.3\,\text{Mels}$	$B(12) = 1,010.9\,\text{Mels}$	$B(17) = 1,447.5\,\text{Mels}$	$B(22) = 1,884\,\text{Mels}$
$B(3) = 225\,\text{Mels}$	$B(8) = 661.6\,\text{Mels}$	$B(13) = 1,098.2\,\text{Mels}$	$B(18) = 1,534.8\,\text{Mels}$	$B(23) = 1,971.4\,\text{Mels}$
$B(4) = 312.3\,\text{Mels}$	$B(9) = 748.9\,\text{Mels}$	$B(14) = 1,185.5\,\text{Mels}$	$B(19) = 1,622.1\,\text{Mels}$	$B(24) = 2,058.7\,\text{Mels}$
$B(5) = 399.6\,\text{Mels}$	$B(10) = 836.2\,\text{Mels}$	$B(15) = 1,272.8\,\text{Mels}$	$B(20) = 1,709.4\,\text{Mels}$	$B(25) = 2,146\,\text{Mels}$

Tabla 6.3: Puntos del banco de filtros en escala Mel.

$B^{-1}(1) = 32\,\text{Hz}$	$B^{-1}(6) = 378\,\text{Hz}$	$B^{-1}(11) = 889\,\text{Hz}$	$B^{-1}(16) = 1,640\,\text{Hz}$	$B^{-1}(21) = 2,747\,\text{Hz}$
$B^{-1}(2) = 91\,\text{Hz}$	$B^{-1}(7) = 465\,\text{Hz}$	$B^{-1}(12) = 1,017\,\text{Hz}$	$B^{-1}(17) = 1,829\,\text{Hz}$	$B^{-1}(22) = 3,025\,\text{Hz}$
$B^{-1}(3) = 155\,\text{Hz}$	$B^{-1}(8) = 559\,\text{Hz}$	$B^{-1}(13) = 1,155\,\text{Hz}$	$B^{-1}(18) = 2,032\,\text{Hz}$	$B^{-1}(23) = 3,325\,\text{Hz}$
$B^{-1}(4) = 224\,\text{Hz}$	$B^{-1}(9) = 661\,\text{Hz}$	$B^{-1}(14) = 1,304\,\text{Hz}$	$B^{-1}(19) = 2,253\,\text{Hz}$	$B^{-1}(24) = 3,650\,\text{Hz}$
$B^{-1}(5) = 298\,\text{Hz}$	$B^{-1}(10) = 770\,\text{Hz}$	$B^{-1}(15) = 1,466\,\text{Hz}$	$B^{-1}(20) = 2,490\,\text{Hz}$	$B^{-1}(25) = 4,000\,\text{Hz}$

Tabla 6.4: Puntos del banco de filtros en Hertz.

La resolución en frecuencia requerida para colocar los filtros en los puntos calculados en la tabla 6.4 requiere redondear las frecuencias al intervalo más cercano de la FFT. Este proceso no afecta la eficiencia de las características acústicas. Para convertir las frecuencias a intervalos

$f(1) = 91\,\text{Hz}$	$f(7) = 559\,\text{Hz}$	$f(13) = 1,304\,\text{Hz}$	$f(19) = 2,490\,\text{Hz}$
$f(2) = 155\,\text{Hz}$	$f(8) = 661\,\text{Hz}$	$f(14) = 1,466\,\text{Hz}$	$f(20) = 2,747\,\text{Hz}$
$f(3) = 224\,\text{Hz}$	$f(9) = 770\,\text{Hz}$	$f(15) = 1,640\,\text{Hz}$	$f(21) = 3,025\,\text{Hz}$
$f(4) = 298\,\text{Hz}$	$f(10) = 889\,\text{Hz}$	$f(16) = 1,829\,\text{Hz}$	$f(22) = 3,325\,\text{Hz}$
$f(5) = 378\,\text{Hz}$	$f(11) = 1,017\,\text{Hz}$	$f(17) = 2,032\,\text{Hz}$	$f(23) = 3,650\,\text{Hz}$
$f(6) = 465\,\text{Hz}$	$f(12) = 1,155\,\text{Hz}$	$f(18) = 2,253\,\text{Hz}$	

Tabla 6.5: Frecuencias centrales del banco de filtros.

de la FFT se necesita conocer el tamaño de la FFT, la frecuencia de muestreo y hacer uso de la siguiente ecuación:

$$F(i) = floor((NFFT + 1) \cdot B^{-1}(i)/F_s) \tag{6.6}$$

donde $NFFT$ es el tamaño de la FFT, $B^{-1}(i)$ es el conjunto de puntos del banco de filtros en Hz, y Fs es la frecuencia de muestreo. Con una FFT de 256 puntos y una frecuencia de muestreo, la ecuación 6.6 puede reescribirse como:

$$F(i) = floor((257) * B^{-1}(i)/8000) \tag{6.7}$$

Los resultados de este proceso se listan en la tabla 6.6, donde puede observarse que el último filtro del banco de filtros termina en la sección 128, la cual corresponde a 4 kHz con una FFT de tamaño igual a 256 puntos.

$F(1) = 1$	$F(6) = 12$	$F(11) = 28$	$F(16) = 52$	$F(21) = 88$
$F(2) = 2$	$F(7) = 14$	$F(12) = 32$	$F(17) = 58$	$F(22) = 97$
$F(3) = 4$	$F(8) = 17$	$F(13) = 37$	$F(18) = 65$	$F(23) = 106$
$F(4) = 7$	$F(9) = 21$	$F(14) = 41$	$F(19) = 72$	$F(24) = 117$
$F(5) = 9$	$F(10) = 24$	$F(15) = 47$	$F(20) = 80$	$F(25) = 128$

Tabla 6.6: Puntos de resolución en frecuencia del banco de filtros.

Finalmente, con la información obtenida previamente y el uso de la ecuación 5.1 o alternativamente, con la ecuación 5.2, se crea el banco de filtros. El primer filtro del banco de filtros comienza en el primer punto, alcanza su máximo en el segundo punto y regresa a cero en el tercer punto. El segundo filtro comienza en el segundo punto, alcanza su máximo en el tercer punto y regresa a cero en el cuarto punto, y así sucesivamente con los filtros restantes. La figura 6.8 muestra la gráfica del banco de filtros implementado. Por otro lado, la figura 6.9 muestra varias gráficas de filtros individuales procedentes del banco de filtros y la sección del espectro de una trama de audio sobre la cual actúan.

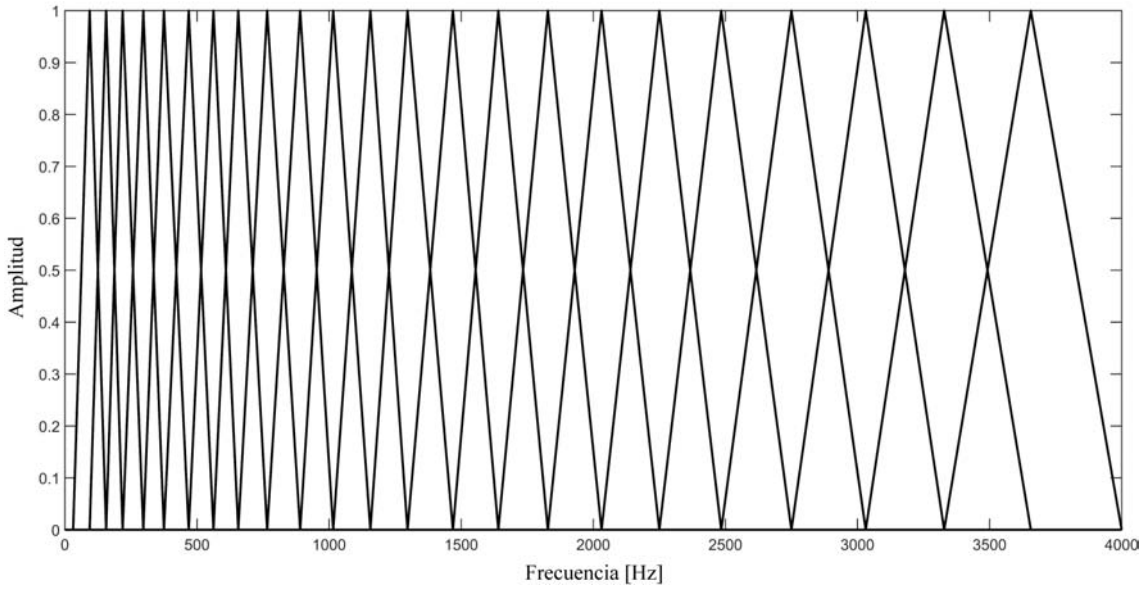


Figura 6.8: Banco de filtros Mel. Cada uno de los filtros triangulares recolecta la energía de un intervalo dado de frecuencias. Los filtros se encuentran espaciados linealmente debajo de 1 kHz y logarítmicamente por encima de 1 kHz.

6.2.2.5. Análisis Cepstral

El último paso en la creación de los coeficientes MFCC consiste en tomar el logaritmo de cada una de las 23 energías y a continuación aplicar la DCT a las 23 energías logarítmicas. Una de las razones principales por las cuales esto se lleva a cabo es porque la superposición de los componentes del banco de filtros ocasiona que las energías se correlacionen unas con otras. La DCT se encarga de decorrelacionar los coeficientes de energía, lo cual permite realizar un modelo acústico mucho más simple. Sin embargo, ya que los coeficientes más altos de la DCT representan cambios rápidos en las energías del banco de filtros y estos cambios rápidos degradan el rendimiento de los sistemas de reconocimiento de voz, se conservan únicamente los coeficientes más bajos. En esta implementación solamente se emplean los primeros 12 coeficientes y los restantes son descartados.

6.2.2.6. Coeficientes Cepstrales Delta Desplazados

La extracción de características acústicas en los sistemas LID se lleva a cabo típicamente construyendo vectores de características constituidos por únicamente coeficientes cepstrales, cuya dimensión varía típicamente entre 10 y 15 coeficientes, o una combinación de éstos y coeficientes cepstrales delta o doble delta, cuya dimensión de los vectores resultantes varía entre 20 y 30 parámetros. El uso de derivadas de primer y segundo orden permite una mejor representación de las propiedades dinámicas de los sonidos del habla, ya que toma en cuenta información dinámica de intervalos de tiempo entre 50 y 80 ms. Se han realizado intentos por derivar coeficientes que logren capturar de una mejor manera la información dinámica contenida en las

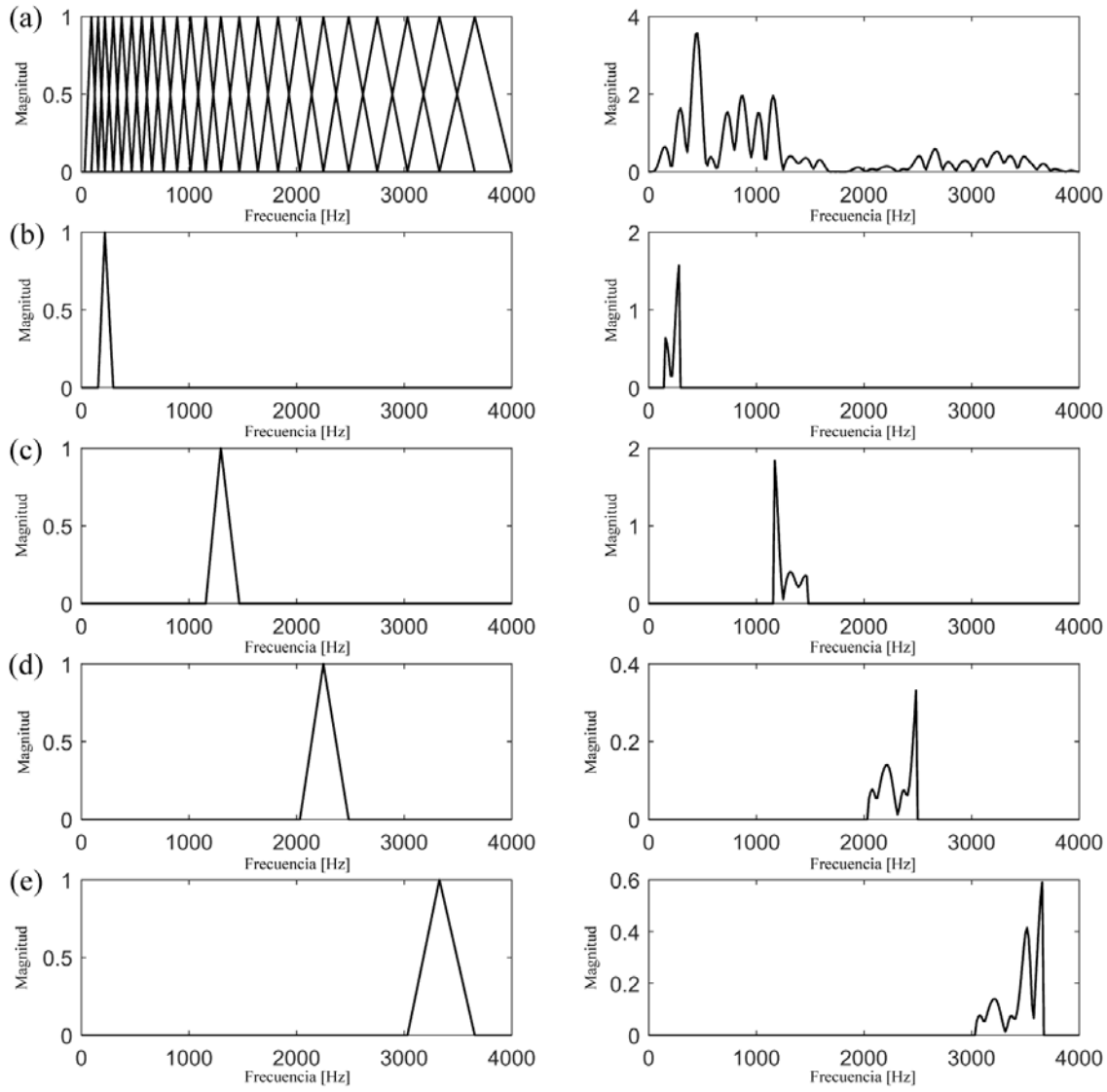


Figura 6.9: Banco de filtros Mel y espectro ventaneado.(a) Banco de filtros completo y espectro de potencia una trama de audio. (b) Filtro 3 del banco de filtros y espectro de potencia ventaneado empleando el filtro 3.(c) Filtro 13 del banco de filtros y espectro de potencia ventaneado empleando el filtro 13.(d) Filtro 18 del banco de filtros y espectro de potencia ventaneado empleando el filtro 18.(e) Filtro 22 del banco de filtros y espectro de potencia ventaneado empleando el filtro 22.

señales de voz [117]. Sin embargo, se ha demostrado en varias investigaciones [82], [83], que el uso de coeficientes SDC mejora notablemente el rendimiento de los sistemas LID. Los coeficientes SDC son una extensión de los coeficientes cepstrales delta: consisten en vectores delta apilados, típicamente de 7 deltas consecutivas. El alcance temporal de dichos vectores es de

aproximadamente 250 ms, lo cual permite incluir en los coeficientes información de al menos una unidad silábica. Estas características de largo alcance permiten tomar en cuenta un modelado implícito de unidades lingüísticas y además resultan ser, no solamente específicas de cada idioma, sino también específicas de cada corpus, siendo las sílabas más frecuentes, las mejores representadas. Por consiguiente, los coeficientes SDC son muy efectivos para capturar información sobre la similitud y las transiciones entre un fonema y otro, lo cual es específico de cada idioma y facilita la discriminación entre los mismos.

Como se mencionó en el capítulo anterior, los coeficientes SDC constan de cuatro parámetros conocidos como $N - d - P - k$. En este trabajo, de acuerdo a los buenos resultados obtenidos en [83], se emplea la siguiente configuración de parámetros:

$$N - d - P - k = 7 - 1 - 3 - 7$$

Estos coeficientes, tal como lo muestra la figura 6.10 son calculados a partir de los coeficientes MFCC. Los vectores resultantes representan coeficientes de 84 dimensiones.

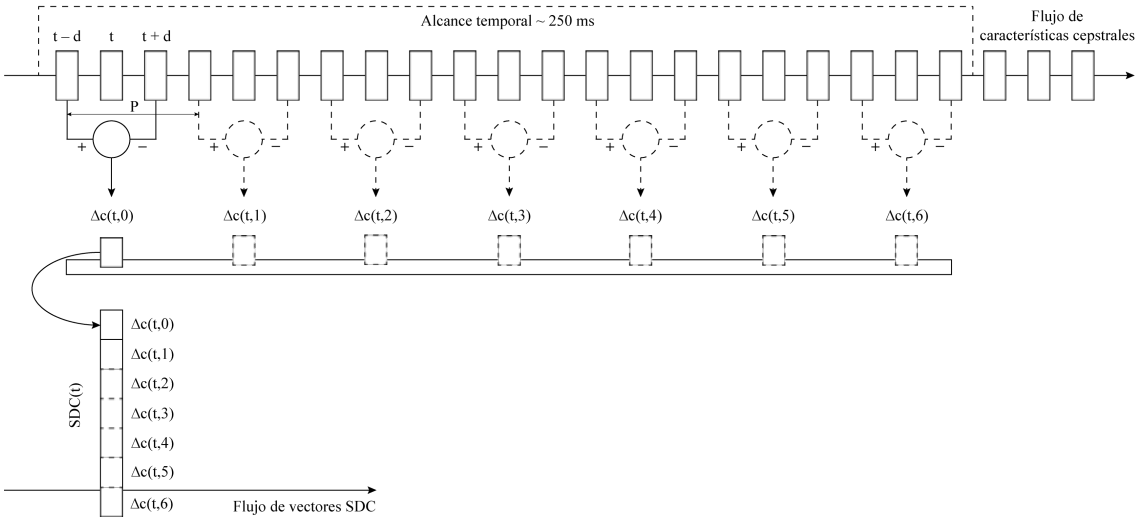


Figura 6.10: Diagrama de la obtención de los coeficientes SDC en el instante t empleando la configuración de parámetros $N - d - P - k = 7 - 1 - 3 - 7$.

6.2.3. Modelado de idiomas

Los sistemas de identificación automática de idiomas que siguen un enfoque puramente acústico solamente requieren de datos de audio específicos de cada idioma considerado para el reconocimiento automático. Esto representa una ventaja ya que no es necesario tener conocimiento específico del idioma, como puede ser una transcripción lingüística. Por consiguiente, el desarrollo de un sistema LID y la extensión de idiomas adicionales es directa. En un inicio, los sistemas LID acústicos utilizaron bancos de filtros y coeficientes LPC, así como también enfoques basados en cuantización vectorial. Actualmente, los modelos de mezclas Gaussianas y las

máquinas de soporte vectorial (SVM, por sus siglas en inglés) son los enfoques más empleados para el modelado acústico.

6.2.3.1. Clasificación Empleando Modelos de Mezclas Gaussianas

En esta etapa se lleva a cabo el entrenamiento de las distribuciones de mezclas Gaussianas. Como se mencionó en el capítulo 5, un modelo de mezclas Gaussianas (GMM) es una suma ponderada de varias componentes Gaussianas, matemáticamente expresado por la ecuación 5.126 como:

$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \Sigma_k)$$

donde K representa el número de componentes, π_k son los pesos para cada componente y cada componente es una distribución Gaussiana con vector de medias μ_k y matriz de covarianza Σ_k.

Ya que los datos se encuentran en un espacio multidimensional de 12 dimensiones en el caso de los coeficientes MFCC o de 84 dimensiones en el caso de los coeficientes SDC, no se sabe cómo es la distribución de los datos, ni cuántos picos (modos) tiene. Por lo tanto, el entrenamiento de las mezclas gaussianas para crear los modelos específicos de cada lenguaje consiste en comenzar con una sola componente Gaussiana para cada uno de los modelos. Los parámetros de los modelos son estimados, y después de eso, cada una de las componentes Gaussianas es dividida en dos y el entrenamiento se repite estimando nuevamente los parámetros de los modelos. La división de las componentes Gaussianas y el entrenamiento para obtener los nuevos parámetros se repite hasta alcanzar el número final de componentes deseadas.

En la figura 6.11 se muestra la distribución de las características acústicas de cada uno de los idiomas elegidos para la implementación del sistema LID. Aunque la representación toma en cuenta únicamente 3 de las 84 dimensiones de los coeficientes SDC, permite visualizar con claridad la forma en que se agrupan los datos en un espacio tridimensional.

Sin embargo, no es posible encontrar la solución para el entrenamiento de un modelo de mezclas Gaussianas con K componentes estimando los parámetros, μ_k y Σ_k con ecuaciones de forma cerrada. Lo que se requiere es hacer uso del algoritmo EM para encontrar los mejores parámetros. La figura 6.12 muestra el diagrama de bloques del algoritmo EM para la estimación de los parámetros de un GMM.

Los modelos de mezclas Gaussianas representan el enfoque acústico más popular para la identificación automática de idiomas, particularmente para la tarea de detección. En esta implementación un GMM es estimado para cada idioma y el único conocimiento previo para entrenar los modelos específicos de cada idioma consiste en la identidad del idioma del corpus de voz. Típicamente el número de componentes en un GMM varía entre 64 y 1024 componentes. En esta implementación se emplearon de 128 a 2048 componentes con pesos uniformes $\pi_k = 1/K$ para la creación de los modelos de idiomas. Como regla general, los modelos de mezclas Gaussianas no tienden a capturar dependencias temporales de una manera eficiente, de ahí la importancia de la introducción de los coeficientes SDC para captura información dinámica de las señales.

La figura 6.13 muestra una representación en dos dimensiones de los modelos de mezclas Gaussianas de cada uno de los idiomas empleados en la implementación del sistema LID. Por otra parte la figura 6.14 muestra las gráficas de los GMMs en un espacio de tres dimensiones.

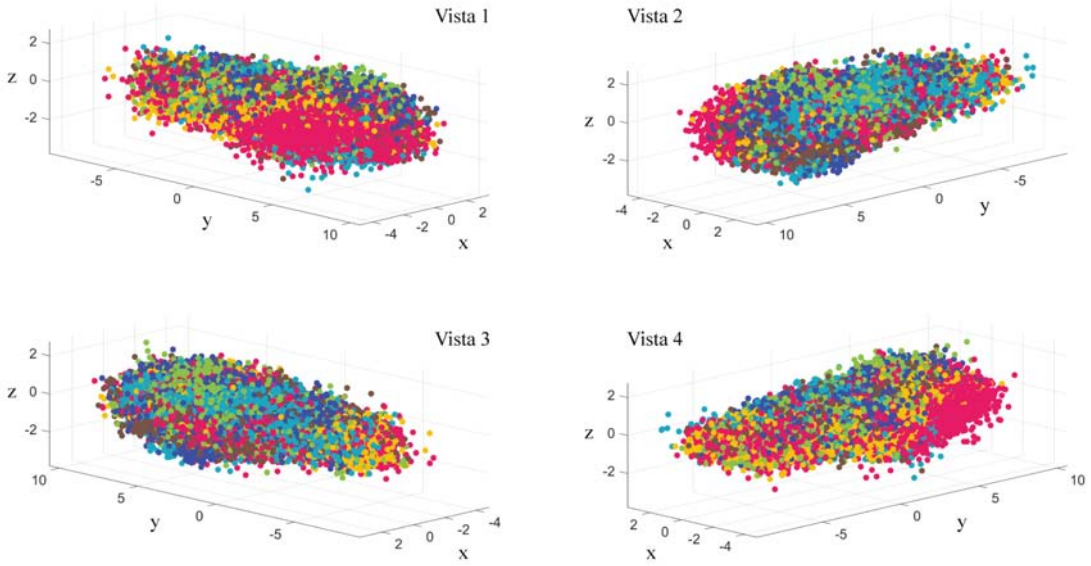


Figura 6.11: Distribución tridimensional de las características acústicas de los idiomas elegidos desde cuatro ángulos distintos. Pueden observarse las zonas donde se agrupan los datos de cada idioma. ●Inglés, ●Español, ●Francés, ●Alemán, ●Ruso, ●Japonés.

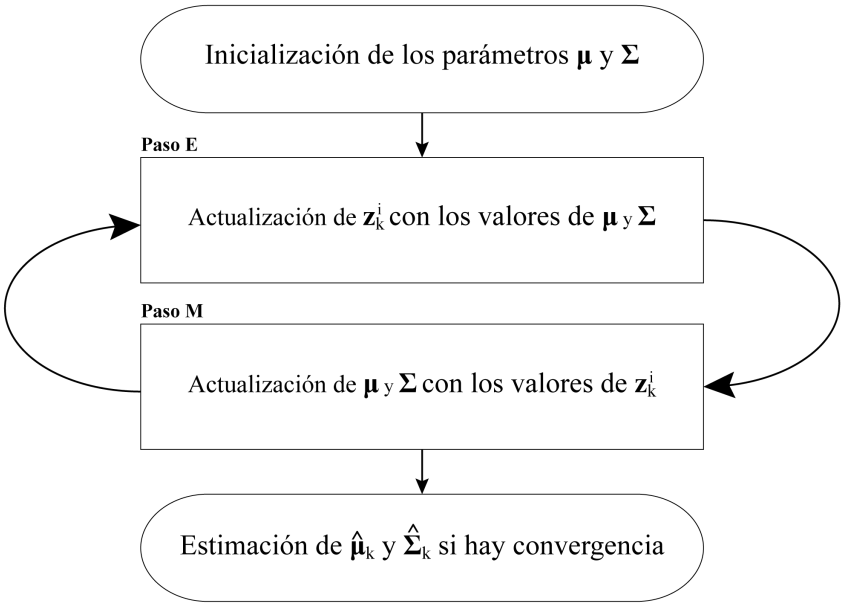


Figura 6.12: Diagrama del algoritmo EM para la estimación de los parámetros de un GMM.

En ambos casos los modelos de los idiomas son muy similares debido a que tres dimensiones es insuficiente para visualizar diferencias más notables de la distribución probabilística de los datos particulares de cada idioma, sin embargo, aún es posible notar similitudes entre cada uno de los modelos.

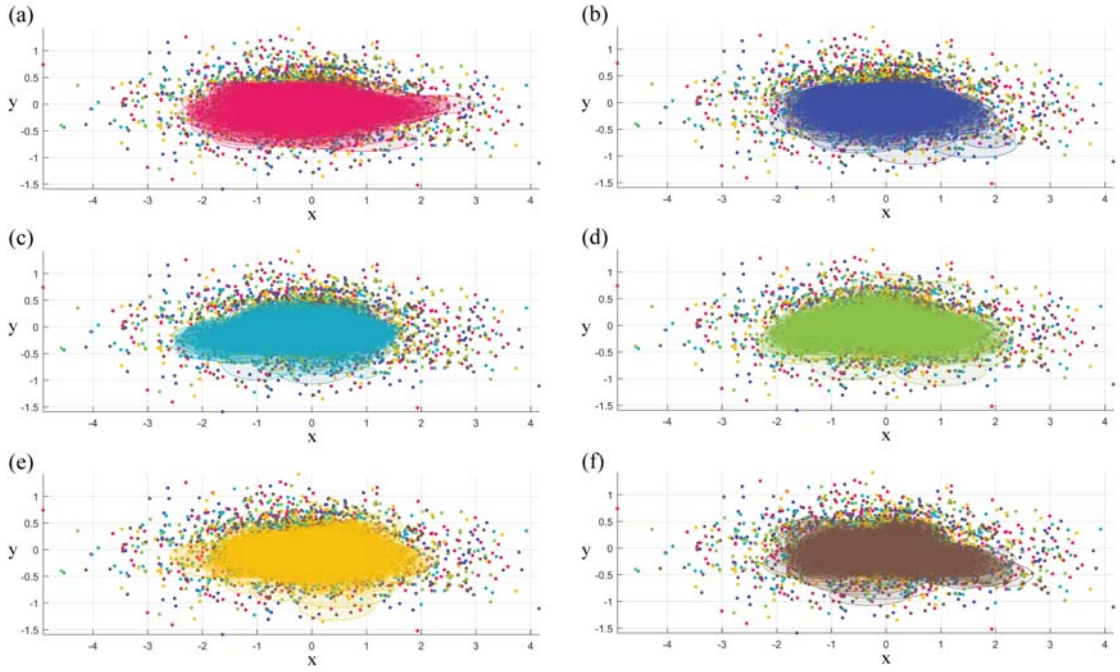


Figura 6.13: Representación bidimensional de los modelos de mezclas Gaussianas para cada idioma. Cada GMM consta de 1024 componentes Gaussianas. (a) GMM del idioma Inglés. (b) GMM del idioma Español. (c) GMM del idioma Francés. (d) GMM del idioma Alemán. (e) GMM del idioma Ruso. (f) GMM del idioma Japonés.

6.3. Etapa de Reconocimiento

La fase de reconocimiento tiene como objetivo presentar al sistema LID una señal de voz sin otorgar referencia alguna sobre el idioma que habla el locutor. En la figura 6.15 se presenta el diagrama de bloques general del sistema LID con énfasis en los bloques involucrados en la fase de reconocimiento. El primer paso hacia la identificación del idioma hablado en el segmento de voz, es la extracción de las características acústicas presentes en la señal. Este proceso genera un vector acústico de coeficientes MFCC o SDC, el cual pasa a un bloque que mediante algoritmos de reconocimiento realiza la comparación de éste con cada uno de modelos de lenguaje generados en la fase de entrenamiento. El modelo λ_N con el valor de probabilidad más alto de haber producido el vector de características corresponde al idioma identificado.

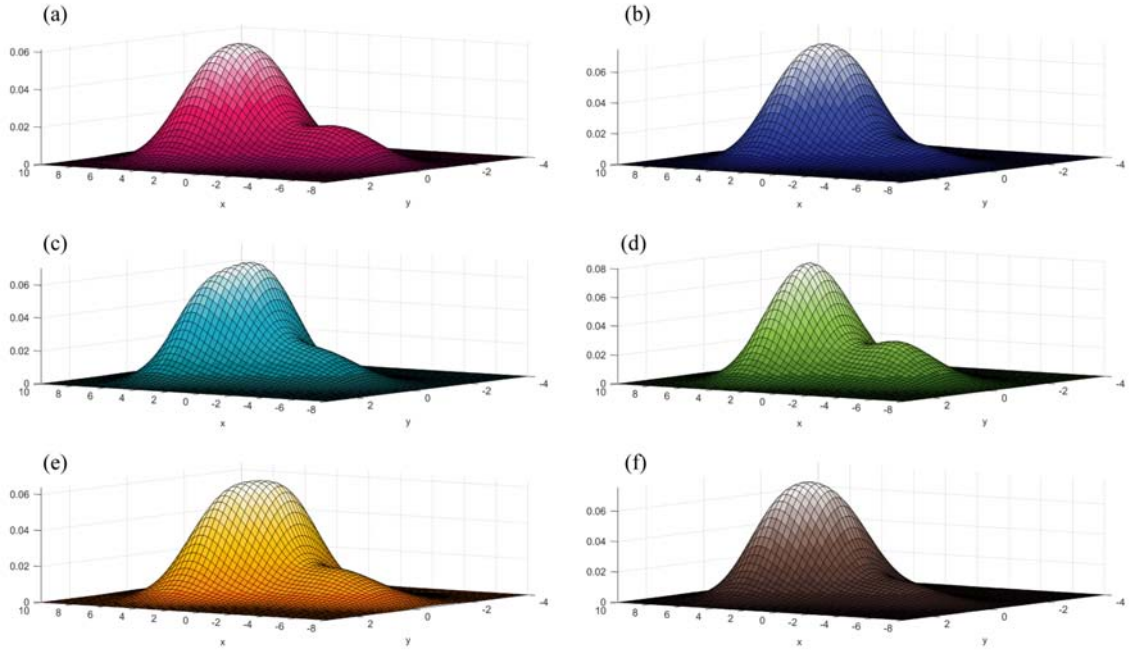


Figura 6.14: Representación tridimensional de los modelos de mezclas Gaussianas para cada idioma. Cada GMM consta de 1024 componentes Gaussianas. (a) GMM del idioma Inglés. (b) GMM del idioma Español. (c) GMM del idioma Francés. (d) GMM del idioma Alemán. (e) GMM del idioma Ruso. (f) GMM del idioma Japonés.

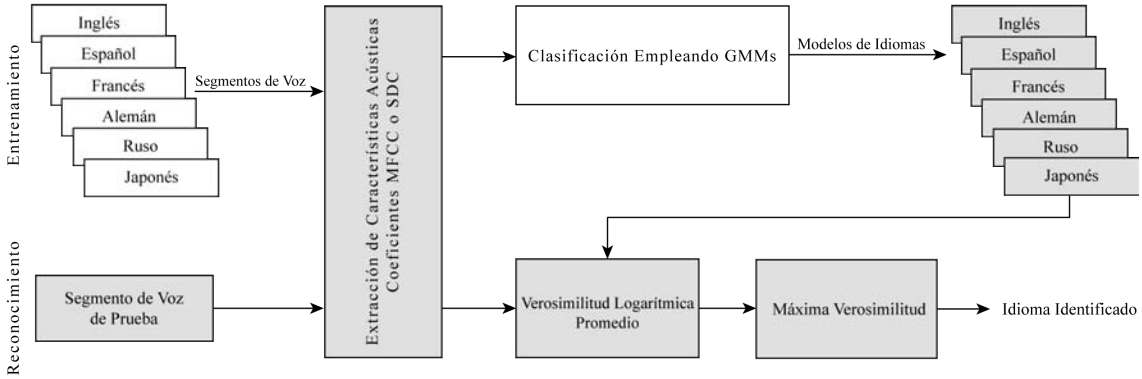


Figura 6.15: Diagrama de bloques del sistema LID con énfasis en los bloques de la fase de reconocimiento.

6.3.1. Identificación de idiomas

Durante la fase de reconocimiento, un segmento de voz de prueba cuyo idioma es desconocido es clasificado convirtiendo primero la forma de onda digital en un vector de características X, el cual está constituido por un conjunto de elementos $x_1, x_2, x_3, ..., x_T$, y después calculando la

verosimilitud logarítmica promedio que cada uno de los modelos produce al evaluar el segmento de voz de prueba.

La verosimilitud logarítmica promedio L se define como:

$$p(X|\lambda_N) = \frac{1}{T}\sum_{t=1}^{T}\log p(x_t|\lambda_N) \tag{6.8}$$

donde λ_N es el modelo GMM correspondiente a cada uno de los idiomas a identificar. Implícitamente, en la ecuación 6.8 se encuentra la suposición de que el conjunto de elementos x_t son estadísticamente independientes entre sí. Finalmente, el bloque clasificador de máxima verosimilitud hipotetiza uno de los idiomas procedente del conjunto de idiomas como aquél propio del segmento de voz de prueba. Matemáticamente, lo anterior se expresa como la ecuación 6.3:

$$L^* = \arg\max_{L} P(X|L) \tag{6.9}$$

donde L^* corresponde al idioma hipotetizado de la señal de voz de prueba y P(X|L) representa la probabilidad de la secuencia X dado el idioma L procedente del conjunto total de idiomas.

6.4. Resumen

En este capítulo se describieron los pasos necesarios fundamentados en la teoría presentada en los capítulos 2, 3, 4 y 5, para la implementación de un sistema LID de conjunto cerrado que sigue un enfoque puramente acústico. Se presentaron las dos fases de un sistema LID: la *fase de entrenamiento* y la *fase de reconocimiento*. La primera fase tiene como objetivo la creación de modelos que puedan representar de una manera eficiente las características acústicas más importantes de los idiomas elegidos para la identificación. Esto se logra mediante un par de bloques, uno de los cuales se encarga de la extracción de las características acústicas generando vectores de coeficientes MFCC y SDC a partir de las señales de voz procedentes de un corpus específico, y el otro se encarga de modelar la información acústica dinámica presente en los coeficientes empleando modelos de mezclas Gaussianas. Por otra parte, la fase de evaluación tiene como objetivo, dada una señal de voz de entrada al sistema LID, identificar a qué idioma pertenece dentro del conjunto de idiomas modelados. Para llevar a cabo esta tarea, al igual que en la fase de entrenamiento, primero se extraen las características acústicas presentes en la señal de voz. Luego, el vector acústico generado es procesado por un bloque que evalúa sus propiedades estadísticas con la de los modelos de cada idioma. Finalmente, el modelo con el valor de probabilidad más alto de haber producido el vector acústico es aquél que corresponde al idioma identificado.

7

Pruebas y Resultados

En este capítulo se presentan dos configuraciones para el sistema de identificación automática de idiomas, cada una de las cuales emplea elementos descritos en el capítulo anterior. La primera configuración consiste en un sistema LID dependiente del género del locutor y la segunda configuración en un sistema LID independiente del género del locutor. Los motivos principales para implementar dos configuraciones de sistemas automáticos de idiomas es, en primer lugar, la generación de múltiples modelos de distinto número de componentes Gaussianas a partir de cada conjunto de vectores de características acústicas procedentes de los diferentes corpus que integran la corpora de voz. En segundo lugar, realizar una comparación de la eficiencia de todos los modelos individuales generados, de manera que los modelos con las eficiencias más altas sean seleccionados para integrar los modelos de cada una de las configuraciones de los sistemas LID y finalmente, poder realizar una comparación del rendimiento de ambas configuraciones.

7.1. Sistema LID Dependiente del Género del Locutor

La primera configuración del sistema LID consiste en un sistema dependiente del género del locutor. Esta configuración cuenta con los mismos bloques de un sistema LID de enfoque acústico descritos en el capítulo anterior, pero adicionalmente presenta un bloque que permite llevar a cabo la identificación del género del hablante de forma automática previo a la identificación del idioma que éste habla. En la figura 7.1 se muestra el diagrama de bloques del sistema de identificación automática de idiomas dependiente del género del locutor.

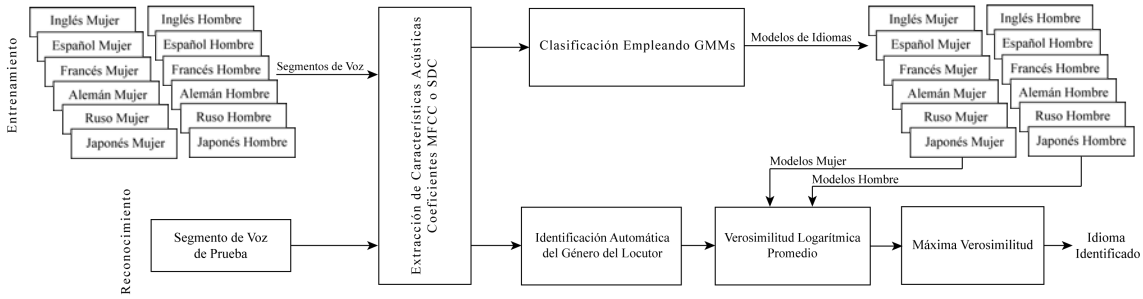


Figura 7.1: Diagrama del sistema LID dependiente del género del locutor.

7.1.1. Subsistema de Identificación Automática del Género del Locutor

El bloque de identificación del género del locutor representa en sí mismo un subsistema dentro del sistema LID. Éste adopta un enfoque puramente acústico para detectar automáticamente el género del hablante, por lo que los procesos que el sistema realiza para dicha tarea siguen el mismo flujo que un sistema LID acústico.

La figura 7.2 muestra el diagrama de bloques del sistema de identificación automática del género del locutor. Al igual que un sistema LID, éste se conforma por una fase de entrenamiento y otra de reconocimiento. El primer paso hacia la identificación del género del hablante consiste en extraer las características acústicas de los archivos de audio de la corpora de voz mediante la generación de coeficientes MFCC. El segundo paso consiste en el entrenamiento de GMMs, por lo que los vectores acústicos generados en el paso anterior son clasificados empleando modelos de mezclas Gaussianas. De la fase de entrenamiento resultan un par de modelos acústicos: uno constituido por las características de la voz masculina y otro integrado por las características de la voz femenina.

Durante la fase de reconocimiento los parámetros estadísticos de los vectores de características son evaluados en cada uno de los modelos acústicos, donde el modelo con mayor probabilidad de generar el segmento de voz representa el género hipotetizado del hablante.

7.1.2. Generación de los Modelos Acústicos

Como se mencionó en el capítulo anterior, el número de componentes Gaussianas de un GMM típicamente varía entre 64 y 1024 componentes. Un mayor número de componentes Gaussianas presentes en un GMM no necesariamente genera un mejor modelo, ya que puede existir

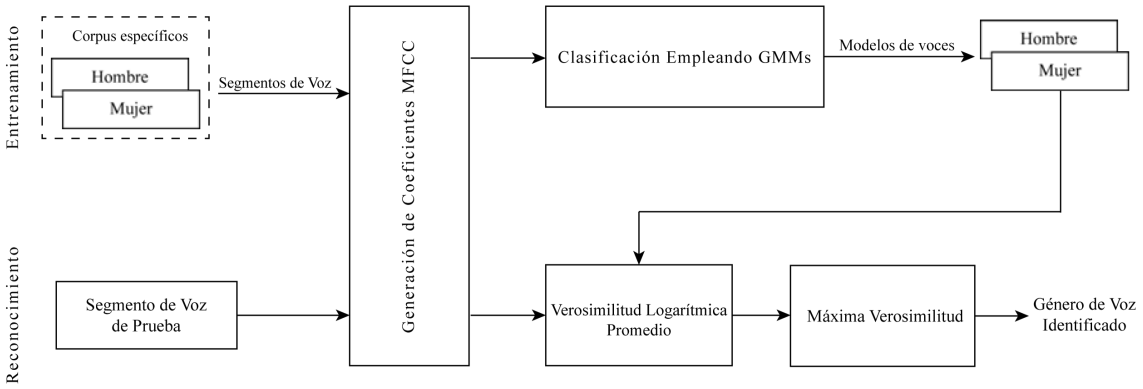


Figura 7.2: Diagrama de bloques del subsistema de identificación automática del género del locutor.

un problema de sobreajuste de datos que derive en una baja eficiencia del modelo durante la fase reconocimiento.

Para los modelos de los géneros de voz, así como de los idiomas, se generaron múltiples conjuntos de GMMs empleando un número variado de componentes Gaussianas. Específicamente, para el sistema LID se generaron 12 conjuntos de modelos de mezclas Gaussianas, de los cuales 6 representan los modelos de los idiomas hablados por hombres y 6 son los modelos de los idiomas hablados por mujeres. Por otra parte, para el subsistema de identificación del género del locutor se generaron 2 conjuntos de modelos; uno para cada género de voz. Cada conjunto de modelos consta de 128, 256, 512, 1,024 y en algunos casos 2,048 componentes Gaussianas. La finalidad de generar modelos con diferente número de componentes Gaussianas es conocer la eficiencia de cada composición de mezclas y poder seleccionar los modelos más eficientes para esta configuración del sistema LID.

7.1.3. Eficiencia de los Modelos del Sistema

El cálculo de la eficiencia de los distintos conjuntos de modelos se llevó a cabo después de presentar a los sistemas una colección de archivos de voz procedentes de los corpus de voz específicos. Los segmentos de voz de prueba fueron exclusivamente utilizados durante la fase de reconocimiento tanto del sistema LID como del subsistema de identificación del género del locutor, por lo que dichos segmentos de voz representan información no considerada para la generación de los modelos acústicos durante las respectivas fases de entrenamiento.

El número total y duración de los segmentos de voz pertenecientes a cada colección de archivos de audio que fue probada en cada sistema para el cálculo de la eficiencia de los modelos se lista a continuación:

- **Colección de archivos de audio para el cálculo de la eficiencia de los modelos de voz:**
  - 7,2000 segmentos de voz de 1 segundo.
  - 2,400 segmentos de voz de 3 segundos.

- 1,200 segmentos de voz de 6 segundos.
- 720 segmentos de voz de 10 segundos.
- 480 segmentos de voz de 15 segundos.
- 240 segmentos de voz de 30 segundos.
- 120 segmentos de voz de 60 segundos.

- **Colección de archivos de audio para el cálculo de la eficiencia de los modelos de idiomas:**

  - 3,600 segmentos de voz de 1 segundo.
  - 1,200 segmentos de voz de 3 segundos.
  - 600 segmentos de voz de 6 segundos.
  - 360 segmentos de voz de 10 segundos.
  - 240 segmentos de voz de 15 segundos.
  - 120 segmentos de voz de 30 segundos.
  - 60 segmentos de voz de 60 segundos.

El porcentaje de eficiencia de cada uno de los modelos se obtuvo para cada duración de los archivos, promediando las eficiencias individuales de cada uno de los objetivos identificables de cada sistema (idiomas o género), las cuales matemáticamente fueron obtenidas mediante el uso de la siguiente ecuación:

$$\%Eficiencia = \frac{segmentos\ identificados\ correctamente}{total\ de\ segmentos} \cdot 100 \tag{7.1}$$

En el anexo A se presenta el conjunto de tablas que muestra las eficiencias de los modelos del subsistema de identificación del género del locutor, los cuales fueron generados a partir de coeficientes MFCC. En estas tablas se puede observar como el porcentaje de eficiencia del subsistema incrementa a medida que el número de componentes Gaussianas en los modelos de mezclas Gaussianas es mayor. Por consiguiente, los modelos seleccionados fueron aquellos con 2048 componentes Gaussianas. Por otra parte, en el anexo B se presenta el conjunto de tablas que muestra las eficiencias de los modelos de idiomas hablados por hombres y mujeres, los cuales fueron generados a partir de coeficientes MFCC y SDC.

7.1.4. Eficiencia del Sistema

La tabla 7.1 muestra las especificaciones de los 12 modelos de idiomas más eficientes para esta configuración del sistema LID de acuerdo a la duración de los segmentos de prueba. Una vez elegidos los mejores modelos, se procedió a calcular la eficiencia de la configuración del sistema LID, al cual se le presentó la misma colección de archivos de voz que la utilizada para la generación de los modelos de idiomas, es decir, un total de 6,180 segmentos de voz cuya duración total es de 70 minutos. La tabla 7.2 presenta la eficiencia de esta configuración del sistema LID,

Duración del Segmento de Voz	Género del Modelo	Tipo de Coeficiente	Número de Componentes Gaussianas
1 s	Hombre	MFCC	2048
	Mujer	MFCC	2048
3 s	Hombre	MFCC	2048
	Mujer	MFCC	2048
6 s	Hombre	SDC	1024
	Mujer	SDC	512
10 s	Hombre	SDC	1024
	Mujer	SDC	512
15 s	Hombre	SDC	1024
	Mujer	MFCC	2048
30 s	Hombre	MFCC	1024
	Mujer	SDC	512
60 s	Hombre	MFCC	512
	Mujer	SDC	512

Tabla 7.1: Especificaciones de los mejores modelos acústicos de idiomas.

Idioma	Duración del Segmento de Voz						
	1 s	**1-3 s**	**3-6 s**	**6-10 s**	**10-15 s**	**15-30 s**	**>30 s**
Inglés	56.41 %	76.00 %	88.50 %	90.00 %	91.25 %	90.00 %	100 %
Español	55.50 %	76.50 %	85.00 %	80.83 %	86.25 %	92.50 %	100 %
Francés	64.50 %	79.25 %	84.50 %	81.66 %	87.50 %	100 %	100 %
Alemán	62.91 %	56.75 %	68.00 %	76.66 %	73.75 %	82.50 %	80.00 %
Ruso	65.16 %	62.75 %	68.00 %	84.16 %	80.00 %	82.50 %	85.00 %
Japonés	61.58 %	85.00 %	90.50 %	86.66 %	85.00 %	82.50 %	90.00 %
Eficiencia	61.01 %	72.70 %	80.75 %	83.32 %	83.95 %	88.33 %	92.50 %

Tabla 7.2: Eficiencia del sistema LID dependiente del género del locutor.

en la cual se toman en cuenta los 2 mejores modelos acústicos del subsistema de identificación del género de locutor, así como los mejores 12 modelos de cada idioma.

La figura 7.3 muestra mediante una gráfica de barras la comparación de las eficiencias de los modelos de idiomas hablados por hombres y mujeres con respecto a la eficiencia del sistema LID, en el cual dichos modelos se integran junto con los modelos de voz.

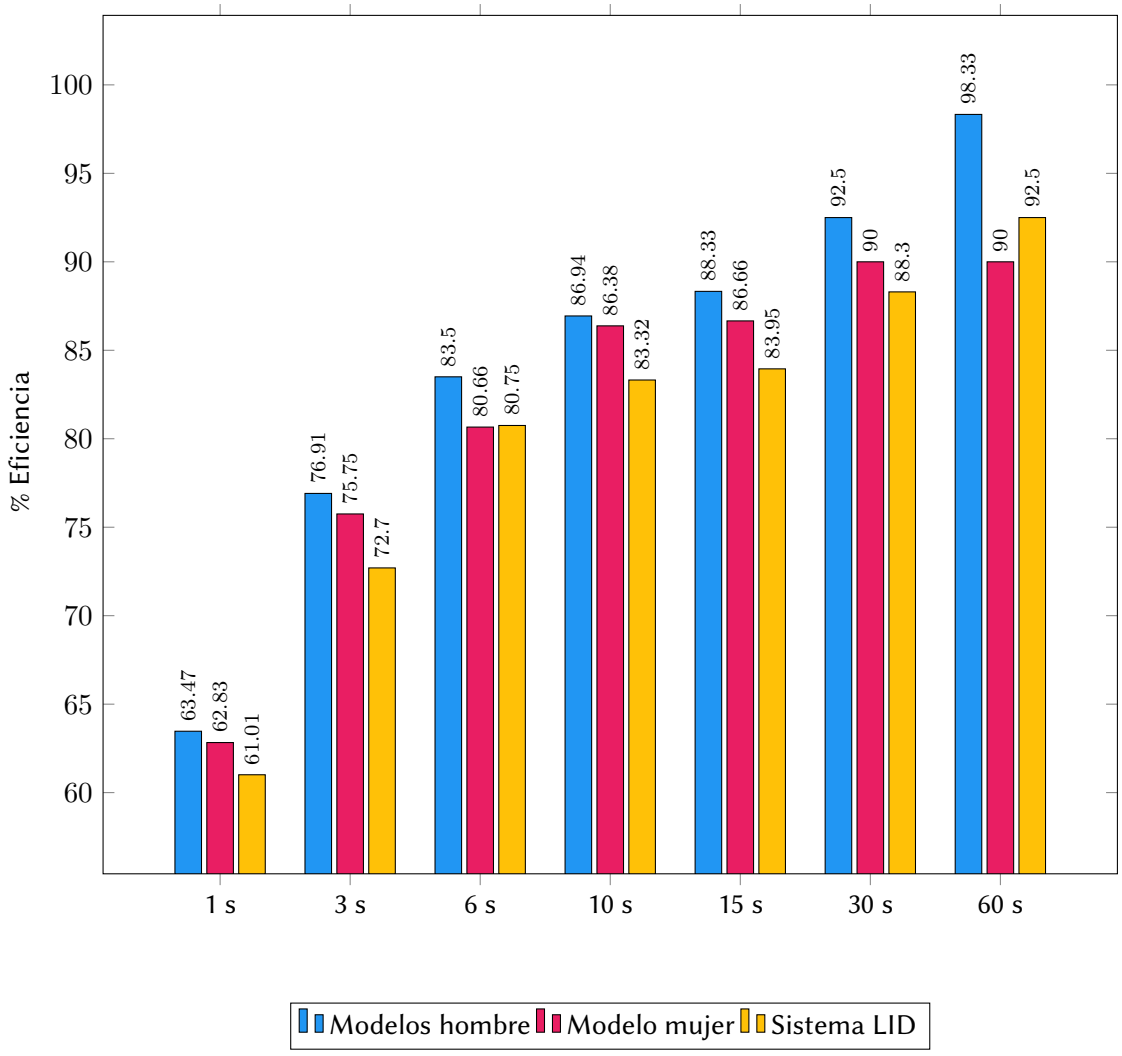


Figura 7.3: Eficiencia de los modelos del sistema LID dependiente del género del locutor

7.2. Sistema LID Independiente del Género del Locutor

La segunda configuración del sistema LID consiste en un sistema independiente del género del locutor. En la figura 7.4 se muestra el diagrama del sistema LID.

El objetivo principal de esta configuración del sistema LID es generar modelos acústicos de

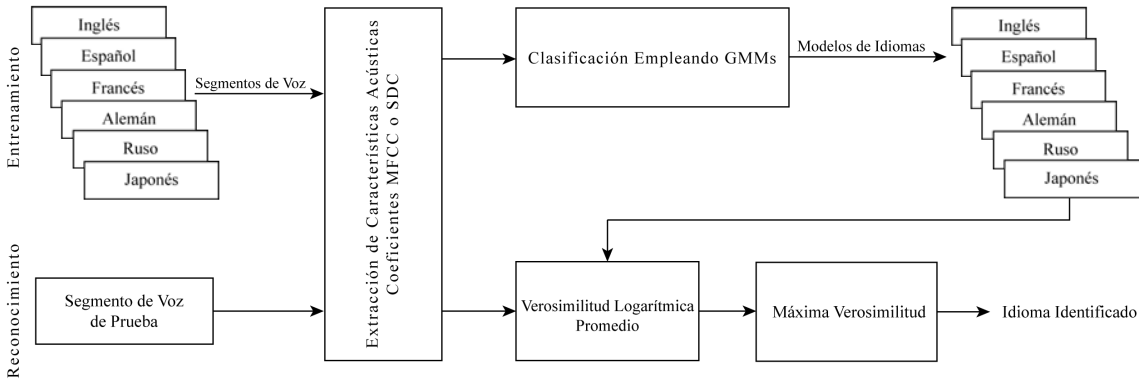


Figura 7.4: Diagrama del sistema LID independiente del género del locutor.

cada uno de los idiomas a partir de vectores acústicos de coeficientes MFCC o SDC procedentes de colecciones de audio mixtas, es decir de corpus de voz específicos que mezclan voces tanto de hombre como de mujer en el mismo idioma. El entrenamiento de modelos de idiomas con voces mixtas permite a esta configuración del sistema LID prescindir completamente del subsistema de identificación del género del locutor. Además, el número de modelos requerido durante la fase de reconocimiento del sistema es mucho menor que el requerido por la configuración del sistema LID dependiente del género del locutor, ya que únicamente requiere de un modelo por cada idioma a identificar.

Para esta configuración del sistema LID se generaron conjuntos de 6 modelos acústicos empleando un número variado de componentes Gaussianas para cada uno de los conjuntos. Al igual que la configuración dependiente del género del locutor, los conjuntos de modelos constan de 128, 256, 512, 1,024 y en algunos casos hasta 2,048 componentes Gaussianas. La finalidad de generar modelos con tales especificaciones es poder seleccionar los modelos más eficientes para el sistema LID.

7.2.1. Eficiencia de los Modelos del Sistema

EL cálculo de la eficiencia de cada modelo se lleva a cabo después de presentar al sistema una colección de 6,180 archivos de voz de prueba procedentes de los corpus de voz específicos. Los segmentos de voz de prueba son exclusivamente utilizados durante la fase de reconocimiento del sistema, por lo que representan información no considerada para la generación de modelos durante la fase de entrenamiento.

El número de segmentos y duración de los audios procedentes de la colección de archivos de prueba se lista a continuación:

- 3,600 segmentos de voz de 1 segundo.
- 1,200 segmentos de voz de 3 segundos.
- 600 segmentos de voz de 6 segundos.
- 360 segmentos de voz de 10 segundos.

- 240 segmentos de voz de 15 segundos.
- 120 segmentos de voz de 30 segundos.
- 60 segmentos de voz de 60 segundos.

En el anexo C se presenta el conjunto de tablas que muestra las eficiencias de los modelos del sistema LID independiente del género del locutor, los cuales fueron generados a partir de coeficientes MFCC y SDC procedentes de la colección de audios mixta. En la tabla 7.3 se listan las especificaciones de los 6 modelos más eficientes para esta configuración del sistema LID de acuerdo a la duración de los segmentos de audio.

Duración del Segmento de Voz	Tipo de Coeficientes	Número de Componentes Gaussianas
1 s	MFCC	2048
3 s	SDC	1024
6 s	SDC	1024
10 s	SDC	1024
15 s	SDC	1024
30 s	SDC	1024
60 s	SDC	1024

Tabla 7.3: Especificaciones de los mejores modelos acústicos de idiomas para el sistema LID independiente del género del locutor.

7.2.2. Eficiencia del Sistema

Al igual que en la configuración del sistema LID dependiente del género del locutor, la eficiencia del sistema se calculó promediando el porcentaje de eficiencia de cada uno de los idiomas, la cual se obtuvo empleando la ecuación 7.1. En la tabla 7.4 se muestra la eficiencia del sistema LID, la cual integra los mejores modelos acústicos.

7.3. Comparación de las Configuraciones del Sistema LID

La figura 7.5 muestra una gráfica de barras, en la que se compara el porcentaje de eficiencia de las dos configuraciones del sistema LID de acuerdo a la duración de los segmentos de voz de prueba. Puede observarse que para audios de duración menor a 3 segundos la identificación de idiomas se lleva a cabo de forma más eficiente por la configuración dependiente del género del locutor, mientras que para segmentos de voz cuya duración es mayor a 3 segundos, la

identificación automática de idiomas se realiza de manera más eficiente por la configuración independiente del género del locutor.

La configuración del sistema LID independiente del género del locutor no solamente es más eficiente que la configuración dependiente para archivos de audio de distinta duración, sino que además realiza la tarea de identificación de idiomas en un tiempo mucho menor, ya que evalúa los parámetros estadísticos de los vectores acústicos de los segmentos de prueba en menos de la mitad del número total de modelos acústicos utilizados en la fase de reconocimiento de la configuración dependiente.

7.4. Resumen

En este capítulo se presentaron dos configuraciones del sistema LID. Se generaron modelos acústicos de idiomas con diferentes especificaciones, de manera que las eficiencias de los modelos pudieran ser comparadas para seleccionar los mejores modelos para cada una de configuraciones del sistema LID implementado. Los experimentos para el cálculo de las eficiencias toman en cuenta archivos de audio procedentes de la corpora de voz y cuya duración es de 1, 3, 6, 10, 15, 30 y 60 segundos. La primera configuración descrita en este capítulo consiste en un sistema LID que es dependiente del género del locutor, por lo cual existe un paso previo a la identificación automática del idioma hablado, siendo éste el de identificar inicialmente si la voz del locutor procede de un hombre o de una mujer. La segunda configuración descrita consiste en un sistema LID independiente del género del locutor. Los modelos de los idiomas empleados en esta configuración fueron generados a partir de corpus de voz constituidos tanto por voces de hombres como de mujeres. Los resultados demostraron que la identificación de idiomas se lleva a cabo de una manera más eficiente empleando la configuración del sistema LID independiente del género del locutor cuando la duración de los audios es mayor a 3 segundos. En tanto que el sistema LID dependiente del género del locutor resultó ser mejor cuando los segmentos de voz tiene una duración menor a 3 segundos.

Idioma	**Duración del Segmento de Voz**						
	1 s	**1-3 s**	**3-6 s**	**6-10 s**	**10-15 s**	**15-30 s**	**>30 s**
Inglés	51.16 %	81.75 %	92.00 %	96.60 %	96.25 %	100 %	100 %
Español	51.66 %	87.50 %	93.50 %	95.00 %	95.00 %	97.50 %	100 %
Francés	65.50 %	77.25 %	88.00 %	89.16 %	92.00 %	90.00 %	95.00 %
Alemán	68.08 %	57.75 %	70.50 %	74.16 %	75.00 %	82.50 %	90.00 %
Ruso	58.00 %	56.25 %	63.50 %	68.30 %	76.25 %	77.50 %	95.00 %
Japonés	58.58 %	75.00 %	82.50 %	88.30 %	88.75 %	95 %	100 %
Eficiencia	58.83 %	72.58 %	81.66 %	85.25 %	87.20 %	90.41 %	96.66 %

Tabla 7.4: Eficiencia del sistema LID independiente del género del locutor

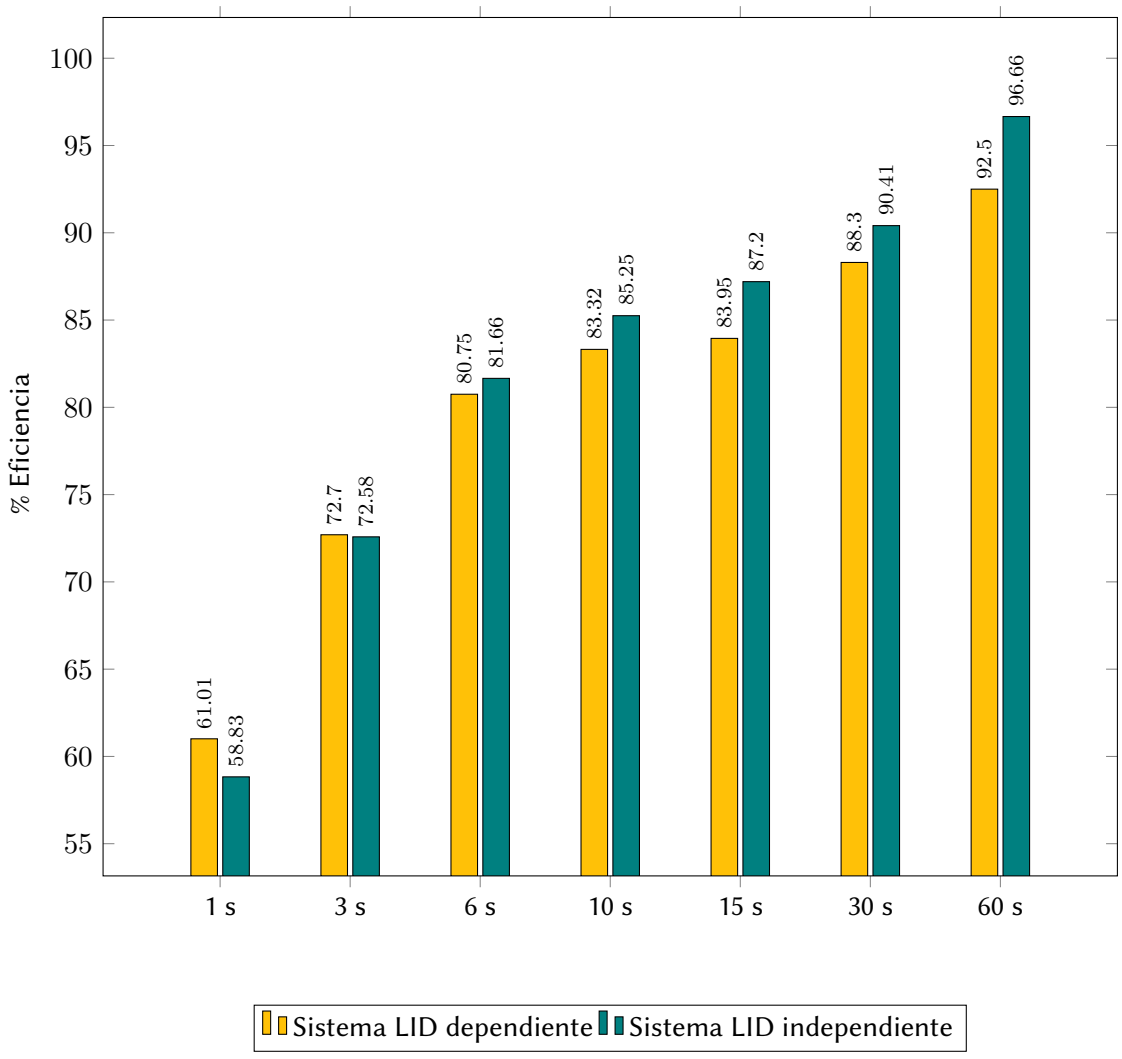


Figura 7.5: Eficiencia de las configuraciones del sistema LID

8

Conclusiones

La tendencia de globalización, la popularidad del Internet y el auge de las tecnologías de la información, ha permitido que la comunicación entre los humanos y las máquinas sea cada vez más usual, particularmente la interacción por voz con una computadora, donde la necesidad de servicios de identificación automática del idioma que se habla en segmentos de voz actualmente es cada vez mayor.

Es por esta razón que en este trabajo se realizó una extensa investigación acerca de la forma en que el ser humano produce y percibe la voz. Para el oído humano el reconocimiento de idiomas es muy claro. En cuestión de segundos una persona puede determinar si en un segmento de voz se habla algún idioma que conoce. Si no es así, entonces la persona suele basar su decisión en juicios subjetivos con base en aquéllos idiomas que le son familiares. Sin embargo, para una computadora la identificación de idiomas es una tarea mucho más compleja. El reto que presenta, es que no existe información previa disponible que le indique el contenido de la palabra o la identidad del hablante. Con el uso de herramientas del procesamiento digital de señales, en conjunto con la teoría de la probabilidad y técnicas de aprendizaje automático, las problemáticas anteriores fueron resueltas exitosamente en el diseño e implementación de un sistema LID de enfoque puramente acústico.

En este trabajo se logró implementar un sistema capaz de identificar automáticamente el idioma hablado en un segmento digital de voz. Los idiomas elegidos para conformar el sistema LID fueron: inglés, español, francés, alemán, ruso y japonés. Ya que el idioma identificado solamente puede ser hipotetizado como uno de los seis idiomas elegidos para conformar el sistema, la implementación consistió en el desarrollo de un sistema LID de conjunto cerrado.

El entrenamiento de los modelos acústicos de los idiomas representó una tarea que demandó muchos recursos computacionales, especialmente para el entrenamiento de modelos de mezclas Gaussianas con un número muy grande de componentes, donde el tiempo promedio de entrenamiento fue mayor a 24 horas. Las especificaciones técnicas más relevantes del equipo de cómputo con el cual se realizó el entrenamiento de los modelos son las siguientes: memoria RAM de 8 GB DDR3 y procesador Intel core i7 4790. Sin el uso de este equipo de cómputo no hubiera sido posible generar modelos acústicos eficientes.

Para la implementación del sistema LID se optó por dos configuraciones distintas: la primera consistió en un sistema LID dependiente del género del locutor, y la segunda en un sistema LID independiente de éste. Los resultados demostraron que sus eficiencias son muy similares, no obstante, la primera configuración es más óptima al identificar el idioma hablado en segmentos de voz de duración menor a 3 segundos, mientras que la segunda configuración presentó un porcentaje de eficiencia mayor para segmentos de voz de duración mayor a 3 segundos. Los porcentajes de eficiencia de la configuración del sistema LID van de 61 % a 72 % para segmentos de voz cuya duración es menor a 3 segundos, y de 81 % a 96 % cuando la duración de los segmentos de voz es mayor.

En cuanto a la velocidad con la cual las configuraciones del sistema LID realizan la tarea de identificación automática, la configuración independiente del género del hablante hipotetiza el idioma hablado en un tiempo menor en comparación con la configuración dependiente. Esto es debido a que en la fase de reconocimiento, el sistema LID independiente del género del locutor evalúa los parámetros estadísticos de los vectores de características en únicamente 6 modelos acústicos de idiomas, mientras que el sistema LID dependiente realiza la evaluación en 14 modelos, incluyendo los modelos acústicos de voz para la identificación del género del hablante.

La creación de modelos acústicos eficientes no es una tarea sencilla. La eficiencia de los modelos de idiomas puede incrementarse, primeramente aumentando el número de segmentos de voz para el entrenamiento de los modelos. En este trabajo se emplearon únicamente 40 minutos para entrenar los modelos acústicos, sin embargo, se sugieren corpus de voz para entrenamiento cuya duración varía entre 1 y 3 horas. Consecuentemente, dadas las especificaciones de los modelos generados en este trabajo, la generación de modelos acústicos a partir de una corpora de voz para entrenamiento más extensa requiere de equipo de cómputo con especificaciones técnicas más robustas.

El conjunto de idiomas del sistema LID puede incrementarse. En este trabajo únicamente se contemplaron 6 idiomas, sin embargo pueden ser integrados más idiomas al conjunto cerrado de identificación. La integración de nuevos idiomas consiste en primer lugar, en la generación de nuevos corpus de voz específicos, y en segundo lugar, de la generación de modelos a partir de los vectores de características acústicas de las señales de voz. En un futuro se buscará realizar la integración adicional de los siguientes idiomas: italiano, portugués, coreano y mandarín.

Bibliografía

[1] L. Rabiner and B.-H. Juang, "Fundamentals of speech recognition," 1993.

[2] P. Roach, *English Phonetics and Phonology Fourth Edition: A Practical Course*. Ernst Klett Sprachen, 2010.

[3] E. Skinner, *Speak with distinction*. Hal Leonard Corporation, 1990.

[4] U. of Manitoba, "Describing consonants (advanced)." `http://home.cc.umanitoba.ca/~krussll/phonetics/ipa/consonant-parameters.html`. [Página Web]; (Consultado 11-Diciembre-2015).

[5] W. T. Fitch and J. Giedd, "Morphology and development of the human vocal tract: A study using magnetic resonance imaging," *The Journal of the Acoustical Society of America*, vol. 106, no. 3, pp. 1511–1522, 1999.

[6] J. Benesty, *Springer handbook of speech processing*. Springer Science & Business Media, 2008.

[7] J. R. Deller Jr, J. G. Proakis, and J. H. Hansen, *Discrete time processing of speech signals*. Prentice Hall PTR, 1993.

[8] R. Scarborough, "Source-filter theory." [Handout 5]. Linguistics 105/205. Lecture on Phonetics. Stanford University, Oct. 11, 2005.

[9] I. Pollak, "Speech processing." EE438. Class notes on Digital Signal Processing. Purdue University, October, 2004.

[10] X. Huang, A. Acero, H.-W. Hon, and R. Foreword By-Reddy, *Spoken language processing: A guide to theory, algorithm, and system development*. Prentice Hall PTR, 2001.

[11] D. Crystal, *The Cambridge encyclopedia of language*. Cambridge Univ Press, 2010.

[12] L. R. Rabiner and R. W. Schafer, "Introduction to digital speech processing," *Foundations and trends in signal processing*, vol. 1, no. 1, pp. 1–194, 2007.

[13] T. Boyce, *Introduction to Live Sound Reinforcement: The Science, the Art, and the Practice*. FriesenPress, 2014.

[14] D. Bernstein, L. A. Penner, A. Clarke-Stewart, and E. Roy, *Psychology (PSY 113 General Psychology)*. Wadsworth Publishing, 9 ed., 2011.

[15] H. Fletcher and W. A. Munson, "Loudness, its definition, measurement and calculation*," *Bell System Technical Journal*, vol. 12, no. 4, pp. 377–430, 1933.

[16] W. A. Yost, "Pitch perception," *Attention, Perception, & Psychophysics*, vol. 71, no. 8, pp. 1701–1715, 2009.

[17] A. N. S. Institute, M. Sonn, and A. S. of America, *American National Standard Psychoacoustical Terminology*. N.: ANSI, American National Standards Institute, 1973.

[18] A. J. Houtsma, *Pitch perception*. Academic Press San Diego, London, 1995.

[19] S. S. Stevens, J. Volkmann, and E. B. Newman, "A scale for the measurement of the psychological magnitude pitch," *The Journal of the Acoustical Society of America*, vol. 8, no. 3, pp. 185–190, 1937.

[20] E. Zwicker, "Subdivision of the audible frequency range into critical bands (frequenzgruppen)," *The Journal of the Acoustical Society of America*, no. 33 (2), p. 248, 1961.

[21] O. Stuhlman Jr, "An introduction to biophysics.," *The American Journal of the Medical Sciences*, vol. 205, no. 6, p. 883, 1943.

[22] D. Havelock, S. Kuwano, and M. Vorländer, *Handbook of signal processing in acoustics*. Springer Science & Business Media, 2008.

[23] J. M. Merino de la Fuente and L. Muñoz-Repiso, "La percepción acústica: física de la audición," 2013.

[24] P. Lieberman and S. E. Blumstein, *Speech physiology, speech perception, and acoustic phonetics*. Cambridge University Press, 1988.

[25] P. Keating and R. Buhr, "Fundamental frequency in the speech of infants and children," *The Journal of the Acoustical Society of America*, vol. 63, no. 2, pp. 567–571, 1978.

[26] J. S. Rubin, R. T. Sataloff, and G. S. Korovin, *Diagnosis and treatment of voice disorders*. Plural Publishing, 2014.

[27] H. Hollien and T. Shipp, "Speaking fundamental frequency and chronologic age in males," *Journal of Speech, Language, and Hearing Research*, vol. 15, no. 1, pp. 155–159, 1972.

[28] M. L. Stoicheff, "Speaking fundamental frequency characteristics of nonsmoking female adults," *Journal of Speech, Language, and Hearing Research*, vol. 24, no. 3, pp. 437–441, 1981.

[29] W. Rappaport, "Über messungen der tonhöhenverteilung in der deutschen sprache," *Acta Acustica united with Acustica*, vol. 8, no. 4, pp. 220–225, 1958.

[30] C. Chevrie-Muller and F. Gremy, "Contribution a l'établissement de quelques constantes physiologiques de la voix parlée de l'adulte," *Journal Français d'Oto-Rhino-Laryngologie, XV*, vol. 1, pp. 149–154, 1967.

[31] Y. Takefuta, E. Jancosek, and M. Brunt, "A statistical analysis of melody curves in the intonation of american english," in *Proceedings of the 7th International Congress of Phonetic Sciences*, pp. 1035–1039, 1972.

[32] G.-t. Chen, "The pitch range of english and chinese speakers," *Journal of Chinese Linguistics*, pp. 159–171, 1974.

[33] L.-J. Boë, M. Contini, and H. Rakotofiringa, "Étude statistique de la fréquence laryngienne," *Phonetica*, vol. 32, no. 1, pp. 1–23, 1975.

[34] P. Kitzing, *Glottografisk frekvensindikering (GFI): en undersokningsmetod for matning av rostlage och rostomfang samt framstallning av rostfrekvensdistributionen.* PhD thesis, PhD dissertation, Malmoe. Diss. Lund: Univ. 90, 1979.

[35] C. Johns-Lewis, "Prosodic differentiation of discourse modes," *Intonation in discourse*, pp. 199–220, 1986.

[36] M. Pegoraro Krook, "Speaking fundamental frequency characteristics of normal swedish subjects obtained by glottal frequency analysis," *Folia Phoniatrica et Logopaedica*, vol. 40, no. 2, pp. 82–90, 1988.

[37] P. Rose, "How effective are long term mean and standard deviation as normalisation parameters for tonal fundamental frequency?," *Speech Communication*, vol. 10, no. 3, pp. 229–247, 1991.

[38] C.-C. Elert and B. Hammarberg, "Regional voice variation in sweden," in *Proceedings of the 12th International Congress of Phonetic Sciences*, pp. 418–420, 1991.

[39] H. Traunmüller and A. Eriksson, "The frequency range of the voice fundamental in the speech of male and female adults," *Consulté le*, vol. 12, no. 02, p. 2013, 1995.

[40] N. T. University, "Introduction to phonetics page by page." `http://homepage.ntu.edu.tw/~karchung/phon1index.htm`. [Página Web]; (Consultado 28-Abril-2016).

[41] D. M. Howard and D. T. Murphy, *Voice science, acoustics, and recording.* Plural Publishing, 2007.

[42] E. C. Zsiga, *The sounds of language: an introduction to phonetics and phonology.* John Wiley & Sons, 2012.

[43] R. Kirchner, "–phonetics and phonology: understanding the sounds of speech,"

[44] C. BaDeNas, "Versión en español del alfabeto fonético internacional, revisada por javier lorenzo," *Extraído el*, vol. 24.

[45] D. Jurafsky, *Speech & language processing*. Pearson Education India, 2000.

[46] I. P. Association, *Handbook of the International Phonetic Association: A guide to the use of the International Phonetic Alphabet*. Cambridge University Press, 1999.

[47] J. Gil Fernández, "Los sonidos del lenguaje," *Síntesis (Lingüística, Textos de apoyo, 3), Madrid*, 1988.

[48] T. U. of Iowa, "Dialectoteca del español." `http://dialects.its.uiowa.edu/`. [Página Web]; (Consultado 31-Mayo-2016).

[49] E. Martínez Celdrán, "Fonética," *Barcelona: Teide*, vol. 1994, 1984.

[50] T. U. of Iowa, "Fonética: Los sonidos del español." `http://soundsofspeech.uiowa.edu/spanish/spanish.html`. [Página Web]; (Consultado 12-Junio-2016).

[51] P. Ladefoged and S. F. Disner, *Vowels and consonants*. John Wiley & Sons, 2012.

[52] W. Holmes, *Speech synthesis and recognition*. CRC press, 2001.

[53] S. Chan, "Signals and systems." Class notes for signals and systems. University of California, San Diego.

[54] O. Alkin, *Signals and Systems: A MATLAB® Integrated Approach*. CRC Press, 2015.

[55] J. Feldman, "Discrete-time linear, time invariant systems and z-transforms." Class notes on mathematical methods for electrical and computer engineering. The University of British Columbia, Spring, 2006.

[56] A. V. Oppenheim, A. S. Willsky, and S. H. Nawab, *Signals and systems*, vol. 2. Prentice-Hall Englewood Cliffs, NJ, 1983.

[57] M. H. Hayes, *Schaum's Outline of Digital Signal Processing*. McGraw-Hill, Inc., 1998.

[58] C. A. Bouman, "Discrete fourier transform." EE438. Class notes on Digital Signal Processing with Applications. Purdue University, October, 2007.

[59] C. Gauss, "Theoria interpolationis methodo novo tractata, vol. 3," *K onigliche Gesellschaft der Wissenschaften, G ottingen*, 1866.

[60] F. Carlini, "Sulla legge delle variazioni orarie del barometro," *Memorie della Società italiana delle Scienze*, vol. 20, 1828.

[61] E. Sabine, "Contributions to terrestrial magnetism.," in *Abstracts of the Papers Printed in the Philosophical Transactions of the Royal Society of London*, vol. 4, pp. 212–213, JSTOR, 1837.

[62] J. D. Everett, "Xviii.—on a method of reducing observations of underground temperature, with its application to the monthly mean temperatures of underground thermometers, at the royal edinburgh observatory," *Transactions of the Royal Society of Edinburgh*, vol. 22, no. 02, pp. 429–439, 1861.

[63] C. Runge, "Ober die zerlegung empirisch gegebener periodischer funktionen in sinuswellen," *Zeitschr. f. Math. u. Phys*, pp. 3–4, 1903.

[64] L. Schrutka, "Tafeln und aufgaben zur harmonischen analyse und periodogrammrechnung," *Monatshefte für Mathematik*, vol. 49, no. 1, pp. A29–A29, 1941.

[65] G. C. Danielson and C. Lanczos, "Some improvements in practical fourier analysis and their application to x-ray scattering from liquids," *Journal of the Franklin Institute*, vol. 233, no. 5, pp. 435–452, 1942.

[66] L. Thomas, "Using a computer to solve problems in physics," *Applications of Digital Computers*, pp. 44–45, 1963.

[67] I. J. Good, "The interaction algorithm and practical fourier analysis," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 361–372, 1958.

[68] J. W. Cooley and J. W. Tukey, "An algorithm for the machine calculation of complex fourier series," *Mathematics of computation*, vol. 19, no. 90, pp. 297–301, 1965.

[69] S. Winograd, "On computing the discrete fourier transform," *Mathematics of computation*, vol. 32, no. 141, pp. 175–199, 1978.

[70] M. Heideman, D. Johnson, and C. Burrus, "Gauss and the history of the fast fourier transform," *IEEE ASSP Magazine*, vol. 1, no. 4, pp. 14–21, 1984.

[71] W. Press, S. Teukolsky, W. Vetterling, and B. Flannery, "Numerical recipes in fortran 77: The art of scientific computing, 933 pp," 1992.

[72] N. Ahmed, T. Natarajan, and K. R. Rao, "Discrete cosine transform," *IEEE transactions on Computers*, vol. 100, no. 1, pp. 90–93, 1974.

[73] R. J. Beerends, *Fourier and Laplace transforms*. Cambridge University Press, 2003.

[74] J. P. Allebach, "Analysis of sampling." EE438. Lecture notes on Digital Signal Processing with Applications. Purdue University, October, 2014.

[75] L. Deng and D. O'Shaughnessy, *Speech processing: a dynamic and optimization-oriented approach*. CRC Press, 2003.

[76] D. O'shaughnessy, *Speech communication: human and machine*. Universities press, 1987.

[77] L. R. Rabiner and R. W. Schafer, *Digital processing of speech signals*. Prentice Hall, 1978.

[78] F. J. Harris, "On the use of windows for harmonic analysis with the discrete fourier transform," *Proceedings of the IEEE*, vol. 66, no. 1, pp. 51–83, 1978.

[79] J. Bernal Bermúdez, J. Bobadilla Sancho, and P. Gómez Vilda, "Reconocimiento de voz y fonética acústica," 2000.

[80] P. Taylor, *Text-to-speech synthesis*. Cambridge university press, 2009.

[81] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE transactions on acoustics, speech, and signal processing*, vol. 28, no. 4, pp. 357–366, 1980.

[82] B. Bielefeld, "Language identification using shifted delta cepstrum," in *Fourteenth Annual Speech Research Symposium*, 1994.

[83] P. A. Torres-Carrasquillo, E. Singer, M. A. Kohler, R. J. Greene, D. A. Reynolds, and J. R. Deller Jr, "Approaches to language identification using gaussian mixture models and shifted delta cepstral features.," in *INTERSPEECH*, 2002.

[84] D. D. Gutierrez, *Machine Learning and Data Science: An Introduction to Statistical Learning Methods with R*. Technics Publications, 2015.

[85] A. L. Samuel, "Some studies in machine learning using the game of checkers," *IBM Journal of research and development*, vol. 3, no. 3, pp. 210–229, 1959.

[86] T. M. Mitchell, "Machine learning. 1997," *Burr Ridge, IL: McGraw Hill*, vol. 45, p. 37, 1997.

[87] X. Zhu and A. B. Goldberg, "Introduction to semi-supervised learning," *Synthesis lectures on artificial intelligence and machine learning*, vol. 3, no. 1, pp. 1–130, 2009.

[88] K. J. Cios, W. Pedrycz, and R. W. Swiniarski, "Data mining and knowledge discovery," in *Data Mining Methods for Knowledge Discovery*, pp. 1–26, Springer, 1998.

[89] N. J. Salkind, *Encyclopedia of measurement and statistics*. SAGE publications, 2006.

[90] S. M. Stigler, *Statistics on the table: The history of statistical concepts and methods*. Harvard University Press, 2002.

[91] D. Yu and L. Deng, *Automatic Speech Recognition*. Springer, 2012.

[92] G. McLachlan and D. Peel, *Finite mixture models*. John Wiley & Sons, 2004.

[93] R. G. Leonard and G. R. Doddington, "Automatic language identification.," tech. rep., DTIC Document, 1974.

[94] A. S. House and E. P. Neuburg, "Toward automatic identification of the language of an utterance. i. preliminary methodological considerations," *The Journal of the Acoustical Society of America*, vol. 62, no. 3, pp. 708–713, 1977.

[95] B. Comrie, *The world's major languages*. Routledge, 2009.

[96] T. Schultz and K. Kirchhoff, *Multilingual speech processing*. Academic Press, 2006.

[97] R. G. Gordon Jr, "Ethnologue: Languages of the world, dallas, tex.: Sil international," *Online version: http://www. ethnologue. com*, 2005.

[98] V. Fromkin, R. Rodman, and N. Hyams, "An introduction to linguistics," *USA: Wadsworth*, 2003.

[99] G. K. Zipf, *Human behavior and the principle of least effort: An introduction to human ecology*. Ravenio Books, 2016.

[100] P. T. Daniels and W. Bright, *The world's writing systems*. Oxford University Press on Demand, 1996.

[101] H. Hammarström, "Ethnologue 16/17/18th editions: A comprehensive review," *Language*, vol. 91, no. 3, pp. 723–737, 2015.

[102] P. Delattre, "Comparing the phonetic features of english, spanish, german and french," 1965.

[103] E. Singer and D. A. Reynolds, "Analysis of multitarget detection for speaker and language recognition," in *ODYSSEY04-The Speaker and Language Recognition Workshop*, 2004.

[104] F. Jelinek, *Statistical methods for speech recognition*. MIT press, 1997.

[105] W. Campbell, T. Gleason, J. Navratil, D. Reynolds, W. Shen, E. Singer, and P. Torres-Carrasquillo, "Advanced language recognition using cepstra and phonotactics: Mitll system performance on the nist 2005 language recognition evaluation," in *2006 IEEE Odyssey-The Speaker and Language Recognition Workshop*, pp. 1–8, IEEE, 2006.

[106] W. Shen, W. Campbell, T. Gleason, D. Reynolds, and E. Singer, "Experiments with lattice-based pprlm language identification," in *2006 IEEE Odyssey-The Speaker and Language Recognition Workshop*, pp. 1–6, IEEE, 2006.

[107] J.-L. Gauvain, A. Messaoudi, and H. Schwenk, "Language recognition using phone latices.," in *INTERSPEECH*, 2004.

[108] P. Matejka, L. Burget, P. Schwarz, and J. Cernocky, "Brno university of technology system for nist 2005 language recognition evaluation," in *2006 IEEE Odyssey-The Speaker and Language Recognition Workshop*, pp. 1–7, IEEE, 2006.

[109] M. A. Zissman, "Predicting, diagnosing and improving automatic language identification performance.," in *Eurospeech*, 1997.

[110] M. A. Zissman, "Automatic language identification using gaussian mixture and hidden markov models," in *Acoustics, Speech, and Signal Processing, 1993. ICASSP-93., 1993 IEEE International Conference on*, vol. 2, pp. 399–402, IEEE, 1993.

[111] P. A. Torres-Carrasquillo, D. A. Reynolds, and J. R. Deller, "Language identification using gaussian mixture model tokenization," in *Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on*, vol. 1, pp. I–757, IEEE, 2002.

[112] P. A. Torres-Carrasquillo, D. A. Reynolds, and J. Deller, "Language identification using gaussian mixture model," in *Tokenization," International Conference on Acoustics, Speech & Signal Processing*, Citeseer, 2002.

[113] E. Wong, J. Pelecanos, S. Myers, and S. Sridharan, "Language identification using efficient gaussian mixture model analysis," in *Australian International Conference on Speech Science and Technology*, vol. 4, pp. 7–6, 2000.

[114] N. Parlangeau, F. Pellegrino, and R. André-Obrecht, "Investigating automatic language discrimination via vowel system and consonantal system modeling," in *Proc. of ICPhS'99*, 1999.

[115] F. Pellegrino and R. André-Obrecht, "Automatic language identification: an alternative approach to phonetic modelling," *Signal Processing*, vol. 80, no. 7, pp. 1231–1244, 2000.

[116] U. Rapajic, "Using phone recognition and language modelling (prlm) for automatic language identification,"."

[117] M. D. Ivan, I. Magrin-chagnolleau, and F. Bimbot, "Language recognition using time-frequency principal component analysis and acoustic modeling," 2000.

Abreviaturas

AFI Alfabeto Fonético Internacional

ANSI (American National Standards Institute) Instituto Nacional Americano de Estándares

API (L'Association Phonétique Internationale) Asociación Fonética Internacional

BIBO (Bounded Input Bounded Output) Entrada Acotada Salida Acotada

DCT (Discrete Cosine Transform) Transformada Coseno Discreta

DFS (Discrete Fourier Series) Serie Discreta de Fourier

DFT (Discrete Fourier Transform) Transformada Discreta de Fourier

DTFT (Discrete-Time Fourier Transform) Transformada de Fourier en Tiempo Discreto

EM (Expectation-Maximization) Esperanza-Maximización

FFT (Fast Fourier Transform) Transformada Rápida de Fourier

GMM (Gaussian Mixture Model) Modelo de Mezclas Gaussianas

HMM (Hidden Markov Models) Modelos Ocultos de Markov

LID (Language Identification) Identificación Automática del Lenguaje Hablado

LPC (Linear Prediction Coefficients) Coeficientes de Predicción Lineal

LPCC (Linear Prediction Cepstral Coefficients) Coeficientes Lineales de Predicción Cepstral

MFCC (Mel Frequency Cepstral Coefficients) Coeficientes Cepstrales de Frecuencia Mel

PDS Procesamiento Digital de Señales

SDC (Shifted Delta Cepstral Coefficients) Coeficientes Cesptrales Delta Desplazados

STDTFT (Short-Time Discrete-Time Fourier Transform) Transformada de Fourier en Tiempo Discreto de Tiempo Corto

SVM (Support Vector Machines) Máquinas de Soporte Vectorial

UBM (Universal Background Model) Modelo de Referencia Universal

VQ (Vector Quantization) Cuantización Vectorial

Anexos

A

Eficiencias de los Modelos Acústicos del Subsistema de Identificación Automática del Género del Locutor

Voz	Duración del Segmento de Voz						
	1 s	**3 s**	**6 s**	**10 s**	**15 s**	**30 s**	**60 s**
Hombre	93.80 %	94.33 %	94.00 %	94.40 %	93.75 %	95.00 %	95.00 %
Mujer	89.22 %	92.83 %	94.50 %	94.40 %	94.58 %	95.00 %	95.00 %
Eficiencia	91.51 %	93.58 %	94.25 %	94.40 %	94.16 %	95.00 %	95.00 %

Tabla A.1: Sistema de identificación por género. Coeficientes MFCC, GMM: 128

Voz	Duración del Segmento de Voz						
	1 s	**3 s**	**6 s**	**10 s**	**15 s**	**30 s**	**60 s**
Hombre	94.13 %	94.33 %	94.16 %	95.00 %	95.00 %	95.83 %	95.00 %
Mujer	90.75 %	94.08 %	95.66 %	95.83 %	96.66 %	97.50 %	96.66 %
Eficiencia	92.44 %	94.20 %	94.91 %	95.41 %	95.83 %	96.66 %	95.83 %

Tabla A.2: Sistema de identificación por género. Coeficientes MFCC, GMM: 256

Voz	Duración del Segmento de Voz						
	1 s	**3 s**	**6 s**	**10 s**	**15 s**	**30 s**	**60 s**
Hombre	94.58 %	95.08 %	94.83 %	95.55 %	95.00 %	95.83 %	95.00 %
Mujer	91.61 %	94.83 %	96.00 %	96.66 %	97.50 %	97.50 %	96.66 %
Eficiencia	93.09 %	94.95 %	95.41 %	96.10 %	96.25 %	96.66 %	95.83 %

Tabla A.3: Sistema de identificación por género. Coeficientes MFCC, GMM: 512

Voz	Duración del Segmento de Voz						
	1 s	**3 s**	**6 s**	**10 s**	**15 s**	**30 s**	**60 s**
Hombre	94.97 %	95.75 %	96.11 %	95.83 %	95.83 %	95.83 %	96.66 %
Mujer	92.75 %	95.91 %	97.22 %	97.50 %	97.50 %	97.50 %	98.33 %
Eficiencia	93.86 %	95.83 %	96.66 %	96.66 %	96.66 %	96.66 %	97.49 %

Tabla A.4: Sistema de identificación por género. Coeficientes MFCC, GMM: 1024

Voz	Duración del Segmento de Voz						
	1 s	**3 s**	**6 s**	**10 s**	**15 s**	**30 s**	**60 s**
Hombre	95.11 %	96.41 %	96.50 %	96.38 %	95.83 %	96.66 %	96.66 %
Mujer	93.94 %	96.58 %	97.50 %	97.50 %	97.50 %	98.33 %	98.33 %
Eficiencia	94.52 %	96.49 %	97.00 %	96.94 %	96.66 %	97.49 %	97.49 %

Tabla A.5: Sistema de identificación por género. Coeficientes MFCC, GMM: 2048

B

Eficiencias de los Modelos Acústicos de Idiomas para del Sistema LID Dependiente del Género del Locutor

Idioma	Duración del Segmento de Voz						
	1 s	**3 s**	**6 s**	**10 s**	**15 s**	**30 s**	**60 s**
Inglés	52.00 %	63.50 %	71.00 %	75.00 %	75.00 %	75.00 %	90.00 %
Español	60.16 %	66.50 %	68.00 %	70.00 %	70.00 %	75.00 %	80.00 %
Francés	67.33 %	75.00 %	79.00 %	80.00 %	82.50 %	85.00 %	90.00 %
Alemán	16.16 %	15.00 %	14.00 %	16.66 %	20.00 %	10.00 %	10.00 %
Ruso	62.50 %	78.50 %	80.00 %	85.00 %	82.50 %	80.00 %	90.00 %
Japonés	58.33 %	67.00 %	68.00 %	70.00 %	77.50 %	80.00 %	90.00 %
Eficiencia	52.74 %	60.91 %	68.33 %	66.11 %	67.91 %	67.5 %	75 %

Tabla B.1: Sistema LID. Voz: femenina, Coeficientes: MFCC, GMMs: 128

Idioma	Duración del Segmento de Voz						
	1 s	**3 s**	**6 s**	**10 s**	**15 s**	**30 s**	**60 s**
Inglés	34.00 %	47.00 %	55.00 %	60.00 %	67.50 %	75.00 %	80.00 %
Español	52.16 %	64.00 %	71.00 %	73.33 %	80.00 %	80.00 %	80.00 %
Francés	57.16 %	70.00 %	72.00 %	78.33 %	80.00 %	80.00 %	90.00 %
Alemán	86.00 %	96.50 %	100 %	100 %	100 %	100 %	100 %
Ruso	52.33 %	65.5 %	73 %	75 %	85 %	90 %	80 %
Japonés	46.16 %	54 %	57 %	61.60 %	65.00 %	70 %	90 %
Eficiencia	54.63 %	66.16 %	71.33 %	74.71 %	79.58 %	82.50 %	86.66 %

Tabla B.2: Sistema LID. Voz: masculina, Coeficientes: MFCC, GMMs: 128

Idioma	**Duración del Segmento de Voz**						
	1 s	**3 s**	**6 s**	**10 s**	**15 s**	**30 s**	**60 s**
Inglés	53.66 %	64.00 %	73.00 %	81.66 %	82.50 %	85.00 %	90.00 %
Español	61.00 %	68.00 %	69.00 %	75.00 %	72.50 %	75.00 %	80.00 %
Francés	64.83 %	72.00 %	79.00 %	78.33 %	82.50 %	75.00 %	90.00 %
Alemán	29.50 %	37.00 %	41.00 %	38.33 %	37.50 %	50.00 %	50.00 %
Ruso	66.16 %	79.50 %	81.00 %	85.00 %	82.50 %	80.00 %	90.00 %
Japonés	63.50 %	70.50 %	73.00 %	78.33 %	77.50 %	85.00 %	90.00 %
Eficiencia	56.44 %	65.16 %	69.33 %	72.77 %	72.50 %	75.00 %	81.66 %

Tabla B.3: Sistema LID. Voz: femenina, Coeficientes: MFCC, GMMs: 256

Idioma	**Duración del Segmento de Voz**						
	1 s	**3 s**	**6 s**	**10 s**	**15 s**	**30 s**	**60 s**
Inglés	31.83 %	43.00 %	53.00 %	53.33 %	67.50 %	70.00 %	60.00 %
Español	55.16 %	68.50 %	79.00 %	85.00 %	85.00 %	95.00 %	90.00 %
Francés	64.83 %	76.00 %	81.00 %	81.66 %	82.50 %	80.00 %	90.00 %
Alemán	81.33 %	92.50 %	98.00 %	98.33 %	100 %	100 %	100 %
Ruso	56.66 %	68.00 %	74.00 %	75.00 %	90..00 %	95.00 %	90.00 %
Japonés	47.66 %	53.00 %	64.00 %	70.00 %	67.50 %	85.00 %	90.00 %
Eficiencia	56.24 %	66.83 %	74.83 %	77.22 %	82.08 %	87.50 %	86.66 %

Tabla B.4: Sistema LID. Voz: masculina, Coeficientes: MFCC, GMMs: 256

Idioma	Duración del Segmento de Voz						
	1 s	3 s	6 s	10 s	15 s	30 s	60 s
Inglés	51.83 %	64.50 %	67.00 %	76.66 %	85.00 %	90.00 %	100.00 %
Español	65.83 %	74.00 %	79.00 %	78.33 %	80.00 %	75.00 %	80.00 %
Francés	66.50 %	76.50 %	83.00 %	81.66 %	82.50 %	80.00 %	90.00 %
Alemán	37.16 %	46.00 %	52.00 %	55.00 %	52.500 %	60.00 %	80.00 %
Ruso	65.66 %	77.500 %	83.00 %	85.00 %	82.50 %	85.00 %	90.00 %
Japonés	65.66 %	74.00 %	78.00 %	80.00 %	85.00 %	85.00 %	90.00 %
Eficiencia	58.77 %	68.75 %	73.66 %	76.10 %	77.91 %	79.16 %	88.33 %

Tabla B.5: Sistema LID. Voz: femenina, Coeficientes: MFCC, GMMs: 512

Idioma	Duración del Segmento de Voz						
	1 s	3 s	6 s	10 s	15 s	30 s	60 s
Inglés	35.66 %	46.00 %	53.00 %	60.00 %	65.00 %	80.00 %	100 %
Español	55.16 %	72.00 %	80.00 %	83.33 %	82.50 %	85.00 %	100 %
Francés	74.33 %	84.00 %	89.00 %	86.66 %	95.00 %	100 %	100 %
Alemán	79.16 %	93.00 %	98.00 %	98.33 %	100 %	100 %	100 %
Ruso	61.16 %	76.00 %	79.00 %	88.33 %	90.00 %	95.00 %	100 %
Japonés	49.16 %	56.50 %	66.00 %	68.33 %	72.50 %	80.00 %	90.00 %
Eficiencia	59.10 %	71.25 %	77.50 %	80.82 %	84.16 %	90.00 %	98.33 %

Tabla B.6: Sistema LID. Voz: masculina, Coeficientes: MFCC, GMMs: 512

Idioma	**Duración del Segmento de Voz**						
	1 s	**3 s**	**6 s**	**10 s**	**15 s**	**30 s**	**60 s**
Inglés	54.50 %	67.00 %	73.00 %	81.66 %	90.00 %	95.00 %	100 %
Español	70.66 %	79.50 %	84.00 %	86.66 %	82.50 %	80.00 %	80.00 %
Francés	70.16 %	80.00 %	86.00 %	85.00 %	85.00 %	85.00 %	90.00 %
Alemán	39.50 %	44.50 %	51.00 %	60.00 %	57.50 %	65.00 %	80.00 %
Ruso	64.50 %	78.00 %	84.00 %	85.00 %	82.50 %	85.00 %	90.00 %
Japonés	67.83 %	78.50 %	79.00 %	86.66 %	87.50 %	95.00 %	90.00 %
Eficiencia	61.19 %	71.25 %	76.16 %	80.83 %	80.83 %	84.16 %	88.33 %

Tabla B.7: Sistema LID. Voz: femenina, Coeficientes: MFCC, GMMs: 1024

Idioma	**Duración del Segmento de Voz**						
	1 s	**3 s**	**6 s**	**10 s**	**15 s**	**30 s**	**60 s**
Inglés	47.83 %	61.50 %	71.00 %	78.33 %	80.00 %	90.00 %	100 %
Español	42.83 %	58.50 %	67.00 %	73.33 %	72.50 %	80.00 %	90.00 %
Francés	73.66 %	87.00 %	94.00 %	93.33 %	95.00 %	100 %	100 %
Alemán	83.50 %	94.50 %	100 %	98.33 %	100 %	100 %	100 %
Ruso	66.33 %	77.50 %	83.00 %	88.33 %	95.00 %	95.00 %	100 %
Japonés	55.00 %	68.00 %	75.00 %	80.00 %	85.00 %	90.00 %	90.00 %
Eficiencia	61.53 %	74.50 %	81.66 %	85.27 %	87.91 %	92.50 %	96.66 %

Tabla B.8: Sistema LID. Voz: masculina, Coeficientes: MFCC, GMMs: 1024

Idioma	Duración del Segmento de Voz						
	1 s	3 s	6 s	10 s	15 s	30 s	60 s
Inglés	59.33 %	76.50 %	84.00 %	93.33 %	100 %	100 %	100 %
Español	74.00 %	82.50 %	83.00 %	85.00 %	87.50 %	90.00 %	80.00 %
Francés	72.00 %	81.50 %	86.00 %	85.00 %	90.00 %	85.00 %	90.00 %
Alemán	43.33 %	57.00 %	62.00 %	66.66 %	72.50 %	80.00 %	80.00 %
Ruso	65.50 %	80.50 %	85.00 %	85.00 %	82.50 %	85.00 %	90.00 %
Japonés	69.66 %	76.50 %	79.00 %	85.00 %	87.50 %	95.00 %	90.00 %
Eficiencia	62.83 %	75.75 %	79.83 %	83.33 %	86.66 %	89.16 %	88.33 %

Tabla B.9: Sistema LID. Voz: femenina, Coeficientes: MFCC, GMMs: 2048

Idioma	Duración del Segmento de Voz						
	1 s	3 s	6 s	10 s	15 s	30 s	60 s
Inglés	53.66 %	68.00 %	76.00 %	81.66 %	80.00 %	90.00 %	100 %
Español	39.33 %	52.00 %	57.00 %	65.00 %	70.00 %	70.00 %	90.00 %
Francés	75.33 %	89.50 %	94.00 %	98.33 %	97.50 %	100 %	100 %
Alemán	84.50 %	95.00 %	100 %	98.33 %	100 %	100 %	100 %
Ruso	66.00 %	80.50 %	84.00 %	88.33 %	95.00 %	95.00 %	100 %
Japonés	62.00 %	76.50 %	81.00 %	83.33 %	85.00 %	90.00 %	90.00 %
Eficiencia	63.47 %	76.91 %	82.00 %	85.83 %	87.91 %	90.83 %	96.66 %

Tabla B.10: Sistema LID. Voz: masculina, Coeficientes: MFCC, GMMs: 2048

Idioma	**Duración del Segmento de Voz**						
	1 s	**3 s**	**6 s**	**10 s**	**15 s**	**30 s**	**60 s**
Inglés	64.66 %	83.00 %	92.00 %	96.66 %	97.50 %	100 %	100 %
Español	54.83 %	78.50 %	93.00 %	95.00 %	95.00 %	100 %	100 %
Francés	51.00 %	73.00 %	85.00 %	95.00 %	97.50 %	100 %	100 %
Alemán	12.83 %	11.50 %	15.00 %	13.33 %	12.50 %	10.00 %	10.00 %
Ruso	29.83 %	39.00 %	44.00 %	43.33 %	32.50 %	35.00 %	30.00 %
Japonés	64.50 %	82.00 %	91.00 %	96.66 %	97.50 %	100 %	100 %
Eficiencia	46.27 %	61.16 %	70.00 %	73.33 %	72.08 %	74.16 %	73.33 %

Tabla B.11: Sistema LID. Voz: femenina, Coeficientes: SDC, GMMs: 128

Idioma	**Duración del Segmento de Voz**						
	1 s	**3 s**	**6 s**	**10 s**	**15 s**	**30 s**	**60 s**
Inglés	47.83 %	61.50 %	74.00 %	83.33 %	85.00 %	85.00 %	90.00 %
Español	43.60 %	47.00 %	43.00 %	50.00 %	47.50 %	50.00 %	70.00 %
Francés	51.00 %	75.50 %	83.00 %	86.60 %	82.50 %	80.00 %	90.00 %
Alemán	26.60 %	44.00 %	54.00 %	58.30 %	57.50 %	60.00 %	80.00 %
Ruso	37.00 %	44.50 %	53.00 %	48.30 %	50.00 %	50.00 %	50.00 %
Japonés	64.16 %	87.00 %	92.00 %	93.30 %	97.50 %	100 %	100 %
Eficiencia	45.03 %	58.19 %	63.83 %	66.80 %	66.83 %	70.83 %	80.00 %

Tabla B.12: Sistema LID. Voz: masculina, Coeficientes: SDC, GMMs: 128

Idioma	Duración del Segmento de Voz						
	1 s	**3 s**	**6 s**	**10 s**	**15 s**	**30 s**	**60 s**
Inglés	60.50 %	79.50 %	91.00 %	98.33 %	100 %	100 %	100 %
Español	63.50 %	88.50 %	97.00 %	96.66 %	100 %	100 %	100 %
Francés	58.16 %	84.50 %	91.00 %	96.66 %	100 %	100 %	100 %
Alemán	20.33 %	26.00 %	32.00 %	31.66 %	42.50 %	45.00 %	60.00 %
Ruso	33.83 %	43.50 %	43.00 %	50.00 %	50.00 %	55.00 %	50.00 %
Japonés	62.66 %	82.50 %	92.00 %	98.33 %	100 %	100 %	100 %
Eficiencia	49.83 %	67.41 %	74.33 %	78.60 %	82.08 %	83.33 %	85.00 %

Tabla B.13: Sistema LID. Voz: femenina, Coeficientes: SDC, GMMs: 256

Idioma	Duración del Segmento de Voz						
	1 s	**3 s**	**6 s**	**10 s**	**15 s**	**30 s**	**60 s**
Inglés	45.83 %	64.50 %	71.00 %	80.00 %	82.50 %	95.00 %	100 %
Español	50.16 %	55.00 %	61.00 %	63.33 %	60.00 %	65.00 %	70.00 %
Francés	59.66 %	79.00 %	86.00 %	88.33 %	82.50 %	90 %	100 %
Alemán	34.50 %	54.00 %	62.00 %	68.33 %	72.50 %	70.00 %	80.00 %
Ruso	47.16 %	59.50 %	72.00 %	70.00 %	72.50 %	65.00 %	90.00 %
Japonés	64.66 %	86.00 %	94.00 %	95.00 %	95.00 %	100 %	100 %
Eficiencia	50.32 %	66.33 %	74.33 %	77.49 %	77.50 %	80.83 %	90.00 %

Tabla B.14: Sistema LID. Voz: masculina, Coeficientes: SDC, GMMs: 256

Idioma	**Duración del Segmento de Voz**						
	1 s	**3 s**	**6 s**	**10 s**	**15 s**	**30 s**	**60 s**
Inglés	60.66 %	84.50 %	94.00 %	96.66 %	100 %	100 %	100 %
Español	70.66 %	89.00 %	97.00 %	100 %	100 %	100 %	100 %
Francés	63.50 %	85.50 %	91.00 %	96.66 %	97.50 %	100 %	100 %
Alemán	26.83 %	41.00 %	52.00 %	65.00 %	55.00 %	70.00 %	70.00 %
Ruso	42.83 %	54.00 %	59.00 %	61.66 %	65.00 %	70.00 %	70.00 %
Japonés	61.50 %	81.50 %	91.00 %	98.33 %	97.50 %	100 %	100 %
Eficiencia	54.33 %	72.58 %	80.66 %	86.38 %	85.83 %	90.00 %	90.00 %

Tabla B.15: Sistema LID. Voz: femenina, Coeficientes: SDC, GMMs: 512

Idioma	**Duración del Segmento de Voz**						
	1 s	**3 s**	**6 s**	**10 s**	**15 s**	**30 s**	**60 s**
Inglés	48.16 %	64.50 %	78.00 %	81.66 %	85.00 %	95.00 %	90.00 %
Español	54.83 %	60.50 %	68.00 %	75.00 %	70.00 %	80.00 %	90.00 %
Francés	68.50 %	85.00 %	92.00 %	91.66 %	95.00 %	100 %	100 %
Alemán	40.50 %	61.50 %	71.00 %	75.00 %	80.00 %	75.00 %	80.00 %
Ruso	49.83 %	67.50 %	73.00 %	76.66 %	80.00 %	80.00 %	90.00 %
Japonés	69.50 %	89.50 %	94.00 %	96.66 %	95.00 %	100 %	100 %
Eficiencia	55.22 %	71.41 %	79.16 %	82.77 %	84.16 %	88.33 %	91.66 %

Tabla B.16: Sistema LID. Voz: masculina, Coeficientes: SDC, GMMs: 512

Idioma	Duración del Segmento de Voz						
	1 s	**3 s**	**6 s**	**10 s**	**15 s**	**30 s**	**60 s**
Inglés	51.66 %	67.00 %	83.00 %	86.66 %	87.50 %	95.00 %	100 %
Español	57.00 %	64.50 %	73.00 %	76.66 %	77.50 %	80.00 %	90.00 %
Francés	72.33 %	87.50 %	92.00 %	95.00 %	97.50 %	100 %	100 %
Alemán	47.83 %	69.50 %	81.00 %	83.33 %	85.00 %	85.00 %	90.00 %
Ruso	54.33 %	71.00 %	77.00 %	83.33 %	85.00 %	90.00 %	100 %
Japonés	76.00 %	93.50 %	95.00 %	96.66 %	97.50 %	100 %	100 %
Eficiencia	59.85 %	75.50 %	83.50 %	86.94 %	88.33 %	91.66 %	96.66 %

Tabla B.17: Sistema LID. Voz: masculina, Coeficientes: SDC, GMMs: 1024

C

Eficiencias de los Modelos Acústicos de Idiomas del Sistema LID Independiente del Género del Locutor

Idioma	Duración del Segmento de Voz						
	1 s	**3 s**	**6 s**	**10 s**	**15 s**	**30 s**	**60 s**
Inglés	43.83 %	51.00 %	56.00 %	56.66 %	52.50 %	55.00 %	60.00 %
Español	42.00 %	48.50 %	51.00 %	50.00 %	52.50 %	50.00 %	50.00 %
Francés	41.00 %	46.50 %	50.00 %	56.66 %	52.50 %	50.00 %	60.00 %
Alemán	58.00 %	69.50 %	78.00 %	76.66 %	80.00 %	85.00 %	90.00 %
Ruso	54.00 %	69.50 %	75.00 %	78.33 %	77.50 %	75.00 %	70.00 %
Japonés	70.66 %	82.50 %	85.00 %	88.33 %	87.50 %	95.00 %	90.00 %
Eficiencia	51.58 %	61.25 %	65.83 %	67.77 %	62.58 %	69.16 %	70.00 %

Tabla C.1: Sistema LID. Coeficientes: MFCC, GMMs: 128

Idioma	Duración del Segmento de Voz						
	1 s	**3 s**	**6 s**	**10 s**	**15 s**	**30 s**	**60 s**
Inglés	13.66 %	10.50 %	6.00 %	5.00 %	5.00 %	0.00 %	0.00 %
Español	56.83 %	74.50 %	78.00 %	81.66 %	87.50 %	90.00 %	90.00 %
Francés	25.33 %	43.00 %	52.00 %	55.00 %	57.50 %	55.00 %	60.00 %
Alemán	29.50 %	38.50 %	44.00 %	50.00 %	47.50 %	60.00 %	60.00 %
Ruso	27.50 %	26.00 %	29.00 %	23.33 %	25.00 %	20.00 %	20.00 %
Japonés	73.66 %	93.00 %	99.00 %	100 %	100 %	100 %	100 %
Eficiencia	37.74 %	47.58 %	51.33 %	61.99 %	53.85 %	54.16 %	55.00 %

Tabla C.2: Sistema LID. Coeficientes: SDC, GMMs: 128

Idioma	Duración del Segmento de Voz						
	1 s	**3 s**	**6 s**	**10 s**	**15 s**	**30 s**	**60 s**
Inglés	46.00 %	54.50 %	60.00 %	60.00 %	60.00 %	60.00 %	60.00 %
Español	47.83 %	53.50 %	57.00 %	55.00 %	60.00 %	60.00 %	60.00 %
Francés	39.83 %	47.50 %	52.00 %	58.33 %	62.50 %	50.00 %	60.00 %
Alemán	58.66 %	71.50 %	78.00 %	83.33 %	87.50 %	90.00 %	90.00 %
Ruso	54.00 %	72.00 %	78.00 %	81.66 %	77.50 %	85.00 %	80.00 %
Japonés	71.33 %	84.50 %	85.00 %	91.66 %	90.00 %	95.00 %	100 %
Eficiencia	52.94 %	63.87 %	68.33 %	71.66 %	72.91 %	73.33 %	75.00 %

Tabla C.3: Sistema LID. Coeficientes: MFCC, GMMs: 256

Idioma	Duración del Segmento de Voz						
	1 s	**3 s**	**6 s**	**10 s**	**15 s**	**30 s**	**60 s**
Inglés	58.16 %	76.50 %	88.00 %	93.33 %	92.50 %	100 %	100 %
Español	63.33 %	86.00 %	88.00 %	93.33 %	95.00 %	95.00 %	90.00 %
Francés	28.50 %	51.00 %	62.00 %	60.00 %	70.00 %	65.00 %	70.00 %
Alemán	26.00 %	42.50 %	50.00 %	58.33 %	62.50 %	80.00 %	80.00 %
Ruso	37.83 %	48.50 %	54.00 %	58.33 %	65.00 %	70.00 %	60.00 %
Japonés	60.33 %	81.50 %	97.00 %	100 %	100 %	100 %	100 %
Eficiencia	45.69 %	64.33 %	73.16 %	77.22 %	80.83 %	85.00 %	83.33 %

Tabla C.4: Sistema LID. Coeficientes: SDC, GMMs: 256

Idioma	Duración del Segmento de Voz						
	1 s	**3 s**	**6 s**	**10 s**	**15 s**	**30 s**	**60 s**
Inglés	49.75 %	61.00 %	70.50 %	73.33 %	73.50 %	80.00 %	80.00 %
Español	41.16 %	47.00 %	49.00 %	50.00 %	53.75 %	55.00 %	70.00 %
Francés	60.33 %	67.00 %	73.50 %	78.33 %	78.75 %	75.00 %	85.00 %
Alemán	67.08 %	79.25 %	83.50 %	86.66 %	85.00 %	90.00 %	90.00 %
Ruso	55.66 %	71.25 %	77.50 %	80.83 %	83.75 %	87.50 %	90.00 %
Japonés	55.50 %	67.75 %	67.00 %	70.00 %	71.25 %	70.00 %	75.00 %
Eficiencia	54.91 %	65.54 %	70.16 %	73.19 %	74.33 %	76.25 %	81.66 %

Tabla C.5: Sistema LID. Coeficientes: MFCC, GMMs: 512

Idioma	Duración del Segmento de Voz						
	1 s	**3 s**	**6 s**	**10 s**	**15 s**	**30 s**	**60 s**
Inglés	57.50 %	78.25 %	89.50 %	94.16 %	97.50 %	100 %	100 %
Español	67.25 %	86.50 %	92.00 %	92.50 %	91.25 %	97.50 %	100 %
Francés	49.41 %	71.00 %	82.50 %	84.16 %	83.75 %	87.50 %	95.00 %
Alemán	30.91 %	46.50 %	57.50 %	60.83 %	65.00 %	70.00 %	65.00 %
Ruso	39.16 %	50.50 %	54.50 %	63.33 %	70.00 %	70.00 %	80.00 %
Japonés	49.58 %	65.00 %	74.00 %	80.83 %	82.50 %	87.50 %	95.00 %
Eficiencia	48.96 %	66.29 %	75.00 %	79.30 %	81.66 %	85.41 %	89.16 %

Tabla C.6: Sistema LID. Coeficientes: SDC, GMMs: 512

Idioma	Duración del Segmento de Voz						
	1 s	**3 s**	**6 s**	**10 s**	**15 s**	**30 s**	**60 s**
Inglés	49.50 %	63.00 %	72.50 %	74.16 %	75.00 %	82.50 %	90.00 %
Español	47.16 %	55.75 %	62.00 %	60.00 %	61.25 %	62.50 %	75.00 %
Francés	63.16 %	70.50 %	78.00 %	81.66 %	81.25 %	82.50 %	85.00 %
Alemán	67.08 %	78.25 %	82.00 %	86.66 %	86.25 %	90.00 %	90.00 %
Ruso	58.33 %	73.75 %	78.00 %	81.66 %	85.00 %	85.00 %	90.00 %
Japonés	56.33 %	64.00 %	70.00 %	69.16 %	72.50 %	75.00 %	70.00 %
Eficiencia	56.92 %	67.54 %	73.75 %	75.55 %	76.87 %	79.58 %	85.00 %

Tabla C.7: Sistema LID. Coeficientes: MFCC, GMMs: 1024

Idioma	Duración del Segmento de Voz						
	1 s	**3 s**	**6 s**	**10 s**	**15 s**	**30 s**	**60 s**
Inglés	59.30 %	81.75 %	92.00 %	96.60 %	96.25 %	100 %	100 %
Español	69.30 %	87.50 %	93.50 %	95.00 %	95.00 %	97.50 %	100 %
Francés	54.00 %	77.25 %	88.00 %	89.16 %	92.00 %	90.00 %	95.00 %
Alemán	38.75 %	57.75 %	70.50 %	74.16 %	75.00 %	82.50 %	90.00 %
Ruso	41.50 %	56.25 %	63.50 %	68.30 %	76.25 %	77.50 %	95.00 %
Japonés	55.30 %	75.00 %	82.50 %	88.30 %	88.75 %	95 %	100 %
Eficiencia	53.02 %	72.58 %	81.66 %	85.25 %	87.20 %	90.41 %	96.66 %

Tabla C.8: Sistema LID. Coeficientes: SDC, GMMs: 1024

Idioma	Duración del Segmento de Voz						
	1 s	**3 s**	**6 s**	**10 s**	**15 s**	**30 s**	**60 s**
Inglés	51.16 %	64.75 %	73.00 %	77.50 %	76.25 %	90.00 %	95.00 %
Español	51.66 %	61.00 %	65.50 %	68.33 %	66.25 %	67.50 %	75.00 %
Francés	65.50 %	72.25 %	79.50 %	82.50 %	82.50 %	80.00 %	85.00 %
Alemán	68.08 %	79.25 %	84.50 %	87.50 %	86.25 %	90.00 %	90.00 %
Ruso	58.00 %	72.00 %	76.50 %	78.30 %	81.25 %	82.50 %	90.00 %
Japonés	58.58 %	65.00 %	69.50 %	70.00 %	71.25 %	75.00 %	70.00 %
Eficiencia	58.83 %	69.04 %	74.75 %	77.35 %	77.29 %	80.83 %	85.83 %

Tabla C.9: Sistema LID. Coeficientes: MFCC, GMMs: 2048

UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

FACULTAD DE INGENIERÍA

Sistema de identificación automática del lenguaje hablado en archivos multimedia de voz

TESIS

Que para obtener el título de

Ingeniero en Telecomunicaciones

P R E S E N T A

Mauricio Michel Olvera Zambrano

DIRECTOR DE TESIS

M.I. Larry Hipólito Escobar Salguero

Ciudad Universitaria, Cd. Mx., 2017