Caracterización de voz empleando análisis tiempo-frecuencia aplicada al reconocimiento de emociones

Trabajo de Grado
Presentado por

Christian Duque Sánchez
Mauricio Morales Pérez

como requerimiento parcial para optar al título de
Ingeniero Electricista

Facultad de Ingenierías Eléctrica, Electrónica, Física y Ciencias de la Computación
Programa de Ingeniería Eléctrica
Línea de Instrumentación y Control
Universidad Tecnológica de Pereira - UTP
Abril 2007

Caracterización de voz empleando análisis tiempo-frecuencia aplicada al reconocimiento de emociones

Aprobado por:

Alberto Ocampo, Director del programa

Julían David Echeverry Correa., Director

Primer Jurado

Fecha de aprobación

Resumen

La voz además de permiter comunicarnos (habla), es también una señal biológica que contiene información extra-lingüística sobre características físicas, estados fisiológicos y emocionales. Se presenta en este trabajo una metodología para la caracterización de la señal de voz aplicada en el reconocimiento de estados emocionales. Los diferentes estados emocionales de un hablante producen cambios fisiológicos en el aparato fonador, lo que se ve reflejado en la variación de dichas características. Las técnicas empleadas en el análisis de la señal de voz se pueden dividir en dos categorías: **Transformadas Tiempo-Frecuencia** y **Análisis Paramétrico**. La primera de estas categorías hace referencia a la representación de la señal en espacios conjuntos del tiempo y la frecuencia, permitiendo conocer la ubicación temporal del contenido espectral, esta técnica es efectiva en el tratamiento de señales no estacionarias como es la señal de voz. El análisis paramétrico busca estimar un modelo matemático que de forma aproximada represente el sistema de producción vocal.

Este documento se divide en la siguiente forma: en el **Capítulo 1** se describe la fisiología del mecanismo de producción de voz y la naturaleza de los sonidos. En el **Capítulo 2** se hace una introducción a la naturaleza de las emociones, su clasificación y efectos en el habla. En el **Capítulo 3** se encuentran las técnicas de análisis empleadas para la extracción de características. En el **Capítulo 4** se hace una descripción de las características extraídas de acuerdo a la técnica de análisis empleada. En el **Capítulo 5** se desarrolla la metodología de evaluación y se muestran los resultados obtenidos.

Abstract

The voice besides contains information that allows comunication (speech) is also a biological signal that contains information about phisycal features, functional and emotional states. This work presents a methodology for the characterization of voice signal applied to the recognition of emotional states. The different emotional states produce physiologyc changes in vocal producction system, wich are reflected in the variation of these features. The techniques used in the voice signal analysis can be divided in two categories: **Time-Frequency Representation** and **Parametric Analysis**. The first one makes reference to the signal representation in a time-frequency joint domain, this alternative let know the spectral content in the temporal domain without losing the resolution in any of the two spaces, it makes good to the analysis of non-stationary signal like the voice. The second category search to estimate mathematic models that represents the vocal production system, reducing the number of parameters used in the analysis. This document is divided in the following way: **Chapter 1** descibes the physiology of the vocal production mechanism and the sound nature. **Chapter 2** makes an introducction of emotion nature and effects in the speech. In **Chapter 3** are found the analysis technique employed in features extraction. **Chapter 4** makes a description of the extracted features according to the analysis technique employed. **Chapter 5** develop the evaluation methodology and saw the obtained results.

Dedicatoria

A Dios por regalarnos un poco de su sabiduría, a nuestros padres por su apoyo, esmero y dedicación para poder salir adelante en ésta etapa de nuestras vidas, a nuestros hermanos por estar ahí cuando más los necesitamos.

A Natalia (cachetona) por entregarme su tiempo, corazon y cada detalle que me hizo fuerte en los momentos más dificiles (in lake-ch).

A Paola por su apoyo, paciencia y entrega incondicional.

Tabla de Contenido

Lista de Tablas

Lista de Figuras

Agradecimientos

A Julian, por todo el tiempo dedicado, la paciencia, pero sobre todo por la excelente persona que es.

A Carlos Andrés por la colaboración para que este proyecto empezara a hacerse realidad

Este trabajo se desarrolló en el marco del siguiente proyecto de investigación:

- "IMPLEMENTACIÓN Y EFECTIVIDAD DE UN SISTEMA BASADO EN INTELIGENCIA ARTIFICIAL COMO HERRAMIENTA PARA EL TRATAMIENTO PSICOLÓGICO DE PERSONAS CON TRASTORNO DE ESTRÉS POSTRAUMATICO", financiado por Colciencias. Codigo 111037019600 y la Universidad Tecnológica de Pereira. Código 511-3-243-08

Notaciones

Notación	Significado
$s(n)$	Señal discreta
Δ_ω	Ancho de banda de duración para frecuencia
Δ_t	Ancho de banda de duración para tiempo
TFR	Transformadas Tiempo-Frecuencia
$\psi(t)$	Wavelet Madre
$W_\psi s(a,b)$	Transformada Wavelet Continua de $s(t)$
$DW_\psi s(j,k)$	Transformada Wavelet Discreta de $s(t)$
W_j	Subespacios de Wavelet
$WV_\psi s(t,f)$	Transformada Wigner Ville de $s(t)$
$R(\tau)$	Función de Autocorrelación
$sinc(f)$	Función sinc
E_n	Error Cuadrático Medio en la muestra n
$\phi_n(i,k)$	Función de Covarianza en la muestra n

Objetivos

Objetivo general

Desarrollar una técnica de extracción de características en señales de voz mediante análisis en tiempo-frecuencia para el reconocimiento de estados emocionales.

Objetivos específicos

- Determinar los metodos matemáticos que extraigan con mayor efectividad los valores de las características o parámetros de una señal de voz.
- Confrontar las características extraídas, con aquellas que se presentan en estudios e investigaciones previos de reconocimiento de estados emocionales.
- Desarrollar un método de evaluación de la metodología propuesta que garantice la efectividad de las técnicas de extracción de caracteristicas, comprobando su funcionalidad sobre la base de datos SES (*Spanish Emotional Speech* de la lengua española en señales de voz.

CAPÍTULO 1

El sistema de producción vocal

La comunicación humana surgió en el momento en el que nuestros antepasados, en su lucha por la supervivencia y en respuesta a sus instintos se vieron obligados a transmitir a quienes les rodeaban, sus impresiones, necesidades, emociones, etc. Para ello se valieron primero del lenguaje biológico: la mímica y las interjecciones, posteriormente por el lenguaje hablado: la voz, hasta llegar a manifestaciones escritas y demás. La comunicación oral, en un sentido más amplio, es la expresión de nuestros pensamientos por medio de la palabra hablada con fines comunicativos, y tiene algunas ventajas prácticas con respecto a la comunicación escrita [1], las cuales son:

- Facilidad: Es el mecanismo natural de comunicación humana.
- Aprendizaje: Es el mecanismo más precoz de comunicación.
- Expresión: Se puede transmitir información extra-lingüística generada por manifestaciones fisiológicas de estados anímicos o patologías.

La voz es el resultado final de un proceso físico que se realiza voluntariamente en el aparato fonador en busca de satisfacer la necesidad de comunicación. Es una señal acústica (sonido), es decir, una onda de presión longitudinal formada por la compresión y expansión de las moléculas de aire que se transmite en dirección paralela a la aplicación de la energía; en la zonas de compresión existe mayor densidad de moléculas las cuales son forzadas por acción de la energía, mientras que en las zonas de expansión existe una menor densidad de moléculas. Usualmente se representa como una onda seno, como se muestra en la figura 1. Las crestas representan los momentos de mayor compresión y los valles los momentos de mayor expansión.

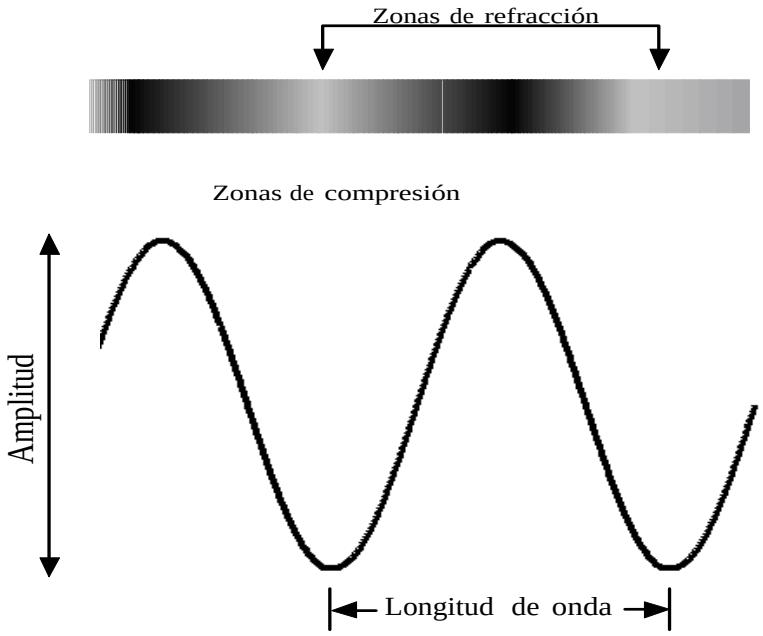


Figura 1: Onda de presión sonora

1.1 El aparato fonador

En el proceso de producción intervienen órganos del sistema respiratorio y digestivo, los cuales son controlados por el sistema nervioso central [2, 3]. Esencialmente es generada por una excitación en las cuerdas vocales, que se propaga a través de la faringe y las cavidades bucal y nasal. Estas cavidades actúan como cavidades resonantes, y su forma determina las características acústicas de la señal de voz [4]. El sistema fonador se puede dividir en tres bloques:

- **Sistema de generación:** Los músculos abdominales y torácicos aumentan la presión en los pulmones produciendo un exceso en la corriente de aire, ésta sale por los bronquios y la tráquea hasta llegar a la laringe donde es excitado el sistema de vibración.
- **Sistema de vibración:** Está conformado básicamente por las cuerdas vocales, las cuales se dividen en dos pares superiores e inferiores, de estas, sólo las últimas participan en la producción de voz. En el caso de la respiración las cuerdas se abren y se recogen a los lados permitiendo el libre paso del aire, si por el contrario se encuentran juntas y tensas el aire choca haciendo que se produzcan los diferentes sonidos.
- **Sistema resonante:** lo componen tres cavidades articulatorias: cavidad faríngea, cavidad oral y cavidad nasal. Los sonidos producidos por el sistema de vibración se desplazan desde las cuerdas vocales hasta los orificios nasales

y la boca, la articulación de las cavidades modifica y amplifica los sonidos que finalmente son expulsados al exterior.

El conjunto de órganos que intervienen en la fonación (ver figura 2) pueden dividirse en tres grupos bastante bien delimitados [4].

- Cavidades infraglóticas.
- Cavidad laríngea.
- Cavidades supraglóticas.

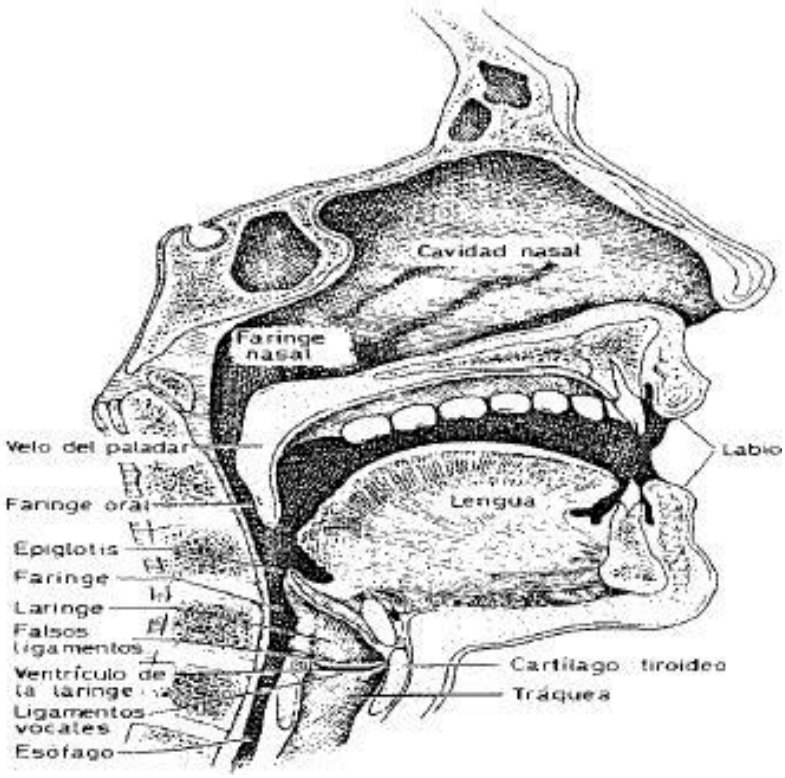


Figura 2: Sistema fonador

1.1.1 Cavidades infraglóticas

Las cavidades infraglóticas constan de los órganos propios de la respiración (pulmones, bronquios, y tráquea), que son la fuente de energía para todo el sistema de producción de voz. En el proceso de inspiración, los pulmones toman aire, bajando el diafragma y agrandando la cavidad torácica. En el momento de la fonación, la espiración, provocada por la contracción de los músculos intercostales y del diafragma, aporta la energía necesaria para generar la onda de presión acústica que atravesará los órganos fonadores superiores.

1.1.2 Cavidad laríngea

La cavidad laríngea es la responsable de modificar el flujo de aire generado por los pulmones y convertirlo o no, en una señal susceptible de excitar adecuadamente las posibles configuraciones de las cavidades supraglóticas. El último cartílago de la tráquea, el cricoides, forma la base de la laringe, cuyo principal órgano son las cuerdas vocales que son dos pares de repliegues compuestos de ligamentos y músculos. El par inferior son las llamadas cuerdas vocales verdaderas, que pueden juntarse o separarse mediante la acción de los músculos crico-aritenoides lateral y posterior, y que están protegidas en su parte anterior por el cartílago tiroides, el más importante de la laringe, abierto por su parte posterior. Finalmente, la parte superior de la laringe está unida al hueso hioides.

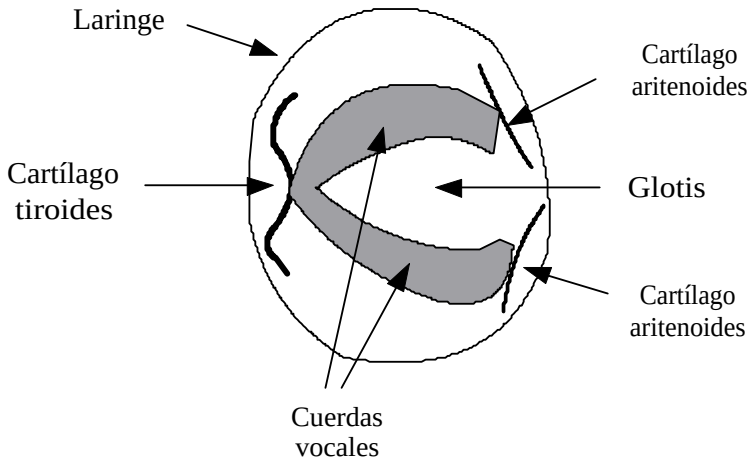


Figura 3: Cuerdas vocales

En la figura 3 se muestra una vista longitudinal simplificada de la zona en la que se encuentran las cuerdas vocales, en su posición extrema abiertas. A la apertura que queda entre las cuerdas vocales se le denomina glotis. La cavidad laríngea está terminada por la epiglotis, un cartílago en forma de cuchara, que permite cerrar la apertura de la laringe en el acto de la deglución.

1.1.3 Cavidades supraglóticas

Las cavidades supraglóticas, también llamadas tracto vocal están constituidas por la faringe, la cavidad oral y la cavidad nasal. Su misión fundamental, de cara a la fonación, es perturbar adecuadamente el flujo de aire procedente de la laringe, para dar lugar finalmente a la señal acústica generada a la salida de la nariz y la boca. La faringe es una cavidad en forma tubular que une la laringe con las cavidades bucal y nasal, y que suele dividirse en tres partes: faringe laríngea, faringe bucal (boca) y faringe nasal, las dos últimas separadas por el velo del paladar. El volumen de la

faringe laríngea puede ser modificado por los movimientos de la laringe, la lengua y la epiglotis mientras que el volumen de la faringe bucal se modifica por el movimiento de la lengua. Dentro del tracto vocal existen lugares de articulación (ver figura 4) donde los sonidos pueden sufrir cambios a nivel temporal estos cambios se relacionan directamente con la salida de la voz y los fenómenos transitorios que los acompañan.

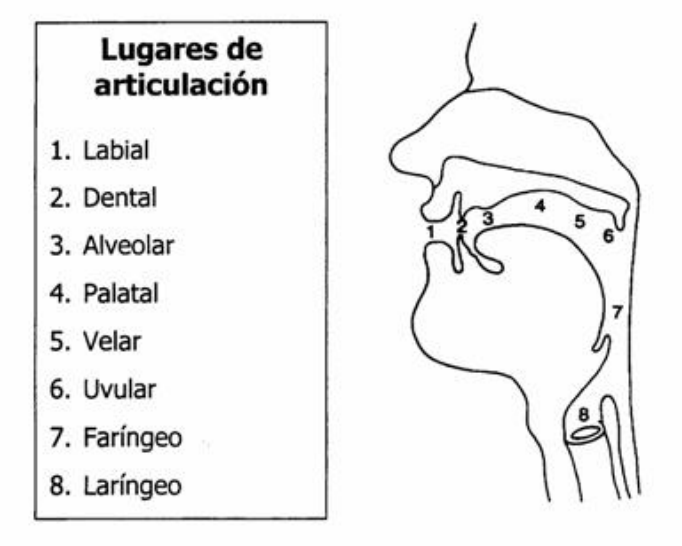


Figura 4: Lugares de articulación

La faringe nasal y las restantes cavidades nasales forman, desde el punto de vista de su acción sobre el flujo de aire procedente de la faringe, un resonador que puede o no conectarse al resonador bucal mediante la acción del velo del paladar. Según el resonador nasal esté o no conectado, el sonido será nasal u oral, respectivamente. En una descripción de la cavidad bucal se pueden señalar las siguientes partes:

- Labios
- Dientes
- Zona alveolar (entre dientes y paladar duro)
- El paladar (paladar duro y paladar blando o velo)

La raíz de la lengua forma la pared frontal de la faringe laríngea, y sus movimientos le permiten modificar la sección de la cavidad bucal (movimiento vertical), adelantar o retrasar su posición frente a la de reposo (movimiento horizontal), así como poner en contacto su ápice o la parte trasera con alguna zona del paladar. El movimiento de los labios también interviene en la articulación, pudiendo ser de apertura o cierre y de protuberancia, alargando en este último caso la cavidad bucal [1].

1.2 Producción de voz

El mecanismo de producción de la voz se inicia en los pulmones; el aire sale expulsado de ellos hacia la laringe (atravesando la tráquea y la glotis) a diferente presión en función del sonido que se desea generar. La glotis separa las cuerdas vocales y se mantiene abierta mientras se respira, pero en el momento de producir sonidos se va estrechando de manera intermitente. La velocidad con la que las cuerdas vocales se abren y se cierran está ligada con lo que se conoce como frecuencia fundamental. Tras superar la glotis, el aire se acerca al tracto vocal, el cual varía su forma dependiendo de los sonidos a generar. El tracto vocal es una caja de resonancias, cuya forma, y por lo tanto su respuesta, varían de acuerdo a la posición de los órganos articuladores (lengua, labios, mandíbula, velo del paladar). Las resonancias producidas tienen su energía concentrada alrededor de determinadas frecuencias del espectro, a las que se conoce como formantes [5].

1.2.1 Modelo general de producción de voz

En la figura 5 se muestra una representación simplificada del mecanismo fisiológico completo de producción de voz. La función primaria es la inhalación, posible gracias a la expansión de la cavidad torácica, mediante la cual desciende la presión en los pulmones y entra el aire a través de las fosas nasales o bien por vía bucal. La energía necesaria para expulsar el aire reside en los músculos torácicos y abdominales (representados en la figura por un pistón). Cuando la cavidad torácica se contrae aumentando la presión en los pulmones, el aire sale expelido, pasa a través de los bronquios y de la tráquea, y actúa como excitación del tracto vocal. En función de lo que ocurra después hay dos tipos elementales de sonido; sonoros y sordos. Para la voz sonora, las cuerdas vocales son tensadas y forzadas a vibrar por el paso de un flujo de aire. Dicho flujo es troceado en pulsos cuasi periódicos que son, entonces, modulados en frecuencia al pasar por la faringe, la cavidad bucal, y, en ocasiones, la cavidad nasal, generando voz sonora. En la voz sorda, los fonemas se producen por una excitación debida a un flujo de aire que, en algún punto del tracto vocal (normalmente cerca de la abertura bucal), por la acción de una obstrucción parcial o total se convierte en turbulento.

El modelo general de producción de voz se basa en la idea de modelar el tracto vocal como una concatenación de tubos (modelo de tubos) de sección variable para obtener una función de transferencia del mismo. El desarrollo de este modelo se presenta en el apéndice A.

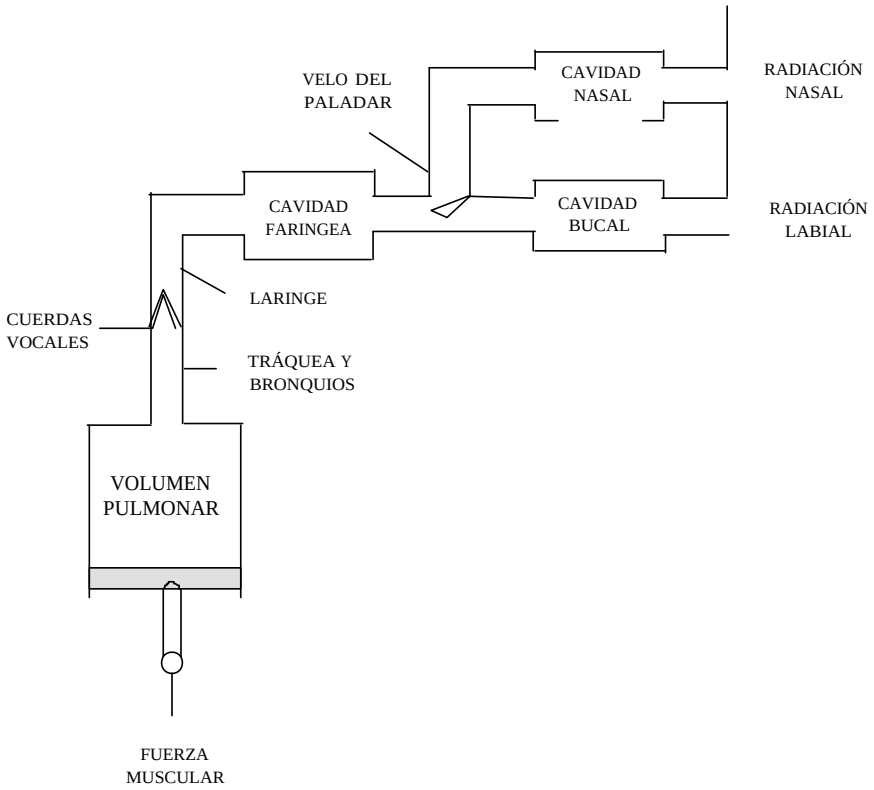


Figura 5: Esquema del mecanismo de producción de voz

1.3 Los sonidos

Desde el punto de vista de su forma de generación podemos clasificar los sonidos en:

- Sonoros: para los sonidos denominados sonoros (como las vocales), la excitación está formada por un tren de pulsos cuasi-periódico generado por la vibración de las cuerdas vocales.
- Sordos: también llamados no-sonoros (como las consonantes fricativas /s/ /z/), la excitación es una señal aleatoria generada por el flujo turbulento del aire a través de las cuerdas vocales en relajación.
- Plosivos: también llamados oclusivos (como las consonantes /p/, /t/, etc.) la excitación se genera mediante una oclusión del tracto vocal seguida de un súbito aumento de presión y una relajación, lo que genera una excitación transitoria.

La sección transversal de las cavidades bucal y nasal varía dependiendo de la posición de los elementos articulatorios (lengua, labios, etc.), y modifica las características acústicas de la señal generada [2, 4].
La mínima unidad lingüística capaz de producir cambios de significados se llama fonema. Tanto los fonemas vocálicos como los fonemas consonánticos pueden ser

clasificados según el estado vibratorio de las cuerdas vocales en fonemas sonoros y fonemas no sonoros.

1.3.1 Las vocales

Existen varias teorías con respecto a la producción de las vocales [6]. Una de ellas se relaciona a cada abertura de la glotis. Una cierta cantidad de aire penetra en la cavidad bucal, la cual excita el aire de la cavidad en una vibración libre amortiguada; cuando ésta vibración se repita periódicamente e irradia fuera de la cavidad como un sonido, el resultado es una vocal. Otra teoría se asocia a la idea de cómo un tono laríngeo resuena con la cavidad bucal y aquellos de sus tonos parciales armónicos cuyas frecuencias naturales de la cavidad son reforzadas, mientras que las otras son debilitadas; el resultado es una vocal. Ambas teorías no se excluyen, sino se complementan.
Las vocales son fonemas sonoros los cuales pueden distinguirse entre sí mismos por la radiación de salida y sus formantes, donde las dos primeras formantes bastan para caracterizar el timbre de las vocales, por lo tanto, si a los fonemas vocálicos se les asocia una articulación, que consiste en la modificación de la acción filtrante de los diversos resonadores, lo cual depende de las posiciones de la lengua (tanto en elevación como en profundidad o avance), de la mandíbula inferior, de los labios y del paladar blando, se pueden clasificar con base a dos parámetros (ver figura **??**):

- **Apertura del tracto vocal**:
  - Abierta: lengua totalmente separada del paladar /a/
  - Media: lengua a una distancia intermedia del paladar /e/o/
  - Cerrada: lengua muy cerca del paladar /i/u/
- **Grado de elevación del dorso de la lengua**:
  - Anterior: la lengua se aproxima a la región delantera del paladar /e/i/
  - Central: la lengua se encuentra en la parte central del paladar /a/
  - Posterior: la lengua de aproxima a la zona velar /o/u/

1.3.2 Las consonantes

Las consonantes pueden ser fonemas sonoros o no sonoros. Son producidas con una estrecha constricción en alguna región del flujo de aire arriba de la laringe en donde para las vocales el flujo de aire es mayor. De acuerdo a la articulación y a sus características pueden ser clasificadas con base a cuatro grupos:

1. **Lugar de articulación** Son los lugares de estrechamiento en las cavidades del tracto vocal controlados por los órganos móviles contra los órganos inmóviles de su misma zona. Una clasificación de las consonantes respecto a los lugares de articulación es la siguiente:

- Bilabial: labios superior e inferior en contacto durante la producción: /p/b/m/
- Labiodental: incisivos superiores con labio inferior: /f/
- Linguodental: ápice de la lengua con incisivos superiores: /t/d/
- Linguointerdental: ápice de la lengua se sitúa en posición interdental: /z/
- Linguoalveolar: ápice de la lengua contacta con los alvéolos: /s/n/l/r/rr/
- Linguopalatal: ápice de la lengua contacta con el paladar: /ch/ll/y/ñ/
- Linguovelar: el post-dorso de la lengua contacta con el velo del paladar: /j/ga/k/

2. **Modo de articulación** Según el modo de articulación las consonantes se pueden clasificar en:
   - Oclusiva: el sonido se produce en dos fases, cierre del tracto seguido de apertura súbita (explosión): /p/b/t/d/k/g/m/n/ñ/
   - Fricativa: el aire encuentra un cierre parcial o total en algún punto del tracto, provocando una turbulencia: /f/s/z/j/
   - Africada: composición de oclusiva seguida de fricativa: /ch/
   - Vibrante: el ápice de la lengua se pone en vibración simple o múltiple: /r/rr/
   - Lateral: el aire sale por uno o ambos lados de la lengua: /l/ll/
3. **Función de las cuerdas vocales** Se asocia a la idea de si hay presencia o ausencia de las cuerdas vocales al momento de producir un fonema.
   - Sordas: no vibra la cuerda vocal. Turbulencias. Baja energía, alta frecuencia y poca estabilidad a corto plazo: /p/t/k/f/z/s/j/ch/
   - Sonoras: vibra la cuerda vocal. Periódico. La frecuencia fundamental es el llamado tono. Alta energía y estabilidad a corto plazo: /b/d/g/y/m/n/ñ/l/ll/r/a/e/i/o/u/
4. **Posición del velo del paladar** Se asocia a la idea de si existe o no radiación nasal.
   - Nasales: el velo del paladar está separado de la pared faríngea: /m/n/ñ/
   - Orales: el velo del paladar está unido a la pared faríngea y no permite el paso de aire hacia la cavidad nasal: /p/t/k/f/z/s/ch/j/b/d/g/y/ll/r/rr/a/e/i/o/u/

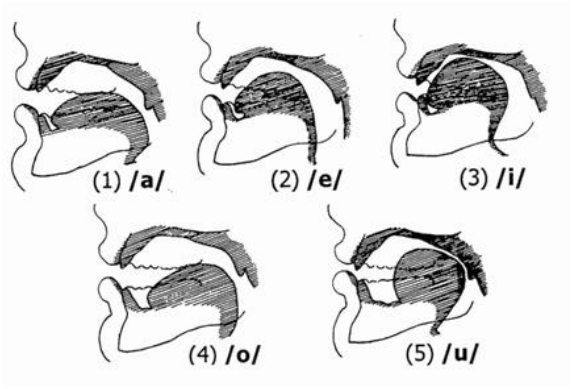


Figura 6: Relaciones articulatorias orales de los sonidos vocálicos

CAPÍTULO 2

El habla y las emociones

La voz es el principal modo de comunicación entre los hombres; además de contener información sobre el mensaje que se desea transmitir, contiene información sobre las emociones y estados fisiológicos del hablante; consecuentemente se ha estudiado los mecanismos de producción de voz humana y se han creado sistemas capaces de simular y reconocer voz electrónicamente. La voz no es otra cosa que un sonido y como tal, se caracteriza por una serie de elementos:

- La intensidad: Es equivalente al volumen. El aire al salir de los pulmones golpea la glotis y produce vibraciones. Cuanto más amplias sean, mayor será la fuerza. Se mide en decibelios (dB) y para tener una referencia, una conversación normal ronda entre los 50 dB. Tiene efectos en el oyente porque transmite emociones. Un volumen de voz alto se asocia a la agresividad, nerviosismo, tensión y lejanía. Al contrario, un volumen bajo puede sugerir depresión, cansancio y proximidad.

- El tono: Está relacionado con la cantidad de vibraciones que posee una onda de sonido. A mayor número más aguda será la voz. Estas vibraciones se producen en el ser humano en la laringe y se miden en Hertzios o Hertz (Hz). Las voces masculinas oscilan entre los 75Hz y los 200Hz. Las femeninas entre los 150 Hz y los 300Hz (ver figura 7).

- El timbre: Es lo que permite que distingamos entre dos sonidos de igual intensidad y tono. El aire que sale de los pulmones, recorre y choca con la laringe, labios, dientes y lengua; tiene peculiaridades únicas dependiendo de la morfología de cada persona. Esta característica es el carnet de identidad de cualquier voz. Aporta mucha información real o imaginaria sobre la edad, la apariencia física e incluso una especie de retrato de la personalidad del hablante.

Figura 7: Relación tonos-emociones

Uno de los mayores problemas encontrados en los estudios sobre el habla ha sido el de la variabilidad en ésta. Se ha demostrado que varios aspectos del estado físico y emocional del locutor, incluyendo edad, sexo, inteligencia, apariencia y personalidad pueden identificarse solamente por la voz [7, 8]. Todos estos factores, que son diferentes para cada interlocutor, contribuyen a la variabilidad del habla.

2.1 Los efectos de las emociones en el habla

Las primeras investigaciones sobre cómo afectaban las emociones al comportamiento y al lenguaje de los animales fueron descritas brevemente por Darwin [9]. Más recientemente, los efectos de las emociones en el habla han sido estudiados por investigadores acústicos que han analizado la señal de voz, por lingüistas que han estudiado los efectos léxicos y prosódicos y por psicólogos. Gracias a estos esfuerzos se ha conseguido identificar muchos de los componentes del habla que se utilizan para expresar emociones, dentro de los cuales se consideran los más importantes [10]:

- El pitch o frecuencia fundamental.
- La duración.
- La calidad de voz.

2.1.1 Pitch

El *pitch* es la frecuencia fundamental a la que las cuerdas vocales vibran, también llamada frecuencia fundamental o F_0. Se considera que las características de la frecuencia fundamental son una de las principales portadoras de la información sobre las emociones.

- El valor medio del pitch expresa el nivel de excitación del locutor. Una media elevada de F_0 indica un mayor grado de excitación.
- El rango del *pitch* es la distancia entre el valor máximo y mínimo de la frecuencia fundamental. Refleja también el grado de exaltación del locutor. Un rango más extenso que el normal refleja una excitación emocional o psicológica.
- Las fluctuaciones en el *pitch* descritas como la velocidad de las fluctuaciones entre valores altos y bajos y si son abruptas o suaves son producidas psicológicamente. En general, la curva de tono es discontinua para las emociones consideradas como negativas (miedo, enfado) y es suave para las emociones positivas (por ejemplo la alegría).

2.1.2 Duración

La duración es la componente de la prosodia descrita por la velocidad del habla y la situación de los acentos, y cuyos efectos son el ritmo y la velocidad. El ritmo en el habla deriva de la situación de los acentos y de la combinación de las duraciones de las pausas y de los fonemas.
Las emociones pueden distinguirse por una serie de parámetros que conciernen a la duración, como son:

- **Velocidad de locución:** generalmente un locutor en estado de excitación acortará la duración de las sílabas, con lo que la velocidad de locución medida en sílabas por segundo o en palabras por minuto se incrementará.
- **Número de pausas y su duración:** un locutor exaltado tenderá a hablar rápidamente con menos pausas y más cortas, mientras que un locutor deprimido hablará más lentamente, introduciendo pausas más largas.
- Cociente entre el tiempo de locución y el de pausas.

2.1.3 Calidad de la voz

La intensidad, las irregularidades en la voz, el cociente entre energías a baja y alta frecuencia, *breathiness* y la laringerización son algunas de las características que diferencian la calidad de la voz.

- **Intensidad:** Está relacionada con la percepción del volumen y se refleja en la amplitud de la forma de onda

- **Irregularidades vocales:** Abarcan un gran rango de características vocales. El *jitter* vocal refleja las fluctuaciones de un pulso glotal al siguiente (como se observa en el enfado) o la desaparición de voz en algunas emociones como la pena, en la que el habla se convierte en un simple susurro.
- **El cociente entre energía de alta y baja frecuencia:** Gran cantidad de energía en las frecuencias altas se asocia con agitación (enfado), mientras que baja concentración de energía en las frecuencias altas se relaciona con depresión o calma (pena).
- ***Breathiness* y laringerización:** reflejan las características del tracto vocal que están más relacionados con la personalización de cada voz. *Breathiness* describe la generación de ruido respiratorio de forma de que la componente fundamental tiende a ser más fuerte, mientras que las frecuencias altas son reemplazadas por ruido aspiratorio. La laringerización se caracteriza por una vibración aperiódica de las cuerdas vocales, con un pulso glotal estrecho y *pitch* bajo, lo que se traduce en una voz chirriante.

2.2 Naturaleza de las emociones

Darwin vinculo la expresión sonora de las emociones al instinto o, cuanto menos, a conductas heredadas y, por tanto, que pueden aparecer en el individuo de modo totalmente involuntario [9]. Desde ese momento, la expresión emocional ha tendido a ser considerada como el carácter del habla humana más claramente universal y transcultural. De hecho, el carácter transcultural del habla emocional ha sido contrastado en diversas investigaciones muy posteriores [11, 12, 13, 14, 15]. En todos estos estudios se explora ese supuesto carácter universal del sonido oral de las emociones y, efectivamente, se encuentra y se constata. No obstante, en estos trabajos se detectaron también elementos de influencia cultural que han desdibujado parcialmente los planteamientos darwinianos iniciales.

Scherer considera que la expresión oral de las emociones se constituye como un sistema analógico-vocal conectado con los mecanismos biológicos y fisiológicos del individuo, al cual se ha sobrepuesto después, a lo largo de la evolución el sistema simbólico de la lengua [16]. Así, se trabaja desde la doble perspectiva hipotética siguiente:

1. la voz sufre cambios acústicos que están motivados directamente por las alteraciones fisiológicas que se producen en el cuerpo humano cuando el individuo experimenta una emoción.
2. los rasgos acústicos de la voz producidos por las emociones resultan parcialmente modificados durante una locución por la influencia específica de la lengua que está utilizando el hablante emocionado.

La emoción no es un fenómeno simple, sino que muchos factores contribuyen a ello. Izard declaró que una definición completa de emoción debe tener en cuenta el sentimiento consciente de la emoción, los procesos que ocurren en el sistema nervioso y en el cerebro y los modelos expresivos observables de emoción [17]. Las emociones se experimentan a veces cuando algo inesperado sucede y los efectos emocionales empiezan a tener el control en esos momentos. La emoción puede describirse también como la interfaz del organismo con el mundo exterior, señalando tres funciones principales de las emociones:

1. Reflejan la evaluación de la importancia de un estímulo en particular en términos de las necesidades del organismo, preferencias, intenciones...
2. Preparan fisiológica y físicamente al organismo para la acción apropiada.
3. Comunican el estado del organismo y sus intenciones de comportamiento a otros organismos que le rodean.

Emoción y estado de ánimo son conceptos diferentes: mientras las emociones surgen repentinamente en respuesta a un determinado estímulo y duran unos segundos o minutos, los estados de ánimo son más ambiguos en su naturaleza, perdurando durante horas o días. Las emociones pueden ser consideradas más claramente como algo cambiante y los estados de ánimo son más estables. Aunque el principio de una emoción puede ser fácilmente distinguible de un estado de ánimo, es imposible definir cuando una emoción se convierte en un estado de ánimo; posiblemente por esta razón, el concepto de emoción es usado como un término general que incluye al de estado de ánimo. Más allá de emociones y estados de ánimos está el rasgo a largo plazo de la personalidad, que puede definirse como el tono emocional característico de una persona a lo largo del tiempo.

Muchos de los términos utilizados para describir emociones y sus efectos son necesariamente difusos y no están claramente definidos; Esto es atribuible a la dificultad en expresar en palabras los conceptos abstractos de los sentimientos, que no pueden ser cuantificados. Por ello, para describir características de las emociones se utilizan un conjunto de palabras emotivas, siendo seleccionadas la mayoría de ellas por elección personal en vez de comunicar un significado estándar.

2.3 Clasificación de las emociones

Muchas teorías sobre emociones usan el concepto de emociones básicas, las cuales son fundamentales, siendo todas las demás emociones modificaciones o combinaciones de estas emociones básicas. Sin embargo, no hay consenso sobre cuáles constituyen las emociones básicas.
Las emociones se definen como un mecanismo flexible de adaptación a un ambiente

cambiante [7] . Pueden distinguirse los siguientes tipos fundamentales de emoción [18] .

- **Emociones extremas:** Este término denota una emoción totalmente desarrollada, la cual típicamente es intensa e incorpora la mayoría de los aspectos que se consideran relevantes en el síndrome de la emoción.

- **Emociones subyacentes:** Denotan el tipo de colorante emocional que es parte de la mayoría o de todos los estados mentales. De cara a seleccionar un subconjunto de emociones que permita realizar una tarea experimental, se han tenido en cuenta las siguientes clasificaciones teóricas:

  1. **Categorías emocionales:** Emplea palabras a la hora de clasificar las emociones. Esta teoría engloba a otras tres:
     - **Emociones básicas:** El número de emociones básicas suele ser pequeño (en los primeros estudios menos de 10, en los más recientes entre 10 y 20)

     - **Emociones súper ordinarias:** Existen una serie de categorías que son más fundamentales que otras en el sentido de que incluyen en sí mismas a las otras [19, 20] .

     - **Emociones esenciales del día a día:** La teoría se ejemplifica en el trabajo de [18] . Comenzando con una lista de términos emocionales de la literatura se insta a los sujetos a seleccionar un subconjunto que represente de manera apropiada las emociones relevantes de la vida diaria.

  2. **Descripciones basadas en psicología:** Según esta teoría el aspecto esencial de una emoción es el estado del cuerpo que tiene asociado [21, 22] .

  3. **Descripciones basadas en la evaluación:** Estas teorías describen las emociones desde el punto de vista de las evaluaciones [20] .

  4. **Dimensiones emocionales:** Las dimensiones emocionales son una representación simplificada de las propiedades esenciales de las emociones. Evaluación (positiva / negativa) y activación (activa / pasiva) son las dimensiones mas importantes, en algunas ocasiones se complementan con la dimensión poder (dominante / sumiso) [23] . Joel Davitz y Klaus Scherer clasificaron las emociones y sus efectos utilizando los ejes o dimensiones de un espacio semántico figura 8:

– **Potencia o fuerza(eje independiente):** corresponde a la atención - rechazo, distinguiendo entre emociones iniciadas por el sujeto a aquellas que surgen del ambiente (desde el desprecio al temor o la sorpresa).

– **Valencia, agrado o valoración (eje horizontal):** según lo placentero o desagradable de la emoción (desde la alegría hasta el enfado).

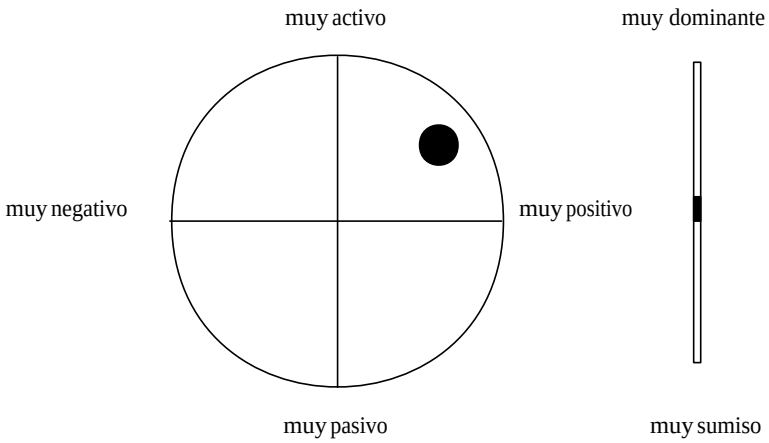


Figura 8: Representación de las emociones en el espacio semántico

– **Actividad (eje vertical):** presencia o ausencia de energía o tensión.

Marc Schroder habla sobre este mismo mapa de las emociones sugiriendo una representación de todos los estados emocionales en tres dimensiones [24] y aunque el enfoque del autor está dirigido hacia el desarrollo de mejores sistemas sintetizadores de voz, su propuesta da luces a las investigaciones enfocadas a identificar estados de emocionales.

Se ha encontrado en las descripciones de estados emocionales en tres dimensiones, correlación con las variables acústicas. En [25] se habla de estas correlaciones hechas a partir de una base de datos de emociones, encontrando por ejemplo que en el eje de la potencia correspondiente a un nivel alto va acompañado por un valor medio bajo de la frecuencia fundamental F_0 *pitch*.
En otros estudios se ha descubierto que se confunden más entre sí las emociones con un nivel similar de actividad (como por ejemplo la alegría y el enfado) que las que presentan similitud en términos de valencia o de fuerza.

También están relacionados el ritmo y la valencia de forma que los sentimientos "positivo" son expresados con un ritmo más regular que los sentimientos "negativo".
Esto lleva a la conclusión que la dimensión de la actividad está más correlacionada con las variables auditivas relativamente más simples de la voz, como pueden ser el tono y la intensidad, mientras que la valencia y la fuerza son probablemente comunicados por modelos más sutiles y complejos.
El habla neutra suele caracterizarse por un tono con un rango de variación estrecho y unas transiciones de F_0 suaves, además de una velocidad de locución alta.

2.3.1 Emociones primarias

Se conocen nuestras emociones gracias a su intromisión en las mentes conscientes, sea ésta bienvenida o no. Pero las emociones no evolucionaron como sentimientos conscientes. Evolucionaron como resultado de especializaciones de la conducta y fisiológicas: respuestas físicas controladas por el cerebro que permitieron sobrevivir a organismos antiguos en entornos hostiles y procrear [26].
Aún cuando se ha avanzado mucho en materia de definir las emociones, hasta hoy, la psicología continúa estudiando si alguna reacción en particular se identifica con una emoción específica. Por ejemplo, si los escalofríos son una reacción exclusiva del miedo.
Pese a esto, la mayoría de los especialistas está de acuerdo en trazar una línea en el conjunto de las emociones humanas y distinguir aquellas que son primarias (evolutivamente) de las que son secundarias.
Para Theodore Ribot, la personalidad envuelve en su profundidad el origen de la gran trinidad afectiva constituida por el miedo, la cólera y el deseo: son los tres instintos nacidos directamente de la vida orgánica: instinto defensivo, instinto ofensivo, instinto nutricio.
Desde este punto de partida se mantuvieron el miedo y la cólera, y se agregaron la alegría y la tristeza, cuatro emociones que poseen también los mamíferos superiores, y quedó entonces conformado un cuadro de cuatro emociones primarias, con su respectiva variedad de manifestaciones:

1. **Cólera:** enojo, mal genio, atropello, fastidio, molestia, furia, resentimiento, hostilidad, animadversión, impaciencia, indignación, ira, irritabilidad, violencia y odio patológico.
2. **Alegría:** disfrute, felicidad, alivio, capricho, extravagancia, deleite, dicha, diversión, estremecimiento, éxtasis, gratificación, orgullo, placer sensual, satisfacción y manía patológica.
3. **Miedo:** ansiedad, desconfianza, fobia, nerviosismo, inquietud, terror, preocupación, aprehensión, remordimiento, sospecha, pavor y pánico

patológico.

4. **Tristeza:** aflicción, autocompasión, melancolía, desaliento, desesperanza, pena, duelo, soledad, depresión y nostalgia.

Charles Darwin planteó que "los principales actos de expresión que manifiestan el hombre y otros animales inferiores son innatos o heredados, es decir, el individuo no los ha adquirido" [9]. Como prueba de que las emociones son innatas, señaló la similitud de las expresiones en una misma especie y entre diferentes especies. A Darwin le impresionó bastante el hecho de que las expresiones corporales del hombre que tienen lugar cuando se producen las emociones, sobre todo las faciales, son las mismas en todo el mundo, con independencia de los orígenes étnicos o culturales. También indicó que estas mismas expresiones están presentes en personas que han nacido ciegas y que, por tanto, carecen de la posibilidad de haber aprendido los movimientos musculares viéndolos en los demás, y que también están presentes en los niños que tampoco han tenido mucho tiempo para aprender a imitarlas.

Las emociones primarias suelen estar acompañadas de claros indicios físicos. Cuando se está deprimido, el cuerpo se moviliza (o se desmoviliza) para desconectarse. Y cuando se es feliz, el cuerpo se moviliza para asumir compromisos y acciones positivas. Se activan determinados músculos para apoyar ciertas acciones, y el cerebro envía mensajes especiales a sus glándulas endocrinas (que controlan la producción y la liberación de hormonas) y a su sistema nervioso autónomo (que regula los órganos sobre los cuales no se ejerce control voluntario, como el corazón y el estómago).

Scherer en sus trabajos realiza una descripción acerca de la relación entre los parámetros de la voz y las emociones [10]:

- **Enfado:** El enfado ha sido ampliamente estudiado en la literatura sobre emociones. Hay contradicciones entre los efectos recogidos en estos escritos, aunque esto puede ser debido porque el enfado puede ser expresado de varias maneras.

  El enfado se define como "la impresión desagradable y molesta que se produce en el ánimo". El enfado se caracteriza por un tono medio alto (229 Hz), un amplio rango de tono y una velocidad de locución rápida (190 palabras por minuto), con un 32% de pausas.

- **Alegría:** Se manifiesta en un incremento en el tono medio y en su rango, así como un incremento en la velocidad de locución y en la intensidad.

- **Tristeza:** El habla triste exhibe un tono medio más bajo que el normal, un estrecho rango y una velocidad de locución lenta.

- **Miedo:** Comparando el tono medio con los otras cuatros emociones primarias estudiadas, se observó el tono medio más elevado (254 Hz), el rango mayor, un gran número de cambios en la curva del tono y una velocidad de locución rápida (202 palabras por minuto).

- **Disgusto/odio:** Se caracteriza por un tono medio bajo, un rango amplio y la velocidad de locución más baja, con grandes pausas.

2.3.2 Emociones secundarias

Actualmente, para la mayoría de los autores existen ocho emociones básicas, de las cuales cuatro son primarias y otras cuatro son secundarias [26].

Las secundarias, con su respectiva variedad de manifestaciones, son éstas:

1. **Amor:** aceptación, adoración, afinidad, amabilidad, amor desinteresado, caridad, confianza, devoción, dedicación, gentileza y amor obsesivo.

2. **Sorpresa:** asombro, estupefacción, maravilla y shock.

3. **Vergüenza:** arrepentimiento, humillación, mortificación, pena, remordimiento, culpa y vergüenza.

4. **Aversión:** repulsión, asco, desdén, desprecio, menosprecio y aberración.

Otros teóricos consideran emociones básicas a las 8 mencionadas hasta ahora (primarias y secundarias), y postulan que las emociones secundarias serían el resultado de fusiones o mezclas de las más básicas. Izard, por ejemplo, describe la ansiedad como la combinación del miedo y de dos emociones más, que pueden ser la culpa, el interés, la vergüenza o la agitación [**?**].

Plutchik ha expuesto una de las teorías mejor desarrolladas sobre la combinación de las emociones [27, 28]. Utiliza un círculo de emociones, análogo al círculo cromático en el que la mezcla de colores elementales proporciona otros. Cada emoción básica ocupa un lugar en el círculo. Las combinaciones compuestas por

dos emociones básicas se llaman “díadas”. Las compuestas por emociones básicas adyacentes en el círculo se llaman “díadas primarias”; las compuestas por emociones básicas separadas entre sí por una tercera se llaman “díadas secundarias”, etc.

En este esquema, el amor es una díada primaria resultante de la mezcla de dos emociones básicas adyacentes: la alegría y la aceptación, mientras que la culpa es una díada secundaria formada por la alegría y el miedo, que están separadas por la aceptación. Cuanta más distancia haya entre dos emociones básicas, menos probable será que se mezclen. Y si dos emociones distantes se mezclan, es probable que surja el conflicto. El miedo y la sorpresa son adyacentes y se combinan directamente para dar lugar a un estado de alarma, pero la alegría y el miedo están separadas entre sí por la aceptación, y su fusión es imperfecta: el conflicto resultante es la fuente de la culpa.

Tanto las emociones primarias como las secundarias casi nunca se presentan aisladas, mas bien son una combinación de todas las familias de emociones básicas mencionadas. Por ejemplo, los celos pueden ser una combinación de enojo, tristeza y miedo.

Finalmente, conviene mencionar otra categoría que podría incluir los sentimientos personales que pueden ser de estimación propia o egocéntricos como el orgullo, la vanidad y el narcisismo, contrarios a la simpatía, el amor o la compasión.

Según [10] las emociones secundarias son:

- **Pena:** es una forma extrema de tristeza, generalmente causada por una aflicción. Se caracteriza por un bajo tono medio, el rango de tono más estrecho, la pendiente de la curva de tono más baja, una velocidad de locución baja y un alto porcentaje de pausas.

- **Ternura:** se expresa con un alto nivel de tono que no fluctúa excesivamente.

- **Ironía:** caracterizada por una velocidad de locución baja y una acentuación muy marcada.

- **Sorpresa:** con un tono medio mayor que la voz neutra, una velocidad igual a la neutra y un rango amplio de variación.

Otras emociones secundarias: como el temor, la queja, el anhelo, el aburrimiento, la satisfacción, la impaciencia, el ensueño, la coquetería han sido también objeto de estudio. Algunos investigadores han utilizado otra clasificación, dividiendo las

emociones en:

- **Activas:** Se caracterizan por una velocidad de locución lenta, un volumen bajo, un tono bajo y un timbre más resonante.

- **Pasivas:** Caracterizadas por una velocidad de locución rápida, alto volumen, alto tono y un timbre "encendido".

CAPÍTULO 3

Transformadas tiempo-frecuencia

Los campos de aplicación de las transformadas tiempo-frecuencia conocidas como representaciones tiempo-frecuencia, en adelante, TFR son cada vez más amplios, pues se ha comprobado que mejoran los resultados de los métodos espectrales y temporales clásicos al ser capaces de reflejar cambios en frecuencia con respecto al tiempo (transitorios espectrales), algo que en un análisis espectral clásico no se puede detectar, por lo que la clasificación o detección de determinadas propiedades de la señal analizada se mejora. Análogamente, los métodos basados en características temporales no consiguen detectar características esenciales de la señal que son las que muestran con certeza su naturaleza. Por ello, un uso combinado de ambos dominios resulta en el aprovechamiento de características útiles presentes en ambos dominios para así realizar diagnósticos más fiables. Inicialmente se aplicó en la detección radar y reconocimiento del habla, pero hoy en día se aplica casi en todos los campos del tratamiento digital de señales.

Dado que el tiempo es una variable intrínseca a cualquier señal y la adquisición o generación de señales contiene como paso imprescindible el tratamiento de la señal en el dominio temporal, en la mayoría de los casos es preciso conocer sus características (forma, amplitud, pendientes, cruces por cero, energía, etc.) ya que esto aporta gran cantidad de información. Una propiedad importante de las señales es la estacionaridad, lo que implica que sus propiedades (momentos estadísticos) no varían a lo largo del tiempo, con lo que su análisis resulta más sencillo. En cambio, para señales no estacionarias, que son las que generalmente ocurren en la naturaleza, las propiedades de la señal varían a lo largo del tiempo.

Aunque el concepto de estacionariedad en sentido amplio y ergodicidad [29] permite realizar estimaciones bastante fiables ya sea por métodos paramétricos o no paramétricos, esta suposición restringe el ámbito de aplicación ya que en el caso

de señales no estacionarias la señal varía en el tiempo ya sea en amplitud, rango de frecuencias, forma de onda, etc. Por tanto, al aplicar métodos para señales estacionarias los resultados no son siempre los esperados.

3.1 Señales en el dominio del tiempo

Dos características especiales definen una señal en el dominio temporal: el tiempo medio y la duración. El tiempo medio corresponde al momento en el tiempo en el que a ambos lados de él se concentra la densidad de energía de la señal, viene definido por [30] :

$$E(t) = \int_{-\infty}^{\infty} t\,|s(t)|^2\,dt$$

y

$$E\left(t^2\right) = \int_{-\infty}^{\infty} t^2\,|s(t)|^2\,dt$$

donde:
$|s(t)|^2$ = Energía de la señal.
$E(t)$ = Tiempo medio o valor medio.
$E(t^2)$= Tiempo cuadratico medio o valor cuadrático medio.

La duración se define como la desviación estándar de la señal (momento de segundo orden) [31], es decir, indica el tiempo alrededor del tiempo medio en el cual la señal persiste:

$$\sigma_t^2 = E\left(t^2\right) - \left(E(t)\right)^2$$

Asi, en un tiempo $2\sigma_t$ la mayor parte de la señal se habrá extinguido. En general, la desviación estándar nos dice si una serie de datos están más alejados respecto a la media de esa serie.

3.2 Señales en el dominio de la frecuencia

En cuanto al dominio frecuencial, se puede decir que una de las razones principales es su simplicidad ya que una sencilla suma de sinusoides que pueden aparentar una gran complejidad analizando su señal temporal, se reducen a elementos frecuenciales puntuales mediante la transformada de Fourier. Como es ampliamente conocido, la transformada de Fourier consiste en la descomposición de una señal en suma de

señales sinusoidales de diferentes frecuencias [32].

$$S(\omega) = \int_{-\infty}^{\infty} s(t)e^{-j\omega t}dt$$

Así, cada sinusoide de frecuencia ω contribuye a la formación de $s(t)$ en una cantidad $S(\omega)$. Al conjunto $S(\omega)$ se le conoce como el espectro de la señal. Análogamente al dominio temporal, se define también una densidad de energía para conocer la localización de las frecuencias más relevantes contenidas en la señal; de la misma forma, también se define una media frecuencial así como su desviación estándar llamada "ancho de banda".

3.3 Principio de Incertidumbre de Heisemberg.

Aunque el principio de incertidumbre para el análisis de señales no está relacionado con el tema para el que originalmente fue desarrollado (imposibilidad de determinar la posicion y el momento de una partícula en el campo de la mecanica cuántica), se toma el mismo nombre debido a la analogía entre ambos, pues en definitiva se trata de dos variables relacionadas entre sí donde existe alguna propiedad que no se puede cumplir al mismo tiempo para ambas, y en efecto la mejora en la primera variable implica forzosamente un empeoramiento en la segunda, en terminología matemática, se dice que los operadores asociados a dichas variables no conmutan.
En este caso, las variables son el tiempo y la frecuencia. Este principio supone que una mejora en la resolución que se obtiene para una de las variables empeora la resolución de la segunda variable, y viceversa, es decir, existe un compromiso entre buena resolución en el tiempo o buena resolución en frecuencia. Para obtener la estacionaridad se elige una ventana lo suficientemente estrecha en la cual la señal sea estacionaria. Cuanto más estrecha sea la ventana se obtiene mejor resolución en el tiempo y por lo tanto una mejor representación de la estacionaridad y peor resolución en frecuencia. Por lo tanto, el problema consiste en la selección de una ventana para el análisis, dependiendo de la aplicación. En la figura 9 se muestran dos posibilidades, dependiendo de la resolución deseada en tiempo y frecuencia; de lo anterior se puede decir que una TFR llega un punto límite en el que el ancho de banda de duración, para frecuencia y tiempo respectivamente, no puede ser mejorado simultáneamente. Este límite viene impuesto por:

$$\Delta_t.\Delta_\omega \geq \frac{1}{2}$$

En [33] se muestra la demostración detallada.

Esta propiedad supone una restricción importante para las representaciones tiempo-frecuencia, ya que no se podrán obtener resultados totalmente ajustados tanto en el dominio temporal como espectral. Por ello se debe decidir en cada caso la representación más apropiada.
El principio de Heisemberg se puede resumir asi:

- **Ventana estrecha:** Buena resolución en el tiempo y pobre resolución en el dominio de la frecuencia.

- **Ventana ancha:** Buena resolución en el dominio de la frecuencia y pobre resolución en el dominio del tiempo.

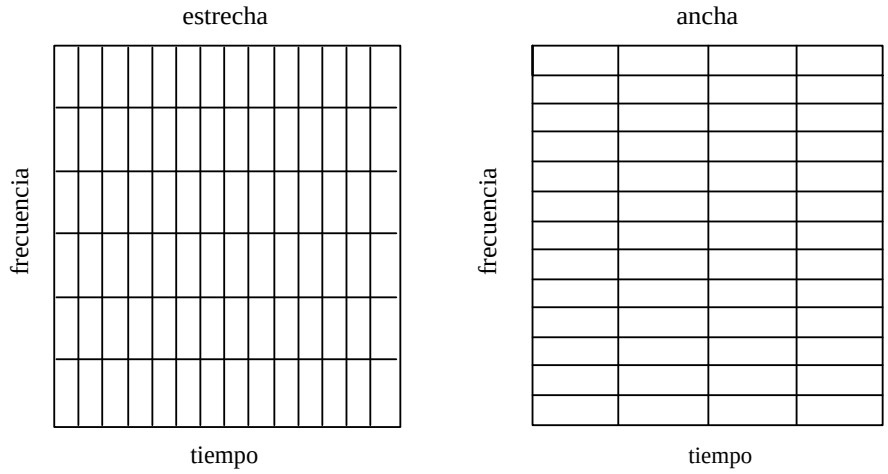


Figura 9: Enrejado resultante en el plano tiempo-frecuencia, resolución de banda estrecha y banda ancha

3.4 Definición y propiedades de las TFR

Una TFR es una distribución en el espacio conjunto tiempo frecuencia, la cual se puede representar por medio de una superficie tridimensional cuyos ejes son el tiempo y la frecuencia, y en la que para cada par (t, f) se dispone de un valor de amplitud al que se le puede llamar "energía" del punto, pero sin que ello implique que su significado sea el clásico ya que en muchos casos los valores de amplitud obtenidos no se corresponden con la definición teórica de energía, pero sirve para conocer la forma de la señal en el plano tiempo-frecuencia y obtener el valor de la concentración de intensidad en cada punto en al distribución global para apreciar la evolución de dicha amplitud tanto en el tiempo como en la frecuencia. Así, en una

representación tiempo-frecuencia se aprecia cómo han evolucionado las componentes frecuenciales a lo largo de la trama de tiempo analizada.
Existe una diferencia principal entre las TFR lineales [34] y las cuadráticas, se puede decir que mientras en las primeras se muestra la señal descompuesta en el plano $t-f$ basados en al amplitud de la señal temporal, en las segundas, la descomposición se realiza basándose en la energía, y lo que se distribuye en el plano $t-f$ es la energía de la señal, y dado que la energía de la señal requiere de la señal al cuadrado para obtenerse, se dice que estas representaciones son cuadráticas.

3.5 TFR Lineales

Este tipo de TFR son ampliamente utilizadas debido a su sencillez y similitud con los métodos espectrales clásicos en cuanto a concepto y estrategia de cálculo. Reciben este nombre ya que en su obtención, la señal a analizar no se multiplica por si misma como ocurre con las TFR cuadráticas o bilineales, y además cumplen el principio de superposición. También se les llama descomposiciones atómicas, ya que proyectan la señal en el plano de la forma que se obtienes celdas (cuyo tamaño depende del tipo de representación, frecuencia de muestreo, número de puntos elegidos, etc.) que conforman la representación global. Como representantes principales se encuentran la transformada corta de Fourier (Short Time Fourier Transform), la Transformada de Gabor y la Transformada Wavelet.

3.6 TFR Cuadráticas (Bilineales)

Las representaciones cuadráticas como se dijo son aquellas en que la dependencia con respecto a la señal es cuadrática. Este tipo de representaciones resultan apropiadas en muchos casos ya que intuitivamente nos permite asumir que se trata de una distribución energética dado que la energía es una representación cuadrática de la señal, por esto, a las representaciones cuadráticas también se les llama "representaciones energéticas" en muchas ocasiones, y tratan de combinar los conceptos de potencia instantánea y densidad espectral de energía, es decir, para una señal $s(t)$:

$$p_s(t) = |s(t)|^2$$

y

$$P_s(f) = |S(f)|^2$$

Aunque no siempre pueden ser interpretados como energía, ya que para ello se deben cumplir las propiedades marginales (ver apéndice de propiedades generales de la TFR), en cualquier caso, a través de la representación tiempo-frecuencia, siempre se puede hacer una idea de la distribución energética aproximada de la señal analizada.

Uno de los principales inconvenientes de las representaciones tiempo-frecuencia bilineales son los términos interferencia. Estos términos se generan debido a la presencia cuadrática de la señal que conviene varias componentes frecuenciales (señal multicomponente). Para el caso más sencillo de una señal formada por la suma de dos señales monofrecuenciales:

$$s(t) = s_1(t) + s_2(t) = e^{j2\pi.f_1.t} + e^{j2\pi.f_2.t}$$

La transformada tiempo-frecuencia de $s(t)$ esta formada por dos términos llamados "auto términos" correspondientes a la TFR de cada señal monofrecuencial, más dos términos cruzados [35] :

$$TFR_x(t,f) = |c_1|^2 TFR_{x_1}(t,f) + |c_2|^2 TFR_{x_2}(t,f) + c_1 c_2^* TRF_{x_{1,2}}(t,f) + c_2 c_1^* TFR_{x_{2,1}}(t,f)$$

Estos términos cruzados generan contribuciones en la representación que son inexistentes, de hecho, puede ocurrir que aparezcan términos de energía en zonas donde resulta incongruente tenerlos. Además, cualquiera que sea la señal, bien sea ruido, picos espúreos, etc, se genera su correspondiente término cruzado, lo que supone la presencia de un gran número de componentes añadidas que puede perjudicar el análisis de la señal útil. Para una señal que contenga N componentes frecuenciales su representación tiempo-frecuencia estará constituida por N términos de señal más $N(N-1)/2$ términos cruzados, por lo que su número crece de forma cuadrática con el número de componentes.
Si no existe atenuación alguna, para representaciones que proporcionan resultados reales se tiene que la amplitud de estos términos es el doble que la de los términos propios (o auto términos), lo que supone un serio problema para poder realizar un análisis sobre las representaciones tiempo-frecuencia ya que los términos ficticios pueden enmascarar términos propios. Por esta razón, se utilizan los llamados "Kernel" [36] , que en muchas ocasiones consisten en filtros bidimensionales que atenúan los términos cruzados, proporcionando mas intensidad a los términos propios.

3.7 Transformada Gabor

Una clase de representaciones tiempo-frecuencia ampliamente difundida en el ámbito del procesamiento de señales se basa en el empleo de ventanas temporales, esto es de funciones suaves y bien localizadas en un intervalo. La ventana $g(t)$ enmarca una porción de la señal y permite aplicar localmente la Transformada de Fourier. De este modo, se releva la información en frecuencia localizada temporalmente en el dominio efectivo de la ventana. Desplazando temporalmente la ventana se cubre el

dominio de la señal obteniéndose la completa información tiempo - frecuencia de la misma:

$$\hat{s}_g(\tau,\omega) = \int_{-\infty}^{\infty} s(t)g(t-\tau)e^{-j\omega t}dt$$

Asumiendo que la ventana real $g(t)$ está bien localizada en un intervalo centrado en t = 0, de longitud Δ_t y que su transformada $\hat{g}(\omega)$ está tambien localizada en una banda centrada en $\omega = 0$, de ancho Δ_ω, las ventanas desplazadas y moduladas $g(t-\tau)e^{-j\omega t}$ son funciones elementales bien localizadas en el dominio conjunto tiempo - frecuencia. Cada función elemental se localiza en el rectángulo centrado en el punto (τ,ω) de dimensión $\Delta_t\Delta_\omega$.
Por tanto el conjunto de valores $\hat{s}_g(\tau,\omega)$ nos da un completo mapa en el dominio tiempo-frecuencia que despliega la información de la señal. Más aún, ésta puede recuperarse con la fórmula de inversión:

$$s(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \hat{s}_g(\tau,\omega)e^{j\omega t}d\omega d\tau$$

La misma sintetiza la señal como la superposición integral de las funciones elementales $g(t-\tau)e^{j\omega t}$. El mapeo sobre domino tiempo-frecuencia, bajo las condiciones referidas, se conoce como la *transformada de Gabor* [37] y representa una atractiva generalización de la transformada de Fourier.
Esta transformada se puede reformular considerando ahora el par de ventanas moduladas reales $g(t-\tau)\cos(\omega t)$ y $g(t-\tau)\sin(\omega t)$. Estos pares de ventanas moduladas actúan como *filtros-pasabanda*, con definición de fase. De tal modo la Transformada de Gabor puede entenderse como un tratamiento localizado de la señal mediante filtros - pasabanda deslizantes, de ancho de banda constante.
Surgen dos importantes observaciones. La primera es la cuestión de la implementación numérica de la transformada. Claramente, es necesario discretizar los parámetros τ,ω en una apropiada red que equilibre la eficiencia del mapeo de la información con la complejidad computacional.
Se observa que los rectángulos de localización próximos se superponen. Esto implica la redundancia de la información del mapeo. La superposición y redundancia dependen de las dimensión de los rectángulos $\Delta_t\Delta_\omega$. Justamente, el principio de incertidumbre nos dice que es imposible reducir indefinidamente esta dimensión, existiendo una constante positiva universal tal que: $\Delta_t\Delta_\omega \geq C$. La igualdad se alcanza en el caso de las ventanas gaussianas (ver figura 10).

Por otra parte, se comprende que el esfuerzo computacional se incrementa en función de la redundancia. Por tanto el ideal sería preservar la totalidad de la información del mapeo continuo sin redundancia en una red discreta de rectángulos,

con mínima superposición.

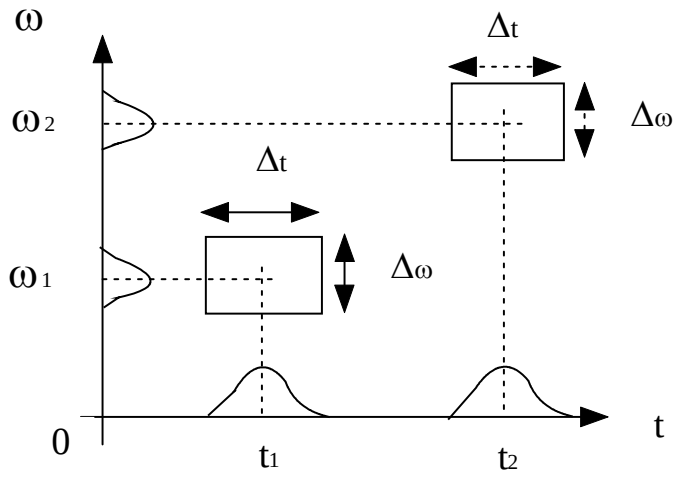


Figura 10: Plano tiempo-frecuencia para la Transformada Gabor

La segunda cuestión la plantea el hecho de que las dimensiones de los rectángulos de localización son constantes. En las altas frecuencias el número de oscilaciones en el dominio temporal de las ventanas es elevado, y la información frecuencial del mapeo es nítida. En contraposición, en las bajas frecuencias, las oscilaciones son relativamente largas y no pueden caracterizarse apropiadamente.

Este último fenómeno se explica por la superposición de los rectángulos asociados a las bajas frecuencias, positivas y negativas. En otras palabras, la redundancia es crítica entorno de $\omega = 0$.

La implementación de la Transformada de Gabor para el procesamiento de señales de emisiones acústicas resulta eficiente cuando se trata de localizar y caracterizar eventos con patrones de frecuencia bien definida, no superpuestos y relativamente largos, respecto de la ventana de análisis. En contraposición, es totalmente inapropiada para detectar detalles de corta duración, oscilaciones largas asociados a las bajas frecuencias, o caracterizar patrones autosimilares presentes en fenómenos a distintas escalas. En la figura 11 se puede ver la aplicacion de la transformada gabor a una señal de voz.

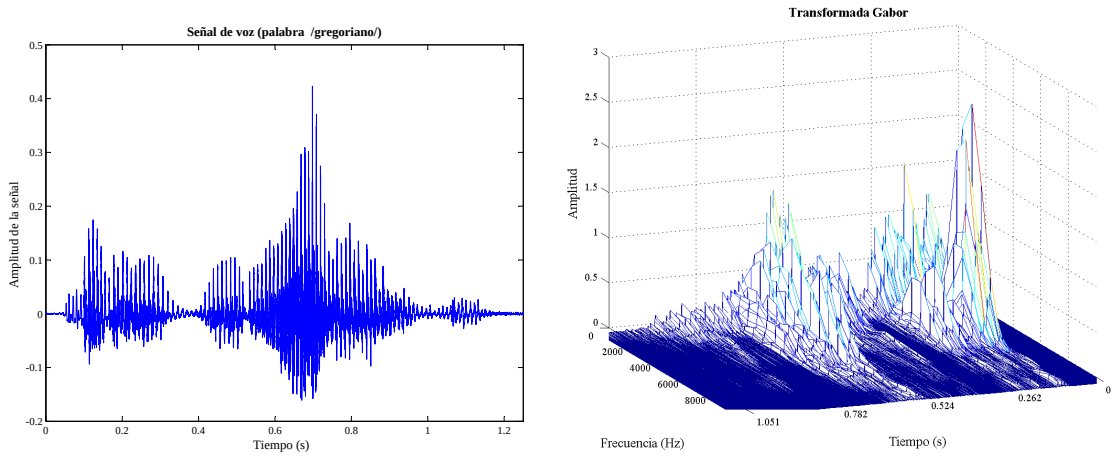


Figura 11: Proyección 3D de la Transformada Gabor de señal de voz

En la aplicación de esta técnica se utilizan los espectrogramas, definidos en [4]. Un espectrograma de una señal en el tiempo es una representación especial bidimensional que muestra el transcurso del tiempo en el eje x y los rangos de frecuencia en el eje y. El objetivo del espectrograma es calcular la transformada de Fourier cada n segundos con el fin de conocer el contenido frecuencial de la señal en ese intervalo de tiempo. En la figura 12 puede verse el espectrograma de una señal de voz. Las zonas más oscuras indican un contenido de altas frecuencias más acentuado en ese intervalo de tiempo.

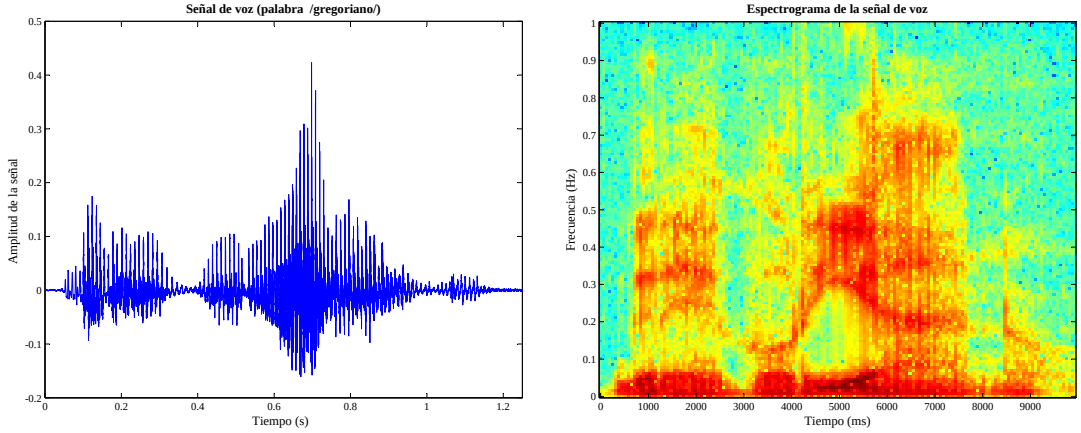


Figura 12: Espectrograma de la señal de voz

Se toman intervalos de aproximadamente de 30 ms, en los cuales, la señal de voz se considera cuasi periódica. Si se toman intervalos mayores se pierde la periodicidad de la señal y por lo tanto, la definición exacta de la transformada de Fourier no puede ser empleada. Los espectrogramas a su vez, se pueden dividir en dos clases: El espectrograma de banda ancha utiliza una ventana $w(t)$ estrecha en

el tiempo ($< 10\ ms$) con una resolución en frecuencia mayor a 200 Hz, mientras que el espectrograma de banda angosta ($> 20\ ms$) busca una mejor localización de la distribución de la energía en el dominio frecuencial, ya que su resolución en frecuencia es aproximadamente menor a 100 Hz [38].

3.8 Transformada Wavelet

3.8.1 Transformada Wavelet Continua

Una alternativa a la Trasformada de Gabor es la de utilizar ventanas moduladas, de dimensión variable, ajustada a la frecuencia de oscilación. Más precisamente, que mantenga un mismo número de oscilaciones en el dominio de la ventana. Esto sugiere, naturalmente, contar con una única ventana modulada y generar una completa familia de funciones elementales mediante sus dilataciones o contracciones y traslaciones en el tiempo [39]:

$$\psi(t) \Rightarrow \frac{1}{\sqrt{|a|}}\psi(\frac{t-b}{a})$$

donde $a \neq 0$ y b son los parámetros de *escala* y de *traslación*. Se preserva la energía de las funciones mediante un factor de normalización figura 13.

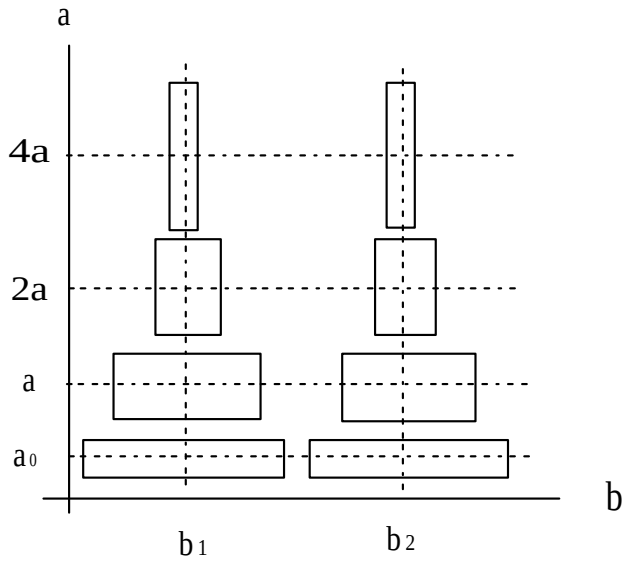


Figura 13: Plano tiempo-frecuencia para la Transformada Wavelet

Donde la escala se define como:

$$Escala = \frac{1}{frecuencia}$$

La función $\psi(t)$, debe verificar ciertas condiciones de admisibilidad y se denomina *wavelet madre* y el resto de las funciones generadas, simplemente, *wavelets*. Se denotan las mismas como:

$$\psi_{a,b}(t) = \frac{1}{\sqrt{|a|}}\psi(\frac{t-b}{a})$$

Las condiciones de admisibilidad, básicamente requieren que la función $\psi(t)$ esté bien localizada en tiempo, de media nula y que la transformada $\hat{\psi}(\omega)$ sea un filtro continuo pasa-banda, con rápido decaimiento hacia el infinito y hacia $\omega = 0$. Entonces, dada la señal $s(t)$, de energía finita la *Transformada Wavelet Continua* de $s(t)$ se define como:

$$W_\psi s(a,b) = \int_{-\infty}^{\infty} s(t)\psi_{a,b}(t)dt$$

para cada par de parámetros reales (a,b), $a \neq 0$. Si la ondita es real, la definición se restringe para valores positivos de a. La transformación así definida preserva la energía de la señal, y posee una fórmula integral de inversión. Si la wavelet madre es real, la reconstrucción se realiza como:

$$s(t) = C_\psi \int_0^{\infty}\int_{-\infty}^{\infty} W_\psi s(a,b)\psi_{a,b}(t)\frac{dbda}{a^2},$$

donde C_ψ es una constante positiva. La fórmula expresa la síntesis de la señal como la superposición integral de las funciones elementales $\psi_{a,b}(t)$. El mapeo sobre domino tiempo-frecuencia, parametrizado por (a,b), esto es la Transformada Wavelet Continua, representa una novedosa alternativa a la Transformada de Fourier por ventanas. Despliega la información de la señal en una estructura radicalmente diferente. Cualquier wavelet real $\psi(t)$, admisible, está bien localizada en un intervalo centrado en un t_0, de longitud Δ_t y que su transformada $\hat{\psi}(\omega)$ está localizada en una banda bilátera $0 < \omega_1 \leq |\omega| \leq \omega_2$, de ancho Δ_ω. Entonces, las wavelets $\psi_{a,b}(t)$ están localizadas en el intervalo centrado en $at_0 + b$, de longitud $a\Delta_t$ y en la banda bilátera $0 < \omega_1/a \leq |\omega| \leq \omega_2/a$, de ancho Δ_ω/a. Se observa que la precisión en tiempo es inversamente proporcional a la precisión en frecuencia, manteniéndose constante la relación $\Delta_t\Delta_\omega$. Esta es la diferencia fundamental con la Transformada de Gabor. Más aún, para cada valor de a, la familia de wavelets parametrizada por b, se comporta como una ventana deslizante, de ancho de banda constante, pero el número de oscilaciones de estas ondas elementales es siempre el mismo, en el marco

efectivo de la ventana. Por otra parte, si la transformada $\hat{\psi}(\omega)$ decae rápidamente a cero en torno de $\omega = 0$, se verifica la propiedad de oscilación:

$$\int_{-\infty}^{\infty} t^k \psi(t) dt = 0$$

para $k = 0, ..., K$. Esta importante propiedad, que no posee la Transformada de Gabor, es trascendente en las aplicaciones de análisis de señales, para la detección de fenómenos puntuales, como discontinuidades o bruscos cambios en las derivadas. En efecto, si se modela apropiadamente la señal física, por una función $s(t)$, de modo que las derivadas reflejen los cambios locales de comportamiento, en torno de cada punto $t_{a,b} = at_0 + b$, de radio Δ_t, el proceso queda caracterizado por el correspondiente polinomio de Taylor, hasta cierto orden $K + 1$. Si el proceso es razonablemente suave en el entorno, la propiedad de oscilación se dice que la magnitud $|W_\psi s(a, b)|$ es no significativa. En contraposición, un brusco cambio en el entorno, que se refleja en la derivada de orden $K+1$, podrá ser bien detectado. Otra propiedad relevante de la transformada continua es su invariancia respecto de las traslaciones o cambios de escala de la señal. Estructura similares, serán detectadas de la misma forma, independientemente de su localización temporal o escala. En suma, por sus propiedades y las razones antes expuestas la Transformada Wavelet Continua, constituye una promisoria y ventajosa alternativa para el procesamiento de señales, en particular las de Emisiones Acústicas.

3.8.2 Transformada Wavelet Discreta

El diseño de una versión discreta de la Transformada Wavelet, esencialmente consiste en definir una apropiada red discreta de parámetros $\{(a_j; b_{jk})\}$, de escalas y traslaciones, respectivamente. De modo que la familia de wavelets $\psi_{a_j,b_{jk}}$ sea admisible. En general, constituye un problema difícil caracterizar aquellas wavelets que definen una Transformada Discreta [40]. Se cuenta con varias clases de wavelets admisibles. Se mencionan entre las más difundidas las wavelets *spline*, las wavelets de *Daubechies* y otras análogas, ampliamente difundidas en la literatura y en el *sofware* actualmente disponible. Entre éstas, se encuentran diversas variantes, y particularmente las que generan *bases ortonormales de wavelets*. En general, esas clases se asocian a la red *diádica*:

$$a_j = 2^{-j} \ ; \ b_{jk} = 2^{-j} k \ j, k \in \ Z$$

Bajo esta elección de los parámetros, se tiene entonces la usual expresión para las wavelets:

$$\psi_{jk}(t) = 2^{j/2} \psi(2^j t - k) \ j, k \in Z$$

Asumiendo que la wavelet madre real y una señal $s(t)$ de energía finita la trasformada discreta asociada se define como:

$$DW_\psi s(j,k) = \langle s, \psi_{jk} \rangle = \int_{-\infty}^{\infty} s(t)\psi_{jk}(t)dt$$

para todos los valores enteros de j y k. Por otra parte, se tienen las fórmulas de síntesis:

$$s(t) = \sum_j \sum_k c_{jk}\psi_{jk}(t) \approx \sum_j \sum_k \langle s, \psi_{jk} \rangle \psi_{jk}(t)$$

para apropiados coeficientes en wavelets c_{jk}.
En el caso de que $\psi(t)$ genere una base ortonormal de wavelets, se tiene que $c_{jk} = \langle s, \psi_{jk} \rangle$ y la fórmula es exacta. En suma, en la práctica se puede considerar que los valores de la transformada o los coeficientes en wavelets resumen la información de la señal, en forma análoga al caso continuo.
El espectro de coeficientes c_{jk} parametrizado por (j,k), reemplaza al mapeo continuo en el dominio.
A partir de las mismas consideraciones que en el caso continuo, se ve que las wavelets $\psi_{jk}(t)$ están localizadas en el intervalo centrado en $(t_0 + k)2^{-j}$, de longitud $2^{-j}\Delta_t$ y en la banda $0 < 2^j\omega_1 \leq |\omega \leq| \leq 2^j\omega_2$, de ancho $2^j\Delta_\omega$. Estas bandas representan una partición en niveles o en *octavas* del dominio de las frecuencias.
Bajo este punto de vista, analizar una señal por medio de la transformada discreta consiste en descomponer la misma en un *banco de filtros analógicos* pasabanda y en cada octava, caracterizar el comportamiento en el tiempo:

$$s(t) = \sum_j w_j(t)$$

$$w_j(t) = \sum_k c_{jk}\psi_{jk}(t) \; para \; cada \; j$$

Por otra parte, realizado el análisis, es posible reconstruir a discreción las componentes relevantes de la señal y caracterizar así diversos fenómenos de interés. A diferencia de las Series de Fourier Locales, al análisis se realiza por octavas o rangos de frecuencia que duplican su dimensión hacia las altas frecuencias, a la vez que se reduce el rango temporal de localización. Es posible entonces localizar por medio del espectro, tanto fenómenos locales como patrones de autosimilaridad, a distintas escalas. Más aún, el apropiado truncamiento de las series de wavelets, realizado en cada nivel j, no desnaturaliza o destruye la señal. En suma, con un espectro finito se puede representar eficientemente la textura y fenomenología temporal de la señal, clasificada por octavas. Existen también desventajas. Por un lado, la discretización no conserva ciertas importantes propiedades de la Transformada Wavelet Continua. Particularmente, la de invariancia respecto de las traslaciones. Por otra parte, análisis mediante wavelets merece la clasificación de técnica tiempo - escala, más

que, propiamente, tiempo - frecuencia. El problema es que las wavelets no poseen una precisa localización en frecuencia, de modo que no se cuenta con precisión en frecuencia dentro de las octavas. Por esta razón, la Transformada Wavelet no es apropiada para caracterizar fenómenos estacionarios, casi monocromáticos. Sin embargo, y esto es fundamental, el empleo de wavelets, es tan flexible, que es posible subsanar estos inconvenientes. Sin mayores detalles se dice que es posible extender la Transformada Discreta y diseñar otras mejor adaptadas para propósitos particulares. Esto revela la riqueza de las wavelets.

3.8.3 Esquema de Análisis Multirresolución

Se considera una wavelet madre $\psi(t)$. Sin pérdida de generalidad, se asume que la familia de wavelets $\psi_{jk}(t)$, por ella generada constituye una base ortonormal de la clase de señales de energía finita [41, 42]. Supongase que está centrada en $t_0 = 0.5$ y que su transformada de Fourier está localizada en la banda $\pi \leq |\omega| \leq 2\pi$. Para cada valor entero del parámetro de escala j las wavelets $\psi_{jk}(t)$ generan un subespacio de señales que comparten una misma octava de localización, esto es, el rango de frecuencias $2^j\pi \leq |\omega| \leq 2^{j+1}\pi$. Denotemos W_j a estos subespacios *de wavelets*. Son ortogonales entre sí y contienen todas las señales representadas como:

$$w_j(t) = \sum_k c_{jk}\psi_{jk}(t),$$

donde los coeficientes son de cuadrado sumable:

$$\sum_k |c_{jk}|^2 \ < \ \infty,$$

Uniendo los subespacios de Wavelets, desde $-\infty$ hasta cada nivel $j-1$, se obtiene los subespacios, indexados por j:

$$V_j = \bigcup_{l=-\infty}^{j-1} W_l$$

Estos últimos subespacios se denominan *de escala*: y contienen las señales que resultan de las superposiciones:

$$s_j = \sum_{l=-\infty}^{j-1} w_l(t)$$

donde $w_l \in W_l$. Localizadas ahora en la banda que reune a las octava inferiores. Esto es las relativas bajas frecuencias $|\omega| \leq 2^j\pi$.
Se observa que se verifica la relación:

$$s_j(t) = s_{j-1}(t) + w_{j-1}(t),$$

para cada j, que dice que la información de $s_j(t)$ se desdobla en el componente w_{j-1} retiene los detalles relativos a las altas frecuencias (en adelante coeficientes de detalle d_i) y la componente s_{j-1} que conserva la tendencia asociada a las relativas bajas frecuencias (en adelante coeficientes de aproximación a_i) (ver figura14). Las componentes $s_j(t)$ y $w_j(t)$, son justamente las *proyecciones ortogonales* de la señal $s(t)$ sobre los espacios V_j y W_j, respectivamente.
Se deduce, que el proceso de descomposición, o sucesivas proyecciones, puede continuarse en la misma forma:

$$s_j(t) = s_{j-2}(t) + w_{j-1}(t) + w_{j-2}(t) = s_{j-N}(t) + w_{j-1}(t) + ... + w_{j-N}(t)$$

y en cada paso, se van agregando los detalles específicos de cada octava, mientras que la componente de baja frecuencia conserva el aporte de las frecuencias en torno a $\omega = 0$ o la tendencia general.
En teoría, una señal dada $s(t)$, de finita energía, puede descomponerse como la suma de proyecciones:

$$s(t) = +s_J(t) + \sum_{j=J}^{\infty} w_j(t)$$

para cualquier J arbitrario. Este tipo de análisis despliega la informacion de una señal por octavas o bandas de frecuencia.

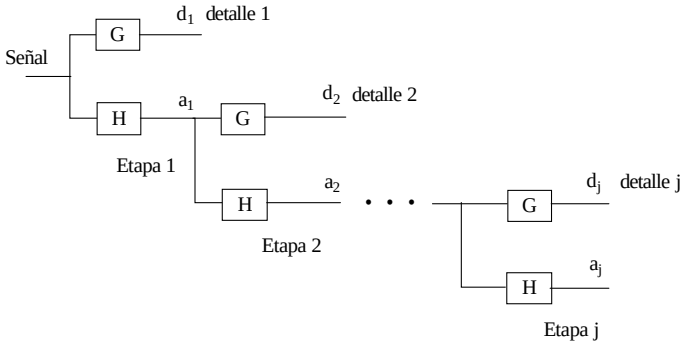


Figura 14: Diagrama piramidal para análisis multirresolución

3.9 Transformada Wigner ville

Según [43, 44, 45] la definición y propiedades de la transformada Wigner Ville contribuyen de forma decisiva a su extensión como herramienta para analizar señales no estacionarias. La distribución de Wigner- Ville correspondiente a una señal $s(t)$ se define como:

$$WV_s(t,f) = \int_{-\infty}^{\infty} s(t+\frac{\tau}{2})s^*(t-\frac{\tau}{2})e^{-j2\pi f\tau}d\tau$$

La obtención de esta fórmula se encuentra en [33]. Esta transformada no es la única distribución posible de energía, ni tampoco la mejor, pero es lo suficientemente genérica como para que a partir de ella se puedan obtener muchas otras.
Una de las formas de llegar a obtener esta fórmula es generalizando la relación entre el espectro de potencia y la función de autocorrelación. Se empieza con el caso estacionario, en el cual la densidad de potencia se calcula a través del espectrograma. Se desarrolla la fórmula del espectrograma para señales estacionarias de la siguiente forma:

$$|S(f)|^2 = \left| \int_{-\infty}^{\infty} s(t) e^{-j2\pi ft} dt \right|^2 = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} s^*(t') s(t) e^{-j2\pi f(t'-t)} dt' dt$$
$$= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} s^*(t-\tau) s(t) e^{-j2\pi f\tau} d\tau dt$$

se llega a que:

$$|S(f)|^2 = \int_{-\infty}^{\infty} R(\tau) e^{-j2\pi f\tau} d\tau \tag{1}$$

siendo $R(\tau)$ de la forma:

$$R(\tau) = \int_{-\infty}^{\infty} s^*(t+\tau) s(t) dt = \int_{-\infty}^{\infty} s^*(t-\frac{\tau}{2}) s(t+\frac{\tau}{2}) dt$$

Recordando que se está trabajando en el caso estacionario. Si ahora se asume la ecuación 1 puede generalizar al caso no estacionario, sin mas que sustituir $R(\tau)$ por una adecuada estimación de dicha función para el caso no estacionario, a lo que se denomina $R(t,\tau)$, entonces:

$$P(t,f) = \int_{-\infty}^{\infty} R(t,\tau) e^{-j2\pi f\tau} d\tau$$

Se han propuesto muchas estimaciones de dicha *función de autocorrelación local*. Si se toma la estimación:

$$R(t,\tau) = s^*(t-\frac{\tau}{2}) s(t-\frac{\tau}{2})$$

Se obtiene para $\epsilon(t,f)$ la expresión correspondiente a la estimación WV. A modo de ejemplo se puede citar que si en vez de esta elección se toma:

$$R(t,\tau) = s^*(t)s(t+\tau)$$

Se habría obtenido otra distribución conocida como distribución de Rihaczek [33]. De momento se ha conseguido plantear una nueva alternativa a la STFT, pero todavía no se ha demostrado si es mejor o peor. En primer lugar se va a usar un ejemplo para comparar ambas técnicas antes de analizar más en profundidad las propiedades de una y otra distribución.
De las propiedades descritas en la tabla 9 del apéndice B, las que más son verificadas por la distribución de Wigner-Ville (WV) son las siguientes:

- Real
- Marginales
- Invariante a los desplazamientos en tiempo y frecuencia
- Soporte finito en tiempo y frecuencia

Por el cotrario, no satisface la condición de *positividad*, tomando valores negativos en algunos puntos. Las demostraciones pueden encontrarse todas ellas en [33, 46, 34]. Existe un teorema debido a Wigner [47] que postula que no puede existir ninguna distribución bilineal positiva que verifique los marginales. Esto significa que si una distribución bilineal verifica los marginales no puede ser positiva y viceversa. Por otra parte, ello no implica que no existan distribuciones positivas que verifiquen los marginales, pero en este caso no podrán ser distribuciones bilineales.
Aclarando el proceso de bilinealidad. Dada una señal $s(t)$, cualquier transformación sobre la $f(s(t))$ verifica el principio de superposición lineal si:

$$f(as_1(t) + bs_2(t)) = af(s_1(t)) + bf(s_2(t))$$

Si no se verifica esta propiedad pero se verifica que:

$$f(as_1(t) + bs_2(t)) = a^2 f(s_1(t)) + b^2 f(s_2(t)) + abf(s_1(t).s_2(t)),$$

se dirá que f verifica el *principio de superposición cuadrática*. Los términos $f(s_1(t).s_2(t))$ *término cruzado*. Se dice que el término $f(s_1(t).s_2(t))$ es *bilineal* por depender del producto de las dos señales. Si en general se operara sobre una señal de N componentes se podría comprobar que se originan N autocomponentes y $N.(N+1)/2$ términos cruzados, uno por cada par de componentes [34].
El concepto mismo de energía está directamente relacionado con el de bilinealidad, dado que la energía es función del cuadrado de la señal, o lo que es lo mismo, del producto de la señal consigo misma. Por este motivo todas las representaciones con las que se va a trabajar verifican el principio de superposiciòn cuadrática.

Otras propiedades que resultan incompatibles son las de positividad y soporte finito. Es posible demostrar que cualquier distribución positiva no posee un soporte finito [46].
Todo esto conduce a una conclusión evidente: no es posible encontrar una distribución $t - f$ que, siendo bilineal, verifique todas las propiedades de la tabla9 del apéndice B. En cada situación concreta resultarán más atractivas unas propiedades que otras y en función de ello se debe optar por la transformaciones tiempo-frecuencia que se adapte mejor a las necesidades.
Hasta el momento se tienen dos candidatos posibles para construir una representación tiempo-frecuencia: el espectrograma y la distribución de WV (siempre y cuando se considere que el uso de un modelado paramétrico de la señal previo al calculo de la transformada de Fourier entre dentro de la denominación de espectrograma). La distribución de WV cumple todas las propiedades de la tabla salvo la de positividad y el espectrograma no cumple con los marginales ni con el soporte finito, pero sí es positivo. Esto lleva a que, desde el punto de vista teórico, las razones para elegir uno u otro no son concluyentes.
Por otra parte, se ha visto un caso práctico para el cual la distribución de WV presentaba una resolución mucho mayor que la del espectrograma, pero introducía el inconveniente de la aparición de términos *ficticios*. De la naturaleza de estos términos se va tratar ahora un poco más a fondo para conocer su origen y en qué medida será posible eliminarlos ya que, de conseguirlo, parece claro que se mejoran las características del espectrograma.

3.9.1 Geometría de los términos cruzados

Desde el punto de vista de las propiedades que cumplen, la comparación entre espectrograma y distribución de WV parece decantarse del lado de esta última. Analizando ahora un aspecto práctico de suma importancia de cara a la interpretación de los resultados de ambas distribuciones: la presencia de términos cruzados en señales multicomponente.
Sea la señal $s(t) = e^{-2\pi f_1 t} + e^{j2\pi f_2 t}$. Una señal estacionaria con dos componentes, dos oscilaciones puras con frecuencias f_1 y f_2. Se trata en primer lugar de calcular el espectrograma de dicha señal. En un caso ideal el espectrograma (SP) será el módulo al cuadrado de la transformada de Fourier (FT), o sea:

$$\begin{aligned} SPs(t,f) &= |FT_s(f)|^2 \\ &= |\delta(f_1 - f) + \delta(f_2 - f)|^2 \\ &= \delta(f_1 - f) + \delta(f_2 - f) + 2\delta(f_1 - f)\delta(f_2 - f) \end{aligned}$$

donde se aprecian claramente los dos términos autocomponentes y el término cruzado $2\delta(f_1 - f)\delta(f_2 - f)$. Excepto en el caso de que $f_1 = f_2$ este término valdría cero y no habría término cruzado.

En la práctica, para una señal genérica no estacionaria no serviría la FT y en su lugar tendría que usarse la STFT. Suponiendo que para este caso se use una ventana rectangular de tamaño T obtendríamos para el espectrograma la expresión:

$$\begin{aligned} SP_s^{\omega}(t,f) &= |STFT_s^{\omega}(t,f)|^2 \\ &= |T(sinc(\pi(f_1-f)T) + (sinc(\pi(f_2-f)T)|^2 \\ &= T(sinc^2(\pi(f_1-f)T) + sinc^2(\pi(f_2-f)T) + 2sinc(\pi(f_1-f)T)(sinc(\pi(f_2-f)T)) \end{aligned}$$

En este caso se aprecian igualmente los dos términos autocomponentes y el término cruzado $sinc(\pi(f_1-f)T) + (sinc(\pi(f_2-f)T)$. Se puede observar que este término alcanza sus máximos cerca de los dos autocomponentes, tendiendo a cero a medida que se alejan de ellos. En consecuencia, siempre que las dos frecuencias estén suficientemente alejadas, la distorsión que introducen es mínima y siempre superpuesta a las propias autocomponentes.
Usando la distribución de WV se obtiene, para el caso de la señal infinita, la expresión:

$$WV_s(t,f) = \delta(f_1-f) + \delta(f_2-f) + 2cos(2\pi(f_1-f_2)t)\delta\left(\frac{f_1+f_2}{2} - f\right)$$

Los dos primeros términos serían las autocomponentes y el tercero es el término cruzado. Si se analiza dicho término cruzado se ve que corresponde a una función δ centrada en $\frac{f_1+f_2}{2}$, o sea, en el punto medio entre las dos autocomponentes, con una forma oscilatoria en la dirección temporal y con una amplitud doble que la de los términos autocomponentes.
En un caso real, con una señal de duración T, se obtiene la expresión:

$$WV_s(t,f) = T(sinc(\pi(f_1-f)T) + (sinc(\pi(f_2-f)T) + 2cos(2\pi(f_1-f_2)T)sinc(\frac{f_1+f_2}{2} - f)$$

que se corresponde con una estructura similar a la obtenida para el caso infinito, pero más suavizada.
El término cruzado para la distribución WV es un problema, pues puede alcanzar magnitudes superiores a las de las propias autocomponentes. La *ventaja* en este caso es que el término cruzado está separado de las propias componentes, de forma que no las distorsiona, siempre y cuando estén lo suficientemente alejadas. Si no es así, el término cruzado podría interferir con las autocomponentes y hacer el espectro totalmente indescifrable.
Si este caso se plantea con una señal bicomponente, que sucedería con una señal real, multicomponente, y a menudo con un espectro de banda ancha?. Está claro que el resultado, en ese caso, sería totalmente inservible.
Para poder solucionar este problema de la distribución de WV se han desarrollado múltiples alternativas, todas ellas con un fin similar: modificar de alguna forma

la expresión de la distribución de WV para disminuir la potencia de los términos cruzados, a costa necesariamente de empeorar otras características de la distribución, tales como la resolución en tiempo o frecuencia.
Aunque cada una de estas alternativas tiene unas características determinadas y suele adaptarse mejor a unos tipos específicos de señales, todas ellas pueden considerarse como casos específicos de una distribución genérica.

3.10 Análisis de predicción lineal

Hasta finales de los años 70 el procesado de señales se basó exclusivamente en la manipulación directa de la señal (representación temporal) o la aplicación de transformaciones que permitieran otro tipo de representación (frecuencia, tiempo-frecuencia). Trabajos tan importantes como los de Akaike, Kalman y Markhoul definen una nueva metodología de procesado llamado procesamiento basado en modelos, en la cual se parte de la obtención de un modelo previo de la señal (generalmente un modelo paramétrico), que incluye información a priori sobre el fenómeno físico (su mecanismo de generación, contenido de ruido etc.) para luego con la estimación de los coeficientes del modelo reproducir la señal y caracterizarla [48].
Una de las técnicas más utilizada para la caracterización de la voz con base a un modelo predeterminado es el Análisis de Predicción Lineal. Esta metodología se basa en el modelo general de producción de voz desarrollado en el apéndice A. El proceso de fonación se modela como un sistema entrada - filtro - salida, en el cual el tracto vocal es el filtro, cuyos parámetros varían en el tiempo en función de la acción consciente que se realiza al pronunciar una palabra. Existen dos posibles señales de entrada para el filtro: sonora o no sonora. Para señales sonoras la excitación es un tren de impulsos de frecuencia controlada, mientras que para las señales no sonoras la excitación será ruido aleatorio, la combinación de estas señales modela el funcionamiento de la glotis. La señal de entrada pasa por el modelo del tracto vocal donde es amplificada y enriquecida de armónicos (formantes), por último, a la salida, se tienen en cuenta los efectos de radiación producidos en los labios. Un esquema de este modelo se presenta en la figura.

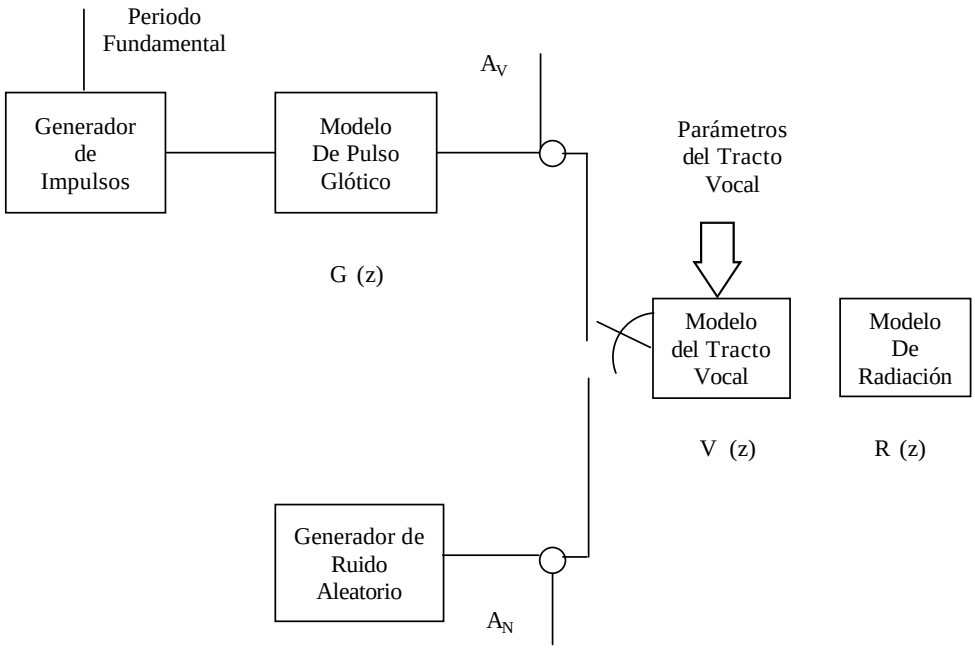


Figura 15: Modelo general de producción de voz

En la figura se pueden distinguir los tres elementos básicos mencionados anteriormente.

- **Tracto vocal:** Generalmente se emplea una función todo polos de la forma

$$V(z) = \frac{G}{1 - \sum_{i=1}^{N} a_i z^{-i}} = \frac{G}{\prod_{i=1}^{N} (1 - p_i z^{-1})}$$

  Donde p_i son los polos de la función de transferencia y G es el factor de ganancia (representa la amplitud de la voz). Los coeficientes del filtro se pueden calcular mediante predicción lineal dando lugar a los coeficientes de predicción lineal.

- **Radiación en los labios:** La técnica que se utiliza para introducir el efecto de la radiación en el modelo digital es la inclusión de un filtro paso alto, típicamente, un diferenciador de primer orden de la forma

$$R(z) = 1 - z^{-1}$$

- **Exitación:** Para voz sonora, el generador produce un tren de impulsos que excita un sistema lineal, $G(z)$, que tiene la forma glótica deseada. Para voz sorda, el modelo es mucho más simple. Todo lo que se requiere es una fuente de ruido aleatoria. Para ambos casos, existen controles de ganancia, A_V y A_N, que controlan la intensidad de la señal de excitación.

Es necesario combinar los modelos de pulso glótico y de radiación conjuntamente con el del tracto vocal a modo de obtener una función de transferencia global.

$$H(z) = G(z)V(z)R(z) = \frac{S(z)}{U(z)} = \frac{G}{1 - \sum_{i=1}^{N} a_i z^{-i}}$$

Las muestras de salida $s(n)$ están relacionadas con la entrada $u(n)$ por la siguiente expresión

$$s(n) = \sum_{i=1}^{N} a_i s_n(n-i) + Gu(n)$$

N, G y a_i son el orden de predicción, la ganancia y los coeficientes de predicción lineal (LPC) respectivamente. Esta función de transferencia equivale a un modelo autorregresivo (AR) de la señal de voz donde una muestra precedente se obtiene a partir de una combinación lineal de las muestras anteriores y la entrada actual.

$$s(n) \approx a_1 s(n-1) + a_2 s(n-2) + ... + a_p s(n-p)$$

Donde los coeficientes $a_1, a_2, .., a_p$ se asumen como constantes sobre la trama de voz analizada. De esta forma, podemos definir un predictor lineal, como un sistema cuya salida es

$$\tilde{s} = \sum_{k=1}^{p} a_k s(n-k)$$

El problema básico del análisis de predicción lineal es determinar un conjunto de coeficientes de predicción a_k directamente de la señal de voz, los cuales minimicen el error cuadratico medio sobre un intervalo corto de la señal
Asumiendo que $s_n(m)$ es un segmento de voz seleccionado en la vecindad de la muestra n, tal que

$$s_n(m) = s(n+m)$$

se minimiza el error cuadrático medio en la muestra n

$$E_n = \sum e_n^2(m)$$

que utilizando la función $e_n(m)$ en función de $s_n(m)$ puede ser expresado como

$$E_n = \sum_m [s_n(m) - \sum_{k=1}^{p} a_k s(n-k)]^2$$

Para minimizar el error se deriva E_n con respecto a cada a_k igualando el resultado a cero

$$\frac{\partial E_n}{\partial a_k}, k = 1, 2, ..., p$$

Con lo que se obtiene

$$\sum_m s_n(m-i)s_n(m) = \sum_{k=1}^{p} \hat{a}_k \sum_m s_n(m-i)s_n(m-k) \tag{2}$$

El término $\sum_m s_n(m-i)s_n(m-k)$ corresponde a la covarianza de $s_n(m)$

$$\phi_n(i,k) = \sum_m s_n(m-i)s_n(m-k)$$

Sustituyendo términos en la ecuación 2 obtiene

$$\phi_n(i,0) = \sum_{k=1}^{p} \hat{a}_k \phi_n(i,k) \tag{3}$$

Esta expresión describe el conjunto de p ecuaciones con p incógnitas. El mínimo error cuadratico medio puede mostrarse como

$$E_n = \sum_m s_n^2(m) - \sum_{k=1}^{p} a_k \sum_m s_n(m)s_n(m-k)$$

Utilizando la ecuación 3 se puede expresar E_n como

$$E_n = \phi_n(0,0) - \sum_{k=1}^{p} a_k \phi_n(0,k)$$

En la expresión anterior el error cuadrático medio comprende un término fijo $\phi_n(0,0)$ y otros términos que dependen de los coeficientes de predicción. Para resolver la ecuación anterior se debe calcular $\phi_n(i,k)$ para $[1 \leq i \leq p]$ y $[0 \leq k \leq p]$ y, entonces, resolver el conjunto de p ecuaciones resultantes. Existen tres métodos básicos para calcular los coeficientes de predicción: el método de la covarianza (basado en la matriz de covarianza), el método de la autocorrelación y el método de enrejado (lattice). En señales de voz se utiliza casi exclusivamente el método de autocorrelación, debido a que se puede calcular mediante algoritmos computacionalmente eficientes y a que los filtros obtenidos son estables. Este método resulta de restringir el intervalo de evaluación al rango $[0, N-1]$ y asumir que los valores fuera de este rango son iguales a cero. Esto es equivalente a asumir que la señal de voz, $s(m+n)$, es multiplicada por una ventana de longitud finita, $w(m)$, la cual es idénticamente cero fuera del rango definido.

$$s_n(m) = \{s(m+n).w(m), \quad 0 \leq m \leq N-1$$

0, *para el resto*

Se utiliza una ventana de tipo Hamming para eliminar los problemas causados por los cambios rápidos de la señal en los límites de cada trama.
En base al enventanado de la señal se expresa la funcion de covarianza de la siguiente forma

$$\phi_n(i,k) = \sum_{m=0}^{N-1-(i-k)} s_n(m)s_n(m+i-k), \quad 1 \leq i \leq p, \quad 0 \leq k \leq p$$

Dado que ésta es sólo función de las variables i y k (es decir solo depende de la ubicación temporal de la muestra), la función de covarianza se reduce a una simple función de autocorrelación

$$\phi_n(i,k) = R_n(i-k) = \sum_{m=0}^{N-1-(i-k)} s_n(m)s_n(m+i-k)$$

Como la función de autocorrelación es simétrica para señales no complejas como la señal de voz $R_n(-k) = R_n(k)$, las ecuaciones del Predicción Lineal quedan

$$\sum_{k=1}^{p} R_n(|i-k|)a_k = R_n(i), \quad 1 \leq i \leq p$$

El calculo de los coeficientes de predicción se realiza eficientemente mediante el algoritmo recursivo de Levinson y Durbin

$$E^{(0)} = R(0)$$

$$k_i = \left[R(i) - \sum_{j=1}^{i-1} a_j^{i-1} R(i-j) \right] / E^{(j-1)} \quad , \quad 1 \leq i \leq p$$

$$a_i^i = k_i$$

$$a_j^i = a_j^{i-1} - k_i a_{i-j}^{i-1} \ , \quad 1 \leq i \leq i-1$$

$$E^{(i)} = (1 - k_i^2) E^{(i-1)}$$

Las ecuaciones del algoritmo Levinson-Durbin se resuelven recursívamenete para $i = 1, 2, ..., p$ y la solución final viene dada por

$$a_j = a_j^{(p)} \qquad , \quad 1 \leq j \leq p$$

Para completar el modelo de coeficientes de predicción lineal es necesario definir el término de ganancia G, que puede expresarse como

$$G = \sqrt{E^N},$$

donde e es la energía en la muestra N, otra forma de calcular la ganancia es

$$G = \sqrt{R_n(0) \prod_{i=1}^{N} (1 - k(i-1)^2)},$$

el término de ganancia permite ajustar en amplitud el espectro del modelo de Predicción Lineal al espectro original de la señal de voz.

CAPÍTULO 4

Extracción de características

La caracterización consiste en la obtención de parámetros, que de acuerdo a su relevancia (es decir su importancia dentro de la señal) permitan de forma completa o parcial la descripción de la misma. Los principales objetivos de la caracterización son obtener una reducción en la dimensionalidad y realzar aspectos de la señal que contribuyan significativamente a realizar procesos posteriores (reconocimiento, segmentación o clasificación). En el análisis de la voz es común el empleo de dos tipos de características: las características acústicas, las cuales poseen un sentido físico determinado; y las características de representación, que corresponden a valores calculados a partir de alguna forma de representación de la voz, y a los cuales, en general, no les corresponde algún sentido físico [38]. En el caso de la voz la extracción de características se realiza en intervalos de tiempo que comprenden entre 20 y 40 ms, tramas de señal donde ésta se puede considerar como cuasi-estacionaria, es decir sus parámetros estadísticos permanecen invariantes dentro de la trama de observación [49].

4.1 Características acústicas

Como se mencionó anteriormente las características o paramétros acústicos son aquellos que poseen un significado físico, por lo que permiten una calificación de las cualidades vocales.
Los paramétros acústicos se pueden clasificar de la siguiente forma:

- *Paramétros cuasiperiódicos*, que reflejan las variadas formas de periodicidad presentes en la señal de voz: Frecuencia fundamental.
- *Paramétros de perturbación*, que reflejan una variación relativa de un paámetro determinado: Jitter, Shimmer, HNR

4.1.1 Frecuencia fundamental

El pitch o frecuencia fundamental (F_0) se determina por la velocidad de apertura o cierre de las cuerdas vocales en la laringe durante la fonación de sonidos sonoros, es el sinusoide de mayor contenido energético cuyo inverso corresponde a su vez al periodo fundamental T_0. En forma general, la definición de F_0 se determina para intervalos infinitos de análisis, sin embargo, para efectos prácticos, su estimación se realiza sobre intervalos finitos, que permitan cubrir varios periodos del *pitch* (en este caso F_0 se calcularía como el valor promedio de todos los F_0 para un determinado intervalo). Otra manera de calcular el pitch es de manera instantánea es decir a partir de la diferencia entre dos *GCI (glotal close instant)* consecutivos [50]. Se presenta el calculo de F_0 mediante la función de autocorrelación.
Se define la autocorrelación de una señal discreta como

$$r_{xx}(i) = \sum_{n=-\infty}^{\infty} s(n)s(n-i), \qquad i = \pm 0, \pm 1, \pm 2, ...$$

Sea $s(n)$ una secuencia periódica de periodo T, entonces la función autocorrelación $r_{xx}(i)$ tambien es una secuencia periódica de periodo T. Si la trama de voz analizada es sonora la función de autocorrelación presenta un máximo central acompañado de picos laterales que disminuyen su tamaño a medida que se alejan del centro. Si la trama de voz analizada es sorda sólo aparecera el máximo central con lo que se comprueba que no existe F_0 para este segmento de señal. En la figura 16 se muestra la función de autocorrelación para un segmento sonoro de señal.

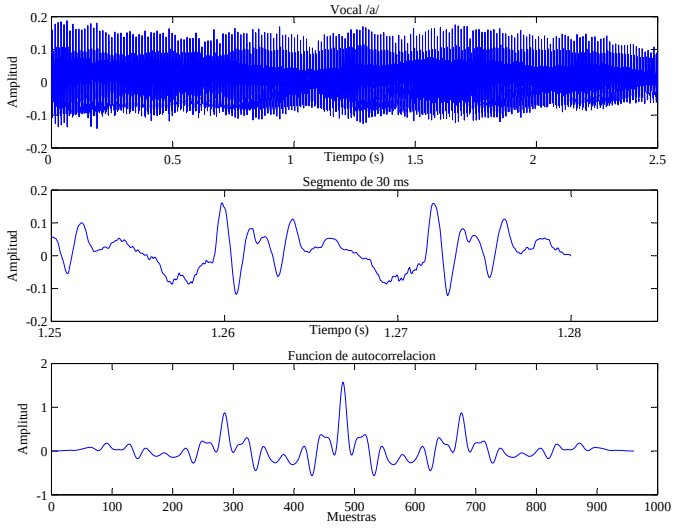


Figura 16: Función de autocorrelación de un segmento de 30 *ms* para la vocal /a/

La F_0 se calcula entonces como el inverso de la distancia entre el máximo central y el máximo adyacente de la de la función de auto correlación $r_{xx}(i)$ [51]. En la figura

17 se pueden observar las variaciones del contorno *pitch* para una misma palabra en diferentes estados emocionales.

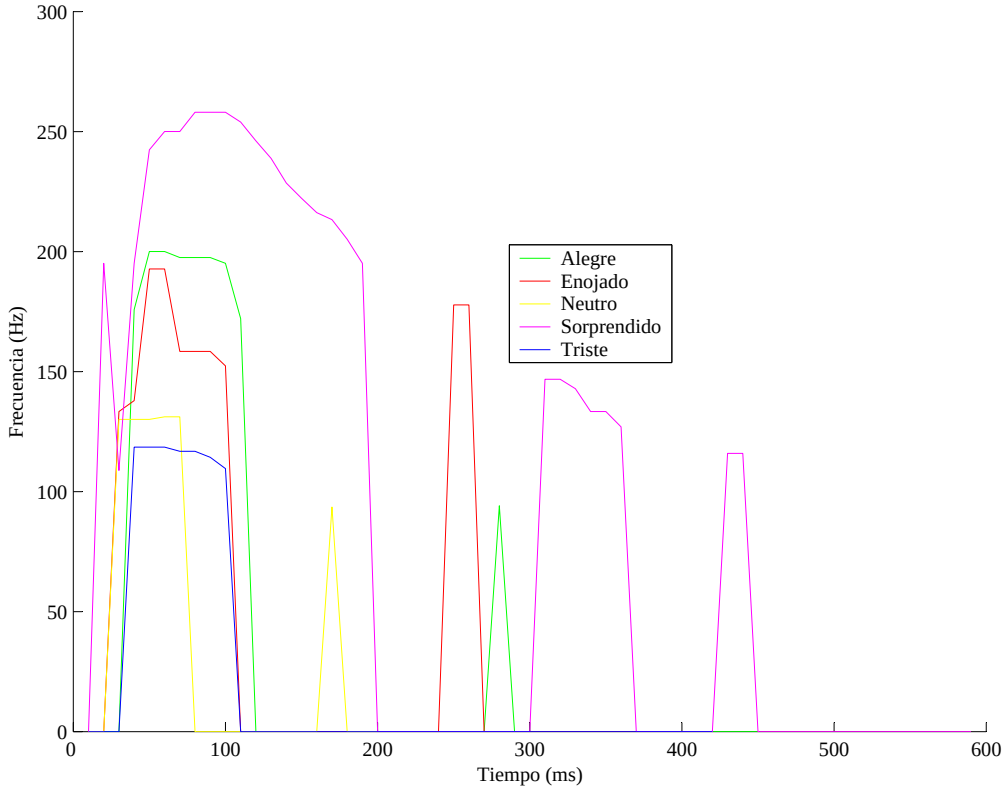


Figura 17: Contorno de F_0 para la palabra */coche/* en diferentes estados emocionales

Obsérvese cómo para este caso el estado emocional sorprendido tiene el mayor contenido de F_0 mientras que en contraste triste tiene el menor.

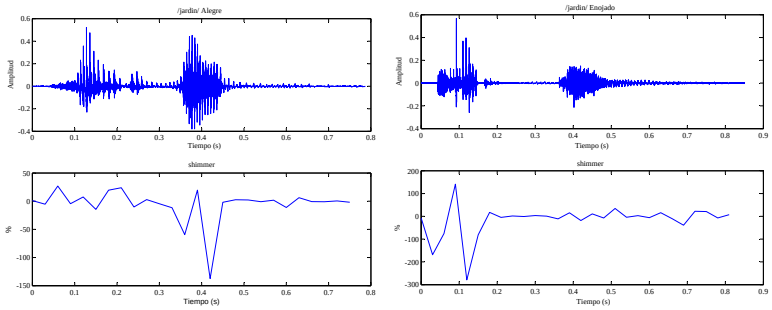
4.1.2 Parámetros de perturbación

Los parámetros de perturbación o variación representan una medida que cuantifica la variación a lo largo del tiempo del valor que toma un parámetro para las distintas tramas. De forma general se puede definir la perturbación de un paramétro como:

$$V_P = \frac{\frac{1}{N-1}\sum_{i=1}^{N-1}|P(i+1)-P(i)|}{\frac{1}{N}\sum_{i=1}^{N-1}P(i)}$$

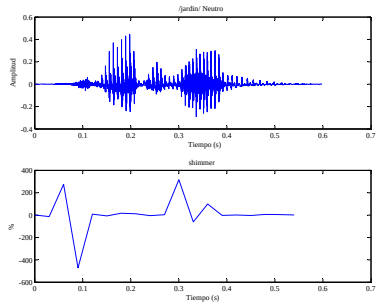
donde P_i es el valor del parámetro para la trama i, N es el número de tramas y V_P es el valor de la perturbación del parámetro [52].

- *Perturabción de amplitud* representa la variación del valor de los picos positivo en cada trama de voz en el dominio temporal.
- *Shimmer* o perturbación de la amplitud maxima representa la variación del valor del pico positivo máximo de cada trama de voz en el dominio temporal. Se determina mediante la variación de la amplitud máxima entre tramas adyacentes. Otros paramétros relacionados con el *shimmer* son: *APQ (amplitude perturbation quotient)*, *SPAQ (smoothed amplitude perturbation quotient)*. La figura 24 muestra la evolución del *shimmer* para la palabra */jardín/* en diferente estados emocionales.

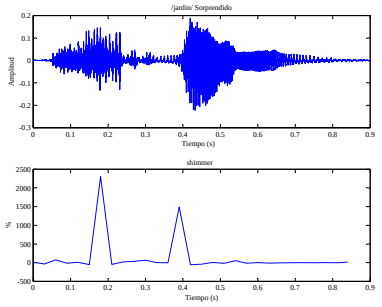


(a) /*jardín*/ Alegre

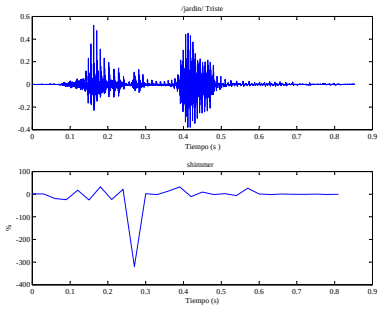
(b) /*jardín*/ Enojado



(c) /*jardín*/ Neutro



(d) /*jardín*/ Sorprendido



(e) /*jardín*/ Triste

Figura 18: Evolución del *shimmer* calculado por ventanas de 30 *ms* para la palabra/*jardín*/ en diferentes estados emocionales

Se puede obseravar para este caso que para los distintos estados emocionales se generan contornos del *shimmer* relativamente dsicriminantes para los diferentes estados emocionales

- *Jitter* o Perturbación de la frecuencia fundamental representa la variación relativa de la F_0 dentro de cada trama de voz en el dominio temporal. Otros paramétros relacionados con el *jitter* son: *RAP (relative average perturbation)*, *PPQ (pitch perturbation quotient)*,
- *HNR (Relación ruido/armónico)* Es una evaluación general de la presencia de ruido en la señal de voz, es útlizado ampliamente por la medicina en la deteccion de patologías o en el filtrado de ruido. Se presenta el cálculo del HNR mediante la función de autocorrelación.

  Si se expresa la señal de audio $y(t)$ como la suma de procesos aleatorios $\xi(t)$ con período T_0 y $\eta(t)$, con las respectivas funciones de autocorrelación $R_\xi(\tau)$ y $R_\eta(\tau)$, tales que puedan ser consideradas como estadísticamente independientes, y por lo tanto, cumplan la condición:

  $$R_y(0) = R_\xi(0) + R_\eta(0)$$

  Si se considera que el proceso $\eta(t)$ es ruido blanco, entonces se encuentra que se cumple la siguiente condición sobre el valor máximo

  $$R_y(\tau_{max}) = R_\xi(\tau_{max}) = R_\xi(0), \forall \tau_{max} = T_0$$

  El valor de $R_y(\tau_{max})$ corresponde a la potencia media del proceso, por lo que el valor de la autocorrelación normalizada en τ_{max} representa la potencia relativa de la componente periódica o armónico del proceso, mientras su complemento es la potencia relativa de la componente del ruido, así:

  $$R_y^{'}(\tau_{max}) = \frac{R_\xi(0)}{R_y(0)}$$

  Luego,

  $$1 - R_y^{'}(\tau_{max}) = \frac{R_\eta(0)}{R_y(0)}$$

  A partir de las ecuaciones anteriores se define la relación armónico/ruido, en forma logarítmica como [5]

  $$HNR = 10log\frac{R_y^{'}(\tau_{max})}{1 - R_y^{'}(\tau_{max})}, (dB)$$

4.2 Características de representación

La características de representación se encargan de describir el comportamiento dinámico de señales, son calculadas a partir de algun método de representación de señales (análisis de prediccion lineal, transformadas tiempo-frecuencia, cepstrum), y generalmente no se les asocia algun sentido físico desde el punto de vista acústico.

4.2.1 Características de representación por medio de la transformada *wavelet*

Utilizando la transformada *wavelet* discreta se obtiene una representación de la señal por medio de los coeficientes de detalle d_i y de aproximación a_i la figura 23 muestra los niveles de descomposición obtebidos al aplicar la *wavelet* discreta de orden cinco sobre la palabra */reina/* en los difirentes estados emocionales.

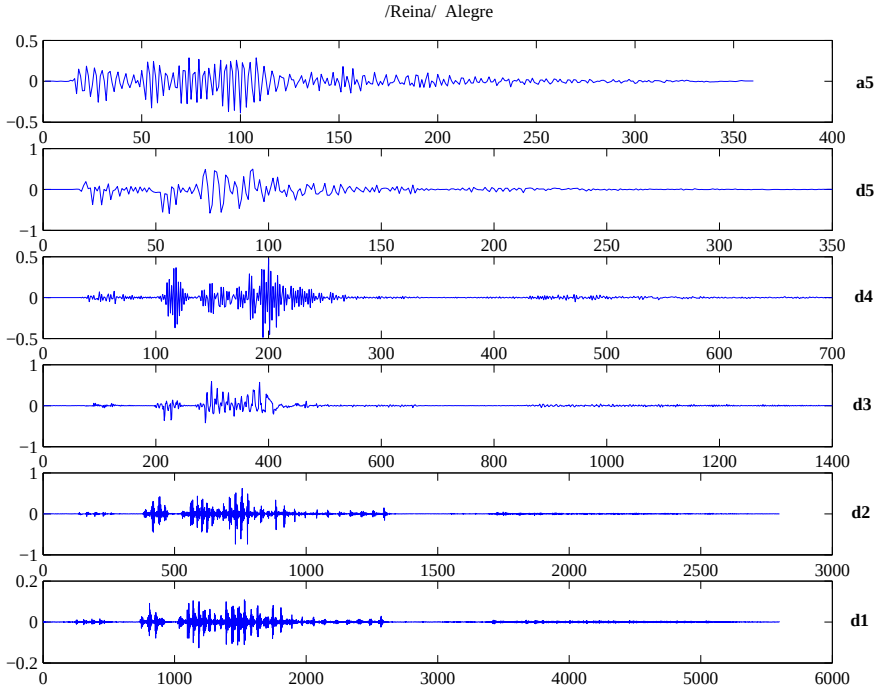


Figura 19: (a) Descomposición en 6 niveles para la palabra */reina/* Alegre

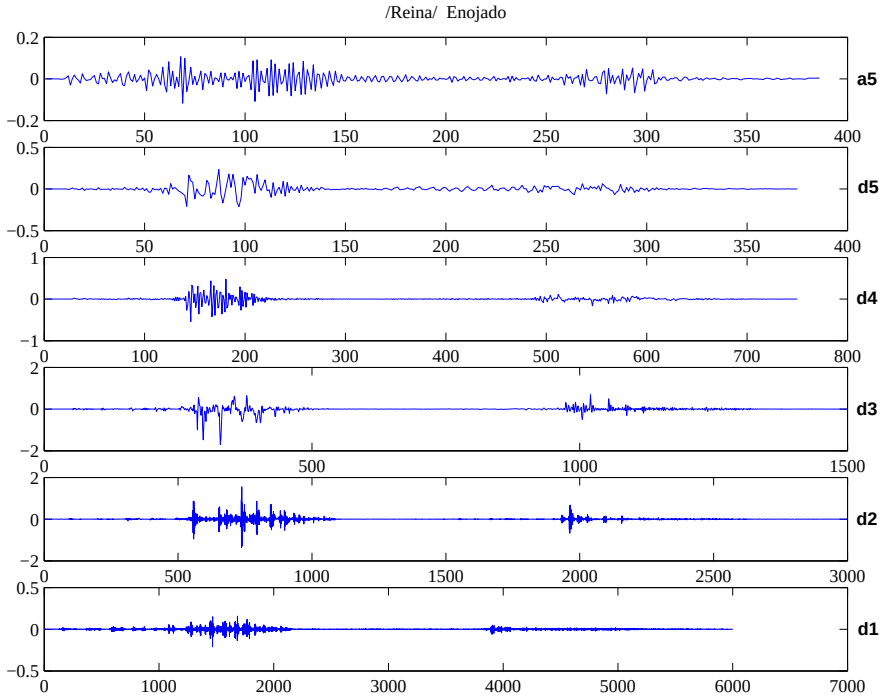


Figura 20: (b) Descomposición en 6 niveles para la palabra /*reina*/ Enojado

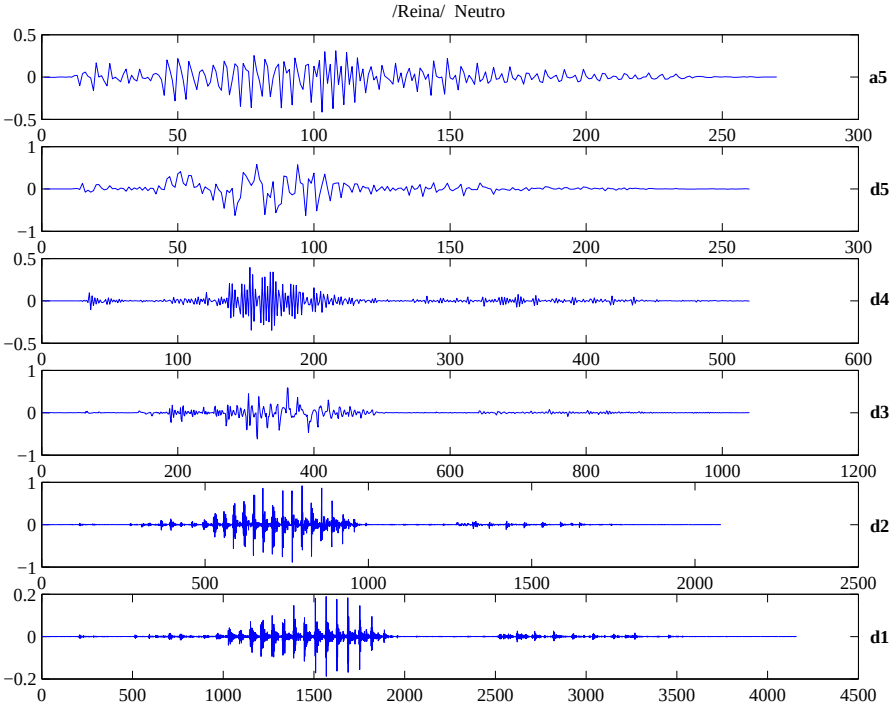


Figura 21: (c) Descomposición en 6 niveles para la palabra /*reina*/ Neitro

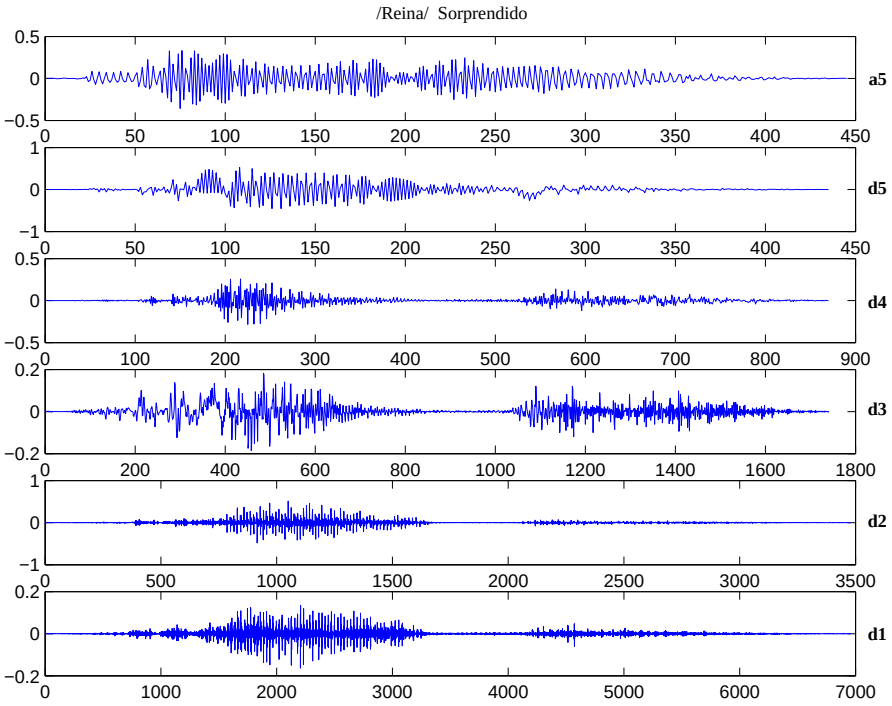


Figura 22: (d) Descomposición en 6 niveles para la palabra /*reina*/ Sorprendido

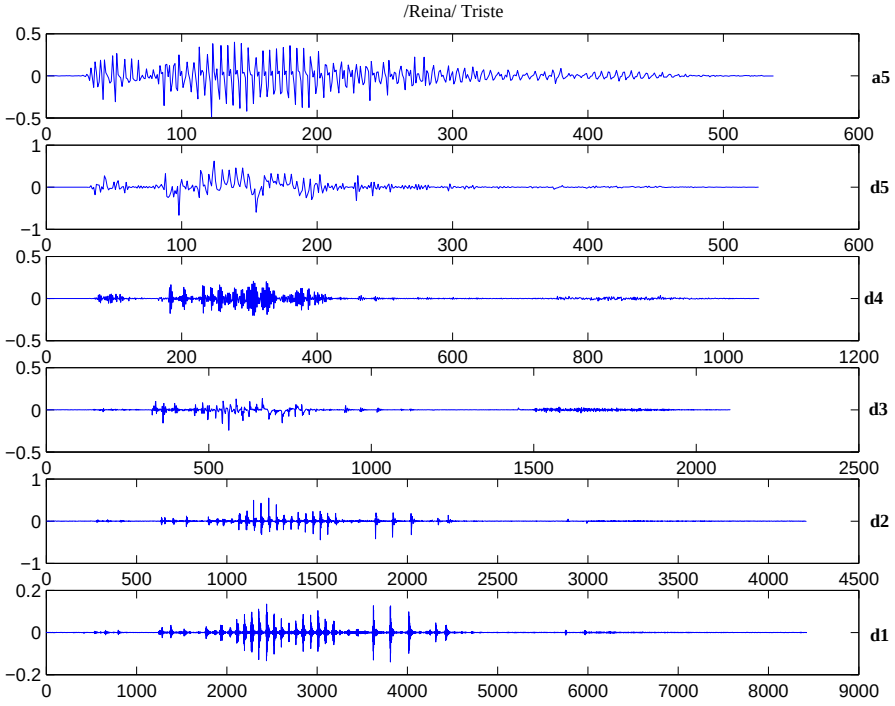
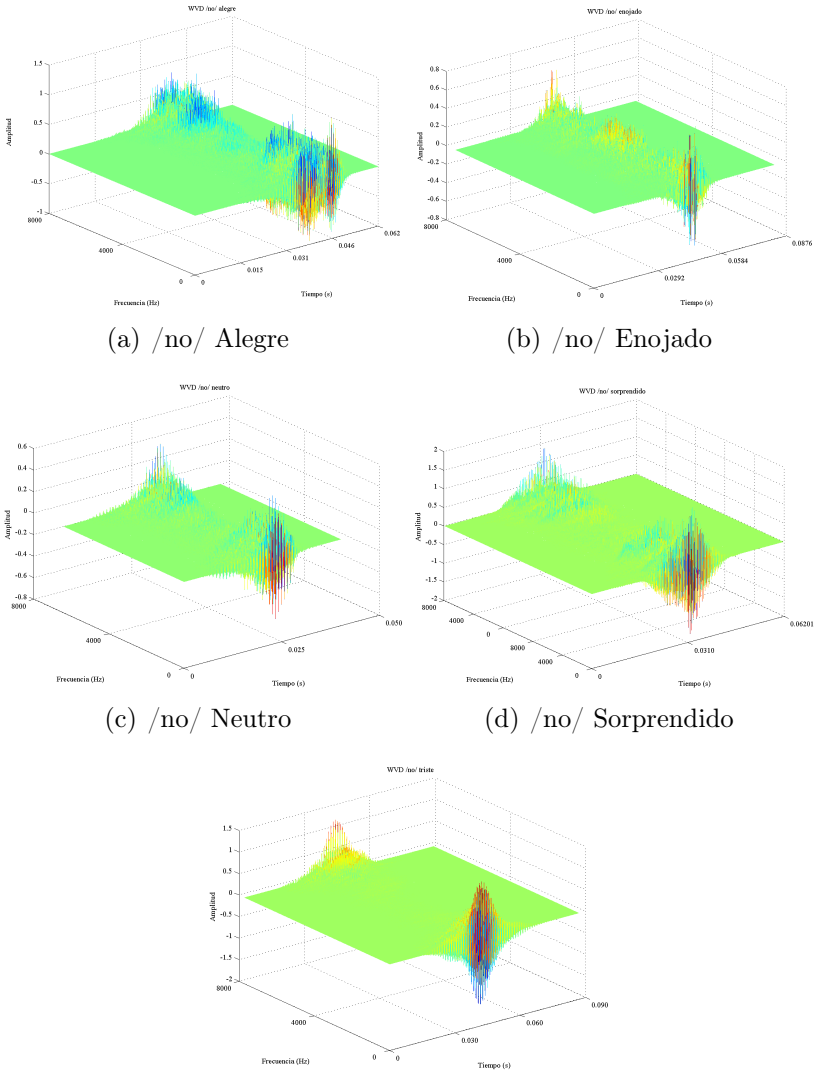


Figura 23: (e) Descomposición en 6 niveles para la palabra /*reina*/ Triste

Se pueden observar diferencias discriminantes entre las bandas de análisis para los diferentes estados emocionales Sobre los coeficientes de la *wavelet* discreta se pueden extraer parámetros estadísticos que finalmente componen un vector de características.

4.2.2 Características de representación por medio de la transformada *Wigner Ville*

Por medio de la transformada *Wigner Ville* se obtiene una distribución de la energía en el espacio conjunto del tiempo y la frecuencia, la cual que se puede representar por medio de una superficie tridimensional.



(a) /no/ Alegre

(b) /no/ Enojado

(c) /no/ Neutro

(d) /no/ Sorprendido

(e) /no/ Triste

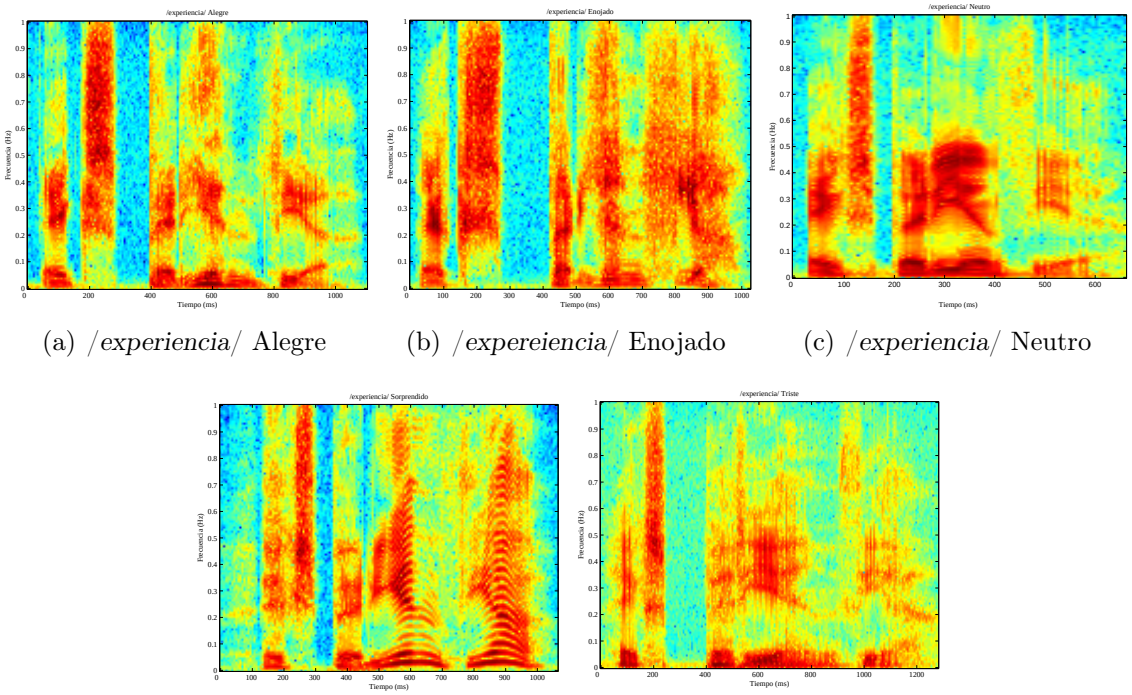
Figura 24: Distribución de *Wigner Ville* para la palabra */no/* en diferentes estados emocionales

Sobre la superficie obtenida se pueden extraer paramétros estadísticos que componen el vector de características.

4.2.3 Características de representación por medio de la transformada Gabor

Por medio de la transformada Gabor se puede obtener una distribución de la energÍa de la señal en un plano tiempo-frecuencia, altos contenidos energéticos están representados por colores calidos mientras que los colores fríos representan bajos contenidos energéticos. La figura 25 muestra los espectogramas obtenidos por

medio de la transformada gabor para la palabra /*experiencia*/ en diferentes estados emocionales.



(a) /*experiencia*/ Alegre (b) /*expereiencia*/ Enojado (c) /*experiencia*/ Neutro

(d) /*experiencia*/ Sorprendido (e) /*experiencia*/ Triste

Figura 25: Espectrograma de la palabra /*experiencia*/ para diferentes estados emocionales

Sobre el espectrograma obtenido se pueden extraer paramétros estadísticos que finalmente componen un vector de caracteristicas.

4.2.4 Características de representación por medio del análisis de predicción lineal

Como se dijo en el capítulo 3 la idea basica del análisis de predicción lineal consiste en determinar los LPC que minimizen el error de predicción para cada trama de la señal, los LPC permiten una representacíon de la señal optima en cuanto a la reducción de parámetros y la simplicidad de su cálculo.
Sobre los datos obtenidos (alguna representación de la señal) y la señal original (*raw datda*) se pueden hallar paramétros estadísticos que permiten una caracterización simple de la misma. Se extrajeron los siguientes paramétros estadísticos.

- Media
- Mediana
- Máximo

- Mínimo
- Varianza
- Desviación estándar
- Asimetría
- Curtosis

Una descripción detallada de estos parámetros y el listado de características se presenta en el apéndice C.

CAPÍTULO 5

Marco experimental y resultados

5.1 Metodología de la evaluación

Los resultados de clasificación y reconocimiento pueden verse afectados por la falta de cuidado en el diseño de los experimentos. Con el fin de garantizar la validez estadística de los resultados, se estimaron los errores de clasificación usando validación cruzada. La validación cruzada del tipo $K - fold$ divide la muestra original en K conjuntos de muestras. De cada una de estos K conjuntos, una solo grupo K se retiene como dato de validación y las restantes $K - 1$ muestras se usan como datos de entrenamiento. El proceso de validación cruzada se repite entonces K veces y cada una de las muestras en los K conjuntos es empleada sólo una vez como dato de validación. Finalmente, se promedian las K estimaciones del error de clasificación para obtener una única estimación.
En la *validación cruzada leave−one−out* se emplea una sola muestra de todo el conjunto muestral como dato de validación y las muestras restantes como datos de entrenamiento. Este procedimiento se repite de tal manera que todas las muestras del espacio inicial sean empleadas como datos de validación una única vez. Esto es similar a tener cross-validación del tipo *K-fold* donde K es igual al número de observaciones en la muestra original. En la etapa de clasificación se utiliza el método de cross-validación *leave one-out*, y un clasificador bayesiano lineal.

5.2 Descripción de la base de datos

Se trabajó sobre la Base de Datos SES [1]. la cual es una base de datos monolocutor de habla emocional en español en la que el locutor, un actor profesional, simula habla triste, alegre, sorprendido, enfadado y neutro. Consta de diversas sesiones de grabación, donde cada sesión contiene diversas palabras, frases o párrafos. Los ficheros de voz son archivos *.PCM, grabados a 16 kHz, 16 bits, sin cabecera y en formato Intel (little endian). En total consta de 30 palabras, 14 frases y 4 párrafos con cada uno de los estados emocionales mencionados.

5.3 Caracterización de señales de voz

La extracción de características en señales de voz se realiza con el fin de detectar estados emocionales en la señal de voz. En este trabajo se emplean dos metodologías: la primera de ellas se basa en el empleo de transformadas tiempo-frecuencia y la segunda en el empleo de análisis de predicción lineal.
Como no existe un conocimiento a *priori* de las características que van a proporcionar un mejor resultado en el reconocimiento de emociones, se consideró adecuado contar un gran número de parámetros diferentes para posteriormente descartar aquéllos que resultaran redundantes.
Como se dijo en el capítulo 4, se plantean dos tipos de características a extraer en señales de voz: *de representación* y *acústicas*. En primer lugar, las características de representación se extraen de las transformaciones que se aplican a la señal. Pueden ser vistas como parámetros estadísticos que aportan información de la naturaleza de la densidad espectral de energía de la señal. Las características acústicas pueden ser vistas como parámetros que aportan información sobre cualidades físicas de la voz (evolución de la tonalidad o variación de la amplitud). Para todos los casos (Excepto en Gabor que la ventana de análisis realiza la segmentación) la señal fue previamente segementada con ventanas de $30ms$ de duración y un traslape del 50%, se promediaron los valores de cada uno de los parámetros extraidos por cada trama con lo que se obtuvo un vector de características promedio para cada señal. Se obtuvieron un total de 104 características agrupadas de acuerdo a la metodología usada en su extracción, para cada grupo se realizaron todas las combinaciones posibles de características a modo de descartar aquellas que presentaran redundancia, bajo un criterio de mejor porcentaje de acierto en la clasificación. El conjunto total de características se detalla en el apéndice C.
Los resultados de reconocimiento de emociones se muestran a continuación:

- Empleando los **Coeficientes de Predicción Lineal (LPC)**. tabla 1:

[1] esta base de datos es propiedad de la Universidad Politécnica de Madrid, Departamento de Ingeniería Electrónica, Grupo de Tecnología del Habla, ETSI Telecomunicación, Ciudad Universitaria, 28040 Madrid España

	Salida					
Entrada	Alegre	Enojado	Neutro	Sorprendido	Triste	Acierto (%)
Alegre	15	3	2	8	2	50.00
Enojado	1	21	5	3	0	70.00
Neutro	1	1	25	0	3	83.33
Sorprendido	1	2	0	24	3	80.00
Triste	5	6	1	6	12	40.00
Acierto Global (%)						**64.66**

Tabla 1: Matriz de confusión empleando LPC

El porcentaje de acierto más alto so obtuvo para neutro con un 83.33% de aciierto, meientras que el más bajo se obtuvo para triste con 40%, el porcentaje global de acierto es bajo con un 64.66% y presenta alta varianza, presentando problemas para el reconocimiento de estados emocionales de bajo contenido enérgetico como es la tristeza.

- Empleando la **Transformada Wigner Ville**. tabla 2:

	Salida					
Entrada	Alegre	Enojado	Neutro	Sorprendido	Triste	Acierto (%)
Alegre	14	6	7	3	0	46.66
Enojado	4	21	4	1	0	70.00
Neutro	4	0	26	0	0	86.66
Sorprendido	3	0	0	26	1	86.66
Triste	4	1	3	8	14	46.66
Acierto Global (%)						**67.33**

Tabla 2: Matriz de confusión empleando Transformada Wigner Ville

El porcentaje de acierto más alto se obtuvo para los estados emocionales enojado y sorprendido con un 86.66% de acierto, mientras que el más bajo para alegre y triste con un 46.66%, se exhibe una tendencia a sobre clasificar el estado emocional neutro en comparación con el resto de estados.El rango entre porcentajes de acierto alto y bajo se debe a la no linealidad que presenta la transformada Wigner Ville, ya que genera términos de inferencia en la matriz de distribución [5], los cuales no permiten una correcta discriminación de los todos estados emocionales.

- Empleando la **Transformada Wavelet Discreta**. tabla 3:

	Salida					
Entrada	Alegre	Enojado	Neutro	Sorprendido	Triste	Acierto (%)
Alegre	22	2	1	2	3	73.33
Enojado	2	18	6	3	1	60.00
Neutro	2	7	20	1	0	66.66
Sorprendido	5	2	1	21	1	70.00
Triste	4	3	1	1	21	70.00
Acierto Global (%)						**68.00**

Tabla 3: Matriz de confusión empleando Transformada Wavelet Discreta

Para este calculo se realizaron pruebas con las *Wavelet* madres *db*3 (*daubechies*) y *sym*6 (*symlets*) con 6 niveles de descomposición. El porcentaje de acierto más alto se obtuvo para el estado emocional alegre con un 73.33% de acierto, mientras que el más bajo se obtiene para enojado con un 60%. Una de las ventajas del resultado global obtenido mediante la transformada *wavelet* es que presenta baja varianza lo que indica una mejor discriminación de los estados emocionales, en comparación con los resultados obtenidos por medio de la transformada de Wigner Ville.

- Empleando la **Transformada Gabor**. tabla 4:

	Salida					
Entrada	Alegre	Enojado	Neutro	Sorprendido	Triste	Acierto (%)
Alegre	22	3	2	2	1	73.33
Enojado	6	21	1	2	0	70.00
Neutro	0	2	25	1	2	83.33
Sorprendido	5	1	0	24	0	80.00
Triste	1	1	4	0	24	80.00
Acierto Global (%)						**77.33**

Tabla 4: Matriz de confusión empleando Transformada Gabor

El porcentaje de acierto más alto se obtuvo para el estado emocional neutro con un 83.33% de acierto, mientras que el más bajo se obtuvo para enojado con un 70.00%, esto se debe a que la transformada Gabor mantiene un compromiso de resolución entre el tiempo y la frecuencia constante, es decir presenta buenos resultados si el contenido de frecuencias premanece parcialmente cónstante como es el caso del estado emocional neutro, además elimina los términos cruzados que producen inferencia en la matriz de distribución permitiendo una mejor discriminación de los estados emocionales.

	Salida					
Entrada	Alegre	Enojado	Neutro	Sorprendido	Triste	Acierto (%)
Alegre	25	3	0	1	2	83.33
Enojado	2	21	0	4	3	70.00
Neutro	3	2	23	0	2	76.66
Sorprendido	1	1	0	28	0	93.33
Triste	3	1	0	2	24	80.00
Acierto Global (%)						**80.66**

Tabla 5: Matriz de confusión empleando *rawdata*

El porcentaje de acierto más alto se obtuvo para el estado emocional sorprendido con un 93.33% de acierto, mientras que el más bajo se obtuvo para enojado con un 70.00%. Aunque no es el mejor porcentaje de acierto, el estado emocional neutro presenta una precisión de 100% (confusión mínima), es decir, es fácilmente discriminable de los demás. La caracterización por *rawdata* representa la variación de la intensidad y duración de las palabras, aunque pareciera poco óptimo analizar directamente la señal de voz (por el elevado numero de datos), se obtienen buenos resultados en la detección de estados emocionales con la aplicación de este método. A modo de obtener mejor resultados se compone una nueva matriz de características a paratir de las características discriminates obtenidas con las técnicas anteriores, de nuevo se realizan todas las posibles combinaciones a modo de descartar las características que presenten mayor redundancia. El proceso de discriminación se realiza en base al mayor porcentaje de acierto en la clasificación.

- Empleando el **Método Combinado**. tabla 6:

	Salida					
Entrada	Alegre	Enojado	Neutro	Sorprendido	Triste	Acierto (%)
Alegre	28	1	0	1	0	93.33
Enojado	1	27	2	0	0	90.00
Neutro	0	0	30	0	0	100.00
Sorprendido	0	0	0	30	0	100.00
Triste	2	0	0	0	27	90.00
Acierto Global (%)						**94.66**

Tabla 6: Matriz de confusión con Método Combinado

El porcentaje de acierto más alto se obtuvo para los estados emocionales neutro y sorprendido con un 100% de acierto, mientras que el más bajo se obtuvo para enojado y triste con un 90%. Este método presenta los mejores resultados con un acieto global de 94.66%, menor varianza y la precisión mas alta en comparación con cualquiera de las metodologías utilizadas en forma independiente.

La combinacíon de características mediante la cual se obtuvo el resultado anterior es la siguiente:

Características	**Descripción**
kuwv	Media de la curtosis de los coeficienters de la Walet Discreta
vamax	*Shimmer* del *raw data*
meshc	Mediana de la perturbación de la amplitud del *raw data*
mec	Media de la mediana del *raw data*
sdc	Media de la desviación estándar del *raw data*
mic	Media del mínimo del *raw data*
vas	Varianza de la función de sonoridad
vaA	Media de la varianza de los LPC
kuA	Media de la curtosis de los LPC

Tabla 7: Listado de la mejor combinación de características

Por ultimo se presentan todos los resultados de acierto con las diferentes técnicas utilizadas

Técnicas de caracterizacíon	**Aciertos (%)**	**Desviación Estándar**
HNR	**35.33**	**9.88**
Contorno de la Energía	**49.33**	**26.86**
Funcion de sonoridad	**52.00**	**19.52**
Contorno del *Pitch*	**53.33**	**32.06**
Coeficientes De Predicción Lineal (LPC)	**64.66**	**18.94**
Transformada Wigner Ville	**67.33**	**20.05**
Transformada Wavelet Discreta	**68.00**	**5.05**
Transformada Gabor	**77.33**	**5.47**
raw data	**80.66**	**8.66**
Método Combianado	**94.66**	**5.05**

Tabla 8: Tabla general de resultados

Conclusiones

- Se deasarrolló una técnica de caracterización de la señal de voz, basada en la utilización de características de representación y acústicas, la cual presentó buenos resultados en el reconocimiento de estados emocionales. Dicha técnica se validó por medio de un clasisficador bayesiano lineal, dada la simplicidad del clasificador y su poca influencia dentro del resultado se garantizó la validez de la metodología propuesta para la extracción de características.

- Los mejores porcentajes de acierto empleando Transformaciones tiempo-frecuencia se obtuvieron con la trnasformada Gabor con un 9.33% por encima de la transformada *Wavelet*, debido a dos factores: el primero de ellos, la eliminación de los términos cruzados por medio de la ventana de suavizamiento. El segundo, el compromiso que existe entre las resoluciones del tiempo y la frecuencia debido al ancho fijo de la ventana de análisis.

- Los resultados de los trabajos de reconocimiento de emociones son difíciles de comparar, pues se utilizan bases de datos muy distintas. Algunos utilizan bases de datos multilocutor y otros monolocutor y el conjunto de emociones básicas consideradas no es el mismo en todos los casos. Un trabajo destacable es el de Luengo y Navas (2005), en el cual se trabajó sobre una base de datos monolocutor de 97 grabaciones de habla emocional en euskara, con un conjunto de emociones basicas similares a las utilizadas en este trabajo. Se extrajeron características espectrales calculadas por medio de MFCC y parámetros prosódicos (con ayuda de un laringógrafo), las características fueron validadas con dos tipos de clasificadores *GMM* (*Gaussian Mixture Models*) y (*SVM*) máquinas de soporte vectorial. Se obtuvo un porcentaje de acierto global de 98.4% utilizando 512 componentes gaussianas y un porcentaje de 92.3% con la aplicación de un discriminador de características. Sin embargo el trabajo mas proximo al nuestro es el de J.M. Montero y J. Macías (2006) los cuales utilizaron la base de datos SES sobre la cual se calcularon características

espectrales por medio de MFCC y parámetros prosódicos (con ayuda del software praat), dichas características fueron validadas por medio de un clasificador bayesiano lineal, obteniendo un porcentaje de acierto global de 94.22%

Apéndice A

Modelo de tubos

Un modelo de producción ampliamente utilizado es el basado en la suposición de que el tracto vocal puede ser representado como una concatenación de tubos acústicos sin pérdidas, como se muestra en la figura 9. Esta suposición está basada en considerar plana la onda propagada a través del tracto vocal. Esto es, a diferencia de la situación de campo libre, en el tracto vocal, para la mayoría de las frecuencias de interés del sonido, la onda sonora se propaga en una sola dimensión, a lo largo de un eje.

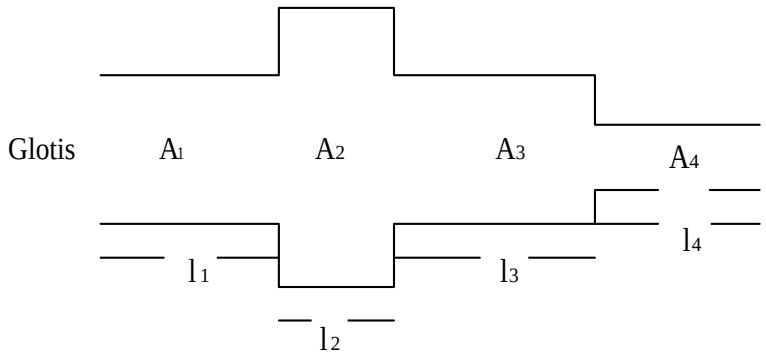


Figura 26: Modelo general de tubos

Esto es válido para frecuencias cuya longitud de onda es grande comparada con el diámetro del tubo. Así, para un tono de 4 kHz, teniendo en cuenta que el diámetro del tracto vocal es como mínimo de $2cm$, esta suposición se cumple, dado que

$$\lambda = \frac{c}{f} = \frac{340ms}{4kHz} = 8.5cm$$

Se asume que la onda viaja a través de un tubo de sección A constante, cuyas paredes son rígidas y en forma de tubo sin pérdidas. Si bien la suposición de que no hay pérdidas no es correcta, un modelo de estas características presenta facilidad computacional y una buena aproximación. Con todas estas suposiciones y aplicando las leyes de conservación de la masa, el momento y la energía, se puede

demostrar que las ondas sonoras en el tubo satisfacen el siguiente par de ecuaciones:

$$\frac{\partial u}{\partial x} = \frac{A}{\rho c^2}\frac{\partial p}{\partial t}$$

$$\frac{\partial p}{\partial x} = \frac{\rho}{A}\frac{\partial u}{\partial t}$$

Donde x se mide desde la glotis hasta el final de cada tubo ($0 \leq x \leq l_i$), p es la presión del sonido, ρ representa la densidad del aire en el tubo y c es la velocidad del sonido. Estas ecuaciones, combinadas con la ecuación diferencial de segundo grado

$$\frac{\partial^2 u}{\partial x^2} = \frac{1}{c^2}\frac{\partial^2 u}{\partial t^2}$$

Tienen la solución para la velocidad volumétrica $u(x,t)$ en el $i-simo$ tubo y la presión de la siguiente forma

$$u_i(x,t) = u_i^+(t - \frac{x}{c}) - u_i^-(t - \frac{x}{c}),$$

$$p_i(x,t) = \frac{\rho c}{A_i}[u_i^+(t - \frac{x}{c}) - u_i^-(t - \frac{x}{c})],$$

En ambas soluciones se evidencia la presencia de dos ondas. Una en el sentido de la propagación onda progresiva ($u_i^+(t - \frac{x}{c}), p_i^+(t - \frac{x}{c})$) y otra en sentido contrario onda regresiva ($u_i^+(t - \frac{x}{c}), p_i^+(t - \frac{x}{c})$). La onda regresiva se produce como consecuencia de que en la unión de dos tubos o secciones no se transfiere toda la energía, debido a que hay una porción de la misma que se refleja. La relación entre ondas progresiva y regresiva viene dada en función de las condiciones de continuidad en la unión de dos secciones, por lo que en el límite de las secciones i e $i+1$, para la velocidad volumétrica, tendremos las siguientes condiciones

$$u_i^+(t - \tau_i) - u_i^-(t - \tau_i) = u_{i+1}^+(t) - u_{i+1}^-(t),$$

$$\frac{\rho c}{A_i}[u_i^+(t - \tau_i) - u_i^-(t - \tau_i)] = \frac{\rho c}{A_{i+1}}[u_{i+1}^+(t) - u_{i+1}^-(t)],$$

donde τ_i es el tiempo necesario para que la onda de sonido de propague a través de la sección de longitud . Resolviendo las ecuaciones anteriores se obtiene

donde

$$r_i = \frac{A_{i+1} - A_i}{A_{i+1} + A_i}$$

$$\beta_i = \frac{2A_{i+1}}{A_{i+1} + A_i} = 1 + r_i$$

$$\phi_i = \frac{2A_{i+1}}{A_{i+1} + A_i} = 1 + r_i$$

A_I es la sección del tubo $i - esimo$. El término r_i es el coeficiente de reflexión entre las secciones i e $i + 1$, e indica qué cantidad de onda incidente es reflejada. La magnitud está limitada a la unidad y sólo será igual a uno cuando una de las áreas de un límite sea cero o infinito, es decir, toda la onda será reflejada cuando encuentre el final de un tubo, como en los instantes de cierre glótico (r_0 =1) o cuando encuentre la unión de los labios (r_N = -1). β y ϕ representan la cantidad de onda propagada que pasa mas allá de un límite para las ondas de ambos sentidos.

Cada unión en un sistema como el de la figura 9 puede ser representada con un sistema como el de la figura 10. En él, los símbolos τ equivalen a retardos.

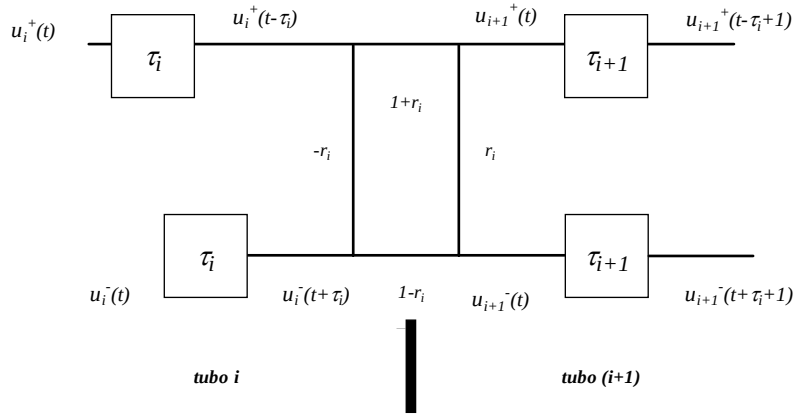


Figura 27: Modelo de tubos

Se considera un modelo de N tubos sin pérdidas, de longitud l_i y sección A_i, donde $i = 1$ en la glotis y N en los labios. Los efectos de vibración, fricción y pérdidas térmicas se incluyen en los modelos de la glotis y los labios. Los coeficientes de reflexión en la glotis y en los labios (r_G y r_L) responden a las siguientes expresiones.

$$r_L = \frac{(\rho c/A_N) - Z_L}{(\rho c/A_N) + Z_L}$$

$$r_G = \frac{Z_G - Z_{01}}{Z_G + Z_{01}}$$

donde Z_L (valor de impedancia labial) se comporta como un diferenciador; muy pequeño a bajas frecuencias y creciente (6 dB/oct) para las frecuencias de interés. Asimismo Z_G (valor de impedancia glotal) no es una impedancia fija, su valor depende de la posición de las cuerdas vocales. En la figura 10 se muestra el modelo completo del tracto vocal utilizando tres tubos. Típicamente N toma valores entre 8 y 12, lo cual permite una simulación más adecuada que con sólo 2 o 3 tubos. Dado que el modelo de tubos contiene solo sumas, productos y retardos se puede transformar facilmente en un modelo discreto.

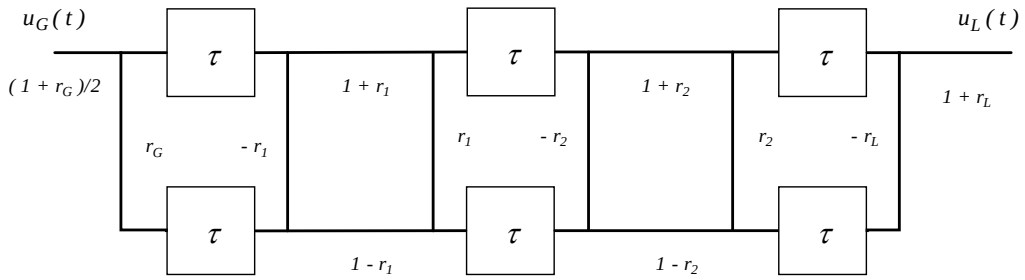


Figura 28: Modelo de tres tubos analógico

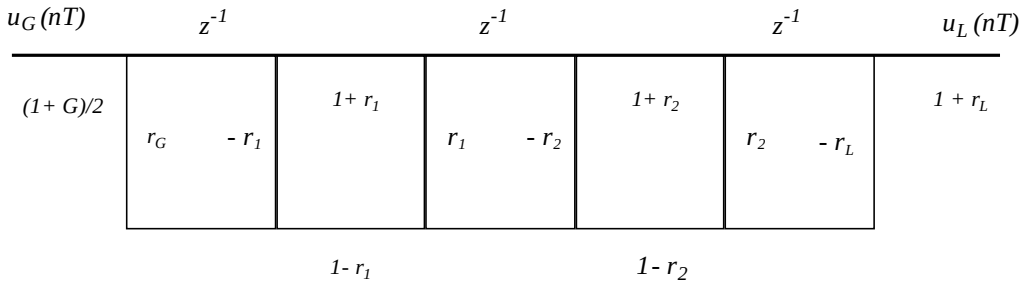


Figura 29: Modelo de tres tubos discreto

Finalmente,ignorandos los efectos de la radiación, la función de transferencia para un modelo de N tubos sería

$$H(z) = \frac{U_L(Z)}{U_G(Z)} = \frac{1 + r_G}{2} * \frac{z^{-N/2} \prod_N^{i=1} 1 + r_i}{1 - \sum_N^{i=1} a_i z^{-i}}$$

donde a_i depende de los coeficientes de reflexión r_i. Los N polos de $H(z)$ representan los formantes del espectro. $H(z)$ no tiene ningún cero, excepto en el origen. Esto se

debe a que esta función de transferencia se obtiene de un modelo en el que sólo hay un camino de propagación para las ondas. En el caso de sonidos nasales o fricativos este modelo quedaría incompleto. Sin embargo, el modelo todo polos (AR) proporciona una buena representación para casi todos los sonidos de voz [49].

Apéndice B

Propiedades generales da las TFR.

Como se ha comentado, las representaciones tiempo-frecuencia muestran la evolución de las componentes frecuenciales de una señal a lo largo del tiempo. Pero en muchas ocaciones, para el caso de las representaciones cuadráticas, se desea que la representación tiempo-frecuencia muestre la cantidad de energía de la señal que existe en cada punto del plano $t-f$, con lo que la distribución se podría tratar como otro tipo de densidad de energía, pudiendo calcular momentos locales y globales. Comprobando el cumplimiento de diversas propiedades, se puede averiguar qué condiciones deben cumplirse para que una TFR se corresponda con una densidad de energía en el sentido estricto de la palabra.

B.1 Marginales

Si se suman todos los términos de la distribución correspondientes a un mismo tiempo o frecuencia, se debe obtener la energía instantánea y el espectro de densidad de energía, respectivamente. Por lo tanto, las condiciones marginales de tiempo y freceuncia establecen:

$$P(\omega) = \int_{-\infty}^{\infty} P(T,\omega)dt = |S(\omega)|^2 \quad P(t) = \int_{-\infty}^{\infty} P(t,\omega)d\omega = |S(t)|^2$$

$P(t,\omega)$ coreesponde con la intensidad en el punto (t,ω) del plano t-f, $|S(t)|^2$ es la intensidad de la señal en el tiempo t, y $|S(\omega)|^2$ la intensidad en la frecuencia ω.

B.2 Energía total

Si la transformada t-f se corresponde con una densidad de energía, se debe cumplir que:

$$\int\int P(t,\omega)d\omega dt = \int_{-\infty}^{\infty} |s(t)|^2 \, dt = \int_{-\infty}^{\infty} |s(\omega)|^2 \, d\omega$$

esta propiedad se cumple automáticamente si los margianles lo hacen, aunque lo contrario no es cierto.

B.3 Invarianza ante desplazamientos temporales y frecuenciales.

Sea una señal $s(t)$, si se tienen desplazamientos en el tiempo:

$$\tilde{s}(t) = s(t - t_0)$$

La representación tiempo-frecuencia correspondiente a dicha señal es invariante ante los desplazamientos temporales cuando se cumple:

$$P_{\tilde{x}}(t,\omega) = P_x(t - t_0, \omega)$$

Si en cambio se tiene que para la señal $s(t)$ existe un desplazamiento en frecuencia:

$$\tilde{s}(t) = s(t)e^{j\omega_0 t}$$

La representación TFR cumple la propiedad de invarianza ante desplazamientos frecuenciales si:

$$P_{\tilde{x}}(t,\omega) = P_x(t, \omega - \omega_0)$$

B.4 Escalado lineal.

Dada la señal $s(t)$, para una constante **a**, la versión escalada de la señal es $s_c(t) = \sqrt{a}.s(at)$,con lo que $s_c(t)$ es expandida o reducida dependiendo de si **a** es mayor o menor que la unidad. Obteniendo el espectro de la señal:

$$S_s(\omega) = \frac{1}{\sqrt{a}} S(\omega/a)$$

Como se aprecia, si la señal se comprime, el espectro se expande y viceversa. Para que esta propiedad se cumpla dentro del ámbito de la distribuciones tiempo-frecuencia, se debe obtener que:

$$P_s(t, \omega) = P(at, w/a)$$

B.5 Soporte finito de la señal.

Algo que parece obvio, pero no siempre se cumple, es la condición de que la TFR no comience mientras la señal no lo haya hecho (comienzo de la señal en el tiempo t_1), y que no se siga manteniendo una vez que la señal ha finalizado (finalización de la señal en el tiempo t_2). Así, se dice que si la TFR no toma valores mientras no lo hace la señal, la condición del soporte finito se cumple: Matemáticamente, la propiedad está expresada de la siguiente forma:

$$P(t, \omega) = 0 \; para \; t \notin \; (t_1, t_2), \; si \; s(t) = 0 \; para \; cualquier \; t \notin (t_1, t_2)$$
$$P(t, \omega) = 0 \; para \; \omega \notin \; (\omega_1, \omega_2), \; si \; S(\omega) = 0 \; para \; cualquier \; \omega \notin (\omega_1, \omega_2)$$

B.6 Distribuciones reales y positivas.

A menudo es importante que los resultados que la distribución tiempo-frecuencia proporciona sean fácilmente manejables, por ello dos de las propiedades que conviene que se cumplan son su valor positivo y su pertenencia al conjunto de números reales:

$$P(t,\omega) = P^*(t,\omega)\ y\ P(t,\omega) \geq 0$$

B.7 Frecuencia instantánea y Retardo del grupo.

una de las formas de describir la evolución frecuencial de una señal a lo largo del tiempo y que para señales con determinadas propiedades puede ser válida para su caracterización, es a través de la frecuencia instantánea y el retardo de grupo, éstos viene definidos por la derivada en el tiempo de la fase de la señal y la derivada en la frecuencia de la parte imaginaria de la transformada de Fourier de señal, respectivamente:

$$\varphi^{'}(t) = \frac{1}{2\pi}\frac{d(t)}{dt} \Rightarrow Frecuencia\ Intantnea$$

$$\psi^{'}(\omega) = -\frac{d(\omega)}{dt} \Rightarrow Retardo\ de\ Grupo$$

como descripción general, se puede decir que la frecuencia instantánea da la idea del valor frecuencial imperante a un determinado momento y que, a su vez, puede ser suma de varias componentes frecuenciales, así, se dice que una TFR cumple la propiedad de la frecuencia instantánea si:

$$\frac{\int \omega P(t,\omega)d\omega}{P(t,\omega)d\omega} = \varphi^{'}(t)$$

es decir, la frecuencia media de la TFR en un determinado momento coincide con la frecuencia instantánea de la señal en ese instante. Por lo tanto, esta magnitud proporciona información útil sólo si la señal contiene en cada instante un rango de frecuencias estrecho.

De forma análoga, pero en el plano temporal, se define el retardo de grupo, que da idea del tiempo medio de llegada de una determinada frecuencia dentro de la señal analizada. Por lo que respecta a las TFR, se dice que cumplen la propiedad de retardo de grupo si:

$$\frac{\int tP(t,\omega)dt}{P(t,\omega)dt} = \psi^{'}(\omega)$$

con lo que la media o "centro de gravedad" en la dirección temporal debe ser igual al retardo de grupo.

De las propiedades anteriormente definidas las más recomendables se pueden citar en la siguiente tabla:

Nombre	**Formulación**
Real	$P_s^* = P_s(t,\omega)$
Positiva	$P_s(t,\omega) \geq 0$
Invariante en t	$P_{\tilde{s}}(t,\omega) = P_x(t-t_0,\omega)\ si\ \tilde{s}(t) = s(t-t_0)$
Invariante en f	$P_{\tilde{s}}(t,\omega) = Px(t,\omega-\omega_0)\ si\ \tilde{s}(t) = s(t)e^{j2\pi\omega_0 t}$
Marginal en t	$\int_{-\infty}^{\infty} P_x(t,\omega)dt = \lvert s(t)\rvert^2$
Marginal en f	$\int_{-\infty}^{\infty} P_x(t,\omega)df = \lvert S(\omega)\rvert^2$
Soporte finito en t	$P_s(t,\omega) = 0\ \notin [t_1,t_2]\ \ si\ x(t) = 0 \notin [t_1,t_2]$
Soporte finito en f	$P_s(t,\omega) = 0 \notin [\omega_1,\omega_2]\ \ si\ S(\omega) = 0 \notin [\omega_1,\omega_2]$

Tabla 9: Propiedades recomendables para una representación tiempo-frecuencia

Apéndice C

Características extraídas

C.1 Características estadísticas

Cuando se aplican nociones estadísticas a la señal de voz, es necesario determinar la función de densidad de probabilidad (fdp), la cual se puede estimar mediante un histograma de las amplitudes sobre un número suficientemente representativo de muestras de señal. De modo que se pueda facilitar el análisis se prefiere extraer una serie de características representativas de la (fdp), estas características reciben el nombre de momentos estadísticos.
También llamados características estadísticas los momentos estadísticos son parámetros o funciones que permiten una caracterización sencilla aunque incompleta de procesos estocásticos. Cada momento estadístico es una medida numérica con una interpretación física clara que proporciona una información relevante sobre una variable aleatoria. Principalmente los momentos los podemos clasificar con respecto al origen y respecto a la media.

C.1.1 Momentos respecto al origen

Definamos U_k como el momento estadístico respecto al origen de orden k y el operador esperanza matemática $E[.]$.

$$U_k = E\left[g(x)^k\right]$$

Donde k es un entero positivo, $g(x)$ es una función medible de la variable aleatoria x y $f_x(x)$ es la función de densidad de probabilidad (fdp)

$$U_k = E\left[g(x)^k\right] = \int_{-\infty}^{\infty} E\left[g(x)^k\right] f_x(x) d(x)$$

C.1.1.1 Media o esperanza

Es el momento de primer orden ($k = 1$), se debe interpretar como el valor central alrededor del cual se dan el conjunto de realizaciones. En otras palabras es el punto que esta más cerca de todos los posibles valores de la variable aleatoria

$$U_1 = E\left[g(x)^k\right] = \int_{-\infty}^{\infty} E\left[g(x)\right] f_x(x) d(x)$$

Los momentos respecto al origen de orden superior son de escaso interés en la mayoría de los casos por lo tanto no serán enunciados.

C.1.2 Momentos respecto a la media

Definamos W_k como el momento estadístico respecto a la media de orden k

$$W_k = E\left[(g(x) - U_1)^k\right]$$

C.1.2.1 Media o esperanza

Para el primer momento ($k = 1$) se obtiene:

$$W_1 = E\left[(g(x) - U_1)\right] = 0$$

Es decir para cualquier variable aleatoria el momento de primer orden respecto a la media es cero.

C.1.2.2 Varianza

Es el momento de segundo orden ($k = 2$), se debe interpretar como una medida de la concentración de los valores de la variable aleatoria en torno a su media.

$$W_2 = E\left[(g(x) - U_1)^2\right]$$

Desarrollando el cuadrado se obtiene

$$W_2 = E\left[(g(x)^2\right] - \left(E\left[(g(x)\right]\right)^2$$

Como se comentó anteriormente, la interpretación de la varianza es la de un promedio que mide la distancia de los valores de la variable a la media de ésta. Si la varianza es pequeña, indica una alta concentración de los valores de la variable en torno a la media, si la varianza es grande, indica alta dispersión de los valores de la variable respecto de la media. El principal problema de la varianza es que se expresa en unidades cuadráticas (lo cual no siempre tiene una interpretación clara), para obviar este problema se define otra medida de la concentración llamada **desviación estándar** la cual se calcula como la raíz cuadrada positiva de la varianza. Es evidente que la desviación típica tendrá las mismas unidades de la variable analizada.

C.1.2.3 Asimetría

Es el momento de tercer orden ($k = 3$) se debe entender como una medida de que tan asimétrica es la distribución de la variable aleatoria, es decir si existen observaciones extremas en algún sentido (altas o bajas) con frecuencias razonablemente altas.

$$W_3 = E\left[(g(x) - U_1)^3\right]$$

Si la asimetría es negativa, la variable toma valores muy bajos con mayor frecuencia que valores muy altos y se dice que tiene una cola a la izquierda. Si por el contrario la asimetría es positiva, la variable toma valores muy altos con mayor frecuencia que valores muy bajos y se dice que tiene una cola a la derecha. Si la asimetría es igual a cero los valores altos y bajos tienen igual probabilidad, es decir la fdp es simétrica respecto al origen, el ejemplo mas simple de este caso es la distribución normal.

C.1.2.4 Curtosis

Es el momento de cuarto orden ($k = 4$) se debe entender como una medida que determina la forma de la distribución de la variable aleatoria.

$$W_4 = E\left[(g(x) - U_1)^4\right]$$

Al representar gráficamente variables con curtosis pequeñas (platicúrticas) se observan distribuciones de aspecto aplanado (mesetas) con colas cortas, por el contrario al graficar variables con curtosis grandes (leptocúrtica) las distribuciones son de aspecto alto con colas largas y pesadas. La curtosis siempre es positiva y se mide en unidades de la variable a la cuarta potencia.

C.2 Listado de características

Las características extraidas se dividieron en grupos de acuerdo a la técnica utilizada en su extracción.

C.2.1 Por medio de los coeficientes de la *wavelet* discreta

La coeficientes de la *wavelet* discreta se extrajeron para tramas de 40 ms sin traslape lo que generó una serie de contornos para cada uno de los parametros estadístiicos a los cuales se extrajo el valor medio.

mwd media
mewd mediana
mawd máximo

miwd mínimo
sdwd desviación estándar
vawd varianza
aswd asimetría
kuwd curtosis

C.2.2 Por medio de la distribución de *Wigner Ville*

mwv media
mewv mediana
mawv máximo
miwv mínimo
sdwv desviación estándar
vawv varianza
aswv asimetría
kuwv curtosis
vamwv perturbación del valor medio
vamaxwv shimmer

C.2.3 Por medio de la transformada Gabor

mg media
meg mediana
mag máximo
mig mínimo
sdg desviación estándar
vag varianza
asg asimetría
kug curtosis
vamg perturbacion del valor medio
vamaxg shimmer

C.2.4 Por medio del Análisis de Prediccion Lineal

Se calcularon las características de representación a partir de los coeficientes de predicción lineal para tramas de la señal de 30 ms y un traslape del 50 % dentro de este conjunto de características se incluyen todos los paramétros calculados por medio de la función de autocorrelación.

C.2.4.1 De los coeficientes de predicción lineal

mA media
meA mediana
maA máximo
miA mínimo
sdA desviación estándar
vaA varianza
asA asimetría
kuA curtosis

C.2.4.2 Del pitch

m media
me mediana
ma máximo
mi mínimo
sd desviación estándar
va varianza
as asimetría
ku curtosis
ji *jitter*
vFo coeficiente de perturbación de la frecuencia fundamental

C.2.4.3 De la derivada del pitch

mdP media
medP mediana
madP máximo
midP mínimo
sddP desviación estándar
vadP varianza
asdP asimetría
kudP curtosis

C.2.4.4 De la sonoridad

m media
mes mediana
mas máximo
mis mínimo
sds desviación estándar
vas varianza
ass asimetría

ku curtosis

C.2.4.5 Del HNR

mh media
meh mediana
mah máximo
mih mínimo
sdh desviación estándar
vah varianza
ash asimetría
kuh curtosis

C.2.4.6 De la energía

mE media
meE mediana
maE máximo
miE mínimo
sdE desviación estándar
vaE varianza
asE asimetría
kuwd curtosis

C.2.5 Directamente de la señal *raw data* (datos crudos)

mc media
mec mediana
mac máximo
mic mínimo
sdc desviación estándar
vac varianza
aswd asimetría
kuc curtosis

C.2.5.1 De la perturbación de amplitud

mshc media
meshc mediana
mashc máximo
mishc mínimo
sdshc desviación estándar
vashc varianza
asshc asimetría

kushc curtosis
vamax perturbación de la amplitud maxima

C.2.5.2 Del coeficiente de la perturbación de amplitud

mvamc media
mevamc mediana
mavamc máximo
mivamc mínimo
sdvamc desviación estándar
vavamc varianza
asvamc asimetría
kuvamc curtosis

REFERENCIAS

[1] J.Bernal, *Reconocimiento de voz y fonética acústica*. Mexico: Alfa Omega, 2000.

[2] C. Hogset, *Técnica vocal*. Editorial Euskal Herrico Abesbatzen Elkartea, 1995.

[3] E. Moulines and J. Laroche, *Speech Communications*, 1995.

[4] X. Huang, *Spoken language processing*. Prentice-Hall Inc, 2001.

[5] J. D. Echeverri, "Caracterización acústica de bioseñales empleando transformadas tiempo frecuencia y modelado paramétrico," Master's thesis, Universidad Tecnológica de Pereira - UTP, 2006.

[6] Q. Antonio, *Fonética General de la Lengua Española*, Madrid España, 1991.

[7] E. Roldan, "Calidad y dinámica de la voz en grupos sociales en la ciudad de valdivia (chile)," *Estudios Fisiólogicos*, vol. 33, pp. 111–118, 1998.

[8] E. S. y F Nuñez y P Cortes y C Suárez, "Indice de incapacidad vocal: factores predictivos," *Acta Otorrinolaringol*, vol. 57, pp. 101–108, 2006.

[9] C. Darwin, *La expresión de las emociones en los animales y en el hombre*. Madrid: Alianza, 1984.

[10] K. R. Scherer, "Personality markers in speech," *Cambridge University Press*, pp. 147–209, 1979.

[11] E. Kramer.

[12] E. C. Beier and A. J. Zautra.

[13] D. C. Albas, K. W. Mccluskey, and C. A. Albas.

[14] K. W. Mccluskey, D. C. Albas, and R. R. Niemi.

[15] R. V. Bezooijen, *Characteristics an recognizability of vocal expressions of emotion*. Dordrecht: Foris Publications, 1984.

[16] Scherer and P. Ekman.

[17] C. Izard.

[18] R. Cowie and R. R. Cornelius.

[19] A. Ortony, G. L. Clore, and A. Collins, *The cognitive structure of emotion*. Londres: Cambridge University Press, 1988.

[20] K. R. Scherer.

[21] K. Alter, E. Rank, S. A. Kotz, U. Toepel, M. Besson, A. Schirmer, and A. D. Friederici.

[22] C. A. Smith.

[23] W. Wundt, *Grundriss der Pshychologie*. Leipzig: Verlag von Wihelm Engelmann, 1896.

[24] M. Schroder.

[25] M. Schroder, R. C. D. E. Cowie, M. Westerdijk, and S. Gielen.

[26] "Inteligencia emocional," Tech. Rep., http://www.inteligencia-emocional.org.

[27] R. Plutchik.

[28] ——, *Las emociones*. México: Editorial Diana, 1987.

[29] B. Porat, *Digital Processing of Random Signals: Theory and Methods*. Prentice-Hall, 1994.

[30] P. Z. Peebles, *Principios de Probabilidad, Variables Aleatorias y Señales Aleatorias*. McGraw Hill,, 2006.

[31] S. P. Roque, "Teoría de las telecomunicaciones," *Departamento De Ciencia y Tecnología*.

[32] J. R. A. Montoya, "Detección de fallas en sistyemas electricos mediante el análisis de transitorios," *Universidad De Mendoza*, 2005.

[33] L. Cohen, *Time-Frequency Analysis*. Prentice Hall Signal Processing Series, 1995.

[34] F. Hlawatsch and G. Boudreaux-Bartels, "Linear and quadratic time-frequency signal representations," in *Signal Processing Magazine*, Abril 1992, pp. 21–67.

[35] L. J. Douglas and T. W. Parks, "A resolution comparison of several time-frequency representations," in *Transactions on Signal Processing*, vol. 40, no. 2, 1992, pp. 413–420.

[36] H. I. Choiy and W. J. Williams, "Improved time-frequency representaion of multicomponent signals using exponential kernels," in *Speech and Signal Processsing*, vol. 35, 1989, pp. 217–250,276–300 y 372–389.

[37] E. Serrano, *Introducción a la transformada wavelet y sus aplicaciones al procesamiento de señales de emisión acústica*. Universidad Nacional de General San Martín: Escuela de Ciencia y Tecnología, 2000.

[38] F. Vargas, "Selección de características en el análisis acústico de voces," Master's thesis, Universidad Nacional de Manizales, 2003.

[39] S. Mallat, *Wavelet Tour of Signal Processing*. Boston: American Press, 1998.

[40] J. J. Benedetto and M. W. Frazier, *Wavelet, Matemathics and Applications*. Boca Raton: CRC Press, 1994.

[41] C. K. Chui, *An Introduction to Wavelet*. Boston: Academic Press, 1992.

[42] G. Strang and T. Nguyen, *Wavelet and Filter Banks*. Boston: Wellesly Cambridge Press, 1996.

[43] T. A. C. M. Claasen and W. F. G. Mecklenbrauker, "The wigner distribution-a tool for time-frequency signal analysis," in *part I: Continuous-time signals*, vol. 35, 1980, pp. 217–250.

[44] ——, "The wigner distribution-a tool for time-frequency signal analysis," in *part II: Discrete-time signals*, vol. 35, 1980, pp. 276–300.

[45] ——, "The wigner distribution-a tool for time-frequency signal analysis," in *part III: Relations with other time-frequency signals transformations*, vol. 35, 1980, pp. 372–389.

[46] P. Flandrin, *Temps-frequence*. Paris: Hermes, 1993.

[47] E. P. Wigner, *Quantum-mechanical distribution functions revisited*. New York: Perspectives in Quantum Theory, 1971.

[48] J. A. Vila, "Análisis de variabilidad de señales fisiológicas , integración de un sistema de monitorización inteligente."

[49] J. L. N. Mesa and P. Q. Morales, *Codificación, Síntesis y Reconocimiento de Voz*. España: Universidad de Las Palmas de Gran Canaria, 1994.

[50] S. Kadambe and G. F. Boudreaux-Bartels, "Application of the wavelet transform for pitch detection of speech signals," *IEEE TRANSACTIONS ON INFORMATION THEORY,*, vol. 38, no. 2, March 1992.

[51] E. S. J. U, H. K. C, F. Watkins, and P. López, "Implementación del un modelo digital de análisis y síntesis de voz empleando técnicas lpc y de autocorrelación," *Universidad de Santiago de Chile.*, 2002.

[52] J. B. A. Hernández, "Sistema de detección automático de disfonías," Master's thesis, Universidad de las Palmas de Gran Canaria (U.L.P.G.C.), 2001.