



MINISTERIO DE CULTURA Y EDUCACION

UNIVERSIDAD NACIONAL DE SAN JUAN

FACULTAD DE INGENIERIA

"Reconocimiento automático de emociones basado en análisis de audio, para diagnóstico complementario en el Trastorno del Espectro Emotivo"

Autor:
BUSTAMANTE, Paola

Asesores:
LÓPEZ, Natalia - PEREZ, Elisa

Julio del 2015 - Bioingeniería



DEPARTAMENTO DE ELECTRONICA Y AUTOMATICA

Av. San Martín (Oeste) 1109
C.P. 5400 – San Juan – ARGENTINA
Tel: 264 4211700 ext: 354
sec_dea@unsj.edu.ar

Agradecimientos

Agradezco enormemente a mis padres, hermanos, hermanas y sobrinos que me dan la confianza suficiente para afrontar estos desafíos y facilitan con su cariño el progreso en mi vida.

Agradezco a la Facultad de Inginería de la Universidad Nacional de San Juan que me ha brindado no solo conocimientos y experiencias necesarias para formarme como profesional sino también la posibilidad de conocer gente interesante, docentes excelentes y principalmente amigos de gran calidad. Agradezco a mis compañeros de labarotario Emanuel, Alejandro y Fernando quienes con paciencia y alegría posibilitaron este camino recorrido y lo tornaron emocionalmente placentero, lleno de amistad y colaboración.

Agradezco a la Universidad EAFIT que permitió llevar a cabo una parte importante del desarrollo de este trabajo, principalmente a Lucía cuyos consejos fueron de gran utilidad para darle un enfoque desde otra perspectiva.

Por último, agradezco a Natalia y Elisa, mis asesoras de tesis, quienes confiaron en mis capacidades, me impulsaron en la ejecución de nuevas ideas y fueron no solo asesoras sino también compañeras en este proyecto de tesis.

Contenidos

CAPITULO 1. INTRODUCCIÓN	11
1.1 Motivación	11
1.2 Objetivos.....	15
1.2.1 Objetivos Generales	15
1.2.2 Objetivos Específicos.....	16
1.3 Estructura del trabajo final	16
1.4 Referencias.....	17
CAPITULO 2: FUNDAMENTOS TEÓRICOS: EMOCIONES	20
2.1 Teoría de las Emociones	20
2.2 Emociones y Neurociencia	22
2.3 Anatomía de las Emociones	28
2.3.1 Sistema Límbico.....	29
2.3.2 Corteza cerebral	35
2.3.3 Formación Reticular	37
2.3.4 Vías Emotivas	39
2.4 Neurofisiología de las emociones básicas	40
2.4.1 Miedo	41
2.4.2 Enojo.....	43
2.4.3 Asco	45
2.4.4 Sorpresa	47
2.4.5 Tristeza	48
2.4.6 Felicidad	50
2.5 Procesos patológicos que involucran las emociones básicas: Trastornos del espectro Emotivo.	51
2.5.1 Depresión	52
2.5.2 Trastorno bipolar.....	55
2.5.3 Trastorno del Espectro Autista.....	55
2.6 Regionalización de las emociones en el espacio bidimensional Arousal-Valencia	56
2.6.1 Modelo de Russell	57
2.7 Referencias.....	62
CAPITULO 3: FUNDAMENTOS TEÓRICOS: ANATOMÍA Y FISIOLOGÍA DE LA VOZ	67
3.1 Introducción.....	67
3.2 Modelo del Aparato Fonador.....	67
3.2.1 Órganos de la fonación	68
3.2.2 Mecanismo Articulatorio	76
3.2.3 Tracto vocal y cavidad nasal	77
3.3 Características fisiológicas de la voz.....	80
3.3.1 Formantes	80
3.3.2 Frecuencia fundamental	84
3.3.3 Sonidos sonoros y sordos	85
3.4 Referencias	87

CAPITULO 4: ANÁLISIS DE LA SEÑAL DE AUDIO Y DESCRIPCIÓN DE CARACTERÍSTICAS	88
4.1 Introducción	88
4.2 Análisis de la señal	88
4.3 Modelado de la señal	88
4.4 Cualidades del sonido	91
4.4.1 Frecuencia Fundamental y Armónicos	91
4.4.2 Intensidad sonora.....	92
4.5 Características la señal de voz.....	93
4.5.1 Enfoque estadístico	93
4.5.2 Enfoque Temporal.....	93
4.5.3 Enfoque prosódico	96
4.5.4 Enfoque frecuencial	101
4.6 Referencias.....	110
CAPITULO 5: DESARROLLO DE UN SEGMENTADOR DE VOZ.....	112
5.1 Introducción	112
5.2 Técnicas estadísticas. Lema de Neyman Pearson	113
5.2.1 Lema de Neyman Pearson.....	113
5.2.2 Aplicación del lema de Neyman Pearson	116
5.3 Desarrollo del segmentador de voz.....	118
5.3.1 Esquema del segmentador.....	118
5.4 Resultados.....	122
5.4.1 Error de detección	123
5.4.2 Comparación con otro detector.....	124
5.5 Conclusión.....	125
5.6 Referencias.....	125
CAPITULO 6: PROCESAMIENTO Y ANÁLISIS DE BASE DE DATOS	127
6.1 Introducción	127
6.2 Materiales	127
6.2.1 Base de datos Berlín	127
6.2.2 Base de datos Enterface.....	128
6.2.3 Base de datos Semaine	128
6.2.4 Base de datos GVEESS	129
6.3 Métodos.....	129
6.3.1 Acondicionamiento de las Bases de Datos.....	129
6.3.2 Pre-procesamiento.....	130
6.3.3 Procesamiento de la señal. Extracción de características.....	134
6.3.4 Técnicas de Selección de características	136
6.4 Resultados.....	137
6.4.1 Experimentación Felicidad-tristeza	137
6.5 Conclusiones	142
6.6 Referencias.....	142
CAPITULO 7: DETERMINACIÓN DE LOS PARÁMETROS INVARIABLES AL IDIOMA	143
7.1 Introducción	143

7.2	Materiales	144
7.3	Procesamiento de la señal	144
7.3.1	Características temporales y prosódicas.....	144
7.3.2	Características frecuenciales	144
7.4	Métodos de análisis.....	144
7.5	Resultados.....	145
7.5.1	Enfoque temporal y prósodico.....	145
7.5.2	Características frecuenciales	147
7.6	Conclusiones	149
7.7	Referencias.....	150
CAPITULO 8: RESULTADOS.....		151
8.1	Introducción	151
8.2	SECCIÓN 1.	152
8.2.1	Reconocimiento de dos emociones.....	152
8.2.2	Reconocimiento de cuatro emociones:	156
8.3	SECCIÓN 2	161
8.3.1	Implementación de técnicas de clasificación en Base de datos Semaine (inglés)	162
8.3.2	Implementación de técnicas de clasificación en Base de datos GVEESS (alemán).....	162
8.3.3	Comparación entre Sistema de Reconocimiento Automático y no Automático.	163
8.4	SECCIÓN 3:	164
8.5	SECCIÓN 4:	165
8.5.1	Regionalización de emociones entrenadas	166
8.5.2	Regionalización de emociones no entrenadas.....	169
8.6	SECCIÓN 5.	173
8.6.1	Entrenamiento en alemán.....	174
8.6.2	Entrenamiento en castellano	175
8.6.3	Entrenamiento mixto	176
8.7	Conclusiones	176
8.8	Referencias.....	178
CAPITULO 9: CONCLUSIONES.....		179
9.1	Introducción	179
9.2	Conclusiones	180
9.3	Líneas futuras	182

Índice de figuras

- Figura 2.1: "Horror" de Le Brun.
- Figura 2.2: "La expresión de la emoción en el hombre y en los animales", Duchenne de Boulogne.
- Figura 2.3: Expresiones faciales.
- Figura 2.4: Sistema Límbico
- Figura 2.5: Sección transversal a través del hipocampo
- Figura 2.6: Interconexión de la formación parahipocampal
- Figura 2.7: Visión medial del hemisferio cerebral. Relación entre giro o circunvolución del cíngulo y el giro o circunvolución parahipocampal
- Figura 2.8: Sección coronal del encéfalo que muestra el Cuerpo Estriado
- Figura 2.9: Sección horizontal del encéfalo. Claustro
- Figura 2.10: Visión lateral derecha. Cuerpo amigdalino
- Figura 2.11: Diencéfalo
- Figura 2.12: Visión Lateral del Hemisferio Cerebral
- Figura 2.13: Visión Lateral del Hemisferio cerebral
- Figura 2.14: Formación Reticular
- Figura 2.15: Formación reticular. Núcleos del rafe y Locus cerúleo
- Figura 2.16: Miedo
- Figura 2.17: Enojo
- Figura 2.18: Asco
- Figura 2.19: Sorpresa
- Figura 2.20: Tristeza
- Figura 2.21: Felicidad.
- Figura 2.22: Concepto de 8 emociones en ordenamiento circular
- Figura 2.23: Escalamiento de ordenamiento circular para 28 palabras de emoción
- Figura 2.24: Escalamiento multidimensional para las 28 palabras de emoción
- Figura 2.25: Escalamiento unidimensional para las 28 palabras de emoción
- Figura 3.1: Modelo del aparato Fonador
- Figura 3.2: Sistema Respiratorio
- Figura 3.3: Patrón Respiratorio durante el habla
- Figura 3.4: Órgano principal para la producción de la voz
- Figura 3.5: Laringe
- Figura 3.6: Cartílago Cricoides
- Figura 3.7: Cartílago Tiroides
- Figura 3.8: Cartílagos Aritenoides
- Figura 3.9: Pliegues Vocales
- Figura 3.10: Patrón de Vibración de las Cuerdas vocales. Ciclo de la glotis
- Figura 3.11: Variación del área glotal y flujo de aire en relación a la salida de los sonidos durante un ciclo glotal
- Figura 3.12: Variación del área glotal y flujo de aire en relación a la salida de los sonidos durante un ciclo glotal
- Figura 3.13: Órganos Articulación del Habla.
- Figura 3.14: Tracto Vocal.
- Figura 3.15: Hipofaringe
- Figura 3.16: Orofaringe
- Figura 3.17: Cavidad Oral- Vestíbulo Oral
- Figura 3.18: Cavidad Nasal
- Figura 3.19: Configuración de las Cavidades Resonantes para la vocal /i/

- Figura 3.20: Espectro de Frecuencias para la vocal /i/
Figura 3.21: Configuración de las Cavidades Resonantes para la vocal /u/
Figura 3.22: Espectro de Frecuencias para la vocal /u/
Figura 3.23: Sonidos Sonoros y Sordos
Figura 3.24: Elocución vocal-consonante-vocal.
Figura 4.1: Modelo de producción del Habla
Figura 4.2: Modelo Simplificado de producción del Habla
Figura 4.3: Curtosis. Comparación entre distribuciones de muestras no Gaussianas y la distribución de Gauss
Figura 4.4: Estimación de la frecuencia fundamental en señal periódica
Figura 4.5: Estimación de la frecuencia fundamental en señal periódica
Figura 4.6: Detección de F0 mediante Técnica del Cepstrum
Figura 4.7: Logaritmo del espectro de la señal con su correspondiente envolvente y análisis cepstral
Figura 4.8: Comportamiento espectral del filtro LPC
Figura 4.9: Relación entre la escala frecuencial en Hertz y la escala en frecuencias Mel
Figura 4.10: Banco de filtros en escala Mel
Figura 4.11: Espectro de las bandas frecuencias del banco de filtros
Figura 4.12: Coeficientes cepstrales en frecuencia Mel
Figura 4.13: Densidad Espectral de Potencia mediante Periodograma de Welch
Figura 5.1: Esquema Básico del segmentador
Figura 5.2: Espectrograma
Figura 5.3: Esquema acondicionamiento de la señal
Figura 5.4: Esquema del cálculo de la envolvente
Figura 5.5: Esquema del cálculo del umbral dinámico
Figura 5.6: Aplicación del segmentador
Figura 5.7: Error absoluto en función de la SNR
Figura 5.8: Detección de la señal a diferentes SNR
Figura 5.9: Aplicación del detector Voicebox a la señal Berlín
Figura 5.10: Error en función de la SNR
Figura 6.1: Acondicionamiento de la señal
Figura 6.2: Señal normalizada a la izquierda y con eliminación continua a la derecha
Figura 6.3: Pre-procesamiento
Figura 6.4: Respuesta frecuencial
Figura 6.5: Ubicación de polos y ceros en el plano Z
Figura 6.6: Señal sin filtrar a la derecha y filtrada a la izquierda
Figura 6.7: Detección de segmentos de actividad de voz
Figura 6.8: Procesamiento de las señales. Parametrización
Figura 6.9: Prueba 1
Figura 6.10: Prueba 2
Figura 6.11: Prueba 3
Figura 6.12: Prueba 4
Figura 7.1: Distancia de Frecuencias fundamentales en G1C Y G2C
Figura 7.2: Formantes de los Grupos 1 y 2
Figura 7.3: Grupo 1
Figura 7.4: Grupo 2
Figura 8.1: Detección de dos emociones mediante LDA
Figura 8.2: Esquema de la red neuronal utilizada
Figura 8.3: Detección de dos emociones mediante RN
Figura 8.4: Esquema de la red neuronal .Prueba 1
Figura 8.5: Detección de cuatro emociones. Prueba 1

- Figura 8.6: Esquema de la red neuronal .Prueba 2
Figura 8.7: Detección de cuatro emociones. Prueba 2
Figura 8.8: Modelo Circumplejo de Russell, J.A. (1980)
Figura 8.9: Diagrama emocional con las emociones de referencia
Figura 8.10: Plano emociones Berlín: enojo (negro), felicidad (magenta)
Figura 8.11: Plano emociones GVEESS: enojo (izquierda), felicidad (derecha)
Figura 8.12: Plano emociones Berlín: miedo (verde), tristeza (azul)
Figura 8.13: Plano emociones GVEESS: miedo (izquierda), tristeza (derecha)
Figura 8.14: Plano de Russell. Aburrimiento
Figura 8.15: Plano emociones Berlín. Aburrimiento
Figura 8.16: Plano emociones GVEESS. Aburrimiento
Figura 8.17: Plano de Russell. Asco
Figura 8.18: Plano emociones Berlín. Asco
Figura 8.19: Plano emociones GVEESS. Asco

Índices de tablas

- Tabla 2.1: Posiciones angulares en el plano Arousal-Valencia, obtenidos de los tres ensayos de Russell.
Tabla 6.1: Matriz de confusión de Prueba 1.
Tabla 6.2: Matriz de confusión de Prueba 2.
Tabla 6.3: Matriz de confusión de Prueba 3.
Tabla 6.4: Matriz de confusión de Prueba 4.
Tabla 6.5: Porcentaje de emoción detectada
Tabla 7.1: Análisis de parámetros temporales y prosódicos.
Tabla 7.2: Análisis de Correlación de Coeficientes cepstrales en Frecuencia Mel
Tabla 8.1: Matriz de confusión para dos emociones mediante LDA
Tabla 8.2: Resultados de entrenamiento mediante RN
Tabla 8.3: Matriz de confusión para dos emociones mediante RN
Tabla 8.4: Características utilizadas en la clasificación de la prueba 1
Tabla 8.5: Resultados del entrenamiento para 4 emociones. Prueba 1
Tabla 8.6: Matriz de confusión para 4 emociones. Prueba 1
Tabla 8.7: Características utilizadas en la clasificación de la prueba 2
Tabla 8.8: Resultados del entrenamiento para 4 emociones. Prueba 2
Tabla 8.9: Matriz de confusión para 4 emociones. Prueba 2
Tabla 8.10: Comparación entre el desempeño humano y el reconocimiento automático
Tabla 8.11: Resultados del entrenamiento para 3 emociones. Semaine
Tabla 8.12: Matriz de confusión para 3 emociones
Tabla 8.13: Resultados del entrenamiento para 4 emociones. GVEESS.
Tabla 8.14: Matriz de confusión para 4 emociones. GVEESS
Tabla 8.15: Comparación entre el desempeño humano y el reconocimiento automático
Tabla 8.16: Porcentajes de aciertos obtenidos de la implementación del clasificador Berlín en las Bases de datos Semaine y GVEES
Tabla 8.17: Posiciones angulares de las emociones de referencia

Tabla 8.18: Porcentaje de detección aplicado a los corpus Berlín y GVEESS completos

Tabla 8.19: Porcentaje de detección de emociones. Aburrimiento

Tabla 8.20: Porcentaje de detección de emociones. Asco

Tabla 8.21: Etiquetas correspondientes a las emociones entrenadas

Tabla 8.22: Resultados del entrenamiento para detección de Tristeza (entrenamiento en alemán)

Tabla 8.23: Detección de Tristeza en Alemán y Castellano con un clasificador Alemán

Tabla 8.24: Resultados del entrenamiento para detección de Tristeza (entrenamiento en castellano).

Tabla 8.25: Detección de Tristeza en Castellano y Alemán, con un clasificador español.

Tabla 8.26: Resultados del entrenamiento para detección de Tristeza (entrenamiento mixto).

Tabla 8.27: Detección de Tristeza en Alemán y Castellano con un clasificador bilingüe

CAPITULO 1

Introducción

1.1 Motivación

La interacción humana con el entorno es básicamente emocional, las características distinguibles de la cognición humana respecto a otros seres vivientes parecen siempre estar definida en el plano emocional. Los sentimientos son indispensables para la toma de decisiones porque nos orientan en la dirección adecuada para sacar el mejor provecho a las posibilidades que nos ofrece la fría lógica.

Todas las emociones son, en esencia, impulsos que nos llevan a actuar, programas de reacción automática con los que nos ha dotado la evolución. La raíz etimológica de la palabra emoción proviene del latín *movere* (que significa «moverse») más el prefijo «e-», significando «movimiento hacia» y sugiriendo, de ese modo, que en toda emoción hay implícita una tendencia a la acción. Cada emoción nos predispone de un modo diferente a la acción; cada una de ellas nos señala una dirección que, en el pasado permitió resolver adecuadamente los desafíos a que se ha visto sometida la existencia humana. Nuestro bagaje emocional tiene un extraordinario valor de supervivencia y esta importancia se ve confirmada por el hecho de que las emociones han terminado integrándose en el sistema nervioso en forma de tendencias innatas y automáticas [1].

Aristóteles considera las pasiones o emociones como afecciones psicofísicas, asociadas con alteraciones fisiológicas, que conllevan sensaciones de dolor y /o placer [2]. Las emociones pueden considerarse como la reacción inmediata del ser humano a una situación que le es favorable o desfavorable.

En efecto, se puede deducir que las emociones son impulsos que están organizados de manera tal que permiten nuestra conservación y de la especie. Estas fuerzas internas nacen de los sentimientos, pensamientos, aprendizaje, estados psicológicos y desencadenan en procesos biológicos que acondicionan un tipo de tendencia a la acción.

Las emociones rigen casi todos los modos de comunicación humana, las expresiones faciales, los gestos, las posturas, el tono de voz, la elección de las palabras, la respiración, la temperatura corporal, etc. Cada emoción evidencia una impronta biológica o sea predispone al cuerpo a un tipo diferente de respuesta.

Por ejemplo el enojo aumenta el flujo sanguíneo en las extremidades, aumenta el ritmo cardíaco y la tasa de hormonas, tales como la adrenalina. Fisiológicamente se genera la cantidad de energía necesaria para acometer acciones vigorosas [3]. En el caso del miedo la sangre se retira del rostro, lo que explica la palidez y la sensación de frío, y fluye a la musculatura esquelética (miembros inferiores) favoreciendo así la huida. Las conexiones nerviosas de los centros emocionales del cerebro desencadenan también una respuesta hormonal que pone al cuerpo en estado de alerta general sumiéndolo en la inquietud y predisponiéndolo para la acción. Uno de los principales cambios biológicos producidos por la felicidad consiste en el aumento en la actividad de un centro cerebral que se encarga de inhibir los sentimientos negativos y de aquietar los estados que generan preocupación, al mismo tiempo que aumenta el caudal de energía disponible. La tristeza provoca la disminución de la energía y del entusiasmo por las actividades vitales, especialmente las diversiones

y los placeres y cuanto más se profundiza y se acerca a la depresión, más se enlentece el metabolismo corporal [3].

Las emociones tienen mediadores neuroanatómicos y neurofisiológicos. Estas son estructuras interconectadas que pueden alterarse y generar lo que se conoce como trastornos del espectro emocional, que alteran la expresión emotiva, lo cual se manifiesta mediante cambios fisiológicos, que pueden ser detectados y medidos, desde la variación de la presión arterial, dosajes hormonales, la temperatura de la piel, como así también los cambios en la prosodia¹.

Los trastornos del espectro emotivo engloban a patologías tales como la ansiedad, el estrés, la depresión, el trastorno del espectro autista, y trastorno bipolar entre otros [4].

Las emociones tienen un efecto directo sobre el sistema nervioso autónomo pudiendo ocasionar o empeorar estados patológicos. Pruebas concluyentes en el laboratorio de la Facultad de Medicina y Odontología de la Universidad de Rochester demostraron la existencia de un vínculo fisiológico directo entre las emociones y el sistema inmunológico [5]. Descubrieron la existencia de conexiones directas entre las terminaciones nerviosas del sistema autónomo y las células del sistema inmunológico. Este punto físico de contacto permite a las células nerviosas liberar los neurotransmisores que regulan la actividad de las células inmunológicas.

Otro factor fundamental en la relación existente entre las emociones y el sistema inmunológico está ligado a las hormonas liberadas en situaciones de, por ejemplo, estrés [5]. Las catecolaminas: epinefrina y norepinefrina (llamadas también adrenalina y noradrenalina), el cortisol, la prolactina y los opiáceos naturales (como la endorfina y la encefalina) son algunas de las hormonas liberadas en situaciones de tensión que tienen una gran influencia sobre las células del sistema inmunológico. El estrés por consiguiente, disminuye la resistencia inmunológica, al menos de forma provisional. Pero en el caso de que el estrés sea intenso y prolongado, la inhibición puede terminar convirtiéndose en una condición permanente.

Investigaciones llevadas a cabo en miles de personas confirman hasta qué punto resultan nocivas para la salud las emociones perturbadoras y demuestran que las personas que sufren de ansiedad crónica, largos episodios de melancolía y pesimismo, tensión excesiva, irritación constante, escepticismo y desconfianza extrema son propensas a contraer enfermedades [5]. Las emociones negativas son un factor de riesgo para el desarrollo de la enfermedad. En resumen, las emociones negativas constituyen una seria amenaza para la salud.

En 1993, la revista Archives of Internal Medicine publicó una extensa investigación realizada por el psicólogo Yale Bruce McEwen [6], en la que refería las consecuencias de la relación existente entre el estrés y la enfermedad, una relación que compromete a la función inmunológica hasta el punto de acelerar el cáncer, aumentar la vulnerabilidad ante las infecciones víricas, acelerar la formación de trombos que pueden causar el infarto de miocardio. El estrés también puede contribuir a la ulceración del tracto gastrointestinal y a empeorar los síntomas de la inflamación intestinal. A largo plazo el cerebro también es susceptible a los efectos del estrés sostenido, produciendo lesiones en el hipocampo y afectando en consecuencia a la memoria. Ciertos experimentos han demostrado que el estrés y la ansiedad debilitan la fortaleza del sistema inmunológico [6].

Se ha demostrado que la ansiedad no solo provoca una disminución de la respuesta inmunológica sino que también tiene efectos negativos sobre el sistema cardiovascular. Mientras la irritabilidad crónica y los episodios repetidos de cólera parecen aumentar el riesgo de enfermedad coronaria en

¹Prosodia: Representa aquellos elementos de la expresión oral, tales como el acento, los tonos y la entonación

los hombres, las emociones más letales para las mujeres son la ansiedad y el miedo. Un estudio llevado a cabo en la Facultad de Medicina de la Universidad de Stanford sobre más de mil personas que habían padecido un ataque al corazón demostró que las mujeres que habían sufrido un segundo ataque presentaban un elevado índice de miedo y ansiedad [7].

El trastorno del espectro autista es un trastorno del desarrollo caracterizado por importantes alteraciones en la interacción social y en la comunicación, que se acompañan de comportamientos e intereses restrictivos. Una de las características más sobresalientes de las personas con espectro autista es su dificultad para el contagio emocional, para mostrar empatía, y para reconocer y comprender las emociones de los demás, independientemente de la capacidad general del individuo [8].

El poder comprender las expresiones emocionales es esencial para las interacciones sociales, lo que permite explicar y anticipar las acciones de los otros. A diferencia de otros procesos mentales, las emociones frecuentemente se hacen perceptibles a través de las expresiones faciales y audibles.

En [9] se dio a conocer un estudio en el que se propuso evaluar si era posible enseñar a comprender estados mentales de emoción, creencia y ficción a personas con autismo. Los resultados mostraron que sí era posible enseñarles a que evalúen la comprensión de emociones. Sin embargo, encontraron que los efectos de la enseñanza no se generalizaban a otras tareas de dominios no enseñados específicamente.

Parece ser difícil dar instrucciones explícitas sobre la comprensión de expresiones emocionales en situaciones sociales que son ambiguas, por lo que es un desafío poder facilitar experiencias sociales e interpersonales que hagan que se destaque las expresiones emocionales, promoviendo su reconocimiento y comprensión.

Los trastornos mentales como la depresión exhiben su descripción desde el punto de vista de la neurociencia como la afección que se asocia con disminuciones en las regiones neo corticales dorsal (compartimientos sensorial-cognitivo) y aumento relativo en áreas límbicas y para-límbicas (compartimiento autonómico).

El trastorno depresivo es un desorden emotivo que genera una serie de síntomas tales como irritabilidad (ira) o ánimo depresivo (tristeza), disminución del interés por actividades diarias, variaciones significativa en el peso, insomnio o hipersomnia, agitación o retardo psicomotor[10]. Actualmente una de cada seis personas padece depresión clínica al menos una vez en su vida y un 7% de la población sufre esta enfermedad al cabo del año. Se calcula que afecta a unos 350 millones de personas en todo el mundo, constituyéndose en una de las principales causas de discapacidad [11].

La depresión desempeña un papel relevante en algunas condiciones clínicas. Se ha demostrado que tiene incidencia sobre el empeoramiento de la enfermedad, como el caso de las patologías cardíacas [12]. En una investigación realizada en un hospital de Montreal [13], los pacientes deprimidos que fueron dados de alta después de haber padecido un primer ataque al corazón presentaron un índice de mortalidad muy elevado durante los seis meses siguientes. La tasa de mortalidad de uno de cada ocho pacientes de los más seriamente deprimidos de ese estudio era cinco veces superior a la de otros pacientes aquejados de una enfermedad similar. Un factor de riesgo tan importante como las principales causas de muerte por ataque cardíaco.

De esta manera el reconocimiento de los diferentes estados emotivos tiene importantes aplicaciones en diversas patologías, además de otros usos. Por lo tanto es de gran importancia el estudio, análisis y clasificación de las distintas emociones que se expresan en los seres humanos.

Diversas metodologías de reconocimiento de emociones se han propuesto basándose en el hecho que las emociones se ven reflejadas en diferentes sistemas biológicos que incluyen las expresiones faciales, los músculos, la voz y la actividad del sistema nervioso y endócrino [14].

Existen estudios que han tratado de definir con exactitud las distintas emociones. Los primeros estudios eran subjetivos y se basaban en las expresiones faciales. A un grupo de personas se les presentaban imágenes, luego se les pedía que describieran con gestos qué habían sentido [15]. De esta manera se establecía una correlación en donde se identificaba la gesticulación con el sentimiento. En cuanto a la metodología objetiva, se toma en cuenta el movimiento de los músculos y se lo relaciona con la emoción sentida. Así, un conjunto de características faciales corresponde al enojo, otra a la ira, etc. Eckman y Friesen en [16], perfeccionaron el sistema de codificación de las acciones faciales (FACS).

En [17] se pretende elaborar las llamadas “representaciones continuas”, consistentes en la ubicación de la emoción en coordenadas. Dicha representación se denomina espacio de activación – valencia. Con esto se pueden ubicar las emociones básicas.

Otro método de análisis y detección de emociones que resulta no invasivo, es a partir de la voz. Se pueden determinar características propias de la misma que permita establecer una clasificación de los distintos estados emocionales.

El reconocimiento automático de emociones a partir de la voz es un área de investigación relativamente nueva, sin embargo se tienen trabajos desde el año 1996 [18]. Existen trabajos [19] donde se reporta un estudio experimental en que se trabaja con 4 emociones: enojado, feliz, triste y neutral, utilizando una base de datos de 741 instancias cortas. Posteriormente, en [20] se reporta un trabajo que consiste en asociar los parámetros prosódicos derivados del pitch, duración y energía al eje de la activación y los rasgos de calidad como el timbre de la voz con el eje de valencia, con el objetivo de mejorar la tasa de reconocimiento. En [21], se realiza una clasificación de los estados emotivos dentro de un contexto multilenguaje. El experimento se llevó a cabo usando bases de datos en inglés, eslovenio, castellano y francés. En [22] se reporta una clasificación de emociones usando la base de datos de voz emocional danesa, donde se extrajeron 87 características y se usaron con un criterio de selección secuencial hacia adelante. Al siguiente año, [23] presentaron un artículo donde se realiza minería de datos sobre 1000 rasgos extraídos del pitch, energía y Coeficientes cepstrales en frecuencia Mel (mfcc) usando las bases de datos Berlín y Mago de Oz. En el mismo año [24] hicieron un reporte de detección de emociones usando una base de datos en chino, alcanzando una precisión de 88,7% usando Análisis discriminante Lineal, K-vecinos y Modelos Ocultos de Markov. En el trabajo reportado [25] se hace uso de una base de datos en Euskara, se construyeron dos clasificadores diferentes: uno utilizando características espectrales y Mixturas Gaussianas, otro con parámetros prosódicos y Máquina de Soporte Vectorial. Se extrajeron 86 características prosódicas y posteriormente se aplicó un algoritmo para seleccionar los parámetros más relevantes. El mejor resultado se obtuvo con el primero de los clasificadores, que alcanzó una precisión del 98.4%. En [26] reportan una extendida revisión de las bases de datos orientadas al reconocimiento automático de emociones, los resultados más altos en la clasificación de emociones alcanzan el 80%.

En nuestro país, se realizaron algunos trabajos de investigación relacionados con el procesamiento de voz. En [27] se presenta la aplicación de un método para el análisis del riesgo vocal debido a las alteraciones en la voz. Se obtienen 3 índices: a) un índice de perturbación que agrupa 4 parámetros clásicos como el Jitter, el Shimmer, la relación armónico-ruido y la amplitud del

cepstrum; b) un índice de precisión vocal vinculado con la estabilidad articulatoria y medido como la inversa de la desviación estándar de los primeros 5 formantes, y c) un índice asociado al grado de aprovechamiento de energía que evalúa tanto la coincidencia entre los armónicos con los formantes como las pérdidas de energía que se producen en el tracto vocal. La agrupación de las voces de docentes en normales, con riesgo vocal y alteradas, se presenta en relación a los respectivos diagnósticos laringológicos verificando su utilidad en la evaluación de los profesionales con riesgo vocal.

El ejército de los Estados Unidos de América, después de los atentados ocurridos el 11 de setiembre de 2001 desarrollaron un analizador de estrés en la voz, tecnología VSA (Voice Stress Analysis) [28]. El estrés es una respuesta corporal no específica y mediante la aplicación de tecnologías de voz, puede deducirse la veracidad de lo expresado por una persona. Su principal función es detectar el micro-temblor de la voz involuntario que se registra en los músculos y que va de 8 a 12Hz. Permite determinar irregularidades en la voz del sujeto de estudio. Su grado de confiabilidad es del 96.12%. Principalmente está enfocado a las áreas de: Recursos Humanos, para determinar la confiabilidad de sus empleados y Seguridad: se aplica para casos de robo, fraudes o secuestros.

Uno de los artículos publicados recientemente [29], consiste en una clasificación de las emociones basada en audio-video que imita las condiciones del mundo real, o sea se plantea una adquisición de datos sin las condiciones controladas que se desarrollan en un laboratorio.

En [30] considera la aplicación de Máquinas de Boltzmann restringidas (RBM) y Redes de Creencias Profundas (DBN) para el reconocimiento automático de emociones en español. Estas técnicas son métodos alternativos a los tradicionales y comparables a otros clasificadores.

La problemática que actualmente existe en este tipo de actividades es que se evalúa principalmente desde el punto de vista matemático-computacional con poco interés en el área biológica y en sus aplicaciones en el área de la medicina. Con esta tesina se intentará encontrar, a partir de técnicas de procesamiento de señales un modelo que se ajuste al reconocimiento de voz emotiva, que pueda clasificar automáticamente emociones a partir de señales de voz y que posibilite aplicaciones en el área de salud.

Este tipo de aplicaciones pueden colaborar en el diagnóstico de ciertos desórdenes psicológicos, y en el seguimiento de su evolución, como así también en mejorar la interacción paciente-entorno a través de un mecanismo de traducción personalizada en el caso del trastorno del espectro autista.

El estado emocional se estimará a través de las señales de voz del usuario, mediante un sistema de reconocimiento de patrones basado en aprendizaje supervisado, con el fin de detectar los cambios de valencia emocional que indiquen alteraciones del modelo.

1.2 Objetivos

1.2.1 Objetivos Generales

Desarrollar técnicas automáticas para el reconocimiento de emociones mediante el análisis de señales de audio. El desarrollo de éstas técnicas permitirá la extracción y selección de características que permitan diferenciar estados emotivos y clasificarlos.

1.2.2 Objetivos Específicos

- Estudiar las señales de audio y su aplicación en reconocimiento de estados emocionales mediante un enfoque multidisciplinario.
- Fortalecer y ampliar los conocimientos en el procesamiento de las señales de audio a fin de comprender su esencia y potenciar su desarrollo.
- Desarrollar e implementar algoritmos de selección de características de las señales de voz, obtenidas en una base de datos de domino público.
- Proponer clasificadores para diferenciar los distintos estados de ánimo según los parámetros más representativos y técnicas utilizadas en el procesamiento de voz.
- Realizar la aplicación de los sistemas de clasificación y evaluar su desempeño.
- Divulgación de los resultados obtenidos.

1.3 Estructura del trabajo final

El presente trabajo está organizado en 9 capítulos, los cuales se describen brevemente a continuación.

1. **Capítulo 1. Introducción:** en el presente capítulo se exponen las problemáticas que motivaron la realización de éste trabajo, como así también la importancia de su implementación en el área de salud. Además, se detallan el objetivo final y los objetivos específicos que se pretenden alcanzar. Finalmente, se coloca una descripción general de la estructura del presente trabajo final.
2. **Capítulo 2. Fundamentos teóricos: Emociones:** en el capítulo 2 se da una introducción a los conceptos básicos sobre las emociones. Se explican los mecanismos fisiológicos que desencadenan los distintos estados emocionales. Se describe el concepto de espacio bidimensional para la clasificación de estados emocionales, (Activación-Valencia). Además, se exhiben los procesos patológicos que surgen de las alteraciones de los procesos emocionales.
3. **Capítulo 3. Fundamentos teóricos: Anatomía y Fisiología de la voz:** en el capítulo 3 se explican los mecanismos que generan los sonidos del habla. Se describe el modelo del aparato fonador y la fisiología del órgano fonador. Se detallan los parámetros que representan las señales de audio desde el punto de vista biológico y su implicancia en el discurso.
4. **Capítulo 4. Análisis de la señal de audio y descripción de características:** en el capítulo 4 se realiza el análisis de la señal de audio desde el punto de vista físico-matemático. Se describen las características más importantes de la señal a nivel temporal, frecuencial y prosódico.
5. **Capítulo 5. Desarrollo de un segmentador vocálico:** en el capítulo 5 se describe el desarrollo de un detector de segmentos vocálicos de audio que elimina el silencio con el objetivo de disminuir la tasa de error. Se exponen los fundamentos estadísticos del desarrollo: Teoría de decisión. Se exhiben los resultados de la aplicación y su validación.
6. **Capítulo 6. Procesamiento y análisis de base de datos:** en el capítulo 6 se explican las distintas etapas de procesamiento de la señal de audio. Se desarrollan e implementan

algoritmos de selección de características de las señales de audio, obtenidas en una base de datos de dominio público y bajo la plataforma MATLAB®2013.

7. **Capítulo 7: Determinación de los parámetros invariables al idioma:** en el capítulo 7 se realiza un análisis de características temporales, prosódicas y frecuenciales en varios idiomas y se determinan los parámetros invariables.
8. **Capítulo 8.Resultados:** en el capítulo 8 se presentan los resultados obtenidos en este trabajo. Se exponen las características principales extraídas del procesamiento de los datos. Se presentan los resultados del sistema de clasificación implementado y su desempeño. Se realiza un análisis y comparación de los resultados obtenidos con otras bases de datos. Regionalización de emociones en el plano bidimensional Activación-Valencia. Se exhibe un sistema de clasificación de Tristeza en dos idiomas.
9. **Capítulo 9. Conclusiones:** Finalmente, en este capítulo se presentan las conclusiones que se obtuvieron durante el desarrollo de este trabajo final, como así también los trabajos futuros que se pueden realizar para extender el trabajo realizado.

1.4 Referencias

1. Goleman, D. (2012). *Inteligencia emocional*. Editorial Kairós.
2. Trueba Atienza C. "La teoría Aristotélica de las emociones". *Signos Filosóficos*, vol. XI, núm. 22, julio-diciembre, 2009
3. Levenson, R., Ekman, P., Friesen, W. (1990). Voluntary facial action generates emotion-specific autonomic nervous system activity. *Psychophysiology*, 27(4), 363-384.
4. BanúsLlort S. (2012). *Trastornos emocionales*. Psicodiagnosis.es. Especialistas en Psicología Infantil y Juvenil. Consultada el 21 de mayo de 2014
5. <http://www.psicodiagnosis.es/areaclinica/trastornosemocionales>
6. Friedman H., Klein T., Friedman A. (Eds.) (1995). *Psychoneuroimmunology, stress and infection*, (pp. 1-21).
7. McEwen BS1, Stellar E. (1993). Stress and the individual. Mechanisms leading to disease. *Archives Internal Medicine*. 153, 2093-2101.
8. Low K.G., Thoresen C.E., Pattillo J.R., King A.C., Jenkins C. (1994). Anxiety, depression, and heart disease in women. *International journal of behavioral medicine*, 1 (4), 305-319.
9. Baron-Cohen S., Spitz A., Cross P. (1993). Do children with autism recognize surprise? A research note. *Cognition & Emotion*, 7(6), 507-516.
10. Hadwin J., Baron-Cohen S., Howlin P., Hill K. (1997). Does teaching theory of mind have an effect on the ability to develop conversation in children with autism? *Journal of autism and Developmental Disorders*, 27(5), 519-537.
11. Soriano Pacheco J. (2009) *Marcadores Relacionales en la Depresión Mayor y la Distimia*. Tesis Doctoral Programa de Doctorado en Psiquiatría y Psicología Clínica. Departament de Psiquiatria y de Medicina Legal, UAB. Barcelona.
12. La depresión (2012). OMS (Organización Mundial de la Salud). Nota descriptiva N°369 <http://www.who.int/mediacentre/factsheets/fs369/es/>

13. Frasure-Smith N., Lespérance F., Talajic M. (1993). Depression following myocardial infarction: impact on 6-month survival. *Jama*, 270(15), 1819-1825.
14. Frasure-Smith, N., Lespérance, F., Juneau, M., Talajic, M., & Bourassa, M. G. (1999). Gender, depression, and one-year prognosis after myocardial infarction. *Psychosomatic medicine*, 61(1), 26-37.
15. Jerritta S., Murugappan M., NagarajanR. Wan, K.(2011). Physiological signals based human emotion recognition: a Review. In *CSPA, 2011 IEEE 7th International Colloquium on*, pages 410 –415.
16. DantzerR. (1989). *Las emociones*, Paidós, Barcelona
17. Ekman P., Rosenberg E. L. (Eds.). (1997). *What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS)*. Oxford University Press.
18. CowieR., Cornelius R. R. (2003). Describing the emotional states that are expressed in speech. *Speech communication*, 40(1), 5-32.
19. Dellaert F., Polzin T., Waibel A. (1996). Recognizing emotion in speech. In *Spoken Language, 1996. ICSLP 96. Proceedings, Fourth International Conference on* (Vol. 3, pp. 1970-1973). IEEE.
20. Yu F., Chang E., Xu Y., Shum H. Y. (2001). Emotion detection from speech to enrich multimedia content. In *Proceedings of the second IEEE pacific rim conference on multimedia: Advances in multimedia information processing* (pp. 550-557). Springer-Verlag.
21. Tato R., Santos R., Kompe R., Pardo J. M. (2002). Emotional space improves emotion recognition. In *INTERSPEECH*.
22. Hozjan V., Kačič Z. (2003). Context-independent multilingual emotion recognition from speech signals. *International Journal of Speech Technology*, 6(3), 311-320.
23. Ververidis D., KotropoulosC., Pitas, I. (2004). Automatic emotional speech classification. In *Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP'04). IEEE International Conference on* (Vol. 1, pp. I-593). IEEE.
24. Vogt T., André E. (2005, July). Comparing feature sets for acted and spontaneous speech in view of automatic emotion recognition. In *Multimedia and Expo, 2005. ICME 2005. IEEE International Conference on* (pp. 474-477). IEEE.
25. Pao T. L., Chen Y. T., Yeh J. H., Chang Y. H. (2005). Emotion Recognition and Evaluation of Mandarin Speech Using Weighted D-KNN Classification. In *ROCLING*.
26. Luengo I., Navas E., Hernández I., Sánchez J. (2005). Reconocimiento automático de emociones utilizando parámetros prosódicos. *Procesamiento del lenguaje natural*, 35.
27. AyadiM., Kamel M. S., Karray, F. (2011). Survey on speech emotion recognition: Features, classification schemes, and databases. *PatternRecognition*, 44(3), 572-587.
28. Gurlekian J. A., Molina N. (2012). Índice de perturbación, de precisión vocal y de grado de aprovechamiento de energía para la evaluación del riesgo vocal. *Revista de Logopedia, Foniatria y Audiología*, 32(4), 156-163.
29. Hopkins C. S., Ratley R. J., Benincasa D. S., Grieco,J. J. (2005). Evaluation of voice stress analysis technology. In *System Sciences, 2005. HICSS'05. Proceedings of the 38th Annual Hawaii International Conference on* (pp. 20b-20b). IEEE.

30. Dhall A., Goecke R., Joshi J., Wagner, M., Gedeon T. (2013). *Emotion recognition in the wild challenge 2013*. In *Proceedings of the 15th ACM on International conference on multimodal interaction* (pp. 509-516). ACM.
31. Sánchez-Gutiérrez M. E., Albornoz E. M., Martinez-Licona F., Rufiner H. L., Goddard J. (2014). *Deep learning for emotional speech recognition*. In *Pattern Recognition* (pp. 311-320). Springer International Publishing.

CAPITULO 2

Fundamentos teóricos: Emociones

*Nous devons peut-être aux passions les plus grands avantages de l' esprit.
[«Quizá debemos a las pasiones los logros más grandes del espíritu»]
(Réflexions et maximes, Vauvenargues)*

2.1 Teoría de las Emociones

En la antigüedad, los filósofos creían que las pasiones eran emociones desbocadas que enturbiaban la capacidad de pensar con claridad. Aristóteles considera a las pasiones como afecciones psicofísicas asociadas a alteraciones fisiológicas que conllevan sensaciones de dolor y/o placer [1].

Durante mucho tiempo fue predominante referirse a las emociones y las pasiones como idénticas.

Descartes define las pasiones como sentimientos o emociones del alma que se refieren particularmente a ella y que son motivadas y amplificadas por algún movimiento de los espíritus [2].

Sin embargo Kant repara en la diferencia entre emociones y pasiones [3]. Describe las pasiones como inclinaciones difíciles o invencibles por la razón, por el contrario la emoción es el sentimiento de un placer o desplacer en el estado presente que no permite que se abra paso en el sujeto la reflexión. La emoción según Kant se halla referida al presente, en tanto que la pasión se extiende al futuro, lo que apunta al carácter pasajero de la primera y persistente en el caso de la segunda. La emoción es un ataque por sorpresa de la sensación, lo que la torna precipitada y de una intensidad tal que imposibilita la reflexión. Las emociones son representadas como un estado efímero que involucra una intensidad o sea podemos decir que son impulsos, fuerzas internas que desencadenan en procesos biológicos que acondicionan un tipo de acción.

Como se dijo en el **capítulo 1**, el significado etimológico de la palabra emoción nos indica su propósito. La raíz de “emoción” proviene del latín “moveré” (que significa moverse) más el prefijo “e” que significa “movimiento hacia” y sugiriendo de ese modo que en toda emoción hay implícita una tendencia a la acción [4]. Esta definición permite comprender que toda emoción está ligada a una acción que la representa. Sin embargo, su motivación, esas fuerzas internas que la inducen, nacen de los sentimientos, pensamientos, estados psicológicos y principalmente con el aprendizaje de conductas que han llevado a la supervivencia de la especie. Nuestro bagaje emocional es el resultado de los desafíos a los que se ha visto sometida la existencia humana como parte de un proceso de adaptación y supervivencia.

En el siglo XIX, Darwin elaboró la teoría de la evolución natural, donde explica el modo en que evolucionaron las características físicas de las especies, y donde además sugiere la teoría que la mente y la conducta también vienen determinadas por la evolución natural [5].

Las obras de Charles Le Brun, artista francés, fueron clave en la investigación de Darwin. Sus pinturas explicaban historias y sus personajes expresaban emociones. Le Brun se propuso establecer una clasificación sistemática perfecta de las pasiones y las emociones, definiéndolas morfológicamente [6]. Por ejemplo, el horror es representado como el entrecejo fruncido, la pupila

descendente, la boca entreabierta, pero más cerrada en la línea media que hacia los lados, que deben ser retirados hacia atrás, así, se formarán pliegues en la mejilla, el color de la cara es pálido y, los labios y los ojos, un poco lívidos (figura 2.1). De esta manera, bosqueja todas las pasiones, las positivas, como la admiración y la estimación, hasta las negativas como el horror, el odio, la desesperación, el miedo y la tristeza.

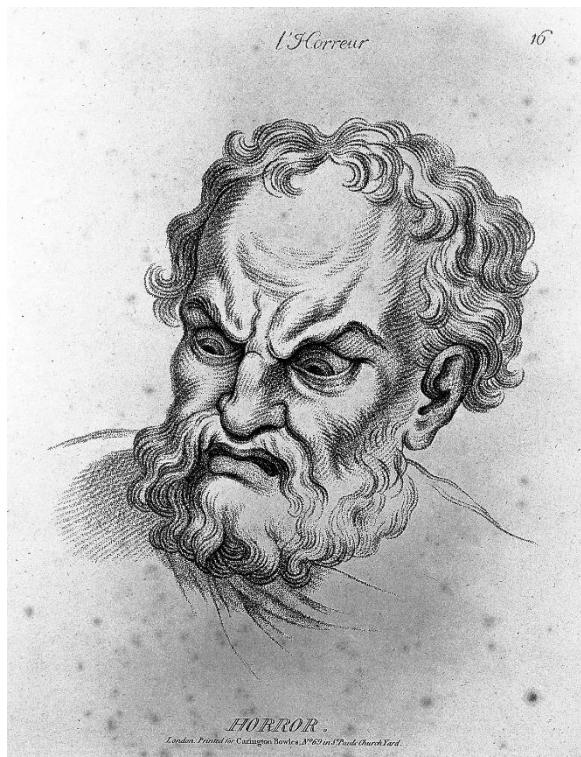


Figura 2.1. “Horror” de Le Brun. Wellcome Images, London

Darwin reconoce que las Conferencias del pintor Le Brun, publicadas en 1667 constituyen el trabajo antiguo mejor conocido y que contiene algunas buenas observaciones. Sin embargo, no tarda mucho en descartar la utilidad de las explicaciones de las raíces fisiológicas que da Le Brun [7]. En 1872, Darwin publica el libro titulado “La expresión de la emoción en el Hombre y los Animales” [8]. Esta obra utiliza la fuente de Le Brun para las descripciones morfológicas, por ejemplo, cuando describe la cólera. Sin embargo, para el caso concreto de las explicaciones fisiológicas, Darwin opta por fuentes más modernas, como los trabajos anatómicos de Charles Bell y de Guillaume Duchenne, en particular los centrados en el rostro humano (figura 2.2).

Darwin ignoró las características y se centró en los cambios visibles pero temporales de la apariencia. Afirmó que los cambios del cuerpo aparecen inmediatamente después de la percepción del acto emotivo, y lo que sentimos al mismo tiempo que suceden los hechos de la emoción. Observó, además, a las expresiones corporales como medio de comunicación y adaptación al entorno físico y a la supervivencia. Sostuvo que existe una coordinación instintiva entre el acto de percepción y las respuestas al organismo [5].

Darwin trató las emociones como entidades discretas separables. Así, definió y clasificó ocho emociones básicas: alegría, angustia, interés, sorpresa, miedo, enojo, disgusto y vergüenza. Su teoría

establece que estas ocho emociones se pueden identificar en humanos y en animales. También propuso que el fenómeno emocional y su expresión están estrechamente relacionados al señalar que la expresión facial y el cuerpo son los medios primarios de expresión emocional [9].



Figura 2.2. "La expresión de la emoción en el hombre y en los animales", Duchenne de Boulogne.

Los biólogos evolucionistas afirman que las reacciones emocionales ante un eventual peligro se han incorporado en nuestro sistema nervioso porque sirvió para garantizarnos la vida durante un periodo largo y decisivo de la prehistoria humana y, más importante aún, porque cumplió con la principal tarea de la evolución, perpetuando las mismas predisposiciones genéticas en la progenie.

2.2 Emociones y Neurociencia

Hasta ahora, las emociones han sido enfocadas desde la perspectiva de la filosofía, psicología y del arte. Para la ciencia, las emociones no eran más que un tabú del alma. Hasta que un visionario de la época, Charles Darwin introdujo un nuevo concepto de las emociones, definiendo su carácter evolutivo y su implicancia con la supervivencia. Este fue el primer abordaje de las emociones desde un enfoque científico.

En 1884 Williams James [10] trató de identificar qué procesos intervienen entre la ocurrencia de un estímulo que despierta la emoción y las emociones conscientes, es decir el sentimiento que evoca. Para esto ofreció la teoría de la existencia del *feedback* mental. Debido a que emociones diferentes tienen respuestas diferentes, el *feedback* hacia el cerebro, será diferente y eso justifica el modo de sentir. Propuso que las emociones no ocurrían a nivel cognitivo, para ir seguidas luego de la respuesta vegetativa, como sugiere la intuición, sino que el proceso ocurría al revés: la experiencia

cognitiva de la emoción sería secundaria a su expresión fisiológica. Cuando el cerebro detecta, por ejemplo, una situación de peligro pone en marcha las reacciones de huida o lucha (actividad motora) a nivel inconsciente y son estas reacciones las que dan lugar a la sensación consciente del miedo en otras zonas del cerebro. De acuerdo con este enfoque, tenemos miedo porque huimos o sufrimos porque lloramos y no a la inversa.

Cannon y Bard (1928) entablaron una nueva teoría de la emoción neurológica [11] [12]. Sus experimentos consistieron en realizar lesiones controladas en animales, que eliminaban los hemisferios cerebrales y parte de los núcleos profundos del cerebro. Ellos encontraron pautas de extirpación, notaron que cuando la lesión preservaba una zona llamada hipotálamo se producía en el animal un cuadro denominado "falsa rabia". Este se caracterizaba porque de manera espontánea o como resultado de un estímulo cutáneo inocuo, el animal desarrollaba el cuadro típico de un estado de cólera: erizamiento del pelo, arqueo del lomo, exhibición de dientes, extrusión de las uñas, midriasis, taquicardia, etc. El nombre de falsa rabia se debió a que pese a la presencia de gestos, el animal no dirigía su agresión a ningún objeto externo en particular. Cuando la lesión afectaba también el hipotálamo, la respuesta de falsa rabia no aparecía. Todo ello sugería que el hipotálamo preservado en el primer caso, era imprescindible para la expresión coordinada de conductas emocionales y que tal expresión era estereotipada e independiente de los elementos cognitivos conscientes de la emoción que serían producidos por estructuras cerebrales más altas, incluyendo la corteza. Esta teoría establecía que unas zonas concretas del cerebro, particularmente el hipotálamo, eran las responsables de las respuestas emocionales integradas, proporcionando a la corteza la información requerida para poner en marcha los mecanismos cerebrales de conciencia de la emoción y que en estas estructuras se organizaban los circuitos neuronales básicos que integran las conductas típicas de las emociones.

James Papez en 1937 [13] propuso que era el hipotálamo el que mandaba y recibía información del cerebro límbico y que el hipocampo actuaba como coordinador entre el hipotálamo y las cortezas cingular y parahipocámbica.

Los experimentos realizados a finales de 1950 por John Downer en la University College de Londres demostraron la importancia de la amígdala, uno de los elementos claves del sistema límbico. La experimentación consistió en extirpar la amígdala de un lado del *maccacus Rhesus*, desconectando al mismo tiempo los dos hemisferios cerebrales, de modo que la información visual se podía hacer llegar separadamente a uno y otro con la premisa que existía una carencia de amígdala en uno de los lados. Cuando el animal veía al mundo a través del hemisferio sin amígdala, por tener el ojo que proyectaba al otro hemisferio tapado, se comportaba de manera plácida. Sin embargo, si veía con el ojo del lado conectado al hemisferio con amígdala, el animal actuaba agresivamente. Estos experimentos pusieron en evidencia que la amígdala, una estructura que contiene núcleos basolaterales que conectan con la corteza cerebral, unos núcleos centrales y anteriores conectados con el hipotálamo y el tronco encefálico y núcleos mediales conectados con el bulbo y corteza olfatorios, sirve de conexión entre corteza y el hipotálamo y es un gran centro de convergencia de información sensorial, cortical y visceral, y cuya actividad varía durante la conducta emocional [14].

En 1957, el psicólogo norteamericano Paul Ekman retomó las ideas de Darwin y demostró que las emociones básicas existen en diferentes culturas, inclusive en aquellas que no han recibido influencias culturales del Occidente [15]. Así estas emociones básicas se asocian con expresiones

faciales distintivas (figura 2.3) y son innatas y comunes en las diferentes culturas del mundo. A partir de estas reflexiones, Ekman postuló que cada emoción básica debería estar asociada a un circuito cerebral en particular.

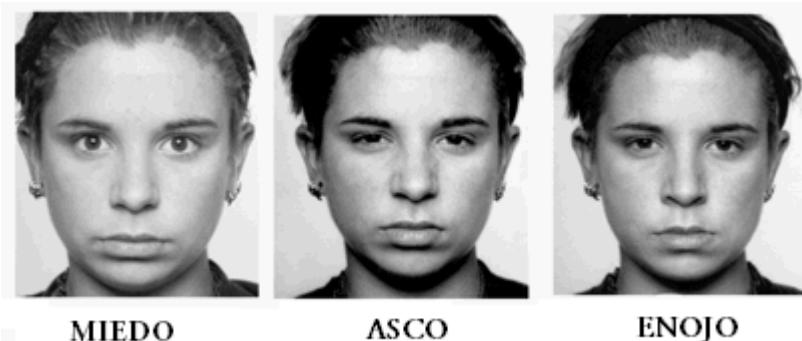


Figura 2.3. Expresiones faciales. Ekman (www.paulekman.com)

En las últimas décadas el estudio de las emociones ha sido abordado, con mayor énfasis, desde la reciente área de las neurociencias. Las neurociencias estudian los fundamentos de nuestra individualidad: las emociones, la conciencia, la toma de decisiones y nuestras acciones socio-psicológicas [16].

La nueva perspectiva de las emociones, desde el área de la neurociencia, la define como procesos influenciados por nuestro pasado evolutivo y personal que se ejecutan a partir de la organización y la interconexión de los elementos del sistema nervioso y que desencadenan un conjunto de cambios fisiológicos (viscerales, del sistema autónomo, del tono de voz, de los gestos) y del comportamiento (memoria, toma de decisiones) [16].

Se ha aceptado considerar que la conciencia no es el único elemento que ocupa la mente, ya que también es el origen de las emociones. O sea, tanto en la emoción como en la cognición subyacen e interactúan toda una serie de mecanismos cerebrales no conscientes, que determinan de manera decisiva las características conscientes del pensamiento y la emoción. Entonces, la emoción no se corresponde con un proceso cerebral separado e independiente, sino el resultado de múltiples mecanismos cerebrales que pueden ser distintos en emociones diferentes.

Damasio (1994), neurólogo y profesor de la Universidad del Sur de California, considera que la esencia de la emoción es la colección de cambios en el estado corporal [4]. Por lo tanto considera que una emoción es la combinación de un proceso mental en el que las respuestas a este proceso son dirigidas principalmente al cuerpo dando como resultado el estado emocional. Sin embargo estas respuestas también son dirigidas al cerebro lo cual produce cambios mentales adicionales. De esta forma, para Damasio una emoción parece ser esencialmente la respuesta corporal de un proceso de evaluación realizada por el cerebro. La razón de que aparezca es que esas respuestas corporales tienen un valor de supervivencia. Señala que el concepto de emoción en las neurociencias afectivas es principalmente neurobiológico.

En 1996, el neurocientífico, profesor de neurociencia y psicología en la Universidad de Nueva York, Joshep LeDoux señala que las emociones son un producto de la evolución y como tal existen debido a que cumplen su función de supervivencia, describe a las emociones como una función biológica del sistema nervioso, indica que las emociones no evolucionaron como sentimientos consciente, sino como una especialización fisiológica y conductual, y las respuestas corporales son controladas por el cerebro, lo cual le permitió a los organismos ancestrales sobrevivir a ambientes

hostiles y procrearse [17]. Explica que cuando una persona está nerviosa, enfada, tiene miedo o está enamorada aumenta el ritmo cardíaco, la velocidad de la respiración y los músculos se tensan [18]. Propone que la experiencia emocional y las activaciones fisiológicas ocurren al mismo tiempo, no una detrás de otra. Es el sistema nervioso quien controla y tiene la capacidad selectiva de que los órganos internos puedan activarse de forma diferente en situaciones diferentes, creyendo también que la hormona más importante en la experiencia emocional es la adrenalina.

Goleman (1996) demuestra que el cerebro pensante o neocorteza evoluciona de centros emocionales surgidos del tronco cerebral [4]. Este hecho, rescata Goleman, es de carácter importante respecto a la relación sentimiento y pensamiento, porque de alguna manera indica que el cerebro pensante nació del emocional y que las primeras expresiones emocionales fueron las responsables de nuestra supervivencia. "El neocórtex del *Homo Sapiens*, mucho mayor que el de cualquier otra especie, ha traído consigo todo lo que es característicamente humano. Es el asiento del pensamiento y de los centros que integran y procesan los datos registrados por los sentidos. A lo largo de la evolución, el neocórtex permitió incrementar la capacidad del individuo para superar las adversidades".

LeDoux (1996), indica que la activación de los sistemas emocionales básicos es más o menos independiente de la conciencia [17]. La información acerca de un estímulo que produce miedo viaja a través de las vías sensoriales y se bifurca en circuitos paralelos cortico-subcorticales en los niveles talámicos y mesencefálicos. En la ruta subcortical que mediaría la respuesta no consciente, la información procedente del tálamo alcanzaría el núcleo lateral de la amígdala, después el basolateral y de aquí pasaría al núcleo central [19]. Mediante las conexiones que la amígdala mantiene con el hipotálamo se produciría la respuesta emocional, sin que hasta el momento la información hubiera alcanzado la corteza y hubiera mediación consciente. En cuanto al aspecto consciente de las emociones, las estructuras involucradas incluyen la porción anterior del cíngulo, la corteza prefrontal orbital y ventromedial, el lóbulo temporal y la ínsula.

LeDoux, ha realizado descubrimientos sobre los circuitos nerviosos del cerebro emocional, asignando a la amígdala un papel central en las emociones [4]. La investigación llevada a cabo por LeDoux explica la forma en que la amígdala asume el control [20], y es en este órgano donde se elaboran las respuestas endocrinas, conductuales y motoras que caracterizan la respuesta emocional integrada. La amígdala, además juega un importante papel en el aprendizaje de las conductas emocionales. Una conducta emocional de gran trascendencia es el llamado condicionamiento de contexto, que se refiere al aprendizaje de las conductas que empujan al animal a ponerse frecuentemente en contacto de aquellos estímulos que son importantes para el mantenimiento de la especie, aprendiendo a aumentar los contactos con el entorno que le proporcionan una recompensa. Entonces, los estímulos que identifican a un entorno en el que se obtienen recompensas, se asocian a esta. Esta asociación tiene lugar en los núcleos basolaterales de la amígdala [20]. Demostró que la primer estación cerebral por la que pasan las señales sensoriales procedentes de los ojos o de los oídos es el tálamo y a partir de ahí y a través de una sola sinapsis, a la amígdala [4]. El científico, revela la existencia de vías nerviosas para los sentimientos que eluden el neocórtex, descubrió la existencia de una pequeña estructura neuronal que comunica directamente el tálamo con la amígdala. Esto explicaría el gran poder de las emociones para desbordar a la razón, ya que los sentimientos que siguen este camino directo a la amígdala son los más intensos y primitivos. Esta vía secundaria y más corta que la vía procedente del tálamo hacia el neocórtex, permite que la amígdala

reciba algunas señales directamente de los sentidos y emita una respuesta antes de que sean registrados por el neocórtex. Las emociones que utilizan esta vía alternativa son calificadas por LeDoux como emociones precognitivas.

En un experimento concluyente, LeDoux extirpó el córtex auditivo de una muestra de ratas y luego las expuso a un sonido que iba acompañado de una descarga eléctrica, o sea creaba miedo asociando un sonido con un choque eléctrico. Tras varias sesiones de acondicionamiento, las ratas no tardaron en aprender a temer el sonido, aun cuando su neocórtex no llegara a registrarlo. En este caso, el sonido seguía la ruta directa del oído al tálamo y desde allí a la amígdala. Las ratas, habían aprendido una reacción emocional sin la menor implicancia de las estructuras corticales superiores. En tal caso, la amígdala percibía, recordaba y generaba las respuestas propias del miedo de una manera independiente de toda participación cortical [4]. Según dijo LeDoux: “anatómicamente hablando, el sistema emocional puede actuar independientemente del neocórtex. Existen ciertas reacciones y recuerdos emocionales que tienen lugar sin la menor participación cognitiva consciente”. Las participaciones inconscientes son recuerdos emocionales que se almacenan en la amígdala.

De este modo, mientras la amígdala prepara una reacción ansiosa e impulsiva, otra parte del cerebro emocional se encarga de elaborar una respuesta más adecuada. El regulador cerebral que desconecta los impulsos de la amígdala son los lóbulos prefrontales que se ponen en funcionamiento cuando alguien tiene miedo o está enojado pero controlan el sentimiento para afrontar de un modo más eficaz la situación. De este modo, el área prefrontal constituye un modulador de las respuestas proporcionadas por la amígdala [4].

La corteza prefrontal, según Manes, se trata de la región de nuestro cerebro que nos hace humanos, pues regula funciones distintivas de nuestra especie: nuestra capacidad de desarrollar un plan y ejecutarlo, para tener un pensamiento abstracto, para llevar a cabo razonamientos lógicos, para tomar decisiones, para inferir los sentimientos y pensamientos de los otros. Se conoce entonces, que el hemisferio izquierdo del cerebro se especializa en el lenguaje y en el pensamiento lógico, mientras que el hemisferio derecho es experto en la percepción visual, en el proceso espacial, en el arte, la creatividad y en el procesamiento holístico de la información [16].

Los neurofisiólogos que han estudiado los estados de ánimo de pacientes con lesiones en el lóbulo prefrontal han llegado a la conclusión de que una de las funciones del lóbulo prefrontal izquierdo consiste en actuar como una especie de termostato neural que regula las emociones desagradables. Así pues, el lóbulo prefrontal derecho es la sede de los sentimientos negativos como el miedo y la agresividad, mientras que el lóbulo prefrontal izquierdo los controla, inhibiendo el lóbulo derecho. En otras palabras, el lóbulo izquierdo parece ser el mecanismo de desconexión del cerebro para las emociones perturbadoras. En una experimentación, los pacientes con lesiones en el córtex prefrontal izquierdo eran propensos a experimentar miedos y preocupaciones mientras que aquellos con lesiones en el córtex prefrontal derecho eran desproporcionadamente joviales [4].

La investigación llevada a cabo por LeDoux y otros neurocientíficos sugiere que el hipocampo, que durante mucho tiempo se había considerado como la estructura clave del sistema límbico, no tiene que ver con la emisión de respuestas emocionales, sino con el hecho de registrar y dar sentido a las pautas perceptivas. La principal actividad del hipocampo consiste en proporcionar una memoria del contexto, algo que es vital para el significado emocional. El hipocampo es el que registra los hechos puros, y la amígdala, por su parte, es la encargada de registrar el clima emocional que

acompaña a estos hechos. Si, por ejemplo, al tratar de adelantar un vehículo en una vía de dos carriles, no estimamos la distancia y tenemos una colisión frontal, el hipocampo registra los detalles concretos del accidente, qué anchura tenía la calzada, quién se hallaba con nosotros y qué aspecto tenía el otro vehículo. Pero es la amígdala la que, a partir de ese momento, desencadenará en nosotros un impulso de ansiedad cada vez que nos dispongamos a actuar en circunstancias similares [4].

Según LeDoux, las emociones se basan en respuestas físicas. La persona además es consciente de los cambios internos y externos, y se siente de forma diferente, porque tiene respuestas físicas diferentes. Por ejemplo, en el acto de huida, el cuerpo humano sufre una sacudida fisiológica, tanto externa que es visible como interna, que es invisible, donde estas respuestas van al cerebro por medio de un feedback que hace única cada sensación. El cuerpo le habla a la mente y no lo contrario, de esta manera el sentimiento es esclavo de su reacción corporal. Habitualmente se interpreta las emociones de las personas descifrando sus expresiones corporales, su tono de voz y su rostro [21].

En la actualidad, a partir de diversos estudios, se ha propuesto que existen diferentes tipos de emociones entre las cuales se destacan las emociones básicas, las cuales son consideradas innatas y están presentes en todas las culturas (ira, miedo, alegría, tristeza, sorpresa y asco) [22]. Y por otro lado, las emociones complejas que son la combinación de las emociones anteriormente mencionadas, las cuales dependen de la evaluación consciente, de la influencia directa del entorno social y que parten o surgen de la interacción con otras personas [23].

Gran parte de la actividad fisiológica implicada en las emociones es regulada por la división simpática (excitación) y parasimpática (relajación) de nuestro sistema nervioso autónomo. Las emociones con niveles de activación similares y misma valencia resultan difíciles de distinguir. Ejemplo (miedo-enojo) [24].

El médico alemán Wilhelm Wundt escribió acerca de las variaciones en las dimensiones de placer y de la actividad o intensidad. Esta conceptualización fue popularizada en 1941 por Schlosberg y luego adoptada por Russell al final del siglo pasado [25].

Fue en 1987 cuando se tuvieron en cuenta los niveles de activación-valencia evidenciando el nivel continuo de las emociones en contrapartida con el esquema discreto de emociones propuesto por Darwin [26]. Un grupo de investigación propuso un modelo dimensional de las emociones que enfatizaba que el conocimiento humano acerca de las emociones está organizado de manera jerárquica e incluye dos dimensiones continuas: valencia (un constructo bipolar que va de agradable a desgradable), activación o arousal (cuyos polos van de relajado a excitado) y que representan la activación metabólica y neuronal [27]. Encontraron que existe una estrecha relación entre la valencia y la activación en diversas respuestas fisiológicas entre las cuales podemos mencionar la actividad electromiografía, frecuencia cardiaca, conductancia de la piel y los potenciales relacionados a eventos.

En la nueva era de la neurociencia se percibe a las emociones como elementos claves en la toma de decisiones [16] [28]. Existe una aceptación general en que los procesos emocionales tienen atributos que incluyen la expresión motora, aspectos sensoriales-perceptuales, autonómicos – hormonales, cognitivos-atencionales y afectivos- sentimientos.

Damasio (2001) estudió a sujetos que eran perfectamente normales en todos los sentidos, salvo en el hecho de que tenían lesiones en las conexiones existentes entre amígdala y lóbulo prefrontal. En consecuencia pese a su apariencia normal eran incapaces de tomar decisiones o funcionar de

manera afectiva en el mundo, no podían decidir dónde vivir, ni que comer, que productos comprar y usar. Estos resultados apoyan la idea que no solo es el pensamiento racional el que ayuda a tomar decisiones [29].

Actualmente, múltiples experimentos demuestran que las emociones tienen una asociación con la memoria y surge el concepto de memoria emocional. La memoria emocional es la capacidad de adquirir, almacenar y recuperar información relacionada con la emoción [16]. El psicólogo suizo Edouard Claparéde describe un caso sobre una mujer que había perdido la capacidad de formar nuevas memorias personales. Una lesión cerebral le impedía recordar cualquier evento ocurrido después de la lesión. Todas las personas que la mujer había conocido después eran olvidadas y cada día Claparéde debía presentarse a su paciente sin que ésta tuviese ningún registro de haberlo visto con anterioridad. Un día Claparéde pensó en implementar una nueva estrategia. Escondió un alfiler en su mano derecha y, cuando saludó a su paciente, ésta recibió un pinchazo. En la siguiente sesión, la paciente seguía sin recordar quién era Claparéde pero la paciente se negaba a estrechar la mano al psicólogo. Ella no recordaba el evento sucedido, pero si la situación emocional. El conocimiento del contexto de las situaciones depende del hipocampo mientras que la memoria emotiva depende de la amígdala. La paciente tenía dañado su hipocampo pero sus amígdalas seguían activas [16].

Las memorias asociadas a una carga emocional intensa logran una mejor consolidación, puesto que dichas emociones disparan cascadas químicas y fisiológicas en nuestro organismo que favorecen la formación de nuevas memorias. Esto ha permitido el desarrollo de líneas de investigación destinadas al tratamiento de pacientes con estrés postraumático [4].

Las emociones desde el área de las neurociencias pueden ser estudiadas como funciones cerebrales biológicas del sistema nervioso, esto permite definirlas desde un enfoque científico. Las emociones tienen un componente cognitivo y uno afectivo, cognitivo porque asigna un significado y afectivo porque le asigna un valor. Los seres humanos son criaturas sociales, preparadas desde el punto de vista biológico para relacionarse e interactuar con la naturaleza. De esa interacción depende la capacidad para comprender el estado de ánimo de los otros. El lenguaje corporal es el resultado automático e indirecto de los estados afectivos y se hayan íntimamente ligados al comportamiento. La organización, funcionamiento del sistema nervioso y la interacción de los elementos que lo componen han dado origen a la conducta de los seres humanos.

Donald Norman (2004) afirma que comprender rápido a las emociones por medio del lenguaje corporal es causa de la evolución y por consecuente es parte de la herencia biológica [30].

2.3 Anatomía de las Emociones

A fin de comprender la naturaleza de las emociones desde un enfoque científico se deben conocer las regiones anatómicas que la involucran. Como ya se ha mencionado, las emociones han sido influenciadas por procesos evolutivos de supervivencia, el cerebro más primitivo (que se encuentra en todas las especies) es el tronco cerebral [4]. A partir de éste se originaron: la paleocorteza (parte de la corteza cerebral que corresponde a las áreas de terminación de las vías olfatorias) y la arquicorteza (compuesta por el hipocampo, la amígdala y otras estructuras límbicas que rodeaban la parte superior del tronco encefálico). La paleocorteza y la arquicorteza en conjunto reciben el nombre de rinencéfalo [31].

El neocórtex o la corteza nueva, es la denominación que reciben las áreas más evolucionadas del córtex. Esta neocorteza es el cerebro racional y constituye la capa neuronal que recubre los lóbulos prefrontales, y frontales de los mamíferos. La neocorteza es la encargada de la interpretación y compresión de los diferentes estímulos y situaciones que le llegan desde los distintos sistemas sensoriales [4]. La arquicorteza se complementa con la neocorteza para dar origen a la interpretación de la emoción.

Los procesos emocionales ocurren principalmente en una región del Sistema Nervioso Central llamado Sistema Límbico.

El sistema anatómico de las emociones consta de:

- Sistema límbico
- La corteza cerebral
- Hipotálamo
- Formación reticular
- Vías del sistema emocional

2.3.1 Sistema Límbico

El sistema límbico, también llamado la corteza emotiva, se encuentra entre la corteza cerebral y el hipotálamo, recibe su nombre por su posición en el borde medial del encéfalo. Consta de varias estructuras con conexiones complejas que se proyectan en el hipotálamo (figura 2.4). Sus funciones principales tienen que ver con las emociones, la memoria, y el comportamiento [32].

La entrada de información sensorial en el sistema límbico se hace directamente en la amígdala o indirectamente en la formación del hipocampo. La amígdala proporciona una connotación afectiva a la experiencia. El flujo de información del hipocampo permite un enlace con experiencias previas ya que la formación del hipocampo es esencial para el recuerdo y el aprendizaje [32].

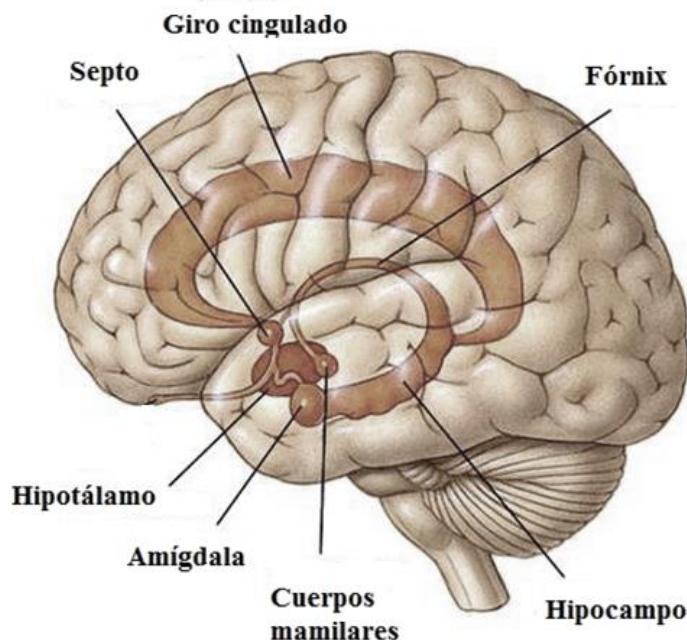


Figura 2.4. Sistema Límbico. http://jralonso.es/files/2011/05/sistema_limbico.jpg

El sistema límbico consta de las siguientes estructuras:

- Formación del hipocampo.
- Giro o circunvolución del cíngulo
- Núcleos basales. Amígdala.
- Tálamo.

2.3.1.1 Formación del hipocampo

La formación del hipocampo incluye el giro dentado, hipocampo y giro parahipocampal [32] (figura 2.5).

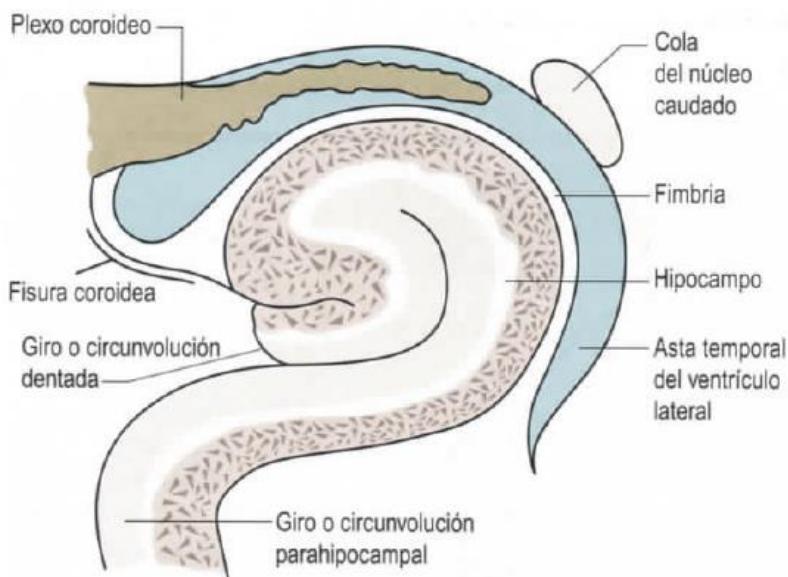


Figura 2.5. Sección transversal a través del hipocampo. (Neuroanatomía. Texto y Atlas en color, Crossman, Neary)

El hipocampo es una elevación curva de sustancia gris que en su parte anterior forma el pie del hipocampo y en la parte posterior y medial se forma el alveo que es un conjunto de fibras que dan lugar a la fimbria, que es donde se forman los pilares del fórnix (figura 2.6).

El fórnix es un fascículo de fibras que enlazan el hipocampo con el cuerpo mamilar del hipotálamo. El cuerpo mamilar se proyecta hacia el grupo nuclear anterior del tálamo, por vía del fascículo mamilotegmental.

Debajo y medialmente del hipocampo se encuentra el giro parahipocampal. El giro o circunvolución dentada se sitúa entre el giro parahipocampal y el hipocampo.

La formación parahipocampal recibe aferentes principalmente de la corteza temporal inferior y del hipocampo contralateral a través del sistema el fórnix y la comisura parahipocampal. La principal vía eferente del hipocampo es el fórnix.

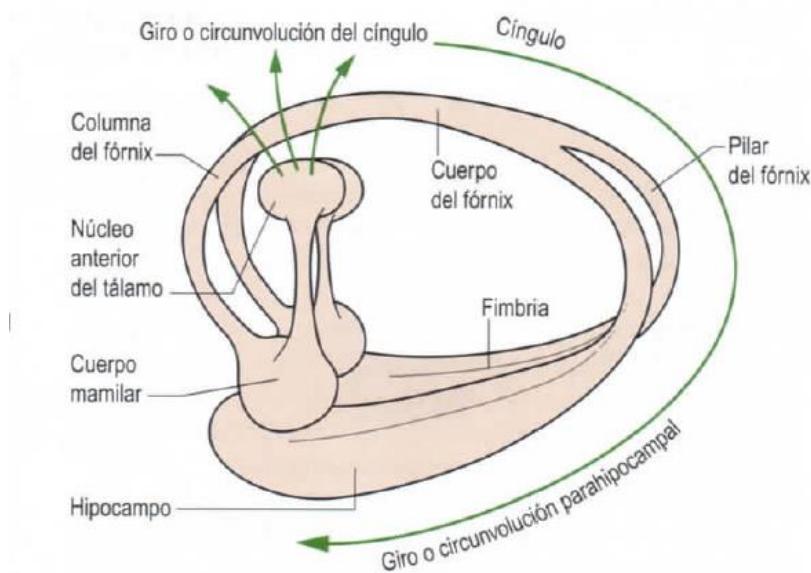


Figura 2.6. Interconexión de la formación parahipocampal. (Neuroanatomía. Texto y Atlas en color, Crossman, Neary)

2.3.1.2 Giro o circunvolución del cíngulo

El giro o circunvolución del cíngulo y el giro o circunvolución parahipocampal se continúan uno con otro alrededor del esplenio del cuerpo calloso (figura 2.7). El giro o circunvolución del cíngulo se proyecta hacia el giro o circunvolución parahipocampal por las vías de las fibras del cíngulo.

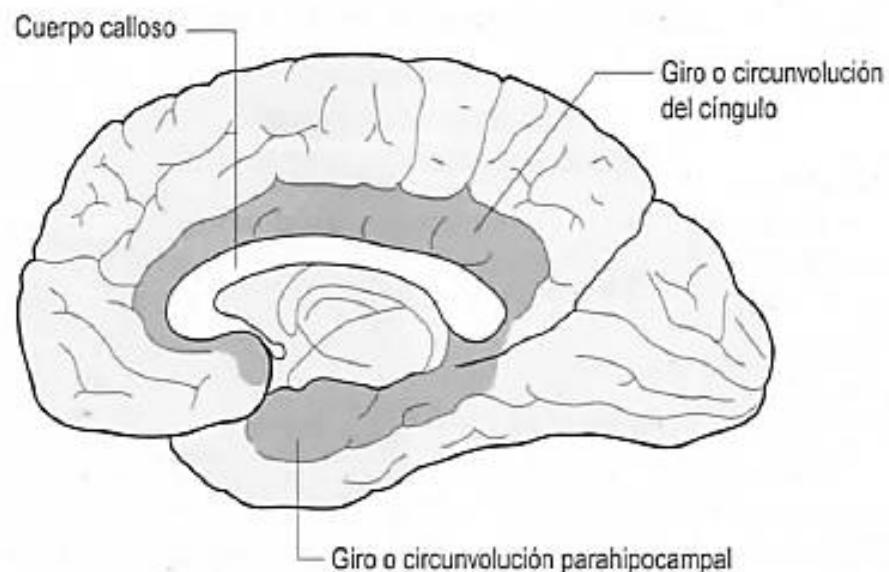


Figura 2.7. Visión medial del hemisferio cerebral. Relación entre giro o circunvolución del cíngulo y el giro o circunvolución parahipocampal. (Neuroanatomía. Texto y Atlas en color, Crossman, Neary).

2.3.1.3 Núcleos Basales.

Los núcleos basales son un conjunto de cuerpos neuronales que forman cúmulos de sustancia gris y que se sitúan dentro de cada hemisferio cerebral.

Las estructuras más generales de los núcleos basales son: el cuerpo estriado, la amígdala y el claustro.

El cuerpo estriado está conformado por el n úcleo lenticular y el caudado (figura 2.8). Su nombre se debe al aspecto estriado que le brindan los haces de sustancia gris que atraviesan la cápsula interna para conectar el n úcleo caudado con el lenticular. El n úcleo lenticular est á conformado por el globo p álido y el putamen. El n úcleo caudado y lenticular se relacionan a trav s de la cápsula interna.

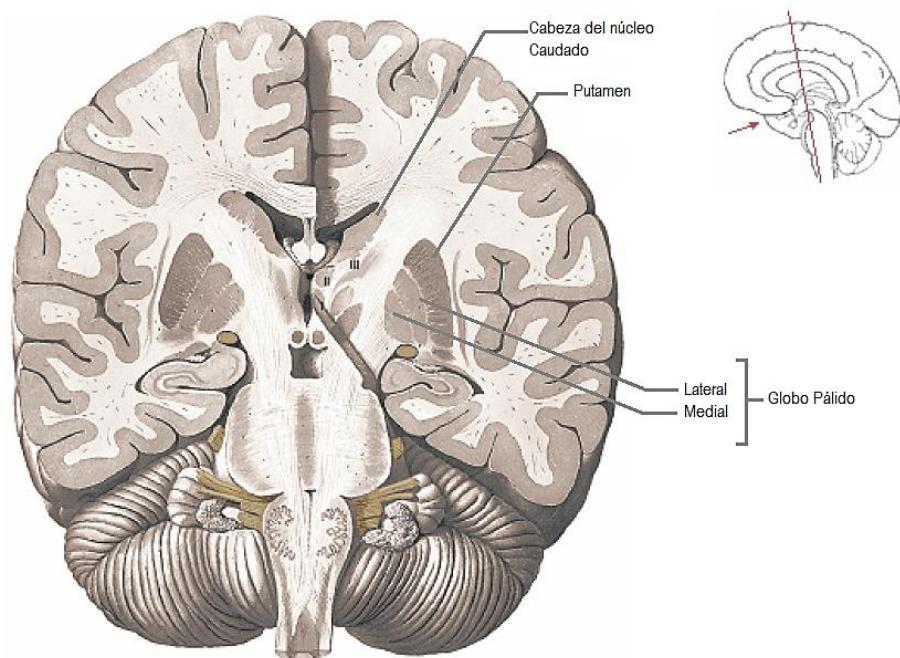


Figura 2.8 .Sección coronal del encéfalo que muestra el Cuerpo Estriado. (Atlas de Anatomía, Sobotta).

El claustro o antemuro es una delgada lámina de sustancia gris, separada de la superficie lateral del n úcleo lenticular por la cápsula externa, y separada de la ínsula por la cápsula extrema (figura 2.9).

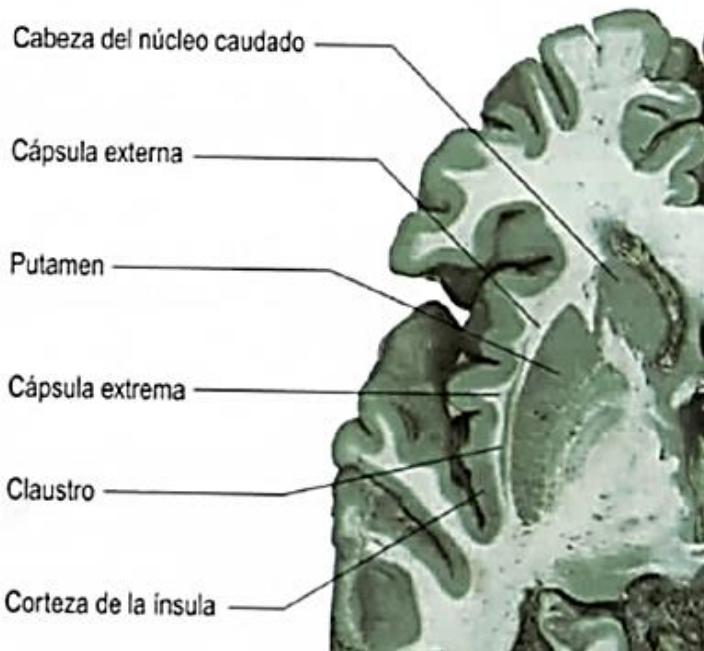


Figura 2.9. Sección horizontal del encéfalo. Claustro. (Neuroanatomía. Texto y Atlas en color, Crossman, Neary)

La amígdala del ser humano es una estructura relativamente grande en comparación con la de los primates. Los individuos poseen dos amígdalas que constituyen un conglomerado de estructuras interconectadas en forma de almendra (de ahí su nombre) [4].

La Amígdala se sitúa cerca del polo temporal, entre el asta inferior del ventrículo lateral y el n úcleo lenticular [32] (figura 2.10).

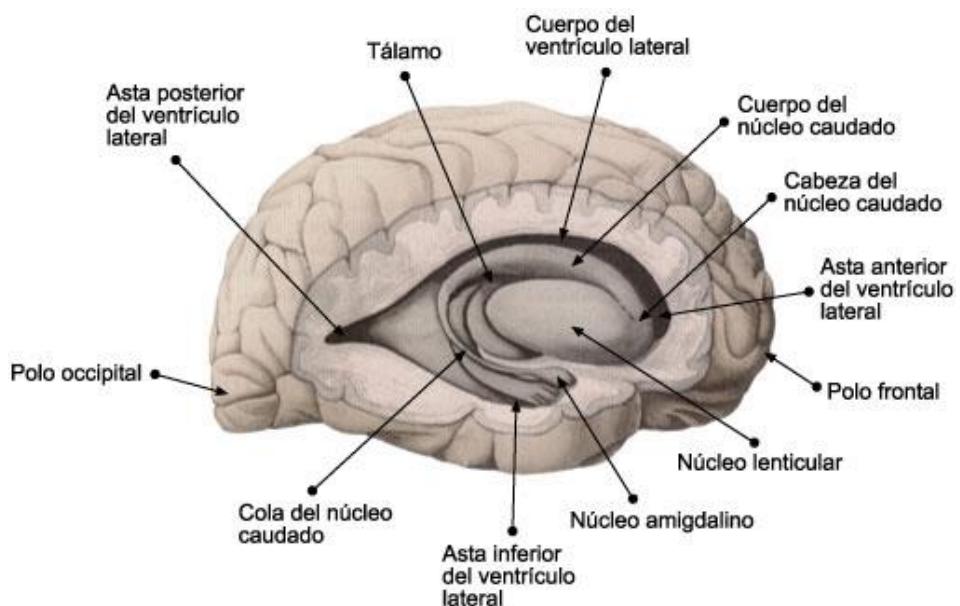


Figura 2.10. Visión lateral derecha. Cuerpo amigdalino. (Neuroanatomía. Texto y Atlas en color, Crossman, Neary)

La amígdala se divide en tres regiones: central, cortico-medial y basolateral [33]. Los aferentes de la amígdala se pueden agrupar en las seis categorías siguientes [34]:

- fibras que se originan en el bulbo olfatorio y la parte olfatoria de la corteza cerebral.
- fibras que surgen del hipotálamo que llevan información visceral.
- aferentes talámicos.
- aferentes directos del tronco encefálico que contienen catecolaminas y serotoninas.
- aferentes hipocámpicos.
- proyecciones de diversas regiones de la neocorteza.

El núcleo cortical del complejo amigdalino está relacionado con el sentido del olfato y el procesamiento de feromonas. Recibe información desde el bulbo olfatorio y la corteza olfatoria. Las aferencias procedentes del tálamo y la corteza llegan principalmente al complejo basolateral. El núcleo central recibe aferencias mayoritariamente del Sistema Autónomo central.

La principal proyección eferente del cuerpo amigdalino es la estría terminal que finaliza en el Hipotálamo, núcleo caudado y putamen. La vía amigdalofuga ventral también se proyecta hacia el hipotálamo [32].

La amígdala envía proyecciones al hipotálamo para incrementar los reflejos de vigilancia, paralización y escape/huida, a los núcleos del nervio trigémino y facial para las expresiones de miedo, al área tegmental ventral, locus cerúleo y al núcleo tegmental lateroventral para la activación de neurotransmisores de dopamina, noradrenalina y adrenalina.

La amígdala regula la producción de respuestas emocionales tanto innatas como aprendidas. Las respuestas innatas vienen determinadas por la aferencia autonómica hipotálamo-troncoencefálica al núcleo amigdalino, desde donde se organizará de manera directa la respuesta somática correspondiente. Por otro lado, la amígdala participa también de los sistemas neurales que subyacen al aprendizaje asociativo, dando lugar a la formación de la memoria implícita [21], al permitir la vinculación de estímulos condicionados (que puedan ser procesados tanto a nivel cortical como solamente a nivel talámico, en este caso permitiendo respuestas cortas, útiles en situaciones de peligro) con respuestas somáticas previamente relacionadas con estímulos no condicionados. Un ejemplo bien estudiado de la participación de la amígdala en el aprendizaje asociativo y la formación de memoria implícita es el miedo aprendido [20].

La función de coordinación emocional que ejerce la amígdala se encuentra regulada asimismo por otros sistemas de control que actúan en paralelo. De este modo, la vía dopamínérgica meso-límbica atenúa la inhibición que la corteza prefrontal ejerce sobre la amígdala, liberando su aferencia sensorial, y con ello la percepción emocional, especialmente de estímulos relacionados con la ira y el miedo. Cuando la amígdala se activa desencadena una serie de cambios fisiológicos que se expresan somáticamente [20].

2.3.1.4 Tálamo

El diencéfalo, estructura cerebral incluída en el prosencéfalo, comprende de dorsal a ventral: el epítálogo, el tálamo, el subtálamo y el hipotálamo (figura 2.11).

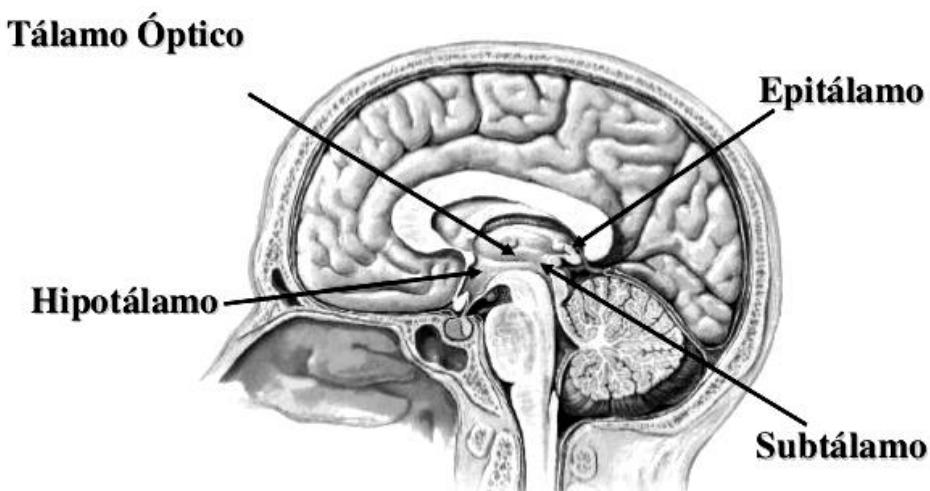


Figura 2.11. Diencéfalo. (Atlas de Anatomía, Sobotta).

El tálamo es el componente más grande del diencéfalo, está situado entre el tronco encefálico y el hemisferio cerebral. Casi todos los núcleos talámicos tienen abundantes conexiones recíprocas con la corteza cerebral. Son de especial interés los núcleos que transmiten información sensitiva hacia las correspondientes regiones de las cortezas cerebrales sensitivas y los núcleos que reciben impulsos desde el cerebelo y desde los núcleos amigdalinos.

El tálamo se divide en regiones anterior, medial y lateral. La región anterior contiene el n úcleo anterior el cual forma parte del sistema límbico. Este participa en el procesamiento de las emociones y en los mecanismos de memoria. La región medial tiene amplias conexiones con la corteza prefrontal y con el hipotálamo, también participa en la integración de aferencias viscerales, olfativas y somáticas, así como en mecanismos que permiten percepciones subjetivas y emotivas. La región lateral juega un rol importante en el procesamiento de la información motora, y medular, proyectando a la corteza premotora y a la corteza motora primaria.

2.3.2 Corteza cerebral

Aunque el estudio neurobiológico de las emociones se haya centrado clásicamente en el sistema límbico, diferentes trabajos experimentales y clínicos han asociado la corteza cerebral (principalmente la corteza prefrontal) con las emociones. En este sentido, a dicha región cerebral se le ha atribuido una función importante relacionada tanto con la experiencia como con la expresión emocional [4] [16] [35] [36].

El hemisferio cerebral está compuesto por:

- La corteza cerebral superficial, replegada para formar giros o circunvoluciones y surcos.
- La sustancia blanca subyacente, que consta de fibras aferentes y eferentes.
- Masas nucleares profundas: los núcleos basales.
- Los dos hemisferios cerebrales, que están separados por la gran fisura longitudinal del cerebro y unidos por el cuerpo calloso.

El hemisferio está dividido en cuatro lóbulos (frontal, parietal, temporal y occipital) sobre la base de la topografía de su superficie (figura 2.12). Los detalles principales que indican las divisiones entre los lóbulos son el surco lateral, el surco parietooccipital y el surco central.

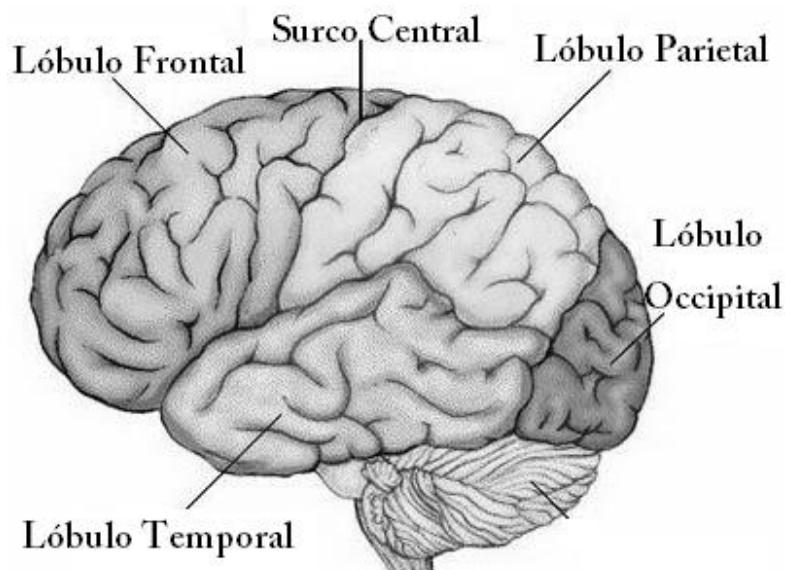


Figura 2.12. Visión Lateral del Hemisferio Cerebral. Anatomía del Cuerpo Humano, Gray.

El surco central marca el límite entre los lóbulos frontal y parietal. El lóbulo frontal constituye toda la región por delante de este surco (figura 2.13), y corriendo paralelo a él se sitúa el giro precentral que es la región motora primaria de la corteza cerebral. Por delante del giro o circunvolución precentral se halla el área premotora, y por delante de ésta se encuentra la corteza prefrontal; esta región es la encargada de la personalidad y la conducta racional.

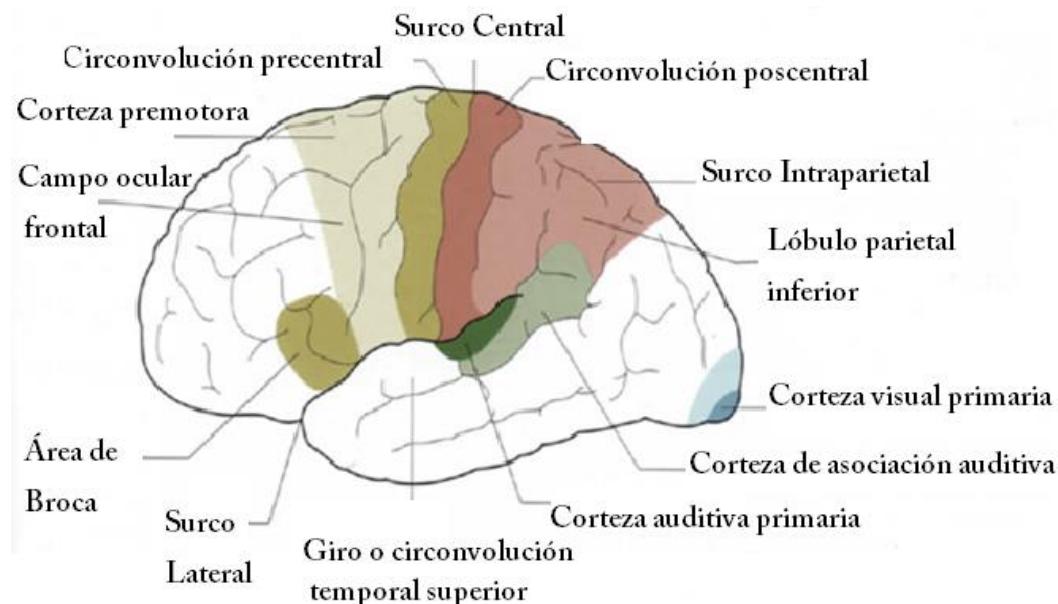


Figura 2.13. Visión Lateral del Hemisferio cerebral. (Neuroanatomía. Texto y Atlas en color, Crossman, Neary)

La corteza prefrontal, tiene numerosas conexiones con las cortezas: parietal, temporal y occipital, a través de fibras de asociación. La corteza prefrontal tiene funciones cognitivas muy superiores, que incluyen facultades intelectuales, de discernimiento y predictivas y la planificación de la conducta. Además existen conexiones recíprocas entre la corteza prefrontal, la amígdala y el hipocampo. De modo que los estímulos con contexto emocional, de acuerdo con circuitos adquiridos por el aprendizaje, actúan sobre la amígdala, pero las conexiones con la corteza prefrontal, el lóbulo temporal anterior y el hipocampo permiten que estos puedan activar esos mismos circuitos sin estímulos externos, por ejemplo a través de la imaginación y la memoria explícita de aquellos. Por ejemplo, si una sombra que nos ha parecido algo amenazador provoca taquicardia y sensación de miedo, la identificación consciente de que tal imagen es inocua, detiene la respuesta automática emocional.

Por detrás del surco central y entre la cara ventral del surco poscentral se localiza el giro poscentral encargado de la sensación general del cuerpo. El área dorsal al giro temporal superior corresponde al área de Heschl, que es el área auditiva primaria, y posterior a ésta el área de Wernicke o de asociación auditiva, que interviene en la compresión del lenguaje hablado. En el giro occipital se encuentra el área visual primaria.

Sin los lóbulos prefrontales la vida emocional desaparecería porque sin compresión de que algo merece una respuesta emocional, no hay respuesta alguna [4].

Hipotálamo

El hipotálamo es la parte más ventral del diencéfalo, se sitúa por debajo del tálamo y ventromedial al subtálamo (figura 2.11).

El hipotálamo es capaz de integrar señales enteroceptivas procedentes de órganos internos y de líquidos corporales y de realizar ajustes apropiados del medio interno en virtud de sus sistemas de entrada y salidas de información. Las entradas de información al hipotálamo son de origen circulatorio y nervioso. Es capaz de generar respuestas frente a estos estímulos por medios también circulatorios y nerviosos. Dirige la síntesis y liberación de hormonas, pero también es capaz de iniciar acciones apropiadas para el comportamiento motor de naturaleza instintiva a través de sus conexiones con el sistema límbico. El hipotálamo tiene la capacidad de influir en los comportamientos adaptativos más complejos debido a su íntima unión con dos estructuras importantes: el sistema límbico y la corteza de asociación del lóbulo frontal.

2.3.3 Formación Reticular

La formación reticular está compuesta por una compleja matriz de neuronas que se extiende a lo largo del tronco encefálico (figura 2.14). Filogenéticamente es una parte relativamente antigua del tronco encefálico, cuyas neuronas efectúan un número importante de funciones, algunas de las cuales son necesarias para la supervivencia. La formación reticular tiene extensas conexiones aferentes y eferentes con otras partes del Sistema Nervioso.

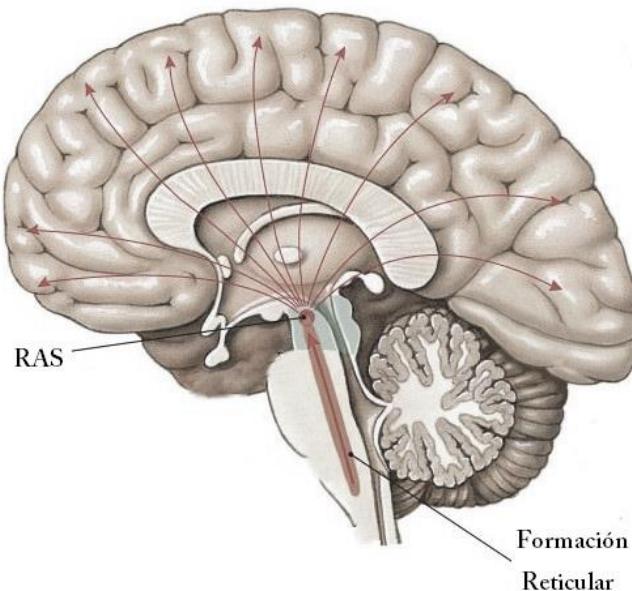


Figura 2.14. Formación Reticular (<http://www.isep.es/wp-content/uploads/2014/06/heminegligencia.pdf.>)

Algunas fibras ascendentes de la formación reticular constituyen el sistema reticular activador (RAS). Estas neuronas reciben información de múltiples fuentes sensitivas. A través de la intermediación de núcleos talámicos, provocan la activación de la corteza cerebral y estimulan el estado de vigilia.

Los núcleos del rafe son un grupo de núcleos en la línea media que se extienden a lo largo de todo el tronco del encéfalo (figura 2.15). Muchas de las neuronas de estos núcleos son serotoninérgicas.

El locus cerúleo (figura 2.15) es un grupo de neuronas pigmentadas involucrado en la respuesta al pánico y al estrés, que se sitúan en el tronco del encéfalo, en la parte dorsal de la protuberancia, bajo el suelo del IV ventrículo. Es el principal grupo de células noradrenérgicas del encéfalo. Las fibras ascendentes se proyectan hacia el cerebelo, hipotálamo, tálamo, estructuras límbicas y corteza cerebral. Las fibras descendentes se proyectan a lo largo del tronco del encéfalo y medula espinal.

EL locus cerúleo está involucrado en muchos de los efectos simpáticos durante el estrés debido al incremento en la producción de la noradrenalina. Estudios recientes sustentan que el locus cerúleo es activado por diversos estímulos estresantes y nociceptivos, y también por estímulos fisiológicos como la hipertensión, la hipoxia y la estimulación visceral, que incrementan la descarga de las neuronas de esta estructura.

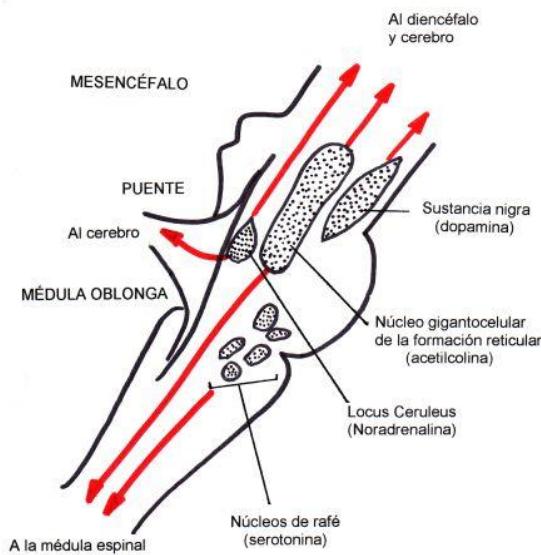


Figura 2.15. Formación reticular. Núcleos del rafe y Locus cerúleo.
[\(http://www.psico.unlp.edu.ar/catedras/neuroanatomia/images/tmp9088.png\)](http://www.psico.unlp.edu.ar/catedras/neuroanatomia/images/tmp9088.png)

2.3.4 Vías Emotivas

Existen dos vías emotivas. Una vía rápida inhibidora del neocortex y una vía lenta inhibidora de movimientos estereotipados

2.3.4.1 La vía rápida e inhibidora de la Neocorteza

Esta vía tiene su efecto en situaciones de peligro y de estrés. La ventaja de esta vía, es que permite las acciones rápidas y los reflejos, sin embargo, éstos son poco precisos.

Los estímulos de emergencia ingresan por los sentidos auditivos y visuales que son captados por la formación reticular que lleva la información a los núcleos intralaminares. Estos núcleos del tálamo dirigen la información al área basolateral de la amígdala. Desde aquí la vía se divide en tres ramas. En una de ellas, la información es dirigida al área central de la amígdala, por medio de interneuronas y desde aquí, mediante un conjunto de fibras, la señal es dirigida: a la formación reticular, para activar la vía retículo espinal, produciendo, de esta manera, movimientos posturales, y al locus cerúleo para que se genere la liberación de noradrenalina. La segunda rama, se dirige al área medial de la amígdala activando la estría terminal que proyecta hacia el n úcleo caudado y el putamen. Éstos activan la vía extrapiramidal, para dar inicio a movimientos automatizados como los necesarios para correr o huir. Debido a que esta vía es instintiva no requiere de un objetivo motriz planeado por el área premotora, sino una actuación inmediata sin la utilización de la corteza prefrontal. La tercera rama es un haz de fibras que sale hacia el hipotálamo con el fin de generar los cambios fisiológicos como por ejemplo elevar la presión arterial, etc. Esta vía es funcionalmente activa en los animales carentes de corteza prefrontal produciendo así los movimientos estereotipados.

2.3.4.2 Vía lenta inhibidora de movimientos estereotipados

La vía lenta es diferente de la vía rápida, ya que ésta sigue un trayecto mucho más corto entre el tálamo y la amígdala.

La vía lenta se encarga de la conducta, dependiendo del entorno, con la ventaja de dar respuestas precisas y apropiadas en el momento, pero con la desventaja de ser muy lenta ya que en su recorrido se involucran las estructuras cerebrales. La vía comienza cuando los órganos sensitivos de la visión y de la audición envían el estímulo que viaja a los núcleos geniculados lateral y medial, respectivamente. De ahí, la información es dirigida e interpretada en las cortezas visual y auditiva. Luego la corteza prefrontal es activada mediante las asociaciones existentes con las cortezas visual y auditiva, a fin de inhibir los movimientos estereotipados y activar el área premotora, para dar origen a los movimientos adecuados dependiendo del estímulo y del entorno. La corteza prefrontal envía información al núcleo dorsomedial del tálamo. De aquí, salen fibras que llegan al área medial del complejo amigdalino, que mediante sinapsis interneuronal estimulan el área basolateral del complejo, activando la vía amigdalofuga ventral. Esta vía eferente contiene fibras que hacen sinapsis al pasar por la sustancia innominada, y desde aquí, envían fibras hacia la corteza prefrontal a través del tálamo. La sustancia innominada es un cúmulo de cuerpos neuronales que corresponde a la porción caudal de los globos pálidos.

La sustancia innominada también envía fibras a la porción medial del núcleo amigdalino con información hacia los núcleos basales a fin de darle armonía a los movimientos voluntarios. Debido a esto, esta vía amigdalofuga ventral solo puede ser perteneciente a la vía lenta de inhibición. Otras fibras no se detienen en la sustancia innominada sino que salen hacia los núcleos septales, del área olfatoria medial. Estos se comunican con: el hipotálamo, para mantener al sistema nervioso autónomo en correcto funcionamiento según su situación y, con la formación reticular a fin de mantener los estados de alerta. Los núcleos septales son formaciones que se encuentran situadas anteriormente a la comisura anterior e inferior de la rodilla del cuerpo caloso. Otras fibras salen directamente del núcleo amigdalino hacia el hipotálamo manteniendo los niveles de metabolismo normales.

2.4 Neurofisiología de las emociones básicas

Darwin trató a las emociones como entidades discretas. Según sus investigaciones definió ocho emociones primarias o básicas que son comunes en humanos y animales: Alegría, Enojo, Miedo, Tristeza o angustia, Interés, Sorpresa, Disgusto o asco, Vergüenza [9].

Sin embargo, una clasificación más reciente a cargo del neurocientífico Antonio Damasio indica que las emociones básicas, aquellas que son fácilmente identificables en los seres humanos de numerosas culturas y también en especies no humanas, son [37]:

- Miedo
- Enojo
- Asco
- Sorpresa
- Tristeza
- Felicidad

Estas emociones van acompañadas de patrones de conducta tales como: respuestas faciales, motoras, vocales, endócrinas y autonómicas, que son reconocibles por encima de diferencias culturales y raciales en los seres humanos.

Existen otras muchas emociones, como la calma, la culpa, la depresión, etc. que se denominan emociones secundarias. Estas presentan un componente cognitivo más alto y además están asociadas a las relaciones interpersonales.

2.4.1 Miedo

"Supongamos que una noche está leyendo tranquilamente un libro en su hogar cuando de repente oye un ruido en otra habitación" [4]

A partir de ese suceso, el primer circuito cerebral implicado se limita a ingresar la información del sonido desde el oído hasta el tallo encefálico y el tálamo. De aquí, se ramifica en dos vías: una de ellas se dirige a las amígdalas y al hipocampo, la otra, más larga, conduce hasta el córtex auditivo, situado en el lóbulo temporal, donde se clasifican y comprenden los sonidos.

El hipocampo, una región clave para el almacenamiento de la memoria, compara rápidamente este ruido con otros sonidos similares que hayan sido almacenados en la memoria, tratando de descubrir si se trata de un sonido familiar o no. Mientras tanto el córtex auditivo realiza un análisis más preciso del sonido intentando comprender su origen. Luego, elabora una hipótesis y envía este mensaje a las amígdalas y al hipocampo, quienes rápidamente lo comparan con recuerdos semejantes. Si la conclusión es tranquilizadora (por ejemplo ruido de la ventana movida por el viento), el estado de alerta general se paraliza. Pero si, por el contrario, la conclusión es dudosa, se pone en marcha una serie de eventos. Las amígdalas lanzan una señal de alarma que activa al hipocampo, el tallo cerebral y el Sistema Nervioso Autónomo.

Las señales procedentes de la amígdala también proyectan a diversas partes del cerebro. Por ejemplo, la rama procedente de las áreas central y cortico medial del complejo amigdalino se dirigen a la región del hipotálamo, encargada de segregar una sustancia que activa la respuesta de urgencia emocional, la hormona corticotrópica, que a través de la liberación de otras hormonas, moviliza la reacción de lucha o huida. Por su parte el área basolateral de la amígdala, envía ramificaciones al cuerpo estriado, que está relacionado con las regiones cerebrales encargadas del movimiento. Otras ramificaciones neuronales de la amígdala envían señales a través del núcleo central hasta la médula, y desde ella, al sistema nervioso autónomo, activando una amplia variedad de respuestas en el sistema cardiovascular, muscular y visceral. Otras ramificaciones procedentes del área basolateral de la amígdala, se dirigen al córtex cingulado (también llamado circunvolución del cíngulo anterior, que corresponde a la parte frontal de la circunvolución del cíngulo) y a otras fibras que regulan la musculatura esquelética. En los seres humanos, estos mismos circuitos son los encargados de tensar la musculatura de las cuerdas vocales responsables del tono de voz agudo propio de quien experimenta el miedo.

Hay otro camino que conduce desde la amígdala hasta el locus cerúleo que genera la liberación de noradrenalina. El efecto de la noradrenalina aumenta la reactividad global de las áreas cerebrales que la reciben, sensibilizando los circuitos sensoriales. La mayor parte de estos cambios fisiológicos ocurren de modo inconsciente, de modo que uno todavía no sabe que experimenta miedo [4].

El miedo produce cambios fisiológicos inmediatos: se incrementa el metabolismo celular, el corazón bombea sangre a gran velocidad para llevar hormonas (adrenalina) a las células, produciendo taquicardia, la sangre fluye a los músculos mayores, principalmente a las extremidades inferiores, aumenta la presión arterial, la glucosa en sangre y la actividad cerebral, se incrementa la sudoración, se detiene el sistema inmunitario y todo aquello que no sea esencial en ese momento, se dilatan las pupilas para incrementar el ingreso de luz y se observan expresiones faciales características (figura 2.16): apertura palpebral (los párpados superiores se elevan al máximo y los inferiores se tensan) para aumentar el campo visual, las cejas elevadas se acercan y los labios se alargan hacia atrás [22].



Figura 2.16. Miedo. (<http://www.artnatomia.net/>)

2.4.1.1 Miedo en la voz

Sherer en sus trabajos, realiza una descripción acerca de la relación entre los parámetros de la voz y las emociones [38]. Según este autor el miedo presenta un tono medio más elevado, un rango del pitch mayor, una velocidad de locución rápida y una voz tensa.

El estado emocional del miedo según Ververidis y Kotropoulos se correlaciona con un alto nivel del tono y un nivel de intensidad elevada. El lapso de tiempo entre segmentos de voz es más corta que en el estado neutral, o sea tiempos de silencio más cortos, por lo tanto una velocidad mayor en la locución [39].

Murray y Arnott (1993) relacionan el miedo con parámetros del discurso tales como la velocidad de locución, indicando que este parámetro se ve acelerado durante este estado emocional. En relación al tono, este estado emocional genera una alta variabilidad de este parámetro y un rango amplio respecto del estado neutro. La intensidad, representada por la energía de la señal, refiere este grupo de investigación, es normal, similar al estado neutro. Estos parámetros describen una calidad de voz irregular cuando la persona presenta miedo [40].

En relación con la curva del tono, un grupo de investigación de Valencia obtuvo mediciones de este parámetro en el estado emocional del miedo mediante laringografía con una resolución temporal de 1ms. Obtuvieron que la emoción con mayor valor medio de pitch era el miedo [41].

En otros trabajos se demuestra que la curva de fluctuación del tono es discontinua para emociones negativas como el miedo. Comparando el tono medio de esta emoción respecto a las demás emociones básicas, este grupo observó que el miedo presentaba un tono más elevado (254Hz), el rango del tono era superior y la velocidad de locución era mayor [42].

Los estudios fisiológicos realizados por Williams y Stevens [43] indican que la existencia de miedo genera una serie de cambios fisiológicos como cambios en la profundidad de los movimientos respiratorios, incremento de la presión subglótica, que se evidencian en el discurso, resultando este, rápido, con un tono promedio más elevado y un rango mayor del pitch.

En general, los autores coinciden en que el estado emocional de miedo, genera cambios fisiológicos que se evidencian con el incremento y la variabilidad del tono, velocidad del discurso superior e intensidad de la señal de voz de orden mayor respecto del neutro.

Ya se ha mencionado que uno de los acontecimientos fisiológicos que se presenta en el estado de miedo es la activación de estructuras cerebrales (tales como el córtex cingulado). Esta estructura se encarga de generar una tensión muscular. Siendo las cuerdas vocales un complejo de base muscular, la tensión muscular genera un estiramiento de las cuerdas que lleva al incremento de su frecuencia fundamental (tono), dando el aspecto de una voz aguda.

Los cambios a nivel de los movimientos respiratorios y el incremento de la presión subglótica controlan el aumento de la energía que atraviesa el tracto vocal, dando mayor intensidad en el discurso.

2.4.2 Enojo

"Supongamos que un conductor se nos acerca peligrosamente mientras estamos circulando por la autopista. Aunque nuestro primer pensamiento reflejo sea exclamar un insulto hacia el irresponsable, lo que realmente resulta decisivo para el desarrollo de la rabia es que el pensamiento vaya seguido de otros pensamientos de irritación y venganza como, por ejemplo ¡podría haber chocado conmigo! ¡No puedo permitírselo!"[44].

En este caso, nuestros nudillos palidecen mientras las manos aprietan firmemente el volante, el cuerpo se predispone a la lucha, tiembla, el corazón late con mayor intensidad y tensamos los músculos del rostro [4].

Es la amígdala la que genera una serie de eventos biológicos que desencadenan en los estados mencionados. Sin embargo, por otro lado el neocórtex tiende a fomentar la ira más calculada, la venganza y los pensamientos que involucran la predicción de tragedias no ocurridas, como ¡podría haber chocado!

Según Dolf Zillmann, psicólogo de la Universidad de Alabama, el detonante universal del enfado es la sensación de hallarse amenazado [4].

Los sentidos activan la amígdala que genera mediante una serie de interconexiones cerebrales la descarga de cortisol y catecolaminas, y su desplazamiento a lo largo de la vía adrenocortical del sistema nervioso, aportando así el estado general adecuado a la respuesta. Esta excitación

adrenocortical generalizada puede perdurar horas lo cual mantiene al cerebro emocional predisposto a la excitación. Además la liberación de cortisol por la corteza suprarrenal influye en las consecuencias fisiológicas. Esto se observa en el aumento del flujo sanguíneo hacia las extremidades superiores, aumento de la temperatura de la piel, incremento del ritmo cardíaco, y dado el aumento en la tasa hormonal en sangre (como la adrenalina) se genera una gran cantidad de energía, que permite acometer acciones vigorosas [4][16][45].

Peter Kaufman, director interino del Behavioral Medicine Branch of the National Heart, Lung, and Blood Institute, indica que la explosión de ira aumenta la frecuencia cardíaca y la tensión arterial, forzando así al corazón a un sobreesfuerzo adicional, concluyendo que los mecanismos desencadenados por el enojo afectan directamente la eficiencia del bombeo cardíaco [4].

Las expresiones faciales (figura 2.17) para el estado de enojo muestran las cejas bajas y contraídas, líneas verticales entre las cejas, parpado inferior tenso, parpado superior tenso, puede estar bajo o no por la acción de las cejas, mirada fija y feroz, labios en una de estas dos posiciones: continuamente apretados, con las comisuras rectas o bajas, o abiertos, tensos y en forma cuadrangular, tendencia a apretar los dientes, las pupilas pueden estar dilatadas [4] [22].



Figura 2.17. Enojo. (<http://www.artnatomia.net/>)

2.4.2.1 Enojo en la voz

Estudios fisiológicos [43] determinan los cambios que se producen durante un estado de ira. Algunos que no resultan evidentes como por ejemplo el aumento de la presión subglótica y los cambios en la profundidad de los movimientos respiratorios generan indirectamente variaciones en la voz. Estas respuestas fisiológicas al estado de enojo generan un flujo de aire rápido no lineal en el tracto vocal que provoca vórtices situados cerca de las cuerdas vocales falsas proporcionando señales de excitación adicionales (armónicos adicionales) distintos al pitch [46]. Por tal motivo, estos armónicos adicionales generan confusión entre el segundo formante (F2) (como se verá en el **capítulo 3 sección 3.2.1**) y el primero (F1) provocando además la interferencia entre el F1 y la

frecuencia fundamental. Además el aumento de presión sub glótica desemboca en niveles superiores de energía y tono en la voz que se evidencian por el aumento en la intensidad del discurso y el tono más agudo [39].

Algunos investigadores han encontrado que la tasa de locución es menor, o sea se habla más lentamente, con algunas diferencias de género. En hombres la expresión de la ira se realiza con un ritmo del habla lento en comparación con el género femenino que emplea una tasa superior en circunstancias similares [47].

Según Sherer el enfado se caracteriza por un tono de voz medio alto y una velocidad de locución rápida [38].

Los parámetros del discurso durante el enojo según Murray y Arnott coinciden con otros autores en el alto nivel de variación del tono, con un rango del mismo amplio e intensidad de locución elevada [40]. A favor de la teoría de Sherer, estos autores también indican una velocidad de discurso ligeramente acelerada.

Los valores de la curva de entonación obtenidos a través de laringografía corroboran el incremento en la variabilidad del tono que se genera en el discurso de ira [41].

Otros autores describen discontinuidad en la curva de entonación durante situaciones de enojo, fisiológicamente esto se representa como una fluctuación en la vibración de las cuerdas vocales.

En general, el enojo se caracteriza por: un tono más elevado, debido a la tensión de los músculos que controlan las cuerdas vocales y por el incremento en la presión subglótica que genera armónicos que distorsionan el valor real del pitch; variabilidad en el tono, lo que implica un rango superior, debido al incremento en el flujo de aire no lineal que atraviesa el tracto vocal. La velocidad de locución puede ser de moderada a rápida sin embargo es un parámetro contradictorio entre los investigadores.

2.4.3 Asco

"Estás en un restaurante. Ya está un poco enfermo y el restaurante parece bastante sucio, pero es el único restaurante en el pueblo, por lo que en realidad no tienen elección.... Cuando finalmente recibe su plato, que es una especie de sopa de fideos. Toma la cuchara, listo para comer. Aunque estás muy hambriento, la sopa no tiene sabor muy bueno. Parece que no es muy fresca... De repente se ve una enorme cucaracha en su plato! .Usted es primero sorprendido y salta hacia atrás de la silla." [44]

Cuando esta situación de asco aparece, tres áreas de interconexión cerebral se activan [48]: la ínsula anterior, los núcleos basales y el córtex prefrontal. La ínsula o corteza insular es una estructura cerebral que está situada en la profundidad de la Cisura de Silvio que separa las cortezas temporales y parietal inferior. La porción anterior de la ínsula está relacionada con el olfato, gusto, sistema nervioso autonómico y la región límbica. La ínsula recibe información de aferencias homeostáticas a través de vías sensoriales por la vía del tálamo y activa, mediante sus vías eferentes, estructuras relacionadas con el sistema límbico tales como la amígdala, cuerpo estriado (núcleos basales) y el córtex prefrontal [49], principalmente el córtex prefrontal derecho, región asociada con los sentimientos negativos [50]. La estimulación de estas estructuras da lugar a las correspondientes respuestas fisiológicas a cargo del sistema nervioso autónomo.

Estas áreas no solo se activan cuando se experimenta la sensación de asco sino también cuando se observa esta expresión en otras personas. En un estudio publicado en la revista Nature

Neuroscience, los investigadores Manes y Calder de la Universidad de Cambridge en Inglaterra, analizaron a un paciente que padecía una lesión en la región insular y, aunque podía reconocer las expresiones faciales de miedo, ira, sorpresa, alegría y tristeza, mostraba imposibilidad selectiva para el reconocimiento del asco [16].

La respuesta fisiológica del asco típicamente se asocia a los procesos nauseabundos y al incremento de la salivación [51]. Investigaciones recientes demuestran que la fisiología del estado de asco está asociada con la respuesta autonómica parasimpática, particularmente la actividad cardíaca [45].

El asco es una emoción primaria que evolucionó en asociación con el rechazo automático y beneficioso de alimentos potencialmente tóxicos o en pudrición o que se sabe han estado en contacto con inductores de asco [37]. Para la sociedad moderna, el asco puede referirse más ampliamente en términos de disgusto, que implica el rechazo emocional a cosas basadas en ideas de cualidades perceptuales (como por ejemplo, racistas, vino barato, corrupción política). Rozin, Haidt y McCauley (2000) cuentan como el disgusto consiguió convertirse en una emoción social que es ante todo una respuesta a los actos impuros de otras personas [52].

A nivel de las expresiones faciales (figura 2.18), el gesto que universalmente transmite el mensaje de que algo resulta literal o metafóricamente repulsivo para el gusto o el olfato corresponde a la elevación del labio superior, frunciendo la nariz lo cual sugiere un intento, según Darwin, de cerrar las fosas nasales para evitar un olor nauseabundo o para expulsar un alimento tóxico, y el estrechamiento de los ojos [48].



Figura 2.18. Asco. (<http://www.artnatomia.net/>)

2.4.3.1 Asco en la voz

Los procesos fisiológicos que ocurren durante el estado de emoción asco se ven reflejados en la actividad del discurso. La activación del sistema nervioso parasimpático genera el aumento de la actividad peristáltica, por la situación nauseabunda, el cierre de la glotis para evitar la salida de aire que en conjunto con la elevación del paladar blando, para bloquear las fosas nasales, (expresión

notoria en la figura 2.18) reducen la velocidad del aire que circula por el tracto vocal, lo cual disminuye la intensidad y genera una tasa de discurso más baja [39] [43].

Es determinante para el estudio de esta emoción el parámetro pitch que se relaciona con la vibración de las cuerdas vocales, las cuales tienen inervación del parasimpático a través del nervio vago. Los expertos coinciden en el bajo tono del discurso cuando se presenta este estado emocional [43] [41] [40].

2.4.4 Sorpresa

Suponga que se encuentra reposando apaciblemente en su habitación, hay calma a su alrededor, solo se oye silencio y quietud. De repente....los vidrios de la ventana vibran, la lámpara se mece sin control alguno, las paredes crujen... todo eso en un solo instante.... Temblor!...y luego....nuevamente la serenidad.

La sorpresa es un estado emocional muy breve, resultante de un evento inesperado que produce una rápida activación de las zonas de proyección sensorial implicadas en la percepción de los desencadenantes emocionales [4]. Los mensajes aferentes pasan por el tálamo, el sistema límbico y el córtex cerebral a fin de dar una respuesta elaborada, en consecuencia, el plan de acción más adecuado.

Las eferencias del sistema emocional generan una aceleración de la frecuencia cardiaca, vasoconstricción periférica, aumento del tono muscular general, cambios en la frecuencia y amplitud de la respiración o inspiración breve y de corta latencia y dilatación pupilar, respuestas propias de la activación del sistema nervioso simpático[43].

A nivel de las expresiones faciales (figura 2.19), la sorpresa genera un arqueo de las cejas a fin de obtener un mayor campo visual, dilatación pupilar para proporcionar más información sobre el acontecimiento inesperado, párpados abiertos, párpado superior levantado y párpado inferior bajado, arrugas horizontales en la frente, la mandíbula se desplaza hacia abajo, de modo que los labios y los dientes quedan separados [4] [48].



Figure 2.19. Sorpresa. (<http://www.artnatomia.net/>)

2.4.4.1 Sorpresa en la voz

La respuesta fisiológica de la sorpresa consiste en un incremento de la velocidad del volumen glotal durante la vibración de las cuerdas vocales, generando un amplio rango de variación del pitch [39]. Las cuerdas vibran a mayor frecuencia lo cual se presenta como un tono de voz mayor que el neutral. La velocidad del discurso no parece estar afectada [38].

2.4.5 Tristeza

"Usted acaba de regresar de un día agotador en el trabajo. Tiene su mente en un estado neutral, cuando de repente suena el teléfono, atiende y se da cuenta que es su novio (novia). Él (ella) le anuncia que no quiere seguir la relación con usted. En primer lugar, no lo cree, pero después de un tiempo se da cuenta lo que acaba de ocurrir. Recuerda los buenos momentos que pasó con él (ella) y asocia estos recuerdos con el hecho que la relación acaba de terminar, su cara está llena de lágrimas, la tristeza lo apodera." [44]

La principal función de la tristeza consiste en ayudarnos a asimilar una pérdida irreparable. La tristeza provoca la disminución de la energía y del entusiasmo por las actividades vitales y cuando se profundiza y se acerca más a la depresión, más se enlentece el metabolismo corporal [4].

Parece ser que la sustancia gris periacueductal (GPA) juega un rol importante en las acciones de esta emoción. La GPA es sustancia gris que rodea el acueducto del mesencéfalo. Se extiende desde cerca de la comisura posterior hasta el nivel del locus cerúleo en la unión del puente mesenfálico. En un experimento realizado en el Hospital de la Salpêtrière en París se obtuvieron resultados interesantes respecto a la neurofisiología de la tristeza. Intentando realizar un tratamiento alternativo con estimulación eléctrica a baja intensidad y elevada frecuencia en los núcleos motores del bulbo raquídeo en una paciente con síntomas parkinsonianos que no respondía al tratamiento farmacológico, se estimularon ciertas zonas del bulbo que generaron en la paciente los signos típicos de un estado de tristeza. Se le implantaron dos electrodos largos orientados verticalmente, y cada uno tenía cuatro contactos. Los científicos determinaron que fue uno de los contactos del electrodo ubicado hacia el lado izquierdo del tallo cerebral, el que inició el estado emocional [37]. Damasio sostiene que la emoción de la tristeza en lugar de originarse en la corteza cerebral, comenzó en una región subcortical.

Otro experimento para determinar qué áreas se activan durante la emoción de tristeza consistió en medir la cantidad de flujo sanguíneo en múltiples áreas cerebrales mediante la Tomografía por emisión de positrones. Se sabe que la cantidad de sangre que fluye hacia cualquier región del cerebro está estrechamente relacionada con el metabolismo de las neuronas en esa región.

El experimento consistió principalmente en evocar un estado emocional preciso haciendo uso de la memoria emotiva. Cada individuo debía pensar en un episodio emocional de su vida.

Los resultados mostraron que las estructuras involucradas durante el estado de tristeza eran la corteza cingulada, cortezas somatosensoriales de la ínsula y los núcleos del tegmento del tallo cerebral. Además se pudo evidenciar la desactivación marcada en las cortezas prefrontales (o sea una reducción de la actividad de toda la región), el hallazgo encaja con el hecho de que el pensamiento lógico, la fluidez de ideas se reduce con la tristeza [37].

Las respuestas fisiológicas provistas por la acción del Sistema Nervioso Parasimpático, evidenciadas en este experimento fueron interesantes, los investigadores advirtieron que los cambios fisiológicos como los cambios en la conductancia dérmica precedían al sentimiento de tristeza. O sea, los monitores eléctricos registraron la actividad de la emoción, antes de que los sujetos evaluados avisaran que estaban sintiendo la experiencia, demostrando que los estados emocionales llegan primero y la conciencia de que está presente llega después [37].

A nivel de las expresiones faciales (figura 2.20), por lo general un individuo con estado de tristeza lo expresa a través de las áreas de la frente, los ojos y la boca. En la tristeza, las esquinas interiores de las cejas se elevan, la piel debajo de la ceja se triangula con el ángulo interno, la esquina superior del parpado inferior se eleva, las comisuras de los labios se reducen [53].



Figura 2.20. Tristeza. (<http://www.artnatomia.net/>)

2.4.5.1 Tristeza en la voz

La respuesta fisiológica cuando se está experimentando la sensación de tristeza se presenta principalmente en una disminución de la energía y del metabolismo en general. A nivel de la voz, esta pérdida de energía implica un decremento en la intensidad del discurso como así también en la tonalidad, lo cual se corresponde con una voz un poco más grave. La locución pierde velocidad a causa de la falta de vigorosidad en los músculos de la fonación.

Williams y Stevens mostraron que la activación del Sistema Nervioso Parasimpático genera en la voz una velocidad de locución baja, con bajo tono y energía en el espectro de frecuencias disminuida, estos resultados son similares a los obtenidos mediante el uso laringografía. [43][41].

Sherer sugiere que durante un estado de tristeza se exhibe un tono medio más bajo que lo normal, con un estrecho rango de variación y una velocidad de locución lenta [38]. Estos resultados se asimilan a los hallazgos de otras investigaciones [40] y corrobora la implicancia del sistema parasimpático en la expresión corporal de la tristeza.

2.4.6 Felicidad

"Esta mañana usted se entera que ha ganado el primer premio de \$ 5.000.000 en la lotería! Usted está en un estado de ánimo muy feliz, por supuesto, porque se da cuenta que muchos de sus sueños se convertirán en realidad. Después de la sorpresa de haber ganado, quiere contarle a sus amigos lo feliz que se siente." [44]

Uno de los principales cambios biológicos producidos por la felicidad consiste en el aumento de la actividad de un centro cerebral que se encarga de inhibir los sentimientos negativos y de aquietar los estados que generan preocupación, al mismo tiempo que aumenta el caudal de energía disponible [4].

La corteza orbitofrontal derecha y la estriada ventral izquierda son regiones implicadas en los estados placenteros [16] [37]. Sin embargo Russell indica que las emociones de placer son causadas principalmente por estructuras cerebrales subcorticales [52].

Utilizando la técnica de tomografía por emisión de positrones (TEP) se midió la cantidad de flujo sanguíneo en áreas cerebrales durante el estado de felicidad y se encontró que las regiones que se activaron fueron el cingulado anterior, cingulado posterior, el hipotálamo y la ínsula [37].

Otros estudios de imágenes funcionales han demostrado que durante el reconocimiento del estado emocional felicidad la amígdala se activa menos que en el reconocimiento de otros estados emocionales [54].

La experiencia de felicidad o tristeza implica una duración relativamente larga, no destellos pasajeros [37]. Los cambios corporales que se generan tienen lugar durante varios segundos y se presentan por la estimulación del sistema nervioso simpático que regula la actividad corporal generando el aumento de la presión arterial, el incremento de la frecuencia cardíaca, cambios en la profundidad de los movimientos respiratorios y el incremento de la presión subglótica [43]. La intensificación emocional se puede producir si las neuronas de las amígdalas son estimuladas [37].

Con respecto a las expresiones faciales (figura 2.21), la característica típica de la felicidad es la sonrisa de alegría llamada de Duchenne, que combina la contracción de una serie de músculos de la cara que dependen de una vía nerviosa motora a cargo de la corteza motora accesoria de la corteza prefrontal [55]. La expresión implica un elevamiento de las mejillas y de las comisuras labiales y la contracción de los párpados oculares. Sin embargo, algunas investigaciones recientes indican que la sonrisa no es una señal necesaria ni suficiente de alegría [56].



Figura 2.21. Felicidad. (<http://www.artnatomia.net/>)

2.4.6.1 Felicidad en la voz

Durante el estado emocional de felicidad la velocidad del volumen glotal aumenta debido al incremento de la presión subglótica, esto genera la aparición de armónicos en la señal que condiciona a que el segundo formante se confunda con el primer formante y por lo tanto, interfiere con la frecuencia del pitch, incrementándola[39].

Las investigaciones que relacionan la alegría con los parámetros del discurso indican que la alegría se manifiesta con un incremento del tono medio y en su rango, así como un incremento en la velocidad de locución y en la intensidad del discurso [38] [39] [40] [43].

2.5 Procesos patológicos que involucran las emociones básicas: Trastornos del espectro Emotivo.

Los trastornos emotivos comienzan en una etapa temprana de la vida, usualmente antes de los 30 años. Padecimientos como la depresión, la ansiedad, el trastorno bipolar y el autismo son las manifestaciones clínicas de sutiles alteraciones en el normal desarrollo del sistema nervioso. Prestigiosos estudios han detectado que el 13% de los chicos entre 8 y 15 años tiene alguna forma de trastorno mental y menos de la mitad recibe tratamiento. Definir estas enfermedades como alteraciones en el neurodesarrollo significa que el proceso que las determina ha ocurrido mucho antes de que se manifestaran los primeros síntomas. La detección temprana se ha convertido en el objetivo primario del trabajo en salud mental. Pese a los grandes avances de las neurociencias, los diagnósticos en psiquiatría se siguen llevando a cabo a partir de conversaciones con el paciente y su familia sobre sus síntomas y su historia. En la medida en que los trastornos mentales son alteraciones cerebrales, podemos esperar que algunos indicadores biológicos o cognitivos sutiles podrían ser detectados antes de la aparición de todos los síntomas de la enfermedad. Esto permitiría

cumplir con la premisa de que cuanto más precoz es el reconocimiento de la enfermedad, mejor es el pronóstico [16].

2.5.1 Depresión

La Organización Mundial de la Salud calcula que para el año 2020, la depresión será la segunda causa de discapacidad en el mundo, superando a los accidentes tránsito, los accidentes cerebrovasculares y la enfermedad pulmonar obstructiva crónica; y prevé que para el 2030 se trasladará al primer lugar. El impacto de las enfermedades mentales en la sociedad mundial es enorme. En la actualidad, una de cada cuatro personas en el mundo sufre un problema de salud mental por año. La depresión es la principal causa de discapacidad entre personas de entre 35 y 50 años. Como patología extrema, la depresión es la principal causa de suicidios que se producen en el mundo (uno cada cuarenta segundos) [4] [16] [57].

Los episodios depresivos se inician a una edad cada vez más temprana y esto parece mostrar una tendencia uniforme a nivel mundial. Según Frederick Goodwin, director del Instituto Nacional de Salud Mental, el núcleo familiar ha experimentado una tremenda erosión, el número de divorcios se ha duplicado, los padres dedican menos tiempo a sus hijos y se ha producido un aumento de la inestabilidad laboral. Goodwin opina además, que la pérdida de una fuente sólida de identificación es la principal causa del aumento de la depresión. Ciertos estudios epidemiológicos han descubierto que la incidencia anual de la depresión en niños comprendidos entre los diez y los trece años, es del orden de un 8% o un 9%. En lo que se refiere a la adolescencia, algunos datos sugieren que este promedio podría casi duplicarse, ya que más del 16% de los adolescentes de entre catorce y dieciséis años han sufrido un brote depresivo [58].

Los brotes benignos de depresión son predictores de episodios más severos durante la vida posterior y esto indica la importancia en detectar y tratar la depresión infantil.

Es importante aclarar que todos los niños y adolescentes se entristecen alguna que otra vez, sin embargo no por eso se trata de un episodio depresivo. La depresión infantil engloba estados de melancolía mucho más graves en los que existe un abatimiento, pesadumbre, desesperación, irritabilidad y la realimentación de los mismos. Las consecuencias de este estado en los niños generan un pobre rendimiento escolar, dificultad en la memoria y en la concentración [59].

Según las neurociencias, la tristeza es una emoción básica para el ser humano e indispensable para superar situaciones de pérdida [4]. Sin embargo, nuestro cerebro puede darnos una señal de tristeza en ausencia de un evento que la justifique, que pueden desembocar en síntomas de depresión. La tristeza común, o simplemente melancolía, en sus manifestaciones extremas, puede llegar a convertirse en una depresión subclínica. Las personas con suficientes recursos internos pueden manejar por si solas este tipo de melancolía pero, por desgracia, algunas de las estrategias frecuentemente empleadas resultan perjudiciales y empeoran la situación.

La diferencia sustancial entre depresión y la tristeza normal ante una situación, viene dada por la intensidad, duración y el nivel de interferencia que producen en nuestro funcionamiento habitual [16]. Uno de los principales determinantes de la duración y la intensidad de un estado tristeza es el grado de obsesión de la persona. Preocuparse por aquello que nos deprime solo contribuye a que la depresión se agudice y se prolongue [4]. Según Nolen-Hoeksma, las mujeres son más proclives que

los hombres a obsesionarse cuando están deprimidas, lo cual explica el hecho de que la cifra de mujeres diagnosticadas de depresión duplique a la de los hombres [60].

William Styron describe a la depresión como una pesadumbre enfermiza que va acompañada de una sombría constrictión y que deja secuelas intelectuales como confusión, imposibilidad de concentrarse y pérdida de memoria. Este estado tiene sus correlatos físicos: insomnio, apatía, disminución de la capacidad de gozar de las situaciones [61].

Actualmente se reconocen como los síntomas típicos de la depresión (pueden estar presentes solo algunos síntomas): el estado de ánimo decaído, tristeza o sensación de vacío en forma persistente, pérdida de interés en las actividades habituales y en la capacidad de experimentar placer, insomnio o muchos deseos de dormir, agitación o el enlentecimiento motor, fatiga y pérdida de energía, falta o exceso de apetito, disminución del interés social y sexual, sentimientos inadecuados de culpa, inutilidad o preocupaciones económicas excesivas, pensamientos sobre la muerte, fallas de memoria y dificultades para pensar y concentrarse [16].

La depresión es una enfermedad que afecta el normal funcionamiento del cerebro. Richard Davidson realizó un experimento donde comparó una serie de voluntarios que presentaban actividad prefrontal predominantemente izquierda con otros 15 sujetos que mostraban una actividad en el lado derecho. Los resultados del estudio mostraron que las personas con actividad del lóbulo prefrontal derecho eran proclives a los estados de ánimo negativos a diferencia las personas alegres o con ánimos positivos donde la actividad cerebral se hacía más intensa en el lóbulo frontal izquierdo [50]. Davidson también descubrió que las personas que tenían un historial de depresión clínica presentaban un menor nivel de actividad cerebral en el lóbulo frontal izquierdo y una mayor activación en el lado derecho, un patrón que también se presentaba en aquellos pacientes a quienes se diagnosticaba una depresión por primera vez.

Las emociones negativas intensas absorben toda la atención del individuo, impidiendo cualquier actividad. Cuando las emociones dificultan la concentración, se dificulta el funcionamiento de la capacidad cognitiva de lo que los científicos denominan “memoria de trabajo”, la capacidad de mantener en la mente toda la información relevante para la tarea que se está llevando a cabo [16]. La memoria de trabajo es la función ejecutiva de la vida mental, la que hace posible cualquier otra actividad intelectual, por ejemplo memorizar un número, o formular una compleja preposición lógica. La región cerebral encargada de procesar la memoria de trabajo es el córtex prefrontal, la misma región que se activa durante las emociones. Es por ello, que la tensión emocional compromete el buen funcionamiento de la memoria de trabajo a través de las conexiones límbicas que convergen en el córtex prefrontal, dificultando de esta manera toda posibilidad de pensar con claridad [4].

La genética confiere predisposición para determinadas enfermedades. Sin embargo, para que éstas se manifiesten es necesario, cierta influencia del ambiente. La mayoría de las enfermedades mentales se corresponden con este tipo de combinación. Avshalom Caspi demostró en sus estudios la relación existente entre la exposición a estrés infantil y el desarrollo posterior de la depresión [16].

Determinadas enfermedades son influenciadas por los trastornos afectivos, que tienden a alterar el normal desarrollo de las mismas.

Por lo menos un 30% de los pacientes con enfermedad cardíaca padecen o van a padecer sintomatología psiquiátrica, particularmente depresión y ansiedad [16].

El desarrollo del estado depresivo aumenta la posibilidad de que la evolución de las enfermedades cardiovasculares no sea tan favorable. En un estudio realizado con personas de mediana edad que fueron sometidos a un seguimiento de doce años, quienes experimentaban una sensación de abatimiento y desesperación presentaban una tasa más elevada de mortalidad debida a enfermedades cardíacas y en el 3% de los casos que correspondían a una depresión mayor, esa tasa era cuatro veces superior. Uno de los posibles mecanismos que explicaría esta situación es que la depresión incide directamente en la variabilidad del latido cardiaco, incrementando así el riesgo de arritmias fatales [4].

La depresión aumenta el riesgo vascular, disminuye la adhesión terapéutica a la medicación, además de aumentar conductas de riesgo como sedentarismo y abuso de alcohol y tabaco. Un informe del Journal of the American Medical Association aseguraba que la depresión quintuplica la posibilidad de muerte tras haber experimentado un ataque cardiaco, además, destacaba que tanto los factores psicológicos como la depresión y el aislamiento social suponen un importante riesgo añadido para los pacientes que padecen enfermedades coronarias [62].

En la depresión mayor la activación autonómica propia de la depresión y la activación del eje hipotálamo-hipofiso-suprarrenal en el árbol vascular, que genera un alto nivel de cortisol, puede agravar el estado general provocando problemas en la memoria.

Muchas investigaciones indican que la depresión desempeña un papel relevante en otras condiciones clínicas, especialmente en lo que concierne a la fase de empeoramiento de la enfermedad. En un estudio realizado en un grupo de 100 pacientes que habían sido sometidos a un trasplante de medula ósea se evidenció que 13 de ellos padecían depresión. De estos 13, 12 fallecieron antes del primer año. De los 87 pacientes restantes sin depresión, 34 seguían con vida dos años después. Otro ejemplo que evidencia la interferencia de la depresión en la clínica, corresponde a un estudio que se realizó en pacientes con Insuficiencia Renal crónica que eran sometidos a diálisis. Los resultados mostraron que los pacientes que habían sido diagnosticados con depresión fallecieron en los dos años siguientes mientras que los que no estaban deprimidos, permanecieron con vida durante varios años más [63]. En este caso según Goleman la vía de la emoción que conecta con la condición médica no es biológica sino actitudinal, esto se refiere a que los pacientes depresivos están menos dispuestos a colaborar con el tratamiento, lo cual los expone a un riesgo todavía mayor [4].

La detección temprana de la depresión podría considerar una ventaja económica de la medicina. El tratamiento de la angustia emocional en pacientes, previene o retarda el comienzo de otro tipo de patologías, al tiempo que aumenta el proceso de recuperación, supondría un considerable ahorro de presupuesto destinado a gastos sanitarios. Un ejemplo es el estudio llevado a cabo en la Facultad de Medicina de Monte Sinaí, de la ciudad de Nueva York y en la Universidad del Nordeste a un grupo de la tercera edad que habían sufrido una fractura de cadera. El estudio mostraba que a los pacientes que recibieron terapia adicional contra la depresión se les daba de alta un promedio de dos días antes que al resto, lo cual determinó un ahorro de 97.361 dólares por cada paciente [64]. Este tipo de atención logra que el enfermo se sienta satisfecho y por otro lado que se reduzcan los costos.

Existe evidencia suficiente que indica que la detección, el tratamiento precoz y adecuado de las emociones negativas puede reducir el número de incidencia de enfermedades y mejorar la evolución de patologías cardiovasculares.

2.5.2 Trastorno bipolar

El trastorno bipolar es un tipo de trastorno emotivo que puede generar incapacidad para la interacción y el desarrollo laboral de quien la padece. Es considerada por la Organización de la Salud como la sexta causa de discapacidad en el mundo [16].

El trastorno bipolar es un conjunto de condiciones psiquiátricas a partir de las cuales se afectan los sistemas cerebrales que regulan el normal fluir de los estados de ánimo que son patológicos por su amplitud y duración o bien se realizan en un contexto inadecuado afectando su capacidad de adaptación y generando conductas inconvenientes.

El desarrollo de la patología engloba episodios maníacos o hipomaníacos. Los episodios maníacos corresponden a sentimientos de bienestar, alegría o estados de ánimo demasiado elevados y expansivos. Los episodios hipomaníacos incluyen estados depresivos, cuya duración es más prolongada que los episodios de manía.

Existe una condición genética para que el trastorno bipolar se desarrolle. Sin embargo, para que la patología se presente, debe existir una fuerte influencia del entorno ambiental

2.5.3 Trastorno del Espectro Autista

El trastorno del espectro autista es un trastorno del desarrollo caracterizado por importantes alteraciones en la interacción social y en la comunicación, que se acompañan de comportamientos e intereses restrictivos. Una de las características más sobresalientes de las personas con espectro autista es su dificultad para el contagio emocional, para mostrar empatía, y para reconocer y comprender las emociones de los demás, independientemente de la capacidad general del individuo. Las personas que padecen autismo tienen una limitada capacidad para reconocer su estado emocional y el de los demás. Sin embargo, pueden tener comportamientos que reflejen su estado emocional. Por ejemplo, si sienten malestar por alguna situación, lo expresan haciendo un berrinche.

Estos niños al no comprender emocionalmente a la sociedad tienen dificultad para establecer una interacción y pueden tener comportamientos desajustados. Fritz, por ejemplo, el primer niño descrito por Asperger, sin ser consciente de los sentimientos que subyacían a esa expresión, hacía enfadar a su profesora simplemente porque le divertía ver cómo expresaba su enfado [65].

El poder comprender las expresiones emocionales es esencial para las interacciones sociales, lo que permite explicar y anticipar las acciones de los otros. A diferencia de otros procesos mentales, las emociones frecuentemente se hacen perceptibles a través de las expresiones corporales.

Diferentes estudios de personas con autismo han reportado una atención reducida a las expresiones de emociones [66]. Observaron que niños autistas prestaban poca atención a las expresiones emocionales de otras personas en relación a un grupo control de niños con desarrollo normal en condiciones que los autores consideraban como neutrales, pero esa diferencia desaparecía si se requería realizar una decisión socialmente relevante. Ello sugiere que la atención a expresiones emocionales es influenciada por factores situacionales.

Otros estudios se han centrado no en el reconocimiento de las emociones, sino en su predicción, para observar en qué medida las personas con autismo comprenden las causas de la emoción. Un estudio evidenció que las personas con autismo, mostraban dificultad en el reconocimiento de

expresiones emocionales de sorpresa, en relación a las expresiones de felicidad o tristeza [67]. Ello parece indicar que las emociones simples se pueden comprender, pero no las emociones cognitivas.

En [68] se dio a conocer un estudio en el que se propuso evaluar si era posible enseñar a comprender estados mentales de emoción, creencia y ficción a personas con autismo. Los resultados mostraron que sí era posible enseñarles a que evalúen la comprensión de emociones. Sin embargo, encontraron que los efectos de la enseñanza no se generalizaban a otras tareas de dominios no enseñados específicamente.

Parece ser difícil dar instrucciones explícitas sobre la comprensión de expresiones emocionales en situaciones sociales que son ambiguas. Por lo que es un desafío poder facilitar experiencias sociales e interpersonales que hagan que se destaque las expresiones emocionales, promoviendo su reconocimiento y comprensión.

2.6 Regionalización de las emociones en el espacio bidimensional Arousal-Valencia

En 1896 Wundt estableció, en contraposición con la teoría de las emociones discretas de Darwin, una estructura dimensional de las emociones, donde las emociones son consideradas continuas y medibles. Wundt argumentó que existen sentimientos subjetivos del cuerpo: agradable-desagradable, excitante-no excitante y tenso-relajado que forman parte del núcleo de la vida, y a partir de estos se pueden establecer un sin número de experiencias. De acuerdo con Wundt la experiencia emocional es continua, con infinitas variaciones en un espacio dimensional, en lugar de, categorías independientes y separadas [69].

Muchos teóricos modernos se limitan a solo dos dimensiones, sugiriendo que las estructuras circulares se adaptan mejor al mapeo de emociones en el espacio bidimensional [70]. Schlosberg (1952) propone que las emociones se deberían organizar en una disposición circular, para que puedan ser representadas con solo dos dimensiones. Examinó los errores que cometían los sujetos cuando clasificaban las expresiones faciales de emociones dentro de los experimentos de categorización. A partir de estos errores, sugirió la representación circular de emociones que incluye las dimensiones de agrado-desagrado y de atención- desatención. Sin embargo, luego sugiere una tercera dimensión de tensión- somnolencia [71].

Como se ha demostrado distintas líneas de investigación del ámbito psicológico sostienen la existencia de un espacio dimensional [70] [71] [72] [73]. El modelo popularizado es el de Russell (1980) con su teoría de placer-activación. De acuerdo con esta teoría, la experiencia emocional puede describirse adecuadamente en dos dimensiones bipolares, continuas y ortogonales, una de placer-displacer y la otra de excitación-relajación, sobre las cuales se posicionarían los estados emocionales concretos [71] [75]. La premisa fundamental de esta aproximación dimensional, es que las emociones varían de forma continua a lo largo de un número limitado de dimensiones afectivas.

Izard (1992) sugiere que concebir las emociones como un espacio de varias dimensiones o bien como discretas e independientes unas de otras, no tiene por qué ser contradictorio, sino todo lo contrario. Según este autor es necesario hacer que ambos sistemas afectivos sean compatibles de manera de obtener un modelo emocional más acabado [76].

2.6.1 Modelo de Russell

El enfoque de Russel (1980) describe las respuestas emocionales del individuo ante el entorno a través de dimensiones: agrado, activación. La dimensión agrado se refiere al estado afectivo positivo/negativo, que es un sentimiento subjetivo de lo agradable/desagradable. La activación es un estado de sentimiento que varía a lo largo de una única dimensión que va desde el estado máximo de relajación hasta un estado de actividad frenética o excitación [71]. Russell indica que el espacio afectivo tiende a ser de dos dimensiones debido a las interrelaciones entre las diversas emociones, pero también señala que pueden encontrarse más de dos dimensiones como el modelo de Mehrabian-Russell [75] que incluía además una dimensión “dominio o control” que se basaba en el grado en que el individuo se siente dominado o libre para actuar.

Russell señala que el modelo circunplejo² de los sentimientos donde se definen las emociones por las dimensiones bipolares, también se pueden dividir en categorías.

La representación más básica (figura 2.22) incluye cuatro emociones que forman dimensiones independientes: la dimensión horizontal Pleasure (agrado)-Misery (pena) y la dimensión vertical Arousal (excitación)- Sleepiness (somnolencia) y otras cuatro emociones: Excitemnt (entusiasmo), Distress (Tenso), Depression (depresión), Contentment (satisfacción) que no forman dimensiones independientes, pero que permiten definir los cuadrantes del espacio. Por ejemplo, Excitemnt (entusiasmo) se define en la región del primer cuadrante, que sería una combinación de alto agrado y alta excitación. Depression (depresión) se define en el tercer cuadrante, como una combinación de desagrado con baja excitación.

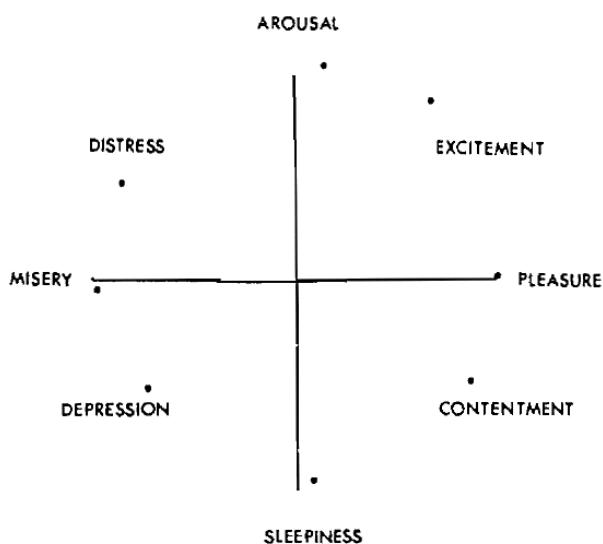


Figura 2.22. Concepto de 8 emociones en ordenamiento circular, (A Circumplex model of affect, Russell)

² Modelo circunplejo: es un modelo desarrollado por David Olson que establece una tipología para clasificar y manejar los distintos perfiles a través de tres dimensiones. Este modelo fue tomado por Russell para describir las emociones

Russell indica que a pesar de que la descripción estructural requiera de dos dimensiones para definir la emoción, el núcleo de la emoción es un solo sentimiento. Esto es análogo a la forma en que se describe un color particular. La descripción esquemática requiere de tres dimensiones (tono, saturación y brillo) pero la sensación de un color particular es una sola sensación. En ambos casos, las dimensiones se combinan de manera integral para formar la experiencia emocional [25].

Russell realiza una serie de experimentos que tienden a conceptualizar la experiencia afectiva. Obtiene la conceptualización de 28 emociones mediante experimentos diferentes. Los experimentos fueron realizados en 36 sujetos de ambos sexos de la Universidad Británica de Columbia. Se utilizaron 28 palabras que representaban diferentes estados emocionales. Estas palabras de estímulos representaban palabras o frases que la gente usa para describir sus estados de ánimo, sentimientos o emociones [71].

Experimento 1: Utiliza la técnica de Ross del ordenamiento circular de variables [76]. (Representación básica).

En la primera experimentación, cada sujeto tenía que realizar una tarea de clasificación y categorización. Cada individuo fue instruido para colocar cada palabra, que fue presentada en orden alfabético, en una de las ocho categorías de la representación básica ya nombrada: Arousal (excitación), Excitement (entusiasmo), Depression (depresión), Distress (Tenso), Contentment (satisfacción), Pleasure (agrado), Misery (pena), Sleepiness (sомнolencia). La segunda tarea consistía en colocar las ocho categorías en orden circular. La instrucción fue *“Su tarea es colocar las palabras sobre un círculo de tal manera que las palabras que describan sentimientos opuestos entre si sean colocadas en posiciones opuestas del círculo, y las palabras que describan sentimientos similares sean colocadas en posiciones cercanas en el círculo.”*

Cada palabra que describe una emoción puede ser considerada como una etiqueta de un conjunto fuzzy (difuso), definido como un grupo o clase con bordes difusos en el que existe una gradual transición entre un miembro y un no miembro. Cada estado emocional podría asociarse con un valor entre 0 a 1 especificando su grado de pertenencia a cada grupo que tiene una etiqueta emocional determinada como Excitement (entusiasmo) o Misery (pena). Por ejemplo el Pleasure (agrado), y Excitement (entusiasmo), están cerca en el ordenamiento circular porque sus bordes fuzzy están superpuestos.

Luego, usando el procedimiento de Ross [77] se calcularon las coordenadas polares para colocar las 28 palabras en el plano. A las 8 categorías de emociones se les asignó una coordenada escalar basada en la teoría del ordenamiento circular (fig. 2.23): Pleasure (agrado) = 0° , Excitement (entusiasmo) = 45° , Arousal (excitación) = 90° , Distress (Tenso) = 135° , Misery (pena) = 180° , Depression (depresión) = 225° , Sleepiness (sомнolencia) = 270° , Contentment (satisfacción) = 315° . Este procedimiento también provee un valor “P” para cada palabra. “P” es una medida de la precisión del ángulo utilizado para cada emoción. “P” puede variar desde cero (lo cual indica que los individuos colocaron las palabras al azar) a uno (lo cual significa que todos los sujetos colocaron correctamente las palabras que correspondían a una determinada categoría).

En el experimento los valores obtenidos de “P” fueron de 0.71 a 0.97, lo cual mostró un alto grado de precisión.

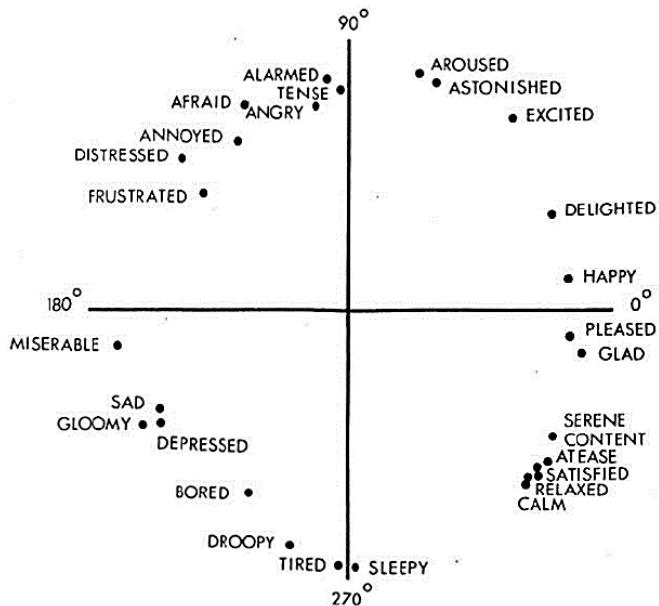


Figura 2.23. Escalamiento de ordenamiento circular para 28 palabras de emoción. (A Circumplex model of affect, Russell)

Experimento 2. Utiliza la técnica de escalamiento multidimensional de manera de obtener coordenadas escalares en un modelo de medición si un orden circular.

El escalamiento multidimensional se basa en la percepción de similitud alrededor de las palabras de emoción.

En este experimento, a cada sujeto se le entregó un conjunto de 28 palabras en tarjetas separadas que indicaban una emoción y se le pidió que ordenara las tarjetas en 4, 7, 10 y 13 grupos y lo debía hacer en ensayo sucesivos. La instrucción era agrupar estados emocionales similares.

La similitud de cada par de palabras por sujeto fue evaluada por el número de ensayos en que ese par fue ubicado en el mismo grupo, con una calificación ponderada por el número de alternativas disponibles. Por ejemplo, una calificación de 13 era dada para el par de palabras colocadas en el mismo grupo cuando el ensayo correspondía a ubicar las palabras en 13 grupos. La calificación de 1 correspondía a colocar las palabras en el mismo grupo en un ordenamiento de 1 grupo. La mínima similitud era 1 y la máxima era 36 que habría ocurrido si los sujetos hubiesen colocado el par de emociones en el mismo grupo en todos los ensayos.

Posteriormente, usando procedimiento de Ross [77] se calcularon las coordenadas polares para la ubicación de los resultados (figura 2.24).

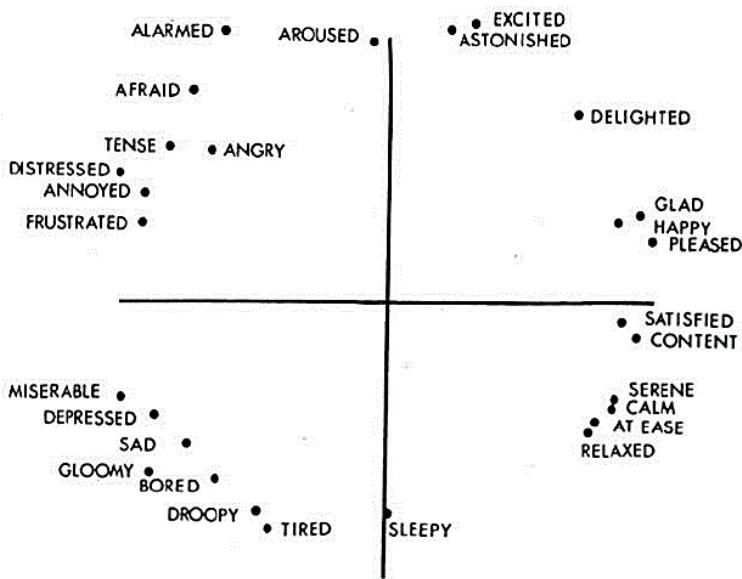


Figura 2.24. Escalamiento multidimensional para las 28 palabras de emoción. (A Circumplex model of affect, Russell)

Experimento 3. Utiliza el escalamiento Unidimensional a lo largo de las dimensiones Pleasure (agrado)-Misery (pena) y grados de excitación.

En este experimento los 28 términos de emociones son definidos en el espacio bipolar de dos dimensiones ya mencionado. Cada una de estas palabras fue ubicada sobre la escala de valencia desde agrado hasta pena y sobre la escala que representa el grado de activación.

Se determinó la regionalización de las palabras de emoción para este experimento, de acuerdo a la frecuencia con la que cada palabra fue clasificada, al criterio fuzzy, ya descripto, y a la técnica de Ros para encontrar las posiciones angulares (figura 2.25).

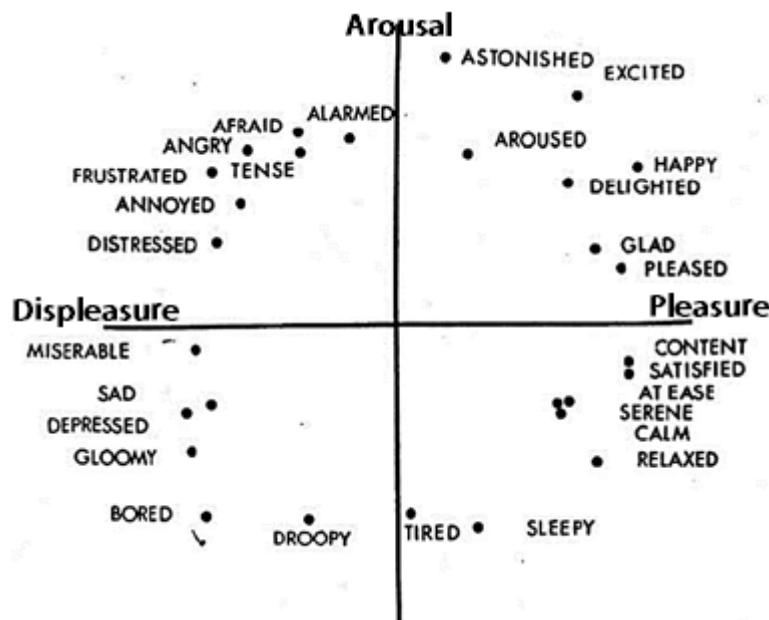


Figura 2.25. Escalamiento unidimensional para las 28 palabras de emoción. (A Circumplex model of affect, Russell)

Posteriormente fue comparado el grado de convergencia de los resultados a través de los tres métodos. Los tres ensayos arrojaron resultados equivalentes. Una evaluación cuantitativa de esa equivalencia fue realizada calculando la redundancia promedio, que es una forma de expresar la variancia, mediante un análisis de correlación entre los diferentes escalamientos.

Las posiciones angulares de los 28 términos emocionales ubicados mediante el procedimiento de Ross en el plano polar y que corresponden a los tres métodos de Russell expuestos se presentan en la siguiente tabla.

Estado emocional	Posiciones angulares		
	Primer ensayo	Segundo Ensayo	Tercer ensayo
<i>Happy (Feliz)</i>	9	20	34
<i>Delighted (Encantado)</i>	25	45	40
<i>Excited (Entusiasmado)</i>	49	73	52
<i>Astonished(Asombrado)</i>	68	77	80
<i>Aroused (Excitado)</i>	72	94	68
<i>Alarmed (Alarmado)</i>	95	115	121
<i>Tenso (Nervioso)</i>	92	120	144
<i>Angry (Enojado)</i>	98	130	139
<i>Afraid (Asustado)</i>	117	117	133
<i>Annoyed (Irritado)</i>	122	142	156
<i>Distressed (Angustiado)</i>	138	154	155
<i>Frustrated (Frustrado)</i>	140	140	162
<i>Miserable (Desdichado)</i>	189	186	200
<i>Sad (Triste)</i>	207	203	216
<i>Gloomy (Melancólico)</i>	209	216	211
<i>Depressed (Depresivo)</i>	210	206	201
<i>Bored (Aburrido)</i>	242	226	224
<i>Droopy (Decaído)</i>	256	239	244
<i>Tired (Cansado)</i>	268	243	274
<i>Sleepy (Somnoliento)</i>	273	270	292
<i>Calm (Calmado)</i>	316	335	332
<i>Relaxed (Relajado)</i>	318	327	326
<i>Satisfied (Satisfecho)</i>	320	356	347
<i>Serene (Sereno)</i>	321	331	335
<i>Content (Satisfecho)</i>	324	352	350
<i>At ease (Estar a gusto)</i>	330	338	333
<i>Glad (Contento)</i>	350	20	20
<i>Pleased (Agradable)</i>	354	14	13

Tabla 2.1. Posiciones angulares en el plano Arousal-Valencia, obtenidos de los tres ensayos de Russell. (A Circumplex model of affect, Russell)

Como puede observarse en la tabla anterior, las posiciones angulares que denotan la regionalización de cada emoción en los tres ensayos de Russel son similares para la mayoría de las emociones. Sin embargo, dado que los últimos dos ensayos son realizados por métodos similares y con mayor exactitud estadística, presentan resultados con un grado de semejanza superior. En efecto, sería adecuado el uso de los datos provistos por cualquiera de los dos últimos ensayos para la implementación de sistemas de regionalización de emociones.

2.7 Referencias

1. Trueba Atienza C. "La teoría Aristotélica de las emociones". *Signos Filosóficos*, vol. XI, núm. 22, julio-diciembre, 2009
2. Descartes, R. (2005). *Las pasiones del alma* (Vol. 290). Edaf.

3. Kant, I., & Gaos, J. (2004). *Antropología en sentido pragmático*. Alianza Editorial.
4. Goleman, D. (2012). *Inteligencia emocional*. Editorial Kairós.
5. Ekman, P. (2009). *Darwin's contributions to our understanding of emotional expressions*. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364(1535), 3449-3451.
6. Bellés, X. (2009). *Les emocions classificades: Darwin i Le Brun*. Mètode: Revista de difusió de la investigació de la Universitat de València, (60), 80-85.
7. Cottegnies, L. (2002). *Codifying the Passions in the Classical Age-a Few Reflections on Charles Le Brun's Scheme and its Influence in France and in England*. *Études Epistème*, 1, 141-158.
8. Darwin, C. (2002). *The expression of the emotions in man and animals*. Oxford University Press.
9. Levav, M. (2005). *Neuropsicología de la emoción. Particularidades en la infancia*. Revista Argentina de Neuropsicología, 5, 15-24.
10. James, W. (1884). II.—What is an emotion?. *Mind*, (34), 188-205.
11. Cannon, W. B. (1928). *Neural organization for emotional expression*.
12. Bard, P. (1928). *A diencephalic mechanism for the expression of rage with special reference to the sympathetic nervous system*. *American Journal of Physiology*.
13. Papez, J. W. (1937). *A proposed mechanism of emotion*. *Archives of Neurology & Psychiatry*, 38(4), 725-743.
14. Purves, D., Augustine, G. J., Fitzpatrick, D., Katz, L. C., LaMantia, A. S., McNamara, J. O., & Williams, S. M. (2001). *The Importance of the Amygdala*.
15. Ekman, P., & Friesen, W. V. (1971). *Constants across cultures in the face and emotion*. *Journal of personality and social psychology*, 17(2), 124.
16. Manes, F.; Niro, M. (2015). *Usar el cerebro: Conocer nuestra mente para vivir mejor*. Editorial Paidos Iberica. ISBN: 9788449330858
17. LeDoux, J. E. (1996). *The emotional brain: The mysterious underpinnings of emotional life*. New York, NY: Simon & Schuster.
18. Armony JL, LeDoux JE. 1999. How danger is encoded: towards a systems, cellular, and computational understanding of cognitive emotional interactions in fear circuits. See Gazzaniga 1999
19. LeDoux JE. (2000). *Emotion Circuits in the brain*. *Annu. Rev. Neurosci.* 2000. 23:155–184
20. LeDoux, J. (2003). *The emotional brain, fear, and the amygdala*. *Cellular and molecular neurobiology*, 23(4-5), 727-738.
21. LeDoux, J. E. (1999). *El cerebro emocional*. Editorial Ariel. ISBN: 84-08-02906-1
22. Ekman, P. & Friesen, W. (1978). *Facial Action Coding System: A Technique for the Measurement of Facial Movement*. Palo Alto, CA: Consulting Psychologists.
23. P N, J. L., & Oatley, K. (2000). *Cognitive and social construction in emotions*. *Handbook of emotions*, 458.
24. La emoción. (2001). *Psicología Básica*. Departamento de Psicología de la salud. Universidad de Alicante. <http://rua.ua.es/dspace/bitstream/10045/4298/26/TEMA%208.La%20emoci%C3%B3n.pdf>
25. Barret L.;Russell J. (2015).*The psychological construction of emotion*. Edit. The Guilford Press. ISBN 978-1-4625-1697-1
26. Shaver, P., Schwartz, J., Kirson, D., & O'connor, C. (1987). *Emotion knowledge: further exploration of a prototype approach*. *Journal of personality and social psychology*, 52(6), 1061.
27. Lang, P. J., Bradley, M. M., & Cuthbert, B. N. (1999). *International affective picture system (IAPS): Technical manual and affective ratings*.
28. Damasio AR (1994) *Descartes' error: emotion, reason, and the human brain*. New York: Grosset/Putnam.

29. Bechara, A., & Damasio, A. R. (2005). *The somatic marker hypothesis: A neural theory of economic decision*. *Games and economic behavior*, 52(2), 336-372.
30. Norman, D. A. (2004). *Emotional design: Why we love (or hate) everyday things*. Basic books.
31. Mesa, J. F. C., & Muñoz, D. I. (2013). *Vías de la emoción y la inhibición del Neocórtex (Ways of emotion and the inhibition neocortex)*. *Revista CES Movimiento y Salud*, 1(1), 52-60.
32. Crossman, A. R., & Neary, D. (2002). *Neuroanatomía: texto y atlas en color*. Masson.
33. SIMS KS, WILLIAMS RS. *The human amygdaloid complex: a cytologic and histochemical atlas using Nissl, myelin, acetylcholinesterase and nicotinamide adenine dinucleotide phosphate diaphorase staining*. *Neuroscience* 1990; 36: 449-472.
34. Nieuwenhuys, R. (2009). *El sistema nervioso central humano*. Ed. Médica Panamericana.).
35. Borod, J.C. (1992). *Interhemispheric and intrahemispheric control of emotion: A focus on unilateral brain damage*. *Journal of Consulting and Clinical Psychology*, 60, 339-348.
36. Damasio, A.R. (1998). *Emotion in the perspective of an integrated nervous system*. *Brain Research Reviews*, 26, 83-86.
37. Damasio, A. (2005). *En busca de Spinoza. Neurobiología de la emoción y los sentimientos*. Barcelona: Crítica.
38. Bänziger, T., & Scherer, K. R. (2005). *The role of intonation in emotional expressions*. *Speech communication*, 46(3), 252-267.
39. Ververidis, D., & Kotropoulos, C. (2006). *Emotional speech recognition: Resources, features, and methods*. *Speech communication*, 48(9), 1162-1181.
40. de Diego, I. M., Serrano, Á., Conde, C., & Cabello, E. (2006). *Técnicas de reconocimiento automático de emociones*. *Teoría de la Educación: Educación y Cultura en la Sociedad de la Información*, 7(2), 7.
41. Luengo, I., Navas, E., Hernández, I., & Sánchez, J. (2005). *Reconocimiento automático de emociones utilizando parámetros prosódicos*. *Procesamiento del lenguaje natural*, 35, 13-20.
42. Resa, C. O. (2009). *Detección de emociones en voz espontánea*.
43. El Ayadi, M., Kamel, M. S., & Karray, F. (2011). *Survey on speech emotion recognition: Features, classification schemes, and databases*. *Pattern Recognition*, 44(3), 572-587.
44. Martin, O., Kotsia, I., Macq, B., & Pitas, I. (2006, April). *The eINTERFACE'05 audio-visual emotion database*. In *Data Engineering Workshops, 2006. Proceedings. 22nd International Conference on* (pp. 8-8). IEEE.
45. Levenson, R. W., Ekman, P., & Friesen, W. V. (1990). *Voluntary facial action generates emotion-specific autonomic nervous system activity*. *Psychophysiology*, 27(4), 363-384.
46. Teager, H. M., & Teager, S. M. (1990). *Evidence for nonlinear sound production mechanisms in the vocal tract*. In *Speech production and speech modelling* (pp. 241-261). Springer Netherlands.
47. Heuft, B., & Porteles, T. (1996, October). *Synthesizing prosody: A prominence-based approach*. In *Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on* (Vol. 3, pp. 1361-1364). IEEE.
48. Rozin, P.; Haidt, J. & McCauley, C.R. (2008). *Disgust*. In M. Lewis, J.M. Haviland-Jones & L.F. Barret (Eds), *Handbook of emotions*, 3rd ed. (pp. 757-776). New York: Guilford Press
49. Husted, D. S., Shapira, N. A., & Goodman, W. K. (2006). *The neurocircuitry of obsessive-compulsive disorder and disgust*. *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, 30(3), 389-399.
50. Davidson, R. J. (1992). *Anterior cerebral asymmetry and the nature of emotion*. *Brain and cognition*, 20(1), 125-151.
51. Angyal, A. (1941). *Disgust and related aversions*. *The Journal of Abnormal and Social Psychology*, 36(3), 393.
52. Russell J.A. (2003). *Pleasure*. Edition published in the Taylor & Francis e-Library. (2005). ISBN 0-203-01061-2

53. Ekman, P., & Friesen, W. V. (2003). *Unmasking the face: A guide to recognizing emotions from facial clues*. Ishk.
54. Breiter, H. C., Etcoff, N. L., Whalen, P. J., Kennedy, W. A., Rauch, S. L., Buckner, R. L. & Rosen, B. R. (1996). *Response and habituation of the human amygdala during visual processing of facial expression*. *Neuron*, 17(5), 875-887.
55. Martínez, C. B. (2007). *Emociones y cerebro*. Revista de la Real Academia de Ciencias Exactas, Físicas y Naturales, 101(1), 59-68.
56. Russell, J. A., & Fernández-Dols, J. M. (Eds.). (1997). *The psychology of facial expression*. Cambridge university press.
57. La depresión (2012). OMS (Organización Mundial de la Salud). Nota descriptiva N°369 <http://www.who.int/mediacentre/factsheets/fs369/es/>
58. Peter Lewinsohn el al.,(1993) "Adolescent Psychopathology: Prevalence and Incidence of Depression in High School Students", en *Journal of Abnormal Psychology*, 102
59. Kovacs, M., & Goldston, D. (1991). *Cognitive and social cognitive development of depressed children and adolescents*. *Journal of the American Academy of Child & Adolescent Psychiatry*, 30(3), 388-392.
60. Nolen-Hoeksem, S. (1993). *Sex differences in control of depression*. *Handbook of mental control*. Century psychology series., (pp. 306-324).
61. Styron, W. (2010). *Darkness visible: A memoir of madness*. Open Road Media.
62. Frasure-Smith, N., Lespérance, F., & Talajic, M. (1993). *Depression following myocardial infarction: impact on 6-month survival*. *Jama*, 270(15), 1819-1825.
63. Burton, H. J., Kline, S. A., Lindsay, R. M., & Heidenheim, A. P. (1986). *The relationship of depression to survival in chronic renal failure*. *Psychosomatic Medicine*, 48(3), 261-269.
64. Snyder, S., Strauss, E., Burton, R., Nuber, G., Abernathy, T., MA, H. S. & Sacks, C. (1991). *Cost offset from a psychiatric consultation-liaison intervention with elderly hip fracture patients*. *Am J Psychiatry*, 148(8).
65. Asperger, H. (1944). "Psicopatía autista" en la infancia. Traducido y anotado por U. Frith. En U. Frith (Ed.) *El autismo y el síndrome de Asperger* (pp. 37-92). Nueva York: Cambridge University Press.
66. Begeer S., Rieffe C., Terwogt M.M., Stockmann L. (2006). *Attention to facial emotion expressions in children with autism*. *Autism*, 10 (1), 37-51.
67. Baron-Cohen S., Spitz A., Cross P. (1993). *Do children with autism recognize surprise? A research note*. *Cognition & Emotion*, 7(6), 507-516.
68. Hadwin J., Baron-Cohen S., Howlin P., Hill K. (1997). *Does teaching theory of mind have an effect on the ability to develop conversation in children with autism* . *Journal of autism and Developmental Disorders*, 27(5), 519-537.
69. Moors, A., Ellsworth, P. C., Scherer, K. R., & Frijda, N. H. (2013). *Appraisal theories of emotion: State of the art and future development*. *Emotion Review*, 5(2), 119-124.
70. Scherer, K. R. (2005). *What are emotions? And how can they be measured?*. *Social science information*, 44(4), 695-729.
71. Russell, J. A. (1980). *A circumplex model of affect*. *Journal of personality and social psychology*, 39(6), 1161.
72. Bradley, M. M., & Lang, P. J. (1994). *Measuring emotion: the self-assessment manikin and the semantic differential*. *Journal of behavior therapy and experimental psychiatry*, 25(1), 49-59.
73. Larsen, R. J., & Diener, E. (1992). *Promises and problems with the circumplex model of emotion*.
74. Tellegen, A., Watson, D., & Clark, L. A. (1999). *On the dimensional and hierarchical structure of affect*. *Psychological Science*, 10(4), 297-303.
75. Mehrabian, A., & Russell, J. A. (1974). *An approach to environmental psychology*. The MIT Press.

76. Izard, C. E. (1992). *Basic emotions, relations among emotions, and emotion-cognition relations.*
77. Ross, R. T. (1938). A statistic for circular series. *Journal of Educational Psychology*, 29(5), 384.

CAPITULO 3

Fundamentos teóricos: Anatomía y Fisiología de la voz

3.1 Introducción

En este capítulo se detallarán las características anatómicas y los mecanismos fisiológicos del aparato fonador que permiten generar la voz. Se describirán parámetros tales como formantes y frecuencia fundamental, que representan las señales de voz desde el punto de vista biológico y que permiten la diferenciación de los sonidos.

3.2 Modelo del Aparato Fonador

El modelo del aparato del habla se divide en órganos de la fonación (que generan la producción de voz) y la articulación (que modulan la voz) (figura 3.1).

Los órganos de fonación (pulmones y laringe) hacen de fuente de energía acústica y modulan la vibración de las cuerdas vocales. Ajustan el tono, el volumen y generan los patrones prosódicos del habla.

Los órganos articulatorios dan resonancia y generan sonidos adicionales. Se componen de la mandíbula, lengua, labios y el velo del paladar. Los músculos constrictores de la faringe y laringe también participan en la articulación, así como en la calidad de voz.

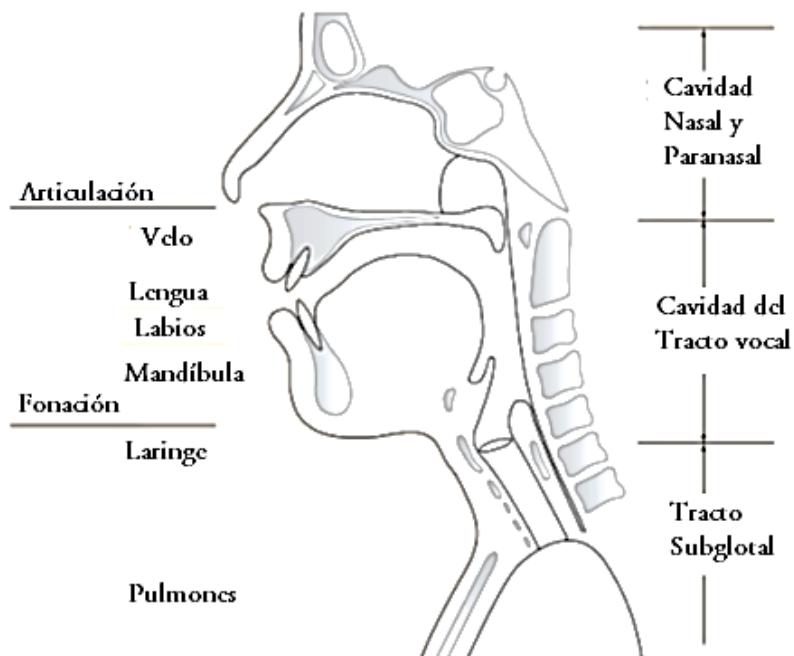


Figura 3.1 Modelo del aparato Fonador (Extraída de Physiological Process if Speech Production, Honda K.)

Los pulmones suministran la columna de aire, ésta se desplaza por la tráquea y es modulada en las cuerdas vocales que vibran haciendo de oscilador. Esta oscilación de los pliegues vocales convierte el aire espirado del flujo de aire en impulsos intermitentes que resultan en un sonido. La presión de aire resulta de las funciones del sistema respiratorio. Las vibraciones de las cuerdas vocales producen la voz en su tono fundamental y sus armónicos. Luego sufre una modificación en la caja de resonancia naso-buco-faríngea en la que se amplifica y se transforma la turbulencia sonora en los sonidos con las funciones lingüísticas conocidas. Finalmente los órganos articuladores generan movimientos para alterar las características de resonancia de la vía aérea supra-laríngea.

3.2.1 Órganos de la fonación

La generación de la voz requiere una configuración adecuada del flujo de aire que proviene de los pulmones y de las características anatómicas de las cuerdas vocales esenciales para la oscilación.

3.2.1.1 Sistema respiratorio

El sistema respiratorio (figura 3.2) se divide en dos segmentos:

- Conducción de las vías respiratorias para la ventilación entre atmósfera y pulmones.
- Intercambio gaseoso a través del tejido respiratorio especializado en los pulmones.

La ventilación (inhalación y espiración) se lleva a cabo por movimientos del tórax, diafragma y abdomen. Estos movimientos implican acciones de los músculos respiratorios y la fuerza de retroceso elástico del sistema.

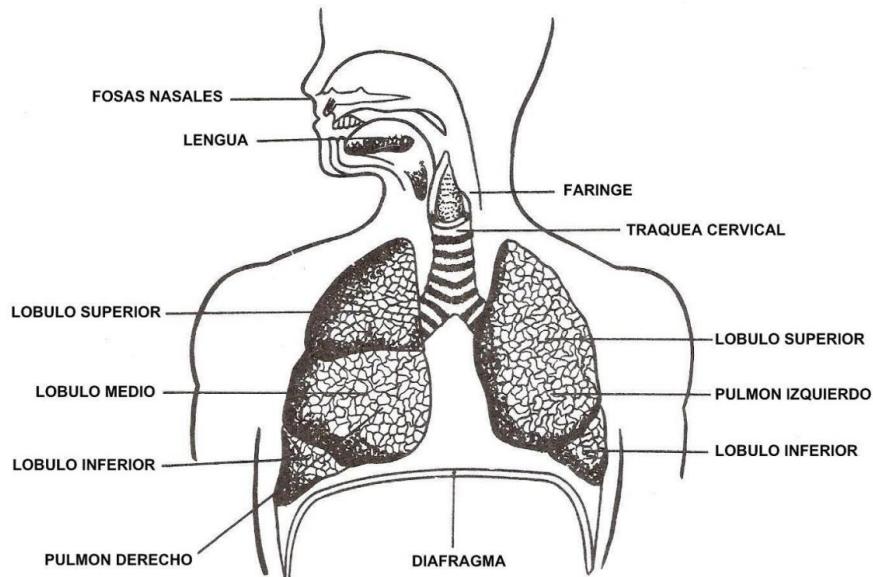


Figura 3.2. Sistema Respiratorio (Obtenido de
http://www.aparatossistema.com/images/sistema_respiratorio.jpg)

Durante la respiración tranquila, los pulmones se expanden para inhalar aire por acción de los músculos inspiratorios (diafragma, músculos intercostales externos) y expulsan el aire por la fuerza de retroceso elástico del tejido pulmonar, diafragma y abdomen.

En la respiración profunda los músculos inspiratorios y espiratorios trabajan alternativamente, haciendo que el tórax se expanda y contraiga.

Durante la producción de habla, el patrón respiratorio cambia a una fase espiratoria más larga y una fase inspiratoria más corta. El tórax se expande por la inspiración antes de iniciar el discurso y luego se comprime por la fuerza de retroceso elástico de los tejidos del sistema respiratorio, llegando así al nivel de la capacidad residual funcional (FRC).

La presión pulmonar durante el habla se mantiene constante a excepción del incremento ocurrido al inicio del discurso. Esta presión pulmonar constante es debida a las acciones de los músculos inspiratorios que previenen el flujo de aire excesivo y mantienen la fase espiratoria larga (fig. 3.3).

A medida que el habla continúa, el volumen pulmonar disminuye gradualmente por debajo de la Capacidad Residual Funcional y la presión pulmonar se mantiene por la acción de los músculos espiratorios que expulsan activamente el aire desde los pulmones (Figura 3.3).

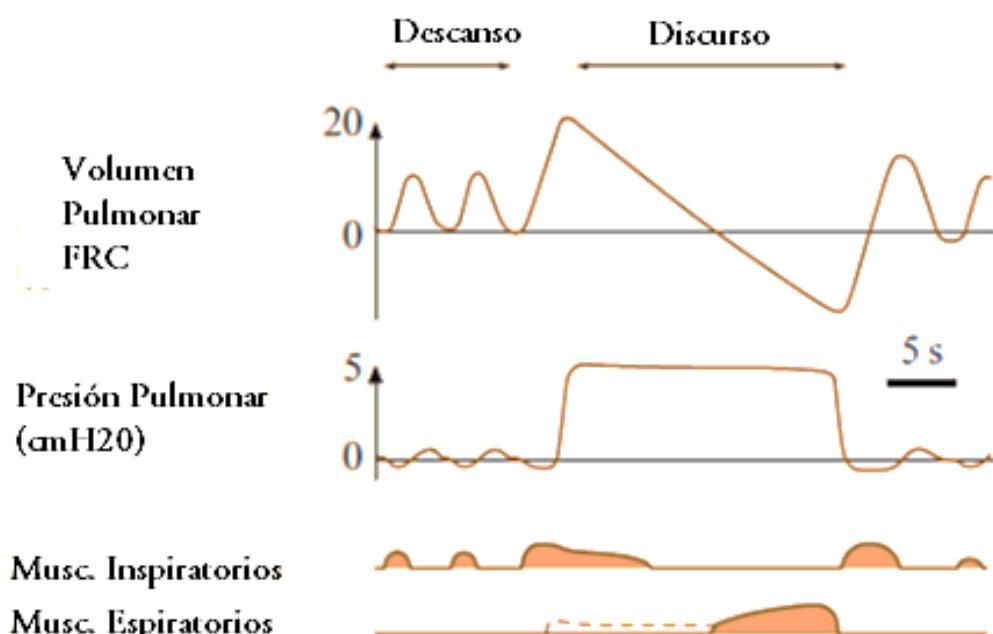


Figura 3.3. Patrón Respiratorio durante el habla. (Extraída de Physiological Process if Speech Production, Honda K.)

Algunos estudios han sugerido que la regulación de la espiración durante el habla no solo es debida al sistema torácico sino también del sistema abdominal.

La visión contemporánea de la respiración hace hincapié en que la espiración de aire durante el habla no es un proceso pasivo sino que está controlado por una co-activación de los músculos espiratorios e inspiratorios [1].

3.2.1.2 Laringe

La laringe es el órgano principal de la producción de la voz (figura 3.4) Se encuentra situada en la porción anterior del cuello. Se relaciona con los cuerpos vertebrales C3-C6. Su función principal es evitar que entre en los pulmones material extraño.

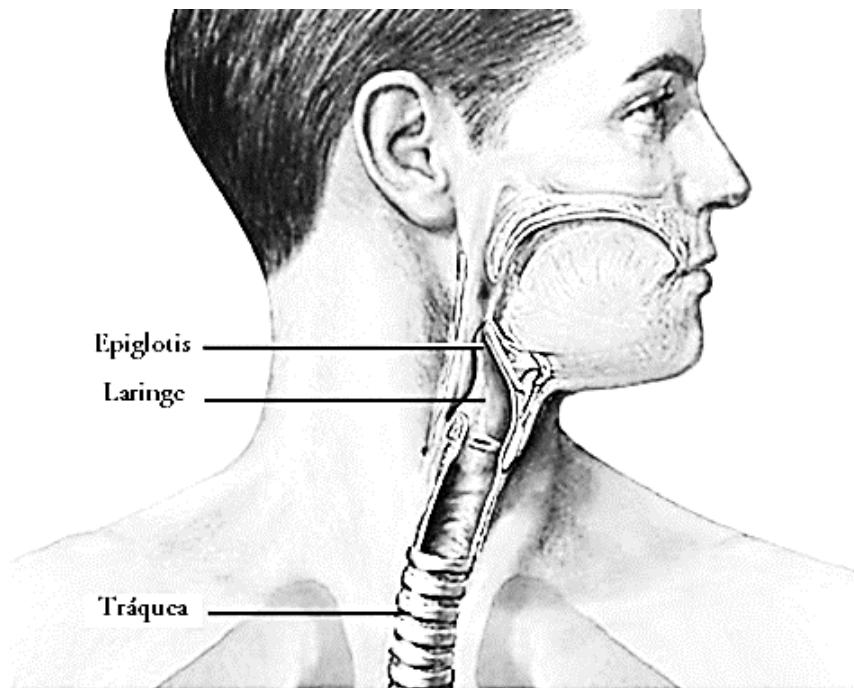


Figura 3.4 .Órgano principal para la producción de la voz: Laringe. (Extraída de Atlas de Anatomía Humana, Adam.)

Su estructura está constituida por piezas cartilaginosas que se articulan entre si y se unen por ligamentos y grupo de estructuras musculares. Contiene varias estructuras rígidas como el cricoides, tiroides, aritenoides, epiglotis y otros cartílagos más pequeños (figura 3.5)

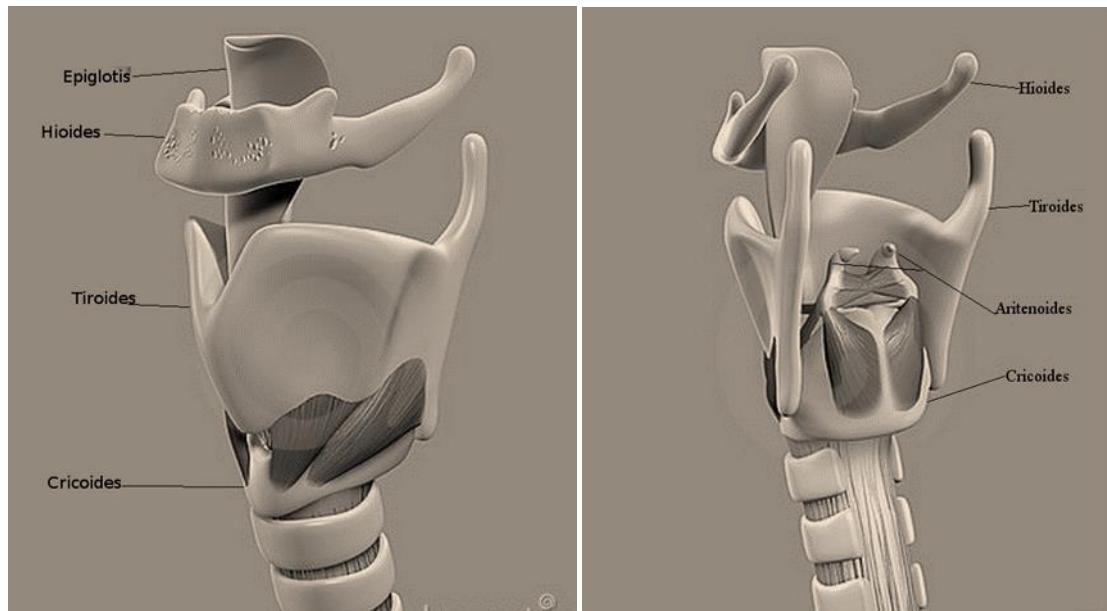


Figura 3.5. Laringe.

(Extraída de <http://es.dreamstime.com/imagen-de-archivo-libre-de-regal%C3%ADas-anatom%C3%A1da-de-la-laringe-image1663326>)

El cartílago cricoides es el primer anillo traqueal modificado para soportar la laringe propiamente dicha (figura 3.6). Ofrece una articulación a ambos lados con el cartílago tiroides: articulación cricotiroidea, y con los aritenoides, conformando las articulaciones cricoraritenoideas.

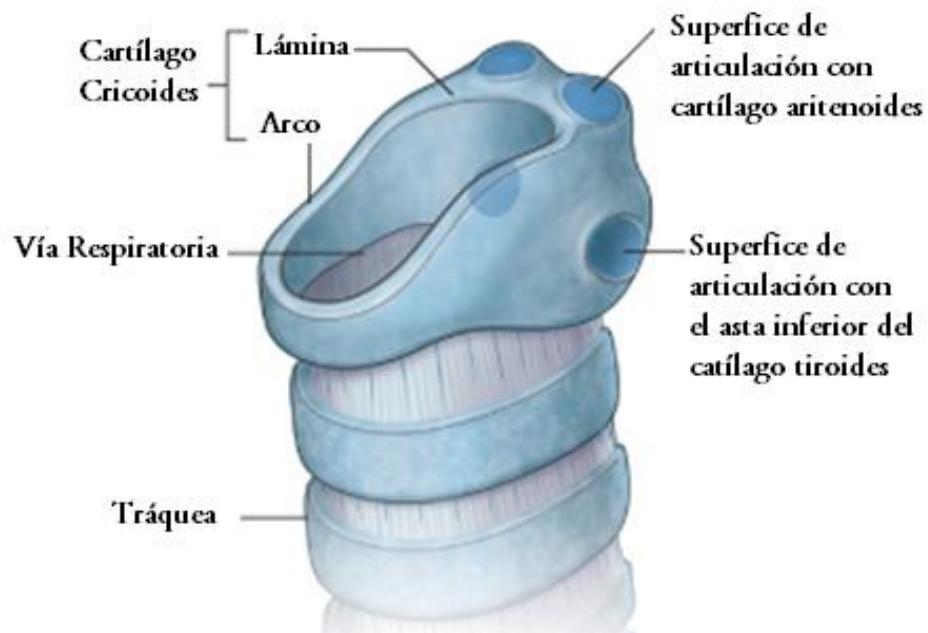


Figura 3.6 .Cartílago Cricoides. (Extraída de Elsevier. Drake et al. Gray's Anatomy for Students)

El cartílago tiroides es un cartílago hialino que limita la laringe anterior y lateralmente (figura 3.7). Consiste en dos láminas que se fusionan anteriormente en la línea media. Sobre el punto de fusión se encuentra la escotadura tiroidea. Desde el borde posterior de cada lámina se proyectan dos astas, uno superior y otro inferior. El asta superior recibe la inserción del ligamento tirohioideo lateral. El asta inferior se articula en su cara interna con el cartílago cricoides.

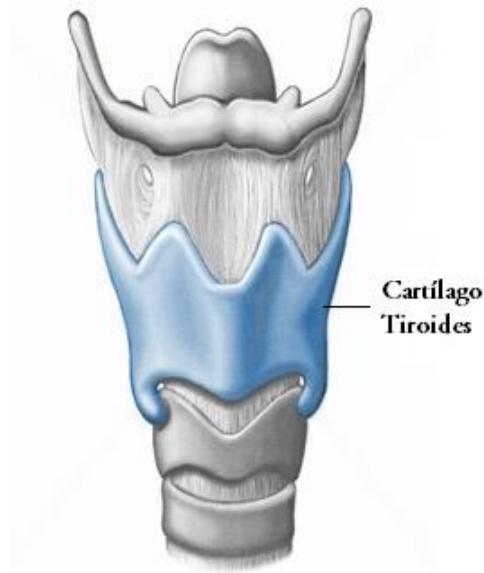


Figura 3.7. Cartílago Tiroides
(Extraída de Elsevier. Drake et al. Gray's Anatomy for Students)

Los cartílagos aritenoides son cartílagos bilaterales tetraédricos que cambian de ubicación y orientación entre la fonación y la respiración (figura 3.8).

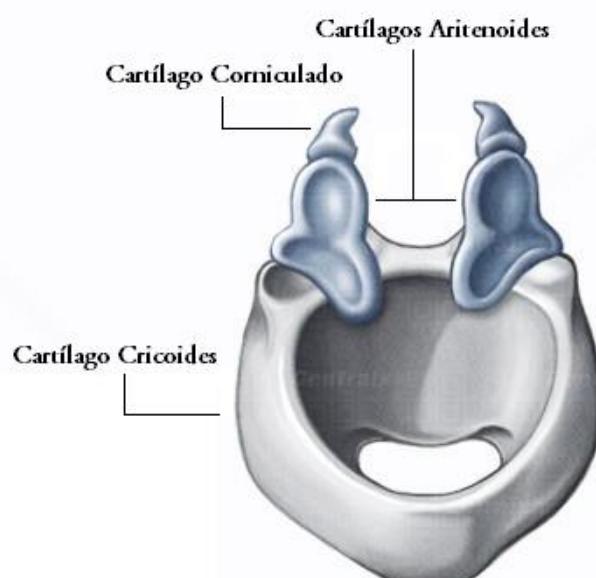


Figura 3.8. Cartílagos Aritenoides. (Obtenida del Atlas del Cuerpo Humano, Centralx Atlas)

Los pliegues vocales, llamados cuerdas vocales, corren en sentido anteroposterior desde los procesos vocales de los cartílagos aritenoides a la superficie interna del cartílago tiroideo (fig. 3.9). El tejido de las cuerdas vocales consiste en: músculo tiroaritenideo, ligamento vocálico, lámina propia y las membranas mucosas. Esta estructura produce esfuerzos aerodinámicos para oscilar. Existen dos repliegues superiores, que son las cuerdas falsas o bandas ventriculares y dos repliegues inferiores que son las cuerdas vocales verdaderas.

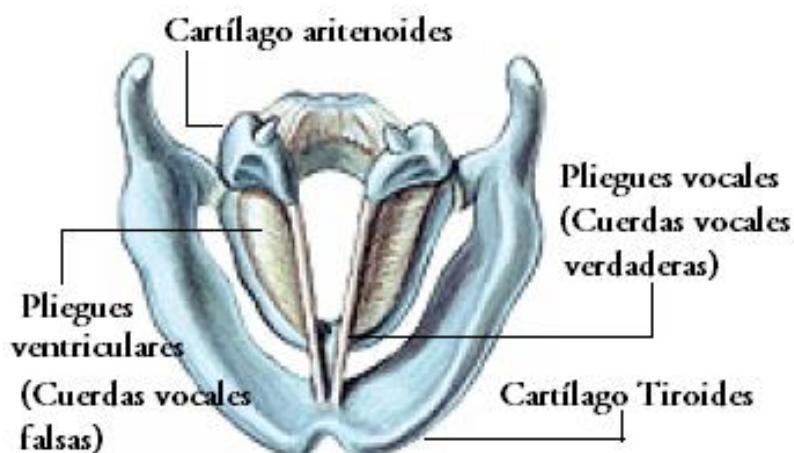


Figura 3.9. Pliegues Vocales. (Obtenido de Atlas Anatomía Humana, Sobotta)

Las cuerdas vocales verdaderas son las que producen las siguientes características del sonido:

- Si dichas cuerdas se aproximan y vibran se origina un “sonido sonoro”, pero si no vibran será un “sonido sordo”.
- La vibración provoca una onda sonora o tono fundamental, y armónicos que filtrados en las cavidades resonantes producen el timbre del sonido.
- Al pasar el aire por las cuerdas vocales con mayor o menor energía, se produce la intensidad de la voz.
- La duración se produce por un impulso psicomotriz a través del nervio recurrente hacia el diafragma. Este comprime los pulmones el tiempo necesario para la duración deseada.

Durante el habla natural las cuerdas vocales realizan un movimiento de oscilación cuasiperiódico. Esto es observado en la forma de onda de la señal del habla y se puede medir mediante el jitter (perturbación de la frecuencia) y el shimmer (perturbación de la amplitud) [2]. Estas irregularidades parecen surgir de factores:

- biomecánicos (por la asimetría de las cuerdas vocales).
- neurogénicos (por las actividades involuntarias de los músculos de la laringe).
- aerodinámicos (por las fluctuaciones del flujo de aire y de la presión subglótica).

En la fonación sostenida de una voz normal, el Jitter es de aproximadamente 1% de la frecuencia y el Shimmer es aproximadamente de 6% de la amplitud.

El espacio entre los bordes libres de las cuerdas vocales se denomina glotis. Este espacio se divide en dos porciones: la porción membranosa, adelante (esencial para la vibración), y la porción cartilaginosa, hacia atrás (esencial para la respiración).

La glotis cambia su forma durante el discurso: se estrecha por la aducción de las cuerdas vocales y se ensancha por la abducción. Este movimiento se lleva a cabo por la acción de los músculos laríngeos intrínsecos que se insertan en los cartílagos aritenoides.

Durante el proceso de emisión de la voz, los pliegues vocales comienzan a vibrar por el aire que circula a una cierta presión, a través de la porción membranosa de la glotis estrecha.

El flujo de aire glotal generado induce al movimiento ondulatorio de la membrana de las cuerdas vocales, que se propaga en forma sinuosa desde la parte inferior a la superior de los bordes de las cuerdas vocales. Cuando este movimiento oscilatorio se acumula, las membranas de las cuerdas vocales entran en contacto, lo que resulta en el cierre y apertura de la glotis. Dentro de un ciclo de la glotis (figura 3.10) se puede observar cuatro fases de la vibración de las cuerdas vocales: fase cerrada, fase de inicio de apertura, fase de apertura y fase de inicio de cierre [1].

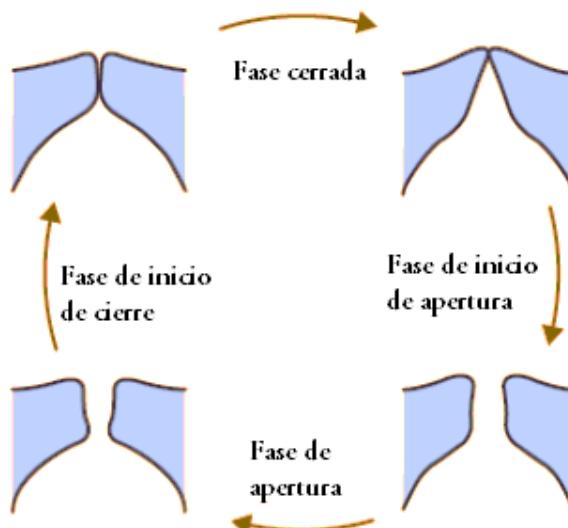


Figura 3.10. Patrón de Vibración de las Cuerdas vocales. Ciclo de la glotis. (Extraída de Physiological Process if Speech Production, Honda K.)

Los factores que determinan la vibración de las cuerdas vocales son:

- la rigidez y masa de las cuerdas vocales.
- el ancho de la glotis.
- la diferencia de presión a través de la glotis.

La regulación de la vibración de las cuerdas vocales viene dada por parámetros aerodinámicos, estos se relacionan con la diferencia de presión transglótica y el flujo de aire glotal.

El aire procedente de los pulmones experimenta un aumento de presión por debajo de la glotis, presión subglótica, cuando ésta se halla cerrada por la acción de la musculatura laríngea. Cuando la presión subglótica es suficiente para vencer la resistencia de los repliegues vocales (fuerza de aducción), abre la glotis y el aire escapa hacia el tracto vocal, provocando un aumento de la presión del tracto vocal, presión supraglótica.

El flujo de aire varía dentro de cada ciclo de la glotis, esto demuestra la variación cíclica del área glotal y de la presión subglótica.

Durante un ciclo glotal, la curva de área glotal muestra un patrón triangular (figura 3.11). La curva del flujo de aire muestra un sesgo del pico a la derecha, esto es debido a la inercia de la masa de aire dentro de la glotis. El cierre de la glotis provoca una disminución de la corriente de aire glotal a cero, esto contribuye a que sea la principal fuente de la excitación de aire del tracto vocal. Cuando el cierre de la glotis es más abrupto, los sonidos de salida son más intensos y con mayor cantidad de componentes armónicos.

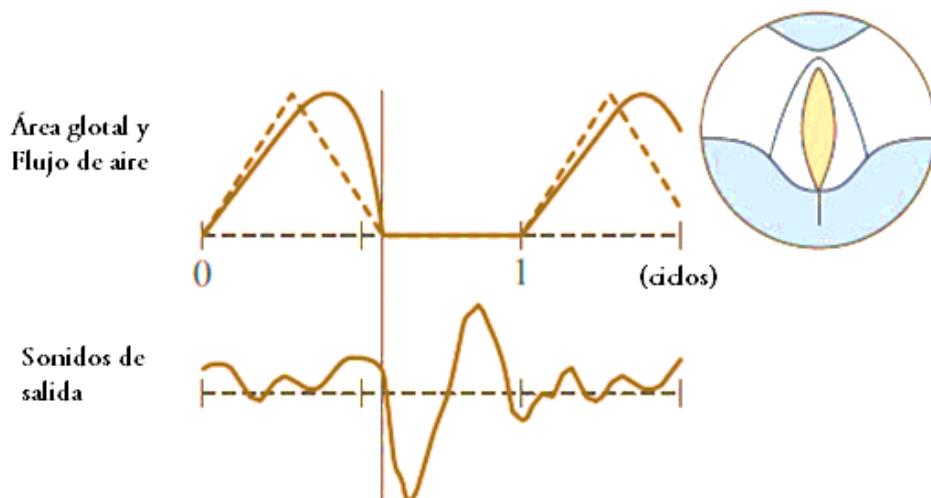


Figura 3.11 .Variación del área glotal y flujo de aire en relación a la salida de los sonidos durante un ciclo glotal. (Extraída de Physiological Process if Speech Production, Honda K.)

Cuando el cierre glotal es incompleto el flujo de aire exhibe una disminución gradual, pero no llega a cero, lo que resulta en una forma de onda más suave y en una menor intensidad de los sonidos de salida (figura 3.12)

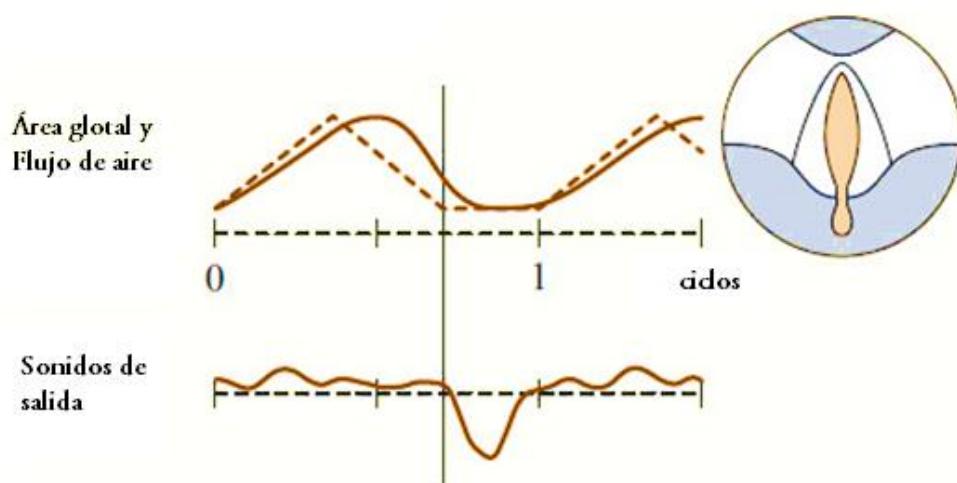


Figura 3.12. Variación del área glotal y flujo de aire en relación a la salida de los sonidos durante un ciclo glotal. (Extraída de Physiological Process if Speech Production, Honda K.)

En una voz aguda, la fase abierta del ciclo de la glotis se hace más corta, mientras que en la voz suave la fase abierta se hace más larga.

La proporción de la fase abierta dentro de un ciclo de la glotis se llama cociente abierto (OQ) y la relación entre la pendiente de cierre y la pendiente deertura en el ciclo glotal se denomina cociente de velocidad (SQ). Estos dos parámetros determinan la pendiente de la envolvente espectral. Cuando la fase abierta es larga (alto OQ) con una fase de cierre larga (lo cual indica menor pendiente y bajo SQ) el flujo de aire glotal se vuelve sinusoidal con componentes armónicos débiles.

Por el contrario, cuando la fase abierta es corta (Bajo OQ), el flujo de aire va formando ondas pulsantes con gran contenido armónico.

3.2.2 Mecanismo Articulatorio

La articulación del habla es una de las actividades motoras más complejas en los seres humanos. Forman parte del aparato articulatorio: la mandíbula, la lengua, los labios, el velo del paladar (figura 3.13). Estos órganos en conjunto alteran la resonancia y generan el sonido de las consonantes en el tracto vocal.

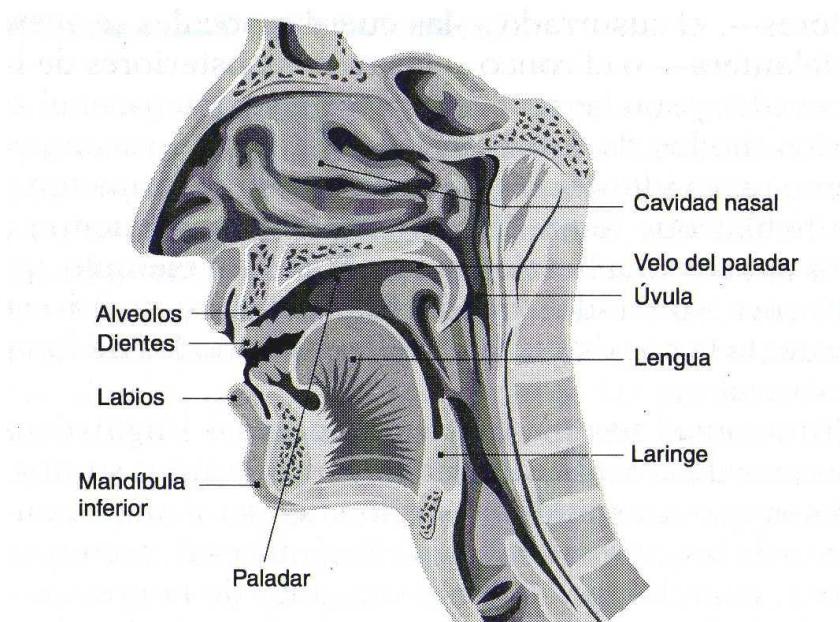


Figura 3.13. Órganos Articulación del Habla.

(Obtenido de <http://salvador707.blogspot.com.ar/2012/09/organos-articulatorios-fijos.html>)

La lengua es el órgano articulatorio más importante. La acción de éste órgano determina la calidad vocálica y produce la consonancia palatina, velar y faríngea. Cuanto mayor sea el tamaño de la lengua respecto a las cavidades oral y faríngea, menor es el espacio articulatorio para generar las vocales.

Los movimientos de la punta y cuerpo de la lengua contribuyen a la diferenciación de las consonantes dentales y alveolares.

La deformación de los labios inducida por los distintos gestos ayuda a la producción de vocales y consonantes labiales.

El velo controla la apertura y cierre velofaríngea y permite distinguir entre sonidos nasales y orales.

La configuración de los órganos articulatorios determina la forma del tracto vocal para la producción vocálica. Por ejemplo, cuando la mandíbula está en una posición alta y la lengua está en una posición frontal alta, el tracto vocálico se prepara para generar la /i/. Mientras que cuando la mandíbula está en una posición baja y la lengua se posiciona hacia atrás y hacia abajo, el tracto vocal toma la forma para generar la /a/.

3.2.3 Tracto vocal y cavidad nasal

El tracto vocal es un espacio acústico. Es el paso de sonidos donde la fuente sonora se propaga y da lugar a las principales resonancias.

La producción de vocales y consonantes se basa en el fortalecimiento o debilitamiento de las componentes espectrales de la fuente sonora por la resonancia de la columna de aire en el tracto vocal.

El tracto vocal incluye todos los espacios de aire donde se lleva a cabo la variación de la presión acústica para la producción del habla (figura 3.14).

El tracto vocal se divide en tres regiones:

- el tracto subglótico o infraglótico
- el tracto supraglótico
- las cavidades nasales.

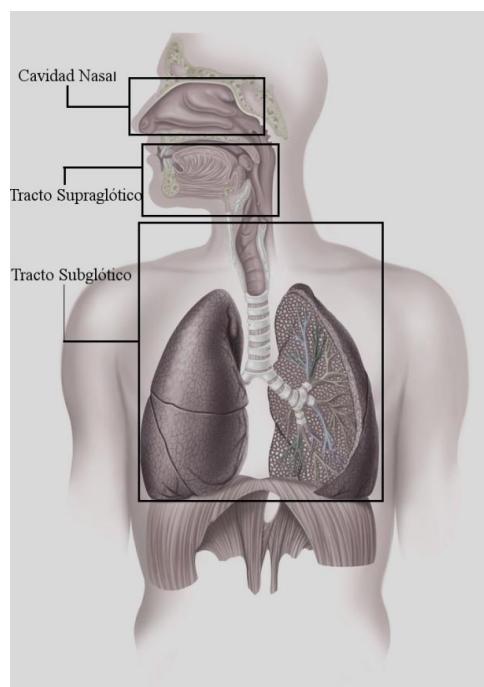


Figura 3.14. Tracto Vocal. (Obtenida de Extraída de Elsevier. Drake et al. Gray's Anatomy for Students)

El tracto subglótico es el tracto respiratorio inferior por debajo de la glotis hasta los pulmones a través de la tráquea y los bronquios. La longitud del tracto de la glotis hasta la carina traqueal es de 10-15 cm en adultos. Este espacio acústico genera resonancia en los sonidos que por él se propagan, modificando de esta manera el espectro de la voz.

Las frecuencias de resonancia de la vía aérea subglótica se estiman en 640, 1400 y 2100 Hz.

Anatómicamente, el tracto supraglótico se divide en cuatro segmentos:

- Cavidades de la hipofaringe
- Oorfaringe
- Cavidad oral
- Vestíbulo oral

La hipofaringe (figura 3.15) consiste en la cavidad supraglótica laríngea (2 cm de longitud) y las cavidades cónicas bilaterales de fosa piriforme (2 cm de longitud). Se extiende desde la línea imaginaria situada a nivel del hueso hioides hasta del borde inferior del cartílago cricoides.

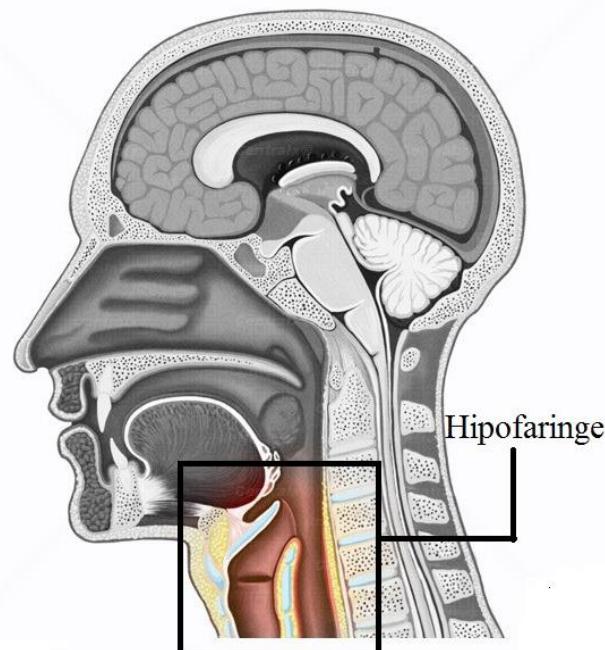


Figura 3.15. Hipofaringe. (Obtenida del Atlas del Cuerpo Humano, Centralx Atlas)

La orofaringe (fig. 3.16) se extiende desde el pliegue ariepiglótico al arco anterior del palatino. En su cara anterior, la orofaringe limita con la cavidad bucal por medio de los pilares anteriores y posteriores y, a cada lado con las amígdalas palatinas.

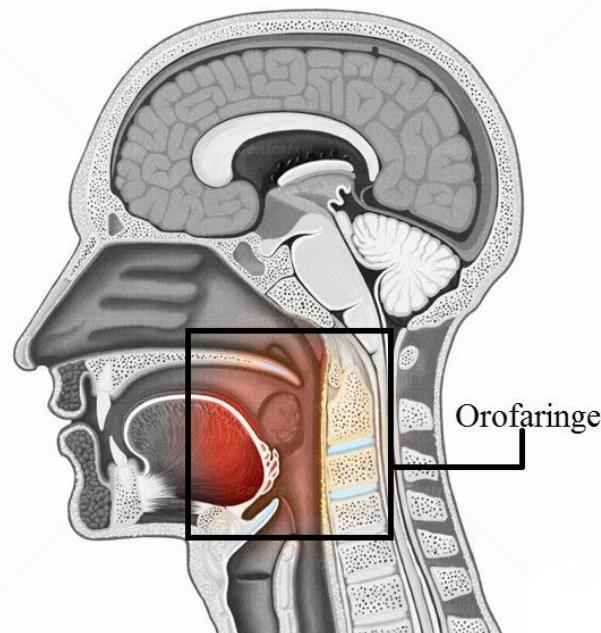


Figura 3.16. Orofaringe. (Obtenida del Atlas del Cuerpo Humano, Centralx Atlas)

La cavidad oral consiste en el segmento comprendido desde arco anterior del palatino a los incisivos. El vestíbulo oral se extiende desde los incisivos a la abertura del labio (figura 3.17).

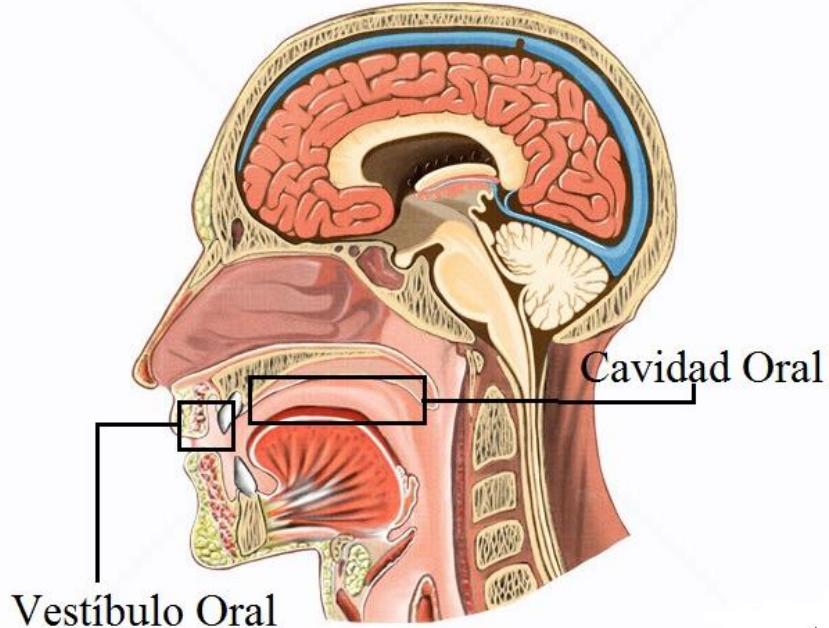


Figura 3.17. Cavidad Oral- Vestíbulo Oral. (Obtenida del Atlas del Cuerpo Humano, Centralx Atlas)

Los valores representativos de la longitud del tracto supraglótico es de 15 cm en mujeres adultas y 17.5 cm en hombres adultos.

En la población infantil, las longitudes del tracto supraglótico son de 14 cm en mujeres y 16.5 cm en varones. El alargamiento del tracto supraglótico durante el crecimiento parece estar ligado a la variación en la longitud de la cavidad faríngea debido a que la longitud de la cavidad oral se mantiene relativamente constante por las estructuras rígidas del cráneo y mandíbula.

La cavidad nasal (figura 3.18) es un canal accesorio al trato vocal principal. La cavidad nasal se puede dividir en el segmento de tubo simple, que comprende la región velofaríngea (que corresponde al área compuesta por el velo del palatino y las paredes laterales y posteriores de la faringe) y nasofaríngea, y el segmento de tubo doble, que incluye la cavidad nasal propiamente dicha y el vestíbulo nasal.

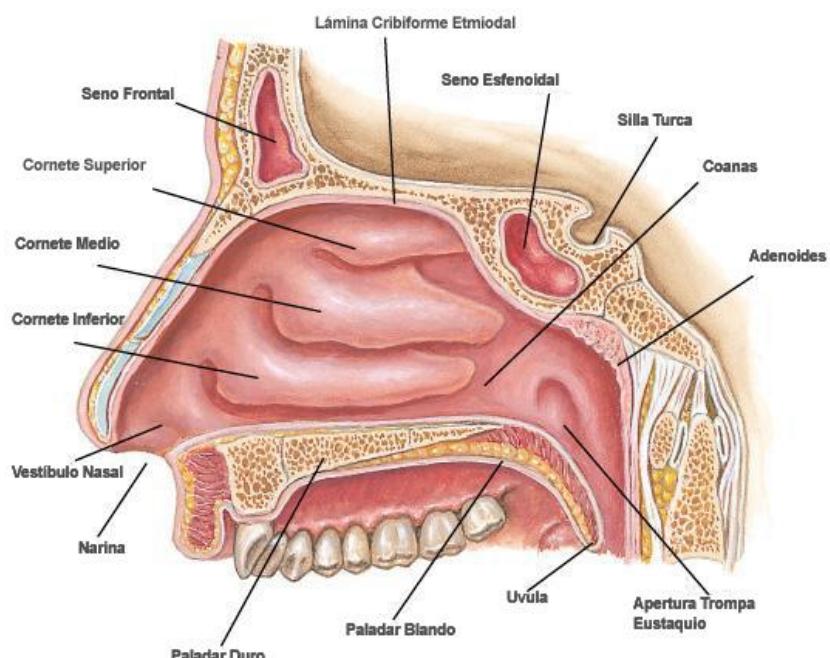


Figura 3.18. Cavidad Nasal. (Extraída de Atlas de Anatomía Humana, Adam.)

La función principal de esta cavidad es construir resonancia nasal para generar las características fonéticas de los sonidos nasalizados. Los senos paranasales también contribuyen a las características acústicas de los sonidos nasales.

3.3 Características fisiológicas de la voz

3.3.1 Formantes

La resonancia es un fenómeno por el cual un cuerpo, denominado resonador y que posee una tendencia natural a vibrar a determinada frecuencia, experimenta una vibración de mayor amplitud cuando es puesto en movimiento por otro cuerpo vibrante a una frecuencia similar [3] [4].

El resonador de Helmholtz consiste básicamente en una cavidad contenida de aire con al menos una abertura en forma de cuello [5]. Un ejemplo de resonancia de Helmholtz es el sonido que se crea cuando se sopla transversalmente sobre el cuello de una botella. Cuando se fuerza al aire a introducirse en una cavidad, la presión interior de la misma crece. Una vez que el agente externo que

fuerza el aire hacia el interior de la cavidad desaparece, el aire comprimido del interior fluye hacia afuera. Este flujo de aire tiende a sobre-compensar la diferencia de presión. Debido a la inercia del aire en el cuello una pequeña porción de aire sale al exterior de modo que la presión de la cavidad disminuye, alcanzando un valor ligeramente menor que la externa y esto hace que el aire de nuevo fluya hacia el interior. Este proceso tiende a repetirse de manera continua. Este comportamiento es similar a las oscilaciones libres amortiguadas de una masa atada al extremo de un resorte. El aire en el cuello de la botella se mueve como un pistón sólido mientras que el aire en el volumen principal del envase se comprime y expande de manera alternativa como si fuera un resorte. Se puede decir que el aire como masa resonante oscila a una determinada frecuencia. De acuerdo con esta teoría, Helmholtz demostró que la fonación es el resultado de flujos de aire emitidos a través de la glotis, que al pasar por las cavidades de las vías aéreas modifican la amplitud de los armónicos debidos a la resonancia de dichas cavidades [6].

El sonido nasal resulta de estas características:

- Resonancia de Helmholtz del tracto nasofaríngeo desde la glotis a las fosas nasales anteriores.
- Resonancia regional de Helmholtz causada por los senos paranasales.

Juntas se caracterizan por un pico de resonancia a 200-300Hz y por un aplanamiento espectral hasta los 2 Khz.

Las funciones de la laringe, como un generador de fuente de sonido, y el tracto vocal desempeñan un papel importante como filtro acústico para modular los sonidos desde la fuente hasta su propagación. Esto es descripto como la teoría de la fuente-filtro [7].

El filtrado que actúa modificando el espectro del sonido, tiene lugar en las cavidades resonantes que enfatizan determinadas bandas frecuenciales, conduciendo al concepto de formantes. Formante es una zona de la escala de frecuencia en la que un sonido presenta una mayor concentración de energía. Aunque, también se puede definir como cada una de las resonancias del conducto vocal [8] [9].

La calidad de las vocales es determinada por los formantes. El formante de frecuencia más baja se denomina F1; el segundo F2; el tercero F3, etc. Normalmente sólo se necesitan los dos o tres primeros para caracterizar una vocal [10].

En la producción vocalica, el tracto vocal forma un tubo cerrado con el extremo cerrado en la glotis y el extremo abierto en la apertura de los labios. Múltiples reflexiones de la onda sonora entre los dos extremos del tracto vocalico dan lugar a los formantes vocálicos (F1, F2, F3). El primer formante tiene una frecuencia más alta cuanto más baja está la lengua; es decir, cuanto mayor apertura tenga una vocal, mayor es la frecuencia en que aparece este formante. El segundo formante tiene una mayor frecuencia cuanto más hacia delante este posicionada la lengua.

Estudios recientes han mostrado evidencia que la resonancia que ocurre en la cavidad hipofaríngea determina la envolvente espectral en altas frecuencias, alrededor de 2.5KHz [8].

La cavidad laríngea por encima de las cuerdas vocales también contribuye a dar forma a las frecuencias más altas. La cavidad laríngea supraglótica forma un tipo de resonador Helmholtz y da lugar a una resonancia a una frecuencia de 3.5Khz. Esta resonancia puede ser considerada como el cuarto formante (F4), aunque en realidad es un extraformante [8].

Cuando la glotis se abre en la fase de apertura de la vibración de las cuerdas vocales, la cavidad supraglótica deja de ser un típico resonador Helmholtz y muestra una fuerte amortiguación de la resonancia que se observa como la desaparición de estos extraformantes. Por lo tanto, la resonancia de la cavidad laríngea muestra una naturaleza cíclica durante la vibración de las cuerdas vocales y es posible que se ausente en la fonación entrecortada o en condiciones patológicas con cierre glotal insuficiente.

Las cavidades hipofaringeas no son estructuras fijas sino que varían debido a los esfuerzos fisiológicos para controlar la frecuencia de vibración de las cuerdas vocales y la calidad de la voz. Un caso típico del ajuste de la cavidad hipofaringea para controlar la calidad de voz, se encuentra en el formante durante el canto. Cuando las notas altas son producidas por cantantes de ópera, la laringe se va hacia adelante debido a la posición avanzada de la lengua que extiende la fosa piriforme para profundizar las antiresonancias de la fosa. Esto resulta en una disminución de la frecuencia del formante adyacente (F5). Cuando la cavidad de la laringe supraglótica es comprimida su formante (F4) se reduce hacia el formante inferior (F3). Consecuentemente, los formantes del tercero al quinto se acercan entre si y generan un pico alto de resonancia observado cerca de 3kHz.

Desde el punto de vista acústico, las vocales se diferencian por la frecuencia a la que se encuentran esas zonas de armónicos de amplitudes reforzadas por acción de la resonancia de las cavidades. Por ejemplo en la vocal /i/ el adelantamiento de la lengua crea una cavidad anterior al punto de constrictión palatina de tamaño pequeño, mientras que la cavidad posterior a este punto presenta un tamaño mayor (figura 3.19). Esto se refleja en un espectro con un primer formante de frecuencia baja, relacionado con la cavidad posterior y un segundo formante de frecuencia alta, relacionado con la cavidad anterior (figura 3.20).

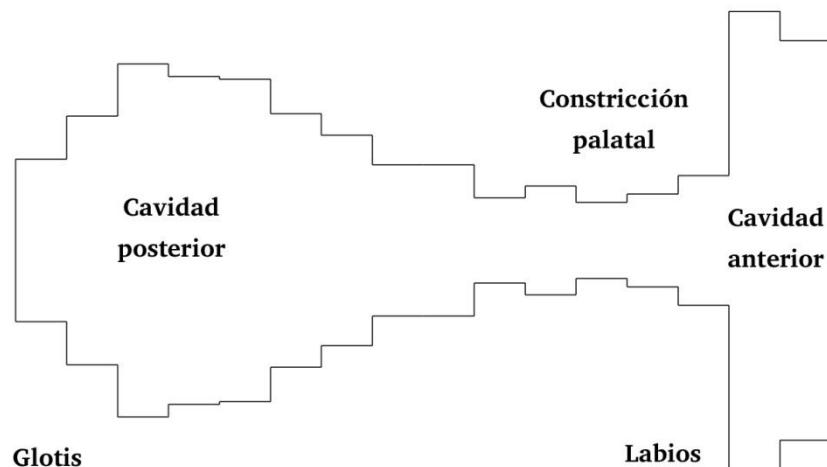


Figura 3.19. Configuración de las Cavidades Resonantes para la vocal /i/. (Obtenida de Los Sonidos del Lenguaje, Gil, Juana)

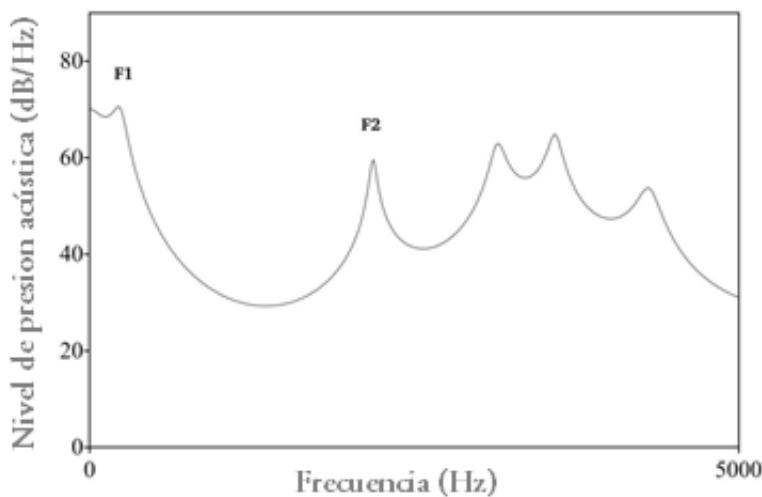


Figura 3.20. Espectro de Frecuencias para la vocal /i/. (Obtenida de Los Sonidos del Lenguaje, Gil, Juana)

En el caso de la vocal /u/, el punto de constricción situado en la zona velar, crea dos cavidades grandes, una anterior a la que contribuye el adelantamiento de los labios y otra posterior (figura 3.21). Esto se refleja en el espectro de frecuencias, donde se evidencian dos formantes de baja frecuencia (figura 3.22).

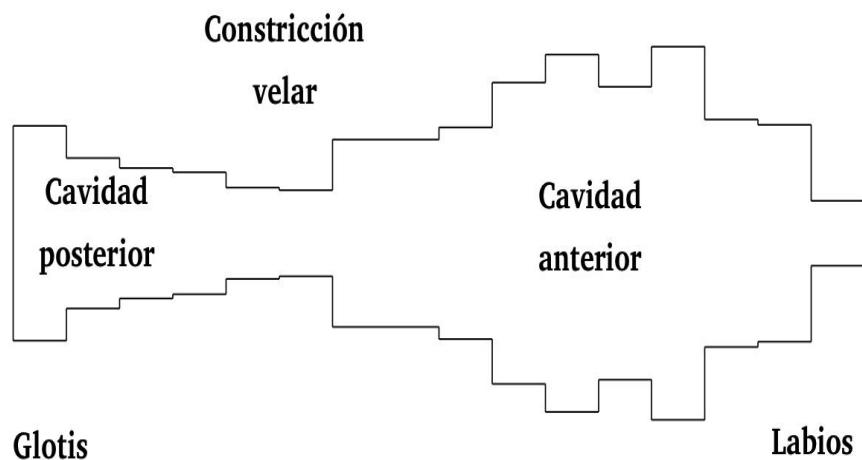


Figura 3.21. Configuración de las Cavidades Resonantes para la vocal /u/. (Obtenida de Los Sonidos del Lenguaje, Gil, Juana)

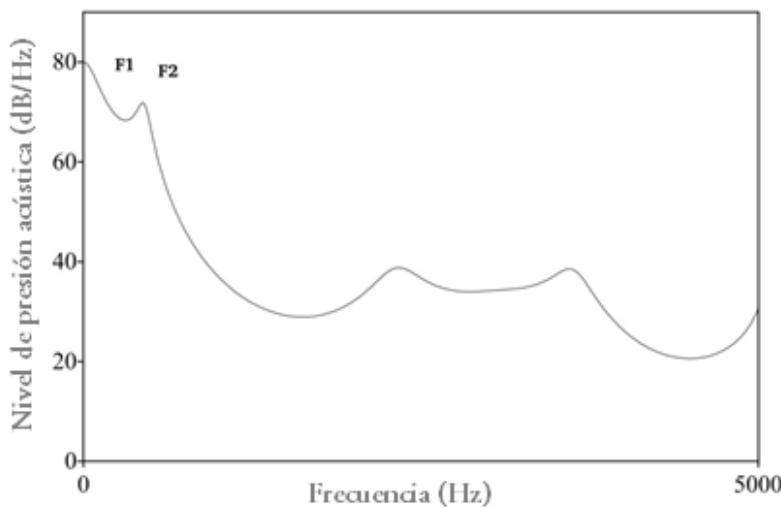


Figura 3.22. Espectro de Frecuencias para la vocal /u/. (Obtenida de Los Sonidos del Lenguaje, Gil, Juana)

3.3.2 Frecuencia fundamental

La frecuencia fundamental (F0) de la voz es el componente armónico más bajo del espectro de los sonidos sonoros. La F0 corresponde a la frecuencia natural de vibración de las cuerdas vocales, y sus variaciones dependen de dos factores:

- longitud de las cuerdas vocales
- factores aerodinámicos

Las frecuencias F0 altas, indican cuerdas vocales delgadas y largas; mientras que las F0 bajas se corresponden con cuerdas vocales gruesas y cortas.

Cuando se estiran las cuerdas vocales, la masa por unidad de longitud del tejido de las cuerdas vocales se reduce mientras que la rigidez del tejido involucrado en la vibración se incrementa.

Por lo tanto, cuando la masa disminuye, la rigidez aumenta y la F0 es mayor.

En adultos varones, el rango de F0 es de 80 a 400 Hz, mientras que el rango en las mujeres es de 120-800 Hz.

La longitud de los pliegues vocales se ajusta mediante el movimiento relativo de los cartílagos cricoides y tiroides. Los cartílagos tiroides y cricoides se ponen en contacto mediante la articulación cricotiroidea. Cualquier fuerza externa aplicada a esta articulación genera en forma conjunta rotación y traslación, esto altera la longitud de las cuerdas.

Los músculos extrínsecos de la laringe también regulan la F0. Por ejemplo, la acción del músculo esternohiodeo disminuye la F0. Cuando se contrae desplaza hacia abajo el hueso hioides al tiempo que desliza en forma vertical toda la laringe. A medida que el cartílago cricoides desciende, va girando debido a la convexidad anterior de la columna cervical, esto acorta las cuerdas vocales.

Las condiciones aerodinámicas generan variaciones en la F0. Por ejemplo, el incremento de la presión subglótica durante un discurso con énfasis o estrés genera un incremento de la F0.

El aumento de la presión de aire subglótica resulta en una mayor tasa de flujo de aire y en una mayor apertura de la glotis, esto causa deformación de las cuerdas vocales con incremento en la rigidez de los tejidos.

Se ha informado que cuando se comprime el pecho externamente, la tasa de aumento de la F0 debido a la presión subglótica es de 2-5Hz/cmH₂O, y cuando se mide entre el inicio y final de un discurso expresivo la tasa de aumento de la F0 es de 5-15Hz/cmH₂O [8].

3.3.3 Sonidos sonoros y sordos

Los sonidos sordos y sonoros son a menudo atribuidos al estado glotal con o sin vibración de las cuerdas vocales.

En vocales sonoras, el tracto vocal forma un tubo cerrado sin estrechamientos significativos, a excepción del estrechamiento de la cavidad laríngea. Por otro lado en las vocales susurradas la glotis membranosa está cerrada, con un estrechamiento moderado de la faringe inferior. Las vocales sordas exhiben una glotis ampliamente abierta y una reducción de la articulación de la lengua (figura 3.23).

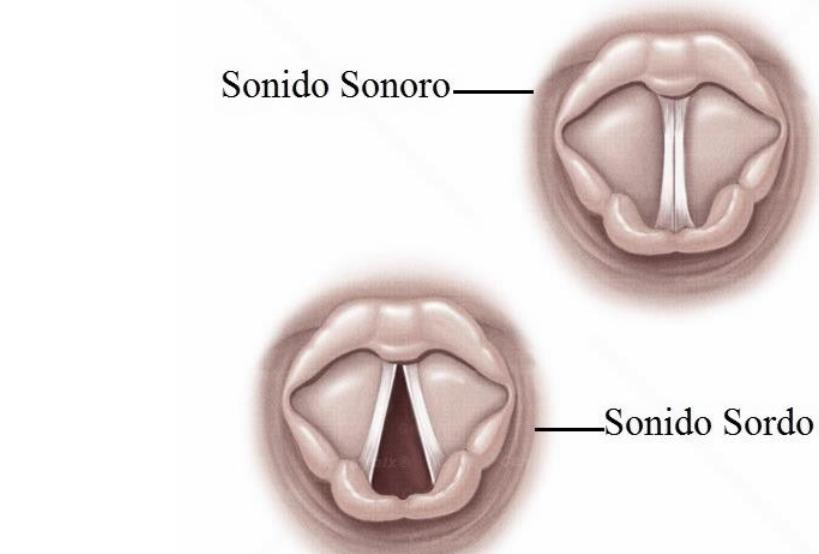


Figura 3.23. Sonidos Sonoros y Sordos. (Obtenida del Atlas del Cuerpo Humano, Centralx Atlas)

Distinciones fonéticas de consonantes sonoras y sordas implican un control temporal preciso de los articuladores de la laringe y supralarínge en tipos específicos de idiomas.

En la producción de consonantes sonoras existe vibración de las cuerdas vocales. En sonidos oclusivos (que se producen por la formación de un obstáculo completo en los órganos bucales y la liberación explosiva del aire, ejemplo: /p/) y fricativos (la emisión de estos fonemas se produce debido a una constrictión parcial, solo es un roce, de los órganos de la boca a la salida del aire, ejemplo: /f/) el cierre o estrechamiento del tracto vocal resulta en la disminución del flujo de aire glotal y la diferencia de la presión transglotal.

El flujo de aire glotal durante el cierre oclusivo se mantiene debido a los incrementos de volumen en el tracto vocal: la expansión de la cavidad oral (descenso de la mandíbula y expansión de la pared

de las mejillas) y la expansión de la cavidad faríngea (expansión de la pared lateral y descenso de la laringe). Durante el período de cierre, las variaciones de presión de aire irradian no sólo desde la pared del tracto vocal sino también desde las fosas nasales anteriores, debido a la propagación trasvelar de la presión acústica intra-oral en la cavidad nasal.

En la producción de las consonantes sordas la vibración de las cuerdas vocales se suprime debido a la rápida reducción de la diferencia de presión transglotal y a la abducción de las cuerdas vocales. Durante el cierre oclusivo la presión intra-oral aumenta hasta alcanzar la presión subglotal, que mejora el flujo de aire después de la liberación de la oclusión.

Entonces, la vibración de las cuerdas vocales se reinicia con un retraso en la liberación, que se observa como un tiempo de latencia largo de la voz (VOT) para oclusivas sordas. El proceso de supresión de la vibración de las cuerdas vocales no es un simple proceso aerodinámico pasivo de las cuerdas vocales sino que también colabora un proceso fisiológico para controlar la rigidez de las cuerdas vocales.

En la oclusión glotal, la vibración de las cuerdas vocales se detiene debido a la aducción forzada de las cuerdas vocales con un cierre forzado de la cavidad laríngea supraglótica.

En la figura 3.24 se observa la evolución temporal durante la elocución vocal-consonante-vocal (VCV). El segmento fricativo sordo (/s/) se inicia con la abducción de la glotis (incremento del área glotal) y el estrechamiento del tracto vocal, la vibración de las cuerdas vocales se desvanece de a poco durante la fase de apertura de la glotis, y a la salida de audio no se observa señal. Despues de alcanzar la abducción máxima, la glotis entra en fase de aducción seguida de la liberación del estrechamiento alveolar. Entonces la glotis se vuelve estrecha y se reinicia la vibración de las cuerdas vocales.

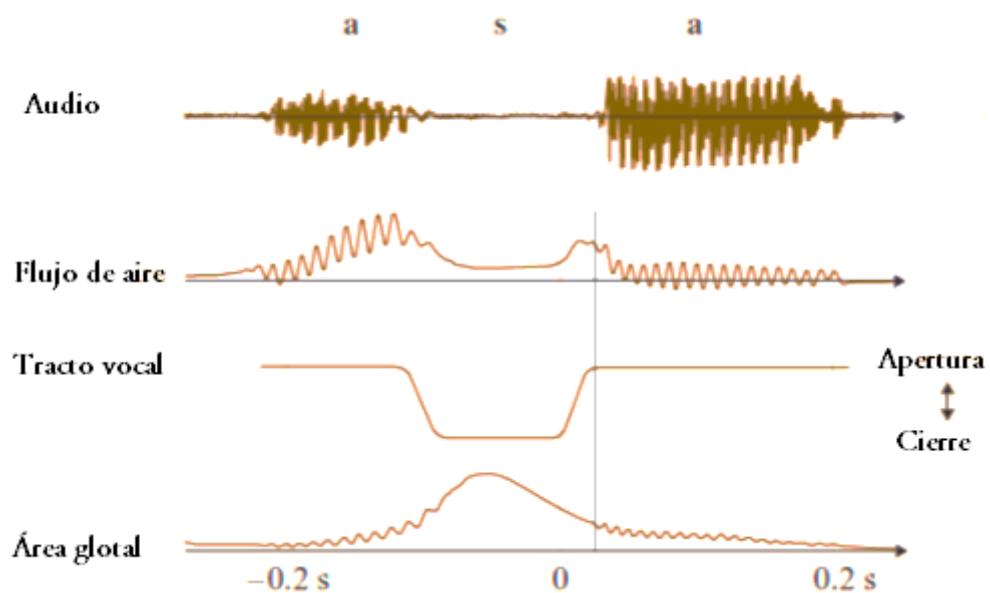


Figura 3.24. Elocución vocal-consonante-vocal. (Extraída de Physiological Process if Speech Production, Honda K.)

3.4 Referencias

1. Honda, K. (2008). *Physiological processes of speech production*. In *Springer Handbook of Speech Processing* (pp. 7-26). Springer Berlin.
2. Haddad, D., Walter, S., Ratley, R., & Smith, M. (2001). *Investigation and evaluation of voice stress analysis technology*. AIR FORCE RESEARCH LAB ROME NY INFORMATION DIRECTORATE.
3. Kane J.W., Sternheim M. M., Vázquez J., Mirabent D.(1989). Física. Editorial Reverte. ISBN: 8429143181, 9788429143188. (Cap.22, pp491-512).
4. Olguín V. (2006). Física: principios con aplicaciones. Editorial Pearson Educación. ISBN: 9702606950, 9789702606956. . (Cap.12, pp322-351).
5. Matar, M., & Welti, R. (2009). *Capturando la física de los resonadores Helmholtz con la ecuación de ondas acústica*. Latin-American Journal of Physics Education, 3(1), 17.
6. Etxebarria, M. (2013). *En torno al vocalismo vasco*. Anuario del Seminario de Filología Vasca "Julio de Urquijo", 635-655.
7. Fant G. (1960). *Acoustic Theory of Speech Production* (Mouton, The Hague).
8. Kollmeier K., Brand T. Meyer B. (2008). *Perception of Speech and Sound*. In *Springer Handbook of Speech Processing* (pp. 7-26). Springer Berlin.
9. Rabiner L. R. y. Schafer R. W. (2007). "Introduction to Digital Speech Processing". *Foundations and Trends® in Signal Processing* Vol. 1, Nos. 1–2.
10. Cobeta, I., Núñez, F., & Fernández, S. (2013). *Patología de la voz*. Marge books. ISBN: 978-84-15340-86-7

CAPITULO 4

Análisis de la señal de audio y descripción de características

4.1 Introducción

En este capítulo se realizará un análisis físico-matemático de la señal de audio a fin de determinar sus características descriptivas, desde los enfoques: temporal, prosódico y frecuencial. El conocimiento de estos parámetros es necesario para detectar, caracterizar y clasificar los estados emocionales.

4.2 Análisis de la señal

Una señal puede ser definida como una cantidad física que varía con el tiempo, el espacio o alguna otra variable independiente. La señal de voz puede ser modelada como una función de la presión acústica variante en el tiempo [1].

Un segmento de la señal de audio se representa como una suma de varias sinusoides de diferentes amplitudes y frecuencias [2].

$$\sum_{i=1}^N A_i(t) \sin[2\pi F_i(t)t + \theta_i t] \quad (4.1)$$

Donde $\{A_i(t)\}$, $\{F_i(t)\}$, y $\{\theta_i(t)\}$ son el conjunto de variables: amplitudes, frecuencias y fases respectivamente de sinusoides. Una forma de interpretar la información contenida en un segmento corto de tiempo de una señal de audio es midiendo las amplitudes, frecuencias y fases contenidas en ella.

Asociado con la naturaleza de la señal están los medios con los que esas señales son generadas. La señal de audio es generada al forzar el aire través del tracto vocal. Por ello, la generación de la señal está usualmente asociada con un sistema que responde a un estímulo o fuerza. En la señal de audio, el sistema consiste de: las cuerdas vocales y el tracto vocal. El estímulo en combinación con el sistema es llamado “Fuente sonora” [2]. Cuando dicha fuente entra en vibración y ésta es transmitida a las partículas de aire adyacentes se originan variaciones en la presión del aire (compresiones y descompresiones). Estas variaciones de presión se propagan en el medio originando las ondas sonoras. El grado de compresión y descompresión del aire es la amplitud de la onda sonora o presión sonora y está relacionada con el nivel sonoro.

4.3 Modelado de la señal

La señal de voz se puede modelar como la salida de un sistema lineal invariante en el tiempo (tracto vocal) excitado por una secuencia de pulsos cuasiperiódicos o ruido blanco, dependiendo del tipo de voz producida, sonora o sorda, respectivamente.

Este sistema lineal se puede dividir en tres bloques:

- *Generador de energía*: Los músculos torácicos y abdominales, al aumentar la presión en las vías respiratorias, generan la energía necesaria para producir la voz.
- *Sistema vibrante*: Lo constituye la glotis y las cuerdas vocales. Los pliegues vocales producen una vibración quasi periódica en el caso de los sonidos sonoros. Esta vibración ocurre a una frecuencia característica para cada individuo.
- *Sistema resonante*: Está constituido por el tracto vocal que se representa como una concatenación de tubos acústicos sin pérdida cuyos límites son las cuerdas vocales en un extremo y los labios por el otro. Posee una sección variable en función de los órganos articulatorios pudiendo variar entre 0 y 20 cm². Esta representación permite considerar que la onda propagada a través del tracto es plana, o sea se propaga en una sola dimensión, a lo largo de un eje.

El modelo de la producción de la voz, se basa en la teoría de Fant [3] del tracto vocal, denominada fuente-filtro, que se representa esquemáticamente en la figura 4.1.

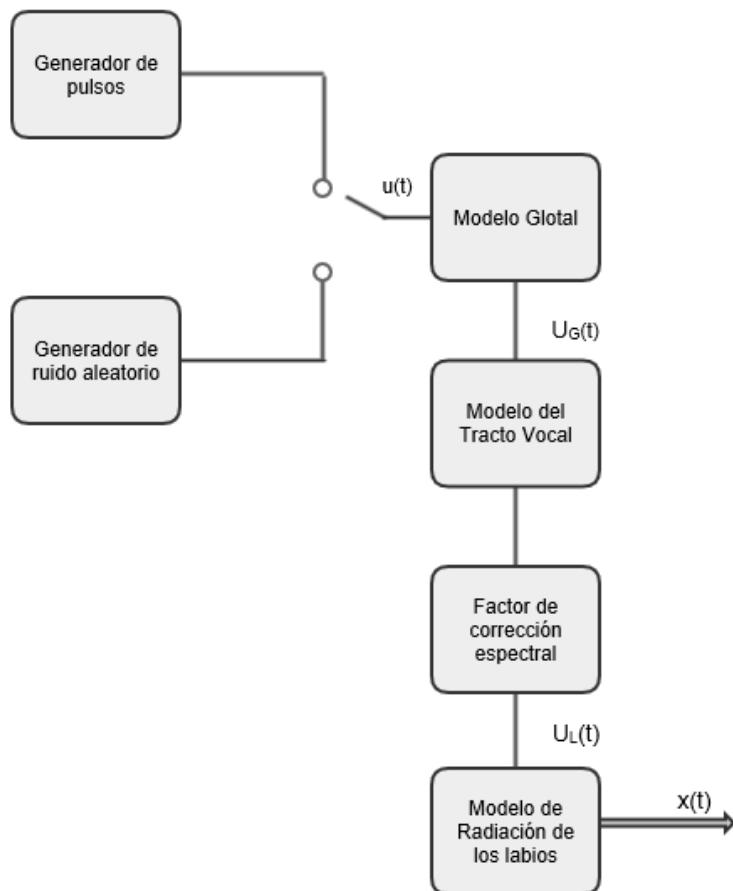


Figura 4.1. Modelo de producción del Habla (Obtenido de Técnica de procesado y representación de la señal de voz para el Reconocimiento del habla en ambientes ruidosos, Pericas)

Esta teoría, considera básicamente tres elementos en la producción de la voz: la excitación (el flujo glotal), la transmisión (condicionada por la configuración y la resonancia del tracto vocal supraglótico) y la radiación (debida a la configuración de la apertura de la boca por la posición de los

labios). La señal de velocidad volumétrica glotal $UG(t)$ se modela como la salida de un filtro pasa bajo. La entrada a este filtro $u(t)$ es un tren de pulsos para sonidos sonoros y ruido aleatorio de espectro plano para el caso de sonidos sordos. El tracto vocal se modela mediante un filtro resonante. Cada resonancia se define como un formante con su frecuencia central y su ancho de banda correspondiente

Un modelo más preciso requiere un número infinito adicional de resonancias que permitiría elevar el nivel del espectro en las bajas frecuencias. Cuando se modela, en forma precisa, el comportamiento de un sistema en bajas frecuencias, se toma en cuenta, un factor de corrección espectral [4]. La señal de velocidad volumétrica en los labios $UL(t)$ se transforma en una señal de presión acústica $x(t)$ a una cierta distancia de los labios, a través del modelo de radiación de los labios.

Este modelo puede describirse en notación de transformada Z para su implementación computacional [4], mediante la siguiente ecuación:

$$X(z) = U(z)G(z)V(z)L(z) \quad (4.2)$$

Donde $X(z)$ y $U(z)$ son la transformada Z de las secuencias discretas $x(n)$ y $u(n)$. $G(z)$; $V(z)$ y $L(z)$ son las funciones de transferencia de los sistemas discretos que modelan los efectos de la glotis, el tracto vocal y los labios, respectivamente.

La representación mediante transformada Z permite eliminar el factor de corrección espectral que figura en el modelo original, simplificando su representación [5].

Una importante reducción de este modelo consiste en combinar los efectos de la glotis, el tracto vocal y los labios, y representarlos mediante una única función de transferencia $H(z)$, es decir:

$$X(z) = U(z)H(z) \quad (4.3)$$

En la práctica, se modela el filtro $H(z)$ como un filtro de todo-polos.

$$H(z) = \frac{G}{1 + \sum_{k=1}^p \alpha_k z^{-k}} \quad (4.4)$$

El modelo de voz simplificado está representado en la figura 4.2. El sistema es excitado por un tren de pulsos en el caso de la voz sonora o por ruido en el caso de la voz sorda. Los parámetros del modelo son: la decisión (sonoro/sordo), el tono, y la ganancia y los coeficientes del filtro $H(z)$.

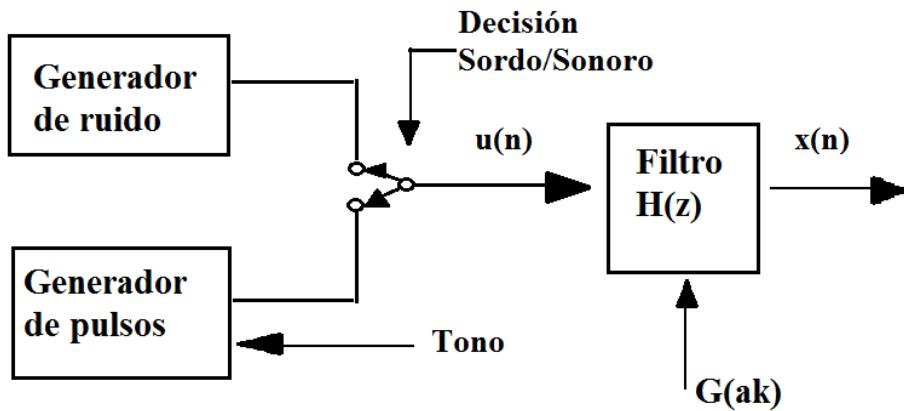


Figura 4.2. Modelo Simplificado de producción del Habla (Obtenido de Técnica de procesado y representación de la señal de voz para el Reconocimiento del habla en ambientes ruidosos, Pericas).

A partir de la expresión (4.3), el espectro de la señal de voz $x(n)$ puede escribirse como:

$$S_{xx}(w) = S(w) \left| H(e^{jw}) \right|^2 \quad (4.5)$$

Donde $S(w)$ es el espectro de excitación $u(n)$ y $H(e^{jw})$ es la respuesta frecuencial del filtro. Mediante algunas técnicas se puede separar el espectro de excitación y la respuesta frecuencial. En el dominio temporal, esto equivale a deconvolucionar la señal de voz $x(n)$ o sea, separar la excitación $u(n)$ y la respuesta impulsional del filtro $H(z)$, $h(n)$.

$$x(n) = u(n).h(n) \quad (4.6)$$

Esto representa separar la información de sonoridad y tono de la estructura de formantes.

4.4 Cualidades del sonido

4.4.1 Frecuencia Fundamental y Armónicos

Las señales sonoras consisten en una frecuencia fundamental y una serie de armónicos. La frecuencia fundamental se define como el número de ciclos por segundo en que las cuerdas vocales completan un ciclo de vibración.

Por definición, las frecuencias de los armónicos son múltiplos de la frecuencia fundamental [6]. Los armónicos representan la forma de onda del sonido. La presencia y el número de armónicos confieren al sonido la cualidad subjetiva de “**Timbre**”. El **timbre** podría definirse como el “color” de un sonido y nos permite reconocer a las personas por su voz. La combinación de la frecuencia fundamental y sus armónicos generan una onda compleja. Sin embargo es importante hacer notar que si los armónicos completaran una cantidad exacta de ciclos en el mismo tiempo que toma un ciclo de la frecuencia fundamental, el resultado es que la señal compleja tendrá un patrón de movimiento repetitivo. El carácter repetitivo del movimiento de la onda indica su naturaleza periódica. Los sonidos generados por la laringe y los modificados por el tracto vocal como el sonido

de una vocal, son sonidos periódicos. Sin embargo los sonidos periódicos pueden diferir en la calidad. Las diferentes estructuras de armónicos con naturaleza periódica generan un patrón característico que es reconocido por el oído y el cerebro. Los sonidos que pertenecen a la clase de sonidos periódicos se les llaman **tonos**. Una característica de los **tonos** es que ellos tienen una frecuencia fundamental que es fácilmente reconocida y esta es una característica del sonido que nos permite distinguir entre sonidos graves y agudos.

Hay sonidos en los que no todos sus componentes son armónicos. En este caso, por cada ciclo de la fundamental, las componentes no armónicas completarán cantidades de ciclos diferentes dado que tienen una fase diferente a la fundamental y a las demás armónicas. En consecuencia, la forma de onda de cada ciclo será diferente y ya no habrá periodicidad. Cada componente en un sonido complejo que no tenga periodicidad será inarmónico. Sin embargo cuando las componentes inarmónicas son de baja amplitud, el oído y el cerebro las toman con poca relevancia y traducen al sonido como un tono.

El término aperiódico es generalmente aplicado a sonidos que no son de base armónica, es decir, sonidos cuyas componentes frecuenciales no están relacionadas entre sí. En este caso se habla de sobretonos, o sea para un determinado sonido existen frecuencias superiores a la fundamental que no están relacionadas con ésta mediante un múltiplo entero.

Los sonidos aperiódicos son aquellos que el oído y el cerebro toman como ruido. La esencia del contraste entre un ruido y un tono es la naturaleza aleatoria del movimiento de las partículas de aire en caso del ruido y la regularidad de la misma en el tono.

Un método por el cual un movimiento aleatorio se puede generar es cuando el aire forzado fluye a través de una abertura relativamente estrecha, lo cual crea turbulencia. Este fenómeno ocurre durante el proceso del habla. Por ejemplo, para pronunciar la /s/ se hace una constrictión entre la lengua y los dientes, lo cual fuerza al aire a fluir rápidamente dentro de la abertura que queda conformada; el resultado es una turbulencia que produce un sonido de silbido característico con una forma de onda aperiódica.

4.4.2 Intensidad sonora

La intensidad es una magnitud que representa la cantidad de energía que está fluyendo en el movimiento de vibración o movimiento de las partículas de aire que constituyen las ondas sonoras [6]. La amplitud y la frecuencia de estas vibraciones están relacionadas con la cantidad de energía suministrada. A partir de esta relación, se define a la intensidad de una onda sonora como una magnitud proporcional al cuadrado de su amplitud y frecuencia e inversamente proporcional a la distancia de la fuente. Las vibraciones de mayor amplitud requieren más energía que aquellas de menor amplitud y el sonido proporcionado por consiguiente, será más fuerte. Sin embargo, si tenemos dos ondas sonoras de igual amplitud con una de ellas a mayor frecuencia que la otra, la de mayor frecuencia necesitaría más energía. De modo que, a sonidos de igual amplitud, el sonido de mayor frecuencia sería el más fuerte a la percepción.

4.5 Características la señal de voz

4.5.1 Enfoque estadístico

El estudio sobre la naturaleza estadística de la señal de voz implica considerar su función de probabilidad y su estacionariedad.

La función de densidad de probabilidad (*pdf*) se puede estimar mediante un histograma de las amplitudes sobre un número suficientemente grande y representativo de muestras de la señal. Se ha demostrado que la señal de voz puede ser representada por una distribución Laplaciana o Gamma [7]. Estas distribuciones son válidas si observamos la señal en tramos muy largos. Si fuese el caso de hacer observaciones en intervalos cortos, estas distribuciones ya no tienen validez.

Entre las características más importantes de la señal de voz está su naturaleza no estacionaria. Un proceso estacionario [8] es un proceso estocástico cuya distribución de probabilidad en un instante de tiempo determinado es la misma para todas las posiciones. Una señal es estacionaria en sentido débil o en sentido amplio si tiene momentos de primer y segundo orden finitos y que no varían en función del tiempo.

La estacionariedad de la señal de voz se puede evaluar al variar la longitud del intervalo de observación. La señal de voz es una señal de evolución lenta en el sentido de que, cuando se examina en intervalos de tiempo suficientemente cortos (típicamente entre 20 y 60 ms), sus características son prácticamente estacionarias. Se habla, entonces, de que la señal es quasi estacionaria. Sin embargo, en intervalos largos las características de la señal cambian para reflejar los diferentes sonidos que se están pronunciando, así da una señal no estacionaria.

La detección de la no estacionariedad resulta estadísticamente fundamental ya que afecta la elección de las técnicas de extracción de características.

Si analizamos la forma en que se genera una señal de voz, es posible advertir que el tracto vocal tiene una velocidad máxima con la que puede alterar su morfología. Esto permite plantear la hipótesis de que la señal de voz permanece estacionaria durante ciertos intervalos de tiempo en relación a la velocidad de variación de la morfología del tracto. De forma precisa, se considera que la señal de voz es estacionaria en periodos de tiempo de 20ms [9].

Desde el punto de vista temporal la señal de voz presenta tres estados definidos, estos son:

- silencio, en el que no hay voz.
- voz sorda, en el que las cuerdas vocales no vibran.
- voz sonora, en el que las cuerdas vocales vibran dando lugar a una señal quasi periódica.

Desde esta perspectiva se pueden analizar las características singulares de la señal de voz estableciendo como premisa la quasi-periodicidad y la quasi estacionariedad.

4.5.2 Enfoque Temporal

Desde el punto de vista temporal se pueden determinar características estadísticas de la señal de voz que permiten dilucidar su comportamiento.

4.5.2.1 Media Aritmética

La Media aritmética coincide con el momento de primer orden respecto al origen de una distribución de variables aleatorias.

$$\bar{X}_n = \frac{1}{N} \sum_{m=0}^N [x(m) \cdot w(n-m)] \quad (4.7)$$

Donde $x(m)$ es la señal de voz y $w(n-m)$ es la ventana de longitud N de n muestras.

4.5.2.2 Varianza

La varianza representa la dispersión de la probabilidad respecto de la media

$$Sn^2 = \frac{1}{N-1} \sum_{m=0}^N ([x(m) \cdot w(n-m)] - \bar{x})^2 \quad (4.8)$$

Donde Sn^2 representa la varianza, \bar{x} representa la media de la muestra, $x(m)$ es la señal, $w(n-m)$ es la ventana temporal y N es el tamaño de las m muestras.

4.5.2.3 Desviación estándar

La desviación estándar Sn permite determinar el promedio aritmético de fluctuación de los datos respecto a su punto central o media

$$Sn = \sqrt{Sn^2} \quad (4.9)$$

4.5.2.4 Cruces por cero

Los cruces por cero se definen como el promedio ponderado del número de veces que la señal de voz cambia de signo en una ventana de tiempo [9].

$$Zn = \sum_{m=\infty}^{\infty} 0.5 | sgn(x(m)) - sgn(x(m-1)) | \cdot w(n-m) \quad (4.10)$$

Donde, $sgn(x): \begin{cases} 1 & x \geq 0 \\ -1 & x \leq 0 \end{cases}$

$x(m)$ es la señal en el instante m , $x(m-1)$ es la misma señal en un instante anterior y $w(n-m)$ corresponde a la ventana temporal.

Por lo tanto,

$$A = 0.5 | sgn(x(m)) - sgn(x(m-1)) | \quad (4.11)$$

De la expresión (4.11) se deduce que A va a resultar 1 cuando $x(m)$ y $x(m-1)$ tengan diferentes signos algebraicos, y va a ser 0 cuando $x(m)$ y $x(m-1)$ tengan el mismo signo. El análisis de cruces por cero es un indicador del comportamiento en frecuencia.

4.5.2.5 Curtosis

El parámetro Curtosis permite estudiar la deformación, en sentido vertical, respecto a la normal, de una distribución. Los valores de la señal se comparan con el valor de curtosis de la distribución de Gauss, que es 3. Un valor mayor implica una distribución con un pico más agudo, o sea los valores están concentrados en un rango pequeño. Un valor menor se corresponde con una distribución aplanada.

Para calcular el coeficiente de Curtosis, se utiliza la siguiente ecuación:

$$k = \frac{\frac{1}{n} \sum (x_i - \bar{x})^4 * n_i}{\left(\frac{1}{n} \sum (x_i - \bar{x})^2 * n_i \right)^2} \quad (4.12)$$

Donde k representa el coeficiente de Curtosis, x_i indica cada uno de los valores de la muestra, \bar{x} la media de la muestra, n_i la frecuencia de cada valor y n es la longitud total de la muestra.

Los resultados de esta fórmula se interpretan de la siguiente manera (figura 4.3):

- $k = 3 \pm 0.5$. La distribución es Mesocúrtica. Indica que la forma de la distribución se asemeja a una curva normal.
- $k > 3$. La distribución es Leptocúrtica. Indica que los datos se encuentran distribuidos en una zona pequeña, la forma de la curva es más “apuntada” que la Gaussiana.
- $k < 3$. La distribución es Platicúrtica. Los datos se encuentran distribuidos en forma uniforme. La curva es menos “apuntada” que la normal.

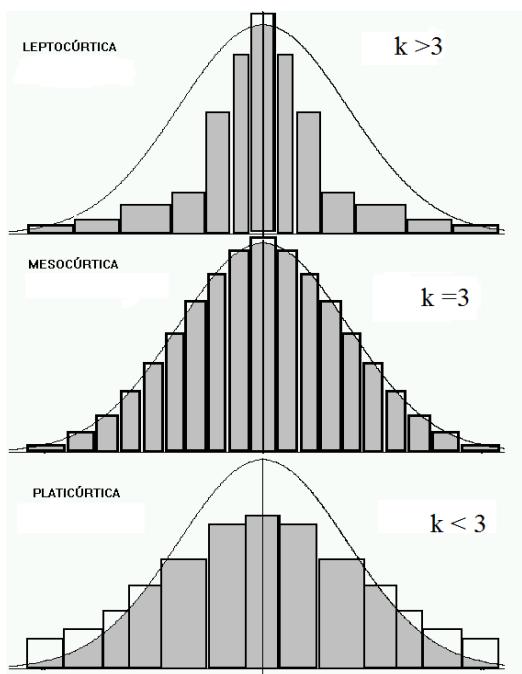


Figura 4.3 Curtosis. Comparación entre distribuciones de muestras no Gaussianas y la distribución de Gauss. (Obtenida de: <http://www.uv.es/ceaces/base/descriptiva/d38>)

4.5.3 Enfoque prosódico

Las características prosódicas de la señal de voz, también llamadas suprasegmentales [10], se refieren a los parámetros físicos que pueden medirse del lenguaje hablado y que aparecen en unidades fonéticas mayores a los segmentos individuales (por ejemplo una sílaba). Abarca la entonación, acentuación, duración y ritmo. Desde el punto de vista físico, las características suprasegmentales ocurren principalmente en la laringe y en la región subglotal. La entonación y el tono son controlados por los músculos de la laringe, mientras que la intensidad depende del flujo de aire emanado por los pulmones, que a su vez es controlado por los músculos respiratorios. Las características prosódicas pueden ser medidas mediante la energía, la duración y el pitch.

4.5.3.1 Energía

La Energía de la señal representada mediante la intensidad (como se menciona en la sección 4.4.2) se obtiene a través del RMS, el cual se utiliza comúnmente como estimador de la energía de una señal, por intervalos.

4.5.3.2 Duración

La duración, en lingüística, corresponde al tiempo que permanece un determinado segmento. Está descripta por la velocidad del habla y el ritmo. Un factor influyente en el ritmo es la combinación de las duraciones de las pausas y de los fonemas. Generalmente un locutor en estado de excitación tenderá a hablar rápidamente con menos pausas y más cortas, mientras que un locutor deprimido hablará lentamente, introduciendo pausas más largas.

4.5.3.3 Entonación

De acuerdo con [11], la entonación se define como la combinación de las estructuras tonales dentro de unidades estructurales más grandes asociadas con el parámetro acústico de la voz, denominado frecuencia fundamental (F0). La percepción de la entonación se denomina pitch. En otras palabras, la entonación es la variación del pitch durante el discurso. La producción de la entonación es regulada por las fuerzas musculares de la laringe que controlan la tensión de las cuerdas vocales, en conjunto con las fuerzas aerodinámicas del sistema respiratorio.

Todos los lenguajes usan la entonación. Por ejemplo para enfatizar, para transmitir sorpresa, para expresar emociones o para plantear una pregunta. En lenguajes tonales, como el chino, la entonación se usa para distinguir palabras [12].

4.5.3.3.1 Frecuencia Fundamental

La mayoría de los sonidos musicales, así como la voz tienen una estructura periódica, cuando son analizados en intervalos cortos de tiempo. Esos sonidos son percibidos por el sistema auditivo como un tono llamado pitch. El pitch [13], al igual que el volumen es un atributo subjetivo del sonido, que está relacionado con la frecuencia fundamental, que es un atributo de la forma de onda acústica.

La estimación de la F0 es también conocida como detección del pitch. Sin embargo se debe tener en cuenta que la F0 es una frecuencia y el pitch es un tono, dada esta distinción la determinación del pitch debe ser presentada en una escala tonal y no en una de frecuencia.

Hay un gran número de métodos que se utilizan para extraer la F0, basados en varios principios matemáticos. Si el análisis se realiza en intervalo de tiempo corto y consideramos periodicidad en la señal, la estimación del período de dicha oscilación debería ser suficiente para encontrar la frecuencia de oscilación (figura 4.4).

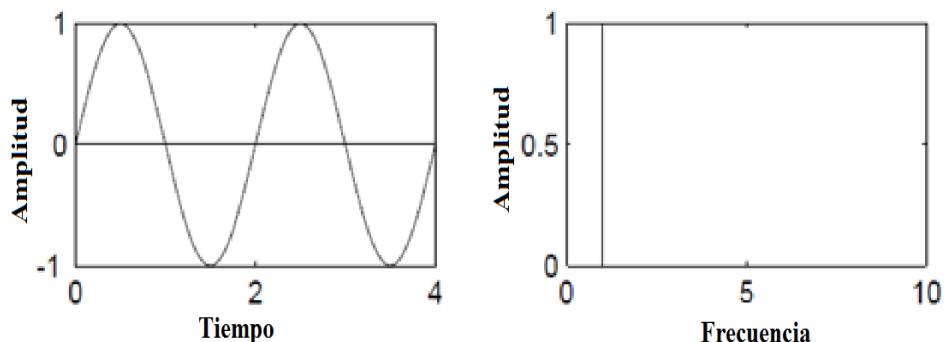


Figura 4.4 Estimación de la frecuencia fundamental en señal periódica. (Obtenida de Pitch Extraction and Fundamental Frequency: History and Current Techniques, Gerhard)

Sin embargo el problema es que la forma de onda no consiste en una simple sinusoida (figura 4.5). A medida que se añaden componentes armónicos a una onda sinusoidal, la determinación del pitch en la señal es más difícil determinar, por lo tanto se intenta estimar la F0.

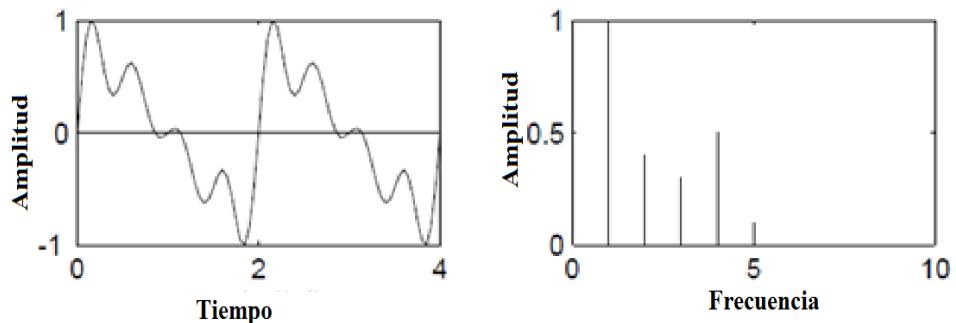


Figura 4.5 Estimación de la frecuencia fundamental en señal periódica. (Obtenida de Pitch Extraction and Fundamental Frequency: History and Current Techniques, Gerhard)

El objetivo de un estimador de la frecuencia fundamental es encontrar esa frecuencia sumergida en los componentes armónicos relacionados.

La dificultad de encontrar la F0 depende de la forma de onda en sí. Si la forma de onda tiene pocos armónicos o con una potencia baja, la F0 es más fácil de detectar.

Si los armónicos tienen más potencia que la F0, entonces el período es más difícil de detectar.

La detección de la frecuencia fundamental se puede realizar por diversos métodos.

Métodos de detección de la Frecuencia Fundamental

- **Análisis Cepstral.**

El análisis del Cepstrum [9] es una forma de análisis espectral donde la salida es la Inversa de la Trasformada de Fourier del logaritmo de la magnitud del espectro de una señal de entrada.

El nombre “Cepstrum” proviene de invertir las cuatro primeras letras de la palabra “spectrum”, que indica un espectro modificado. La variable independiente relacionada con la Transformada del Cepstrum ha sido llamada “quefrency”. Dado que esta variable está estrechamente relacionada con el tiempo, es aceptable referirse a ella como tiempo.

Esta teoría se basa en que la Transformada de Fourier de la señal de voz usualmente tiene un número de picos regularmente espaciados, representando el espectro armónico de la señal. Cuando se toma el logaritmo de la magnitud del espectro esos picos son reducidos, su amplitud es llevada a una escala utilizable. El resultado es una onda periódica en el dominio de la frecuencia, el periodo de esa onda (la distancia entre los picos) está relacionada con la frecuencia fundamental de la señal original. La inversa de la transformada de Fourier de esta onda tiene un pico en el periodo de la onda original. El método del cepstrum asume que la señal tiene la frecuencia de sus armónicos espaciados regularmente. Si este no es el caso, como en los sonidos sordos donde el espectro es inarmónico el método presentara resultados erróneos.

El Cepstrum se puede considerar como un operador matemático que transforma un producto en el dominio temporal en suma en el dominio frecuencial (Transformación Homomórfica). De esta forma puede separar las dos componentes de información de la señal a analizar: la excitación $u(n)$ y la respuesta en frecuencia del tracto vocal $h(n)$ de la ecuación (4.6).

La excitación es producida por las cuerdas vocales y determina la información prosódica de la señal. La respuesta en frecuencia, se corresponde con una determinada configuración del tracto vocal y representa las frecuencias de resonancia (formantes) de la misma.

La excitación se puede modelar mediante una señal con un periodo y amplitud determinada para los sonidos sonoros y mediante una fuente de ruido aleatorio para los sonidos sordos.

La respuesta en frecuencia se modela mediante un filtro, los parámetros de este filtro son los que determinan el timbre de la señal de voz.

El Cepstrum $c(x(n))$ de la señal $x(n)$ se puede representar mediante la siguiente ecuación:

$$c(x(n)) = F^{-1}(\log(|F(x(n))|)) \quad (4.14)$$

Donde $|F(x(n))|$ es la magnitud del espectro de la señal obtenido mediante la aplicación de la DFT (Transformada Discreta de Fourier) y la F^{-1} representa simbólicamente la Inversa de la DFT.

A la señal $x(n)$ expresada en la ecuación (4.6) como el producto del espectro de excitación ($u(n)$) y la respuesta frecuencial ($h(n)$) se le aplica el operador cepstral $c(x(n))$ que transforma el producto del dominio temporal en una suma en el dominio frecuencial.

$$c(x(n)) = c(u(n)) + c(h(n)) \quad (4.15)$$

La componente $c(u(n))$ permite obtener la información prosódica de la señal mientras que la componente $c(h(n))$ incluye información sobre las frecuencias de resonancia del tracto vocal. Con las componentes cepstrales aisladas, se puede utilizar la técnica del Lifting para determinar las características cepstrales.

La técnica del Lifting es similar a la operación de filtrado en el dominio de la frecuencia donde una región deseada se selecciona multiplicando el Cepstrum $c(x(n))$ por una ventana rectangular de longitud deseada.

Existen dos tipos de operación lifting:

- Lifting de tiempo bajo
- Lifting de tiempo alto

Técnica lifting de tiempo alto para la estimación de la FO

Las características de excitación se obtienen a partir del lifting de tiempo alto.

En primer lugar, se elige una ventana temporal:

$$wu(n) = \begin{cases} 1 & Lc \leq n \leq N/2 \\ 0 & 0 \leq n \leq Lc \end{cases} \quad (4.16)$$

Donde $wu(n)$ es la ventana de excitación de tiempo n , de longitud Lc . Por lo general Lc puede tomar valores entre 5 ms a 20 ms de la señal. En el presente trabajo se considerará $Lc=5$ ms, basado en la experiencia de [9].

$N/2$ es la mitad de la longitud total del Cepstrum. Se considera solo la mitad del total de la longitud del Cepstrum porque la Transformada del Cepstrum comparte las mismas propiedades que la Transformada de Fourier en cuanto a la propiedad de Simetría.

Las características de excitación se obtienen multiplicando la ventana temporal $wu(n)$ obtenida de (4.16) con el Cepstrum $c(n)$.

$$ce(n) = wu(n).c(n) \quad (4.17)$$

A partir del análisis de las componentes de excitación se puede estimar la frecuencia fundamental, que corresponde al pico más alto del Cepstrum.

Se analiza ventana a ventana y se determina el pico más alto del Cepstrum por ventana. Luego estos valores encontrados se multiplican por la frecuencia de muestreo. De esta manera, se encuentra el valor de frecuencia fundamental de cada ventana.

Este análisis solo puede ser determinado para señales sonoras. En el caso de las señales sordas estas técnicas no son aplicables porque no presentan periodicidad en la señal (figura 4.6).

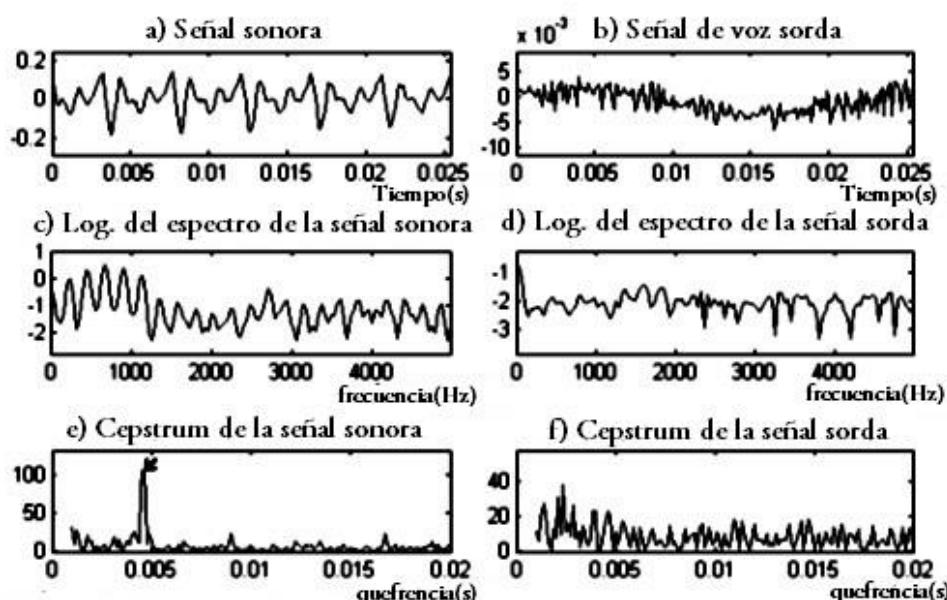


Figura 4.6 Detección de F0 mediante Técnica del Cepstrum. En la columna de la izquierda: Señal sonora, en la columna de la Derecha: Señal sorda. (Obtenida de Estimación de la curva de entonación para aprendizaje de segundo idioma, Schwartz)

Otro ejemplo es mostrado en la figura 4.7, donde se observa que la gráfica de la izquierda representa la secuencia del logaritmo del espectro de la señal y a la derecha su correspondiente Cepstrum. Las secuencias desde la 1 a la 5 de la gráfica de la izquierda corresponden a una señal sorda, los espectros para estos segmentos de voz presentan variaciones rápidas sin estructura periódica. Las secuencias 6 y 7 muestran parte de señal sonora y parte sorda. El resto de las secuencias muestran señales sonoras, sus espectros tienen una estructura periódica debido a la disposición armónica de los segmentos quasi-periódicos. Estas secuencias presentan un pico del Cepstrum en una quefrecuencia de aproximadamente 11-12 ms. La quefrecuencia del máximo global del

cepstrum es una estimación del periodo de la frecuencia fundamental durante el intervalo correspondiente [9].

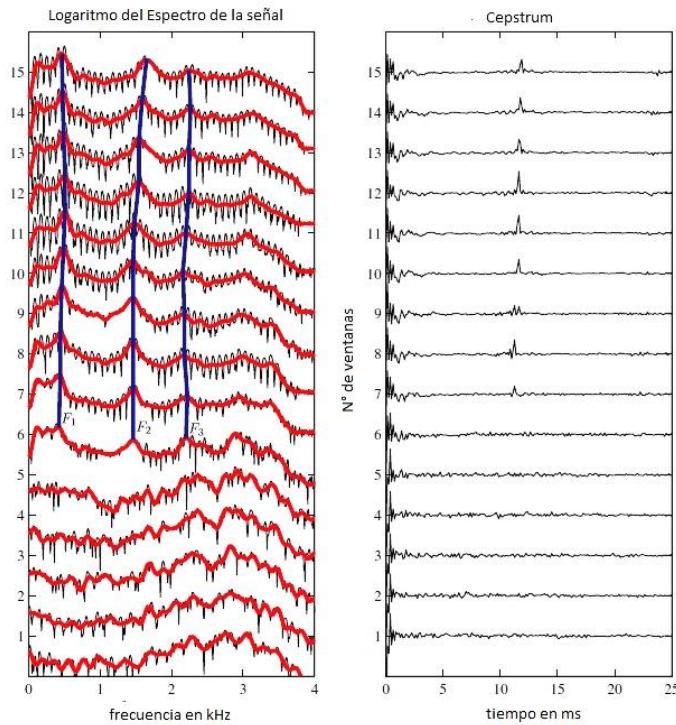


Figura 4.7 Logaritmo del espectro de la señal con su correspondiente envolvente y análisis cepstral. (Obtenida de Introduction to Digital Speech Processing, Rabiner)

4.5.4 Enfoque frecuencial

4.5.4.1 Formantes

Los formantes representan las frecuencias de resonancia que ocurren en el tracto vocal [7]. Dependen de las diferentes configuraciones que puede adoptar el tracto y su propia morfología, dependiente del sexo, la edad y patologías del hablante.

En los espectros de las vocales puede notarse la excitación periódica y la estructura armónica. Cada vocal tiene un conjunto de formantes que la caracterizan. En el idioma español, las vocales son fácilmente identificables debido a que son pocas y sus formantes están bien diferenciados.

Los formantes también pueden dar indicio de los estados emocionales. Se han encontrado resultados que muestran que el primer y segundo formante toman valores más altos que el promedio para emociones con alta excitación y valencia positiva, respectivamente [14].

Si se realiza la operación liftering de tiempo bajo se extraen características del tracto vocal en el dominio de la quefrency.

4.5.4.1.1 Método para extracción de formantes

A partir del análisis cepstral se pueden extraer características del tracto vocal en el dominio de la quefrency.

Aplicando el operador Cepstrum a la ecuación (4.6) se obtienen sus componentes aisladas. Luego mediante la aplicación de la técnica de Liftering de tiempo bajo se extraen las frecuencias de resonancia características del tracto vocal.

Técnica liftering de tiempo bajo para la estimación de formantes

Se elige una ventana temporal, que puede expresarse:

$$wh(n) = \begin{cases} 1 & 0 \leq n \leq Lc \\ 0 & Lc \leq n \leq N / 2 \end{cases} \quad (4.18)$$

Donde N es la longitud total del Cepstrum, $wh(n)$ es la ventana rectangular de la respuesta en frecuencia de tiempo n , de longitud Lc . Las características resonantes del tracto se obtienen multiplicando la ventana temporal ($wh(n)$) con el Cepstrum ($c(n)$).

$$ch(n) = wh(n).c(n) \quad (4.19)$$

A la secuencia del liftering cepstral de tiempo bajo obtenida ($ch(n)$), se le aplica la DFT. De esta manera se obtiene el espectro del tracto vocal.

Posteriormente, a fin de obtener la envolvente del espectro del tracto vocal se aplica un Filtro de Predicción Lineal (LPC).

La técnica de Predicción lineal (LPC) tiene como objetivo representar la envolvente espectral de una señal en una forma comprimida, utilizando la información de un modelo lineal.

Se fundamenta en establecer un modelo de filtro de tipo todos polos, para la fuente de sonido [15]. Este modelo permite describir la función de transferencia de un tubo (tracto vocal) formado por diferentes secciones.

Esta técnica consiste en estimar el valor actual de una señal $x(n)$ como una combinación lineal de las muestras anteriores.

El valor estimado $\hat{x}(n)$ se escribe como

$$\hat{x}(n) = -\sum_{k=1}^p a_k x(n-k) \quad (4.20)$$

Donde p es el orden de predicción y los a_k son los coeficientes de predicción. El problema básico de la predicción lineal consiste en determinar estos coeficientes de forma que la aproximación de $x(n)$ sea suficientemente buena.

El error entre el valor real $x(n)$ y el valor estimado $\hat{x}(n)$ se denomina error de predicción y viene dado por la expresión:

$$e(n) = x(n) - \hat{x}(n) = x(n) + \sum_{k=1}^p a_k x(n-k) \quad (4.21)$$

El filtro de predicción lineal utiliza el método de autocorrelación de modelado Autorregresivo (AR) para encontrar los coeficientes del filtro. Este filtro calcula la solución de mínimos cuadrados.

Del modelo de generación de señal que corresponde a la técnica de predicción lineal, se obtiene el espectro de la señal $S_{xx}(\omega)$.

$$S_{xx}(\omega) = \frac{See(\omega)}{|A(e^{j\omega})|^2} \quad (4.22)$$

Donde $See(\omega)$ es el espectro del error de predicción $e(n)$ y $A(e^{j\omega})$ es la respuesta frecuencial del filtro del error de predicción.

El modelado espectral asociada a la técnica de predicción lineal consiste en aproximar este espectro con la inversa del cuadrado del módulo de la respuesta frecuencial del filtro todos-polos $H(z)$.

$$\hat{S}_{xx}(\omega) = \frac{G^2}{|A(e^{j\omega})|^2} \quad (4.23)$$

Donde $\hat{S}_{xx}(\omega)$ es la aproximación de $S_{xx}(\omega)$.

Comparando las dos últimas ecuaciones se observa que el espectro del error $See(\omega)$ se modela por un espectro plano G^2 . Es decir, la señal $e(n)$ se aproxima por otra señal cuyo espectro es plano, como por ejemplo ruido blanco.

Si $S_{xx}(\omega)$ es el espectro de una señal de voz, $\hat{S}_{xx}(\omega)$ intenta aproximar la envolvente espectral. Es decir tenderá a aproximar $S_{xx}(\omega)$ alrededor de los picos de los formantes.

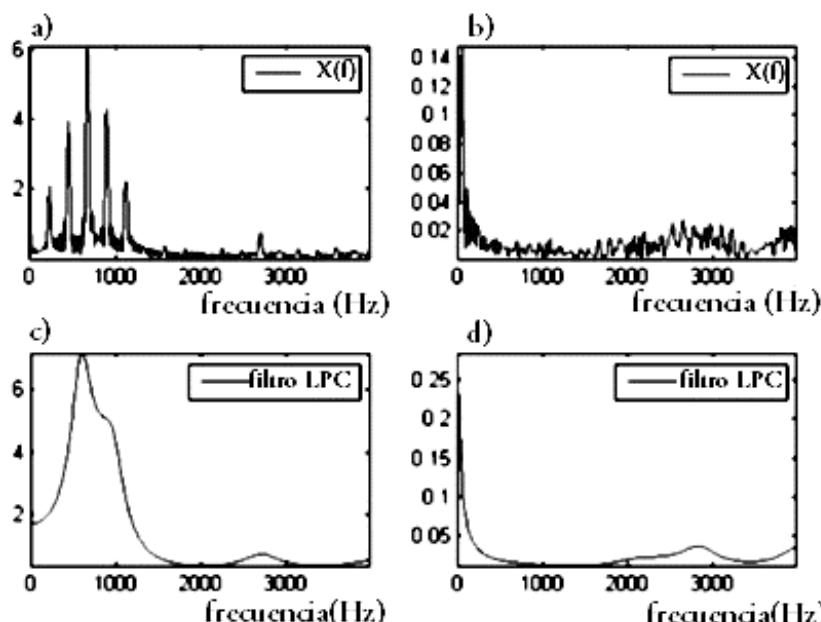


Figura 4.8 Comportamiento espectral del filtro LPC. a) espectro señal sonora, b) espectro señal sorda. c) y d) sus respectivas envolventes espectrales. (Obtenida de Estimación de la curva de entonación para aprendizaje de segundo idioma, Robles)

4.5.4.2 Coeficientes cepstrales en Frecuencia en la escala Mel (mfcc)

Los coeficientes cepstrales en frecuencia Mel (mfcc) son de gran utilidad en la extracción de los parámetros de la señal de voz.

Diversos experimentos muestran que la percepción de los tonos en los humanos no está dada en una escala lineal, esto hace que se trate de aproximar al comportamiento del sistema auditivo.

Los *mfcc* son una representación definida como el Cepstrum de una señal ventaneada en el tiempo, en una escala de frecuencias no lineal, las cuales se aproximan al comportamiento del sistema auditivo[9].

El cálculo de los *mfcc* utiliza una nueva escala, la escala de frecuencias Mel, que es no lineal y que permite representar la percepción de sistema auditivo.

Esta escala se construye equiparando un tono de 1000Hz y a 40dB por encima del umbral de audición del oyente, con un tono de 1000 mels. En consecuencia, cuatro octavas (una octava es el intervalo que separa dos sonidos cuyas frecuencias fundamentales tienen una relación de dos a uno) en la escala lineal de frecuencia medida en Hz se comprimen a alrededor de dos octavas en la escala Mel. En la escala Mel se asigna mayor relevancia a las bajas frecuencias (figura 4.9). Esta escala es diseñada mediante la construcción de un banco de filtros que permite la selección de bandas de frecuencia de la señal bajo análisis.

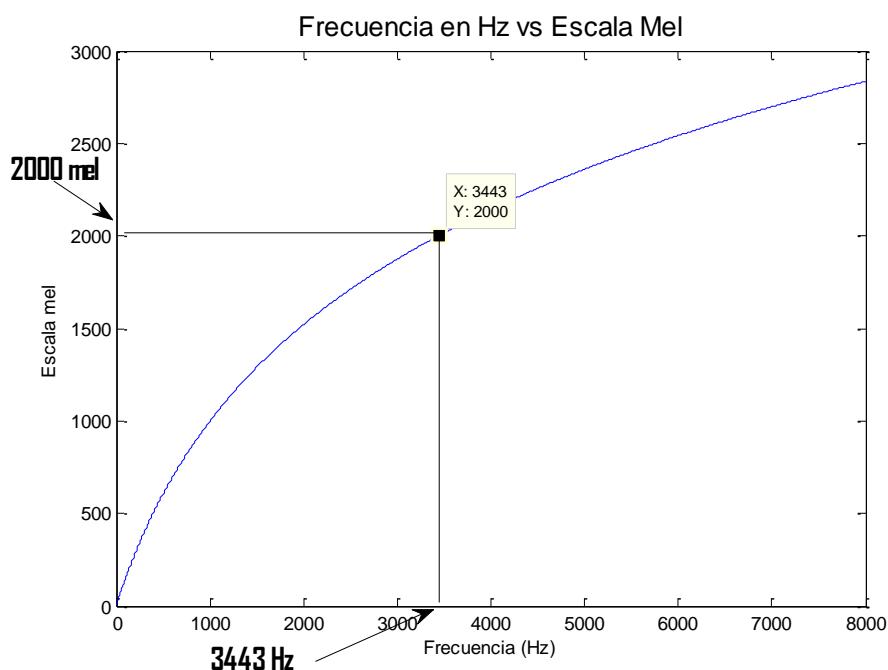


Figura 4.9. Relación entre la escala frecuencial en Hertz y la escala en frecuencias Mel.

Esta técnica además utiliza la Transformada de Fourier para la representación del contenido espectral de la señal. Con la Transformada de Fourier se conoce el contenido frecuencial de la señal y con los filtros diseñados se logra obtener las componentes de frecuencia que aporta cada banda de la señal analizada. El fin es ponderar la energía que aporta cada banda de frecuencia de la señal analizada y luego representarla mediante un número, llamado coeficiente cepstral.

4.5.4.2.1 Diseño de filtros

La obtención del espectro de la señal en la escala Mel requiere un muestreo en la escala Mel, esto puede realizarse a través de un banco de filtros pasabanda con un ancho de banda distribuido en la escala Mel sobre el rango de frecuencias deseado.

El banco de filtros triangulares linealmente espaciados en la escala Mel tiene la forma que se muestra en la (figura 4.10).

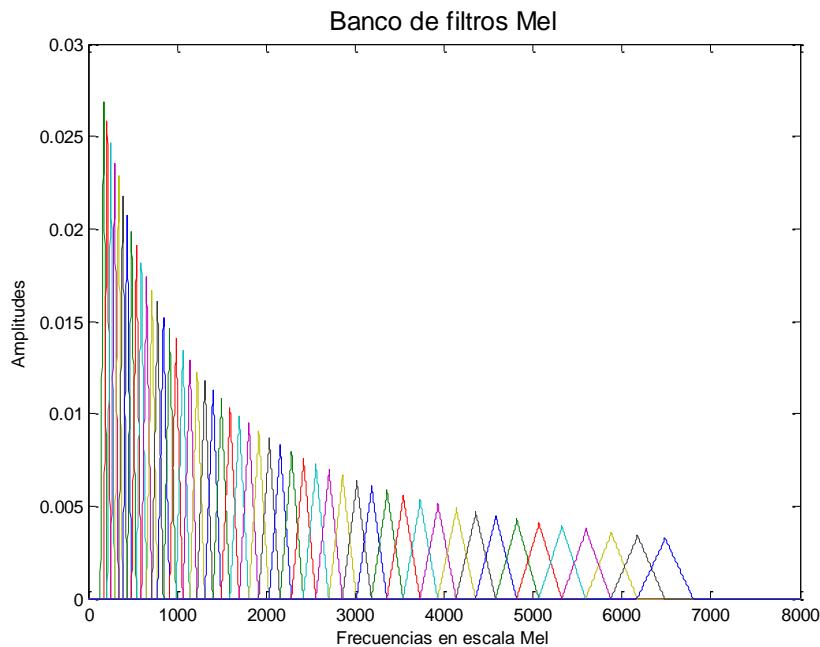


Figura 4.10. Banco de filtros en escala Mel.

4.5.4.2.2 Síntesis del Banco de Filtros

Los filtros triangulares son repartidos en el rango de frecuencias desde cero hasta la frecuencia de Nyquist.

El comportamiento del sistema acústico humano se aproxima mediante la función $\beta(f)$ que permite convertir frecuencias (Hercios) en frecuencias (Mel), teniendo en cuenta que la escala Mel, es lineal por debajo de los 1000Hz y logarítmica por encima de dicha frecuencia. La expresión de esta función, en un principio, se utilizó con el logaritmo decimal (4.24), sin embargo otras aplicaciones utilizan el logaritmo natural (4.25), con resultados similares [16]. En este trabajo se realizó la comparación de ambas ecuaciones y dada la similitud en los resultados, se eligió para la implementación la ecuación (4.24).

$$fmel = \beta(f) = 2595 \log_{10}\left(1 + \frac{f}{700}\right) \quad (4.24)$$

$$fmel = \beta(f) = 1125 \ln\left(1 + \frac{f}{700}\right) \quad (4.25)$$

Donde f es la frecuencia en Hercios y $fmel$ es la frecuencia obtenida en escala Mel. Las constantes que se observan en las ecuaciones 4.24 y 4.25 derivan de desarrollos matemáticos que permiten la aproximación más cercana a la escala Mel.

La ecuación 4.25 es aplicada para determinar la frecuencia de corte más baja (f_l) y la más alta (f_h) en escala Mel.

$$\beta(f_l) = 1125 \ln\left(1 + \frac{f_l}{700}\right) \quad (4.26)$$

$$\beta(fh) = 1125 \ln\left(1 + \frac{fh}{700}\right) \quad (4.27)$$

Siendo $\beta(f_l)$ la frecuencia de corte baja del banco de filtros en escala Mel y $\beta(fh)$ la frecuencia de corte alta del banco de filtros en escala Mel.

La función $f(m)$ es una función espectral espaciada en la escala Mel. Donde m es la variable independiente y M corresponde al número de filtros.

$$f(m) = \beta^{-1}(\beta(f_l) + \frac{\beta(fh) - \beta(f_l)}{M+1}) \quad (4.28)$$

Matemáticamente los filtros Mel se construyen con la siguiente función $Hm[m]$ que corresponde al banco de filtros.

$$Hm[m] = \begin{cases} 0 & k < f(m-1) \\ \frac{2(k-f(m-1))}{[f(m+1)-f(m-1)][f(m)-f(m-1)]} & f(m-1) \leq k \leq f(m) \\ \frac{2(f(m+1)-k)}{[f(m+1)-f(m-1)][f(m)-f(m-1)]} & f(m) \leq k \leq f(m+1) \\ 0 & k \geq f(m+1) \end{cases} \quad (4.29)$$

Para $1 \leq k \leq fs/2$ siendo fs la frecuencia de muestreo de la señal y $1 \leq m \leq M$.

La variable M , correspondiente al número de filtros, puede variar para diferentes implementaciones entre 24 a 40.

CMU Sphinx es un grupo de sistemas de software de reconocimiento de voz desarrollados en la Universidad de Carnegie Mellon. Se han presentado diferentes versiones del grupo Sphinx que incluyen decodificadores de voz, software de entrenamiento de modelos acústicos, diccionario de pronunciación, etc. [17] [18] [19]. En relación a los *mfcc*, este grupo de investigación indica que, para frecuencias de muestro de 16 KHz es óptimo el uso de una frecuencia mínima (f_l) de 130Hz, frecuencia máxima (fh) de 6800Hz y el cálculo de 40 filtros [20]. Este criterio fue utilizado para la implementación del banco de filtros desarrollado en el presente trabajo.

4.5.4.2.3 Cálculo de mfcc

La señal de audio es preprocesada con un filtro de pre énfasis. Este filtro tiene como objetivo compensar las atenuaciones producidas por:

- Procesos fisiológicos del mecanismo de producción del habla.
- Procesos de captura de la señal de voz. La adquisición de la señal de voz se realiza con un dispositivo de captura (micrófono) que se comporta como un filtro pasabajo.

El objetivo es enfatizar las componentes frecuenciales de alta frecuencia, o sea amplificar la zona del espectro (por encima de 1 KHz) sensible a la audición.

El filtro de preéñfasis es un filtro digital de primer orden, cuya expresión es la siguiente

$$y(n) = x(n) - \mu x(n-1) \quad (4.30)$$

Donde $y(n)$ denota la muestra actual de salida del filtro de preénfasis, $x(n)$ es la muestra actual de entrada, $x(n-1)$ es la muestra anterior de entrada y μ es una constante que puede tomar valores entre 0.9 y 1.

Posteriormente, la señal preprocesada se multiplica por una ventana de Hamming de 20ms con un desplazamiento entre ventanas de 10 ms.

Luego se calcula $X[k]$, que es la DFT de la señal $x(n)$, segmentada y discreta en el tiempo, con un número finito de N muestras.

$$X[k] = \sum_{n=0}^{N-1} x(n) e^{-\frac{j2\pi nk}{N}} \quad (4.31)$$

Siendo $k = 0, 1, 2, \dots, N-1$

Se calcula la Potencia Espectral.

$$|X[k]|^2 = S[k] = (\text{real}(X[k]))^2 + (\text{imag}(X[k]))^2 \quad (4.32)$$

Esta Potencia espectral $S[k]$ se multiplica por un banco de filtros triangulares espaciados de acuerdo a la escala de frecuencias Mel. El comportamiento espectral se distribuye en las bandas frecuenciales del banco de filtros (figura 4.11).

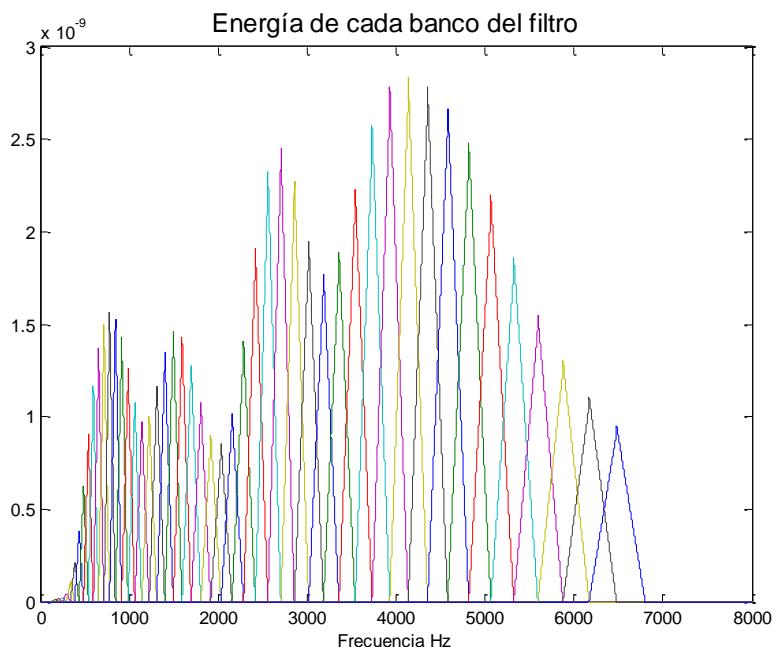


Figura 4.11. Espectro de las bandas frecuencias del banco de filtros

Con estos filtros se calcula el promedio del espectro alrededor de cada frecuencia central. El objetivo es encontrar el espectro normalizado de la energía que aporta cada banda de frecuencia de la señal.

$$Sn(m) = \frac{1}{Am} \sum_{k=0}^{N/2} S[k]Hm(k) \quad (4.33)$$

Siendo $Sn (m)$ el espectro de m bandas frecuenciales mel, $S[k]$ la potencia espectral de la señal $x(n)$, $Hm[k]$ la función del banco de filtros y k es la variable independiente que varía entre 0 y $N/2$, siendo N la longitud de la DFT. Am es el factor de normalización de los m -ésimos filtros mel y se expresa a partir de la siguiente ecuación.

$$Am = \sum_{k=0}^{N/2} |Hm(k)|^2 \quad (4.34)$$

Aplicando el logaritmo natural a la ecuación (4.34) se obtiene $Sm (m)$ que es el logaritmo de la energía $Sn (m)$ de la salida de cada filtro m .

$$Sm(m) = \ln\left(\frac{1}{\sum_{k=0}^{N/2} |Hm(k)|^2} \sum_{k=0}^{N/2} S[k]Hm(k)\right) \quad (4.35)$$

Finalmente, los $mfcc(n)$, con n correspondiente al número de coeficientes mel, se obtienen aplicando la Transformada del coseno a $Sm (m)$, que es la salida del logaritmo del espectro de energía de las m bandas frecuenciales de interés, para un banco de M número de filtros.

$$mfcc(n) = \frac{1}{M} \sum_{m=1}^M Sm(m) \cos\left[\frac{2\pi}{M}(m+0.5)n\right] \quad (4.36)$$

En la figura 4.12 se muestra el resultado de la implementación del algoritmo que permite obtener los $mfcc$ de una señal de voz. En este caso se calcularon solo 13 coeficientes.

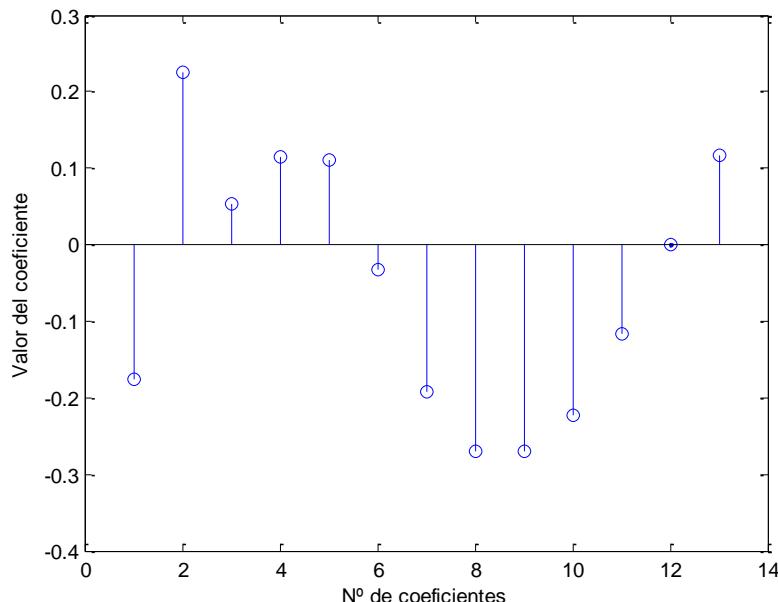


Figura 4.12. Coeficientes cepstrales en frecuencia Mel.

4.5.4.3 Densidad Espectral de Potencia

Considerando la cuasi estacionariedad de la señal, es factible la aplicación de técnicas de estimación espectral no paramétricas.

La elección de la técnica de estimación espectral depende de la cantidad de muestras disponibles. Si el número de muestras es suficiente, se utilizan técnicas no paramétricas. En este caso, la cantidad de muestras disponibles satisface los requerimientos para la utilización de dichas técnicas.

Dentro de las técnicas de estimación espectral no paramétricas se encuentra el Periodograma de Welch. El Periodograma de Welch consiste en dividir la señal en segmentos. A cada segmento de la señal se le aplica la DFT. Posteriormente se promedian las DFTs de los diferentes segmentos.

Analíticamente se puede calcular de la siguiente forma:

Dividir la señal en segmentos $x_w(n)$. Esto implica matemáticamente multiplicar la señal de entrada $x(n)$ por una ventana temporal $w(n)$. En este trabajo, se utilizó una ventana de Hamming. La elección surge de la buena resolución frecuencial (capacidad de diferenciar componentes frecuenciales cercanos entre sí) que ofrece.

$$x_w(n) = x(n).w(n) \quad (4.37)$$

1. Se obtiene la Transformada de Fourier $X_N(k)$ de cada segmento $x_w(n)$ de la señal $x(n)$.

$$X_N(k) = \sum_{n=0}^{N-1} x_w(n) e^{-j2\pi \frac{kn}{N}} \quad (4.38)$$

Donde N es el tamaño de la DFT. Para $0 \leq k \leq N-1$.

2. Se calcula $P_{xx}(k)$, que es la Densidad espectral de cada segmento $x_w(n)$ aplicando la siguiente ecuación (4.39).

$$P_{xx}(k) = \frac{T_s}{N} |X_N(k)|^2 \quad (4.39)$$

Siendo $X_N(k)$ la Transformada de Fourier de cada segmento $x_w(n)$, N el tamaño de la DFT y T_s el periodo de muestreo.

3. Luego se promedian los $P_{xx}(k)$ de cada segmento desde $p=1$ hasta P (cantidad total de segmentos).

$$P_{Welch} = \frac{1}{P} \sum_{p=1}^P P_{xx}(k) \quad (4.40)$$

El periodograma de Welch permite reducir la varianza de la DEP (Densidad Espectral de Potencia) estimada, por lo tanto disminuye la dispersión, lo que se traduce en una reducción de picos espúreos en el espectro.

En la figura 4.13 se muestra la DEP obtenida mediante el periodograma de Welch de un segmento de la señal de voz.

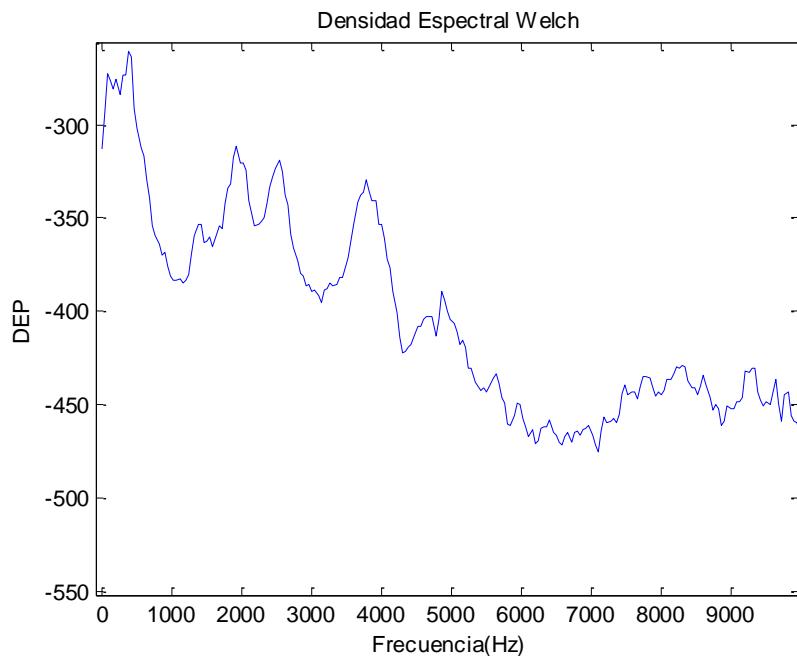


Figura 4.13. Densidad Espectral de Potencia mediante Periodograma de Welch.

4.6 Referencias

1. Oppenheim,A.; Willsky,A. (1998). *Señales y Sistemas*.2da. Edición en español. Editorial Prentice Hall Hispanoamérica S.A. ISBN 970-17-0116-X.
2. Proakis, J.G.; A Manolakis, D.G. (2007). *Digital Signal Processing*. 4ta Edicion. Editorial Pearson Prentice Hall. ISBN 0132287315.
3. Fant, G. (1971). *Acoustic theory of speech production: with calculations based on X-ray studies of Russian articulations* (Vol. 2). Walter de Gruyter.
4. Markel, J. D., & Gray, A. J. (2013). *Linear prediction of speech* (Vol. 12). Springer Science & Business Media
5. Rabiner, L. R. (1968). *Digital-Formant Synthesizer for Speech-Synthesis Studies*. *The Journal of the Acoustical Society of America*, 43(4), 822-828.
6. Fry,D. B.(1979). *The Physics of Speech*. Cambridge Textbooks in linguistics, QP306.F8.
7. Navarro Mesa,J. *Procesador acústico: El bloque de extracción de características*. <http://www2.ulpgc.es/hege/almacen/download/25/25296/apuntesextraccioncaracteristicas.pdf>
8. Mahía R. (1999). *Revisión de los procedimientos de análisis de estacionariedad de las series temporales*. Tesis doctoral. Universidad Autónoma de Madrid.
9. Rabiner L. R. y. Schafer R. W. (2007). "Introduction to Digital Speech Processing". *Foundations and Trends® in Signal Processing* Vol. 1, Nos. 1–2.
10. Albornoz E.; Milone D.; Rufiner H. (2011). "Modelado de estructuras prosódicas para el reconocimiento automático del habla". Tesis doctoral. Universidad Nacional del Litoral.
11. Botinis, A.; Granström, B.; Möbius, B. (2001). "Developments and paradigms in intonation research," *Speech Communication*, vol. 33, no. 4, pp. 263-296.
12. Arias, J.P.; Yoma, N.B.; Vivanco, H. (2009)."Automatic intonation assessment for computer aided language learning".

13. Gerhard, D.(2003). *Pitch Extraction and Fundamental Frequency: History and Current Techniques Technical Report TR-CS 2003-06*
14. Goudbeek,M.; Goldman,J. ; Scherer, K. (2009). "Emotion dimensions and formant position," in *Interspeech 2009*, Brighton, UK, pp. 1575-1578.
15. Pericas,F. ; Camprubi, C. (1993).*Técnicas de procesado y representación de la señal de voz para el reconocimiento del habla en ambientes ruidosos*. Universidad Politécnica de Cataluña.
16. Zheng, F., Zhang, G., & Song, Z. (2001). *Comparison of different implementations of MFCC*. *Journal of Computer Science and Technology*, 16(6), 582-589.
17. Lee, K. F., Hon, H. W., & Reddy, R. (1990). *An overview of the SPHINX speech recognition system*. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 38(1), 35-45.
18. Huang, X., Alleva, F., Hon, H. W., Hwang, M. Y., Lee, K. F., & Rosenfeld, R. (1993). *The SPHINX-II speech recognition system: an overview*. *Computer Speech & Language*, 7(2), 137-148.
19. Lamere, P., Kwok, P., Walker, W., Gouvêa, E. B., Singh, R., Raj, B., & Wolf, P. (2003, September). *Design of the CMU sphinx-4 decoder*. In *INTERSPEECH*.
20. S. Young, G. Evermann, M. Gales et al. *The HTK Book*. Copyright 2001-2005 Cambridge University Engineering Departament, 2005.

CAPITULO 5

Desarrollo de un segmentador de voz

5.1 Introducción

La detección de la actividad de voz (VAD, *Voice Activity Detection*) en aplicaciones de reconocimiento de voz ha tomado relevancia en estas últimas décadas. El problema de determinar su presencia cuando la relación señal a ruido (SNR) es baja ha llevado a motivar a los centros de investigación a desarrollar algoritmos detectores más robustos, ya que demasiados segmentos de no-voz (SNV) erróneamente clasificados como discurso, pueden alterar los modelos acústicos y reducir la precisión del reconocimiento de voz [1]. Según el entorno de adquisición, los SNV pueden ser silencio, ruido o algunas otras señales que no sean de interés tales como papeleo, voces de fondo, puertas, ruidos ambientales propios del medio, etc.

Los VAD constan de dos etapas fundamentales [2]. La primera etapa consiste en la extracción de un conjunto de características de la señal y en la segunda etapa se desarrolla el algoritmo de decisión entre los estados voz o no-voz. Las características de la señal pueden extraerse de diferentes dominios: de dominio espectral, de dominio cepstral (como el pitch), o del dominio temporal como la energía o parámetros estadísticos.

Las características basadas en la energía determinan algoritmos de decisión simples y fáciles de implementar mientras que las características en el dominio espectral o cepstral son robustas a señales con baja SRN [2].

Los mecanismos que incluyen la segunda etapa del detector, suelen basarse en umbrales, modelización estadística o algoritmos de aprendizaje.

En una señal poco ruidosa o que tenga una alta SNR, la detección puede ser implementada usando simplemente como característica la energía y como sistema de decisión la umbralización. Sin embargo cuando la señal se encuentra afectada por ruido, puede ser muy difícil distinguir entre segmentos de voz (SV) y SNV.

En la codificación de voz, las VAD ayudan a evitar la codificación y transmisión innecesaria de SNV, de esta manera se ahorran costos de ancho de banda y cálculo de procesamiento. Las VAD también se han usado para la reducción de ruido en dispositivos de audífonos digitales [3]. En otros casos [4], la VAD se refería como la detección del punto final del habla.

Otros desarrollos de identificadores de SV [5], utilizan el pitch como característica de la primera etapa del detector. En este caso, el sistema consiste en la aplicación de un pre-procesado basado en descomposición empírica de modos. Luego, mediante la evaluación de la periodicidad de la señal se determinan los SV o los SNV.

Otras investigaciones basan su VAD en algoritmos basados en la entropía de la señal [6]. Cada señal es analizada en ventanas, se suaviza mediante un filtro de mediana y posteriormente, se suman los valores de entropía espectral. Se toman umbrales para detectar comienzo y finalidad de cada SV.

Existen otros métodos de VAD que basan sus algoritmos en una estimación no paramétrica del espectro del ruido que utiliza estadísticas mínimas [7]. Su fundamentación se basa en que dada la

naturaleza no estacionaria de la señal de voz, la energía del habla en una banda de frecuencia determinada es probable que caiga a cero en un intervalo posterior. Cuando la energía del habla cae a cero, la energía de la señal es el ruido de fondo.

En el presente trabajo, se utilizaron técnicas estadísticas basadas en el lema de Neyman Pearson intensamente aplicadas en la detección de blancos mediante radares [8] [9] [10] [11]. El sistema de detección trata de maximizar la probabilidad de detección, manteniendo la probabilidad de falsa alarma igual o inferior a un valor determinado. La regla de decisión utiliza un umbral simple fijado por los requisitos de la probabilidad de falsa alarma determinados por el lema de Neyman Pearson. Estos algoritmos fueron acondicionados para el tratamiento de las señales de audio y la detección de los SNV y SV.

5.2 Técnicas estadísticas. Lema de Neyman Pearson

Una regla de decisión estadística puede derivar de la prueba de razón de verosimilitud.

El método considera dos hipótesis:

- H_1 : que indica que el evento que deseamos detectar ha sucedido y
- H_0 : que indica que el evento no ha sucedido.

A veces H_1 se conoce como hipótesis de la señal y H_0 como hipótesis del ruido de base.

Definimos dos sucesos:

- La probabilidad de detección P_d se puede expresar como la probabilidad de que sucedió el evento siendo que efectivamente ha sucedido.

$$P_d = p(H_1 | H_1) \quad (5.1)$$

- La probabilidad de falsa alarma P_{fa} se puede expresar como la probabilidad de detectar el evento siendo que no ha sucedido.

$$P_{fa} = p(H_1 | H_0) \quad (5.2)$$

En función del problema a resolver se elige la optimización de uno de estos sucesos.

En este trabajo, donde se requiere encontrar un umbral óptimo para la detección de SV se priorizan las técnicas estadísticas que permiten maximizar la P_d y minimizar la P_{fa} . Bajo esta premisa, se describe el lema fundamental de Neyman Pearson.

5.2.1 Lema de Neyman Pearson

En estadística, el Lema de Neyman Pearson es un resultado que describe el criterio óptimo para distinguir dos hipótesis [12] [13].

Supongamos que disponemos de una única observación x que bajo H_1 está distribuida como $f(x | H_1)$ y bajo H_0 como $f(x | H_0)$. Consideramos también un test de decisión:

- decide H_1 cuando $x \in R_1$ (el subíndice “1” hace referencia a que es una región que contiene a las muestras x que cumplen con la hipótesis H_1).
- decide H_0 cuando $x \in R_0$ (el subíndice “0” hace referencia a que es una región que contiene a las muestras x que cumplen con la hipótesis H_0).

Siendo $R_0 \cap R_1 = \emptyset$ y $R_0 \cup R_1 = \mathbb{R}$ (conjunto de números Reales).

Teniendo en consideración (5.1) y (5.2). El Criterio de Neyman Pearson es:

Maximizar P_d sujeto a P_{fa}

Siendo la P_{fa} menor o igual a α , parámetro prefijado que suele tomar valores inferiores a 0.1.

La P_d en términos de su distribución se puede expresar como:

$$P_d = \int_{R_1} f(x | H_1) dx \quad (5.3)$$

La P_{fa} se define en función de la P_d mediante la siguiente ecuación.

$$P_{fa} = 1 - P_d = 1 - \int_{R_1} f(x | H_1) dx \quad (5.4)$$

La P_{fa} en términos de su distribución se puede reescribir de la siguiente manera.

$$P_{fa} = \int_{R_0} f(x | H_1) dx = \int_{R_1} f(x | H_0) dx \quad (5.5)$$

Para resolver el problema de optimización, se aplica el método de los multiplicadores de Lagrange que considera una función auxiliar y la minimiza [14]. Este método se basa en encontrar los máximos y mínimos de funciones de múltiples variables sujetas a restricciones. Este método reduce el problema restringido con n variables a uno sin restricciones de $n+k$ variables, donde k es el número de restricciones y cuyas ecuaciones pueden ser resueltas fácilmente. Estas nuevas variables desconocidas, una para cada restricción son llamadas multiplicadores de Lagrange. El método dice que los puntos donde la función tiene un extremo condicionado con k restricciones, están entre puntos estacionarios de una nueva función sin restricciones, construida como una combinación lineal de la función y las funciones implicadas en las restricciones, cuyos coeficientes son los multiplicadores.

Se minimiza la siguiente función auxiliar:

$$J = (1 - P_d) + \lambda(P_{fa} - \alpha) \quad (5.6)$$

Donde $\lambda > 0$ representa el multiplicador de Lagrange y el umbral a calcular y α es una constante que limita la P_{fa} .

La función a minimizar se puede escribir como:

$$J = 1 - \int_{R_1} f(x | H_1) dx + \lambda(P_{fa} - \alpha) \quad (5.7)$$

$$J = \int_{R_0} f(x | H_1) dx + \lambda(\int_{R_1} f(x | H_0) dx - \alpha) \quad (5.8)$$

$$J = \int_{R_0} f(x | H_1) dx + \lambda(1 - \int_{R_0} f(x | H_0) dx - \alpha) \quad (5.9)$$

Agrupando las integrales de (5.9):

$$J = \lambda(1 - \alpha) + \int_{R_0} f(x | H_1) - \lambda(f(x | H_0)) dx \quad (5.10)$$

Si llamamos A a la integral de (5.10):

$$A = \int_{R_0} f(x | H_1) - \lambda(f(x | H_0)) dx \quad (5.11)$$

Para minimizar J , A debe ser lo más negativa posible.

Luego, se elige H_0 si la función de distribución de eventos no detectados es superior a la función de distribución de eventos detectados, tal como se representa en la ecuación (5.12).

$$\lambda f(x | H_0) > f(x | H_1) \quad (5.12)$$

Se elige H_1 si la función de distribución de eventos detectados es superior a la función de distribución de eventos no detectados.

$$\lambda f(x | H_0) < f(x | H_1) \quad (5.13)$$

Relacionando 5.12 y 5.13, se establece el cociente de Verosimilitudes y la regla de decisión clásica de Neyman Pearson.

$$\frac{f(x | H_1)}{f(x | H_0)} > \lambda \quad \text{Elijo } H_1 \quad (5.14)$$

$$\frac{f(x | H_1)}{f(x | H_0)} < \lambda \quad \text{Elijo } H_0 \quad (5.15)$$

En este trabajo, el umbral λ se elige con la premisa de que la P_{fa} , que depende de H_0 , sea igual a la constante α (5.16).

$$P_{fa} = \int_{R_1} f(x | H_0) dx = \alpha \quad (5.16)$$

La regla de decisión no se desarrolla a partir del cociente de verosimilitudes sino que se establece según la aplicación binaria simple del umbral, pero determinando el mismo a partir del lema de Neyman Pearson.

5.2.2 Aplicación del lema de Neyman Pearson

En este trabajo, se aplica el lema de Neyman Pearson para determinar un umbral tal que permita separar los SV de los SNV.

El método para el cálculo de este umbral se desarrolla en función de la condición planteada en la ecuación (5.16).

Se considera el siguiente test binario de hipótesis:

- H_1 : indica que el evento que deseamos detectar ha sucedido. En nuestro caso, el evento a detectar es el SV.
- H_0 : indica que el evento no ha sucedido, lo cual corresponde a un SNV.

Se asume que el SNV de la señal tiene una Distribución Normal por la naturaleza del ruido de base que la compone [15]. A este SNV se le puede extraer su media, y desviación estándar.

La media A_0 de la señal x_n del SNV de longitud N de n muestras, se puede calcular según la siguiente ecuación.

$$A_0 = \frac{1}{N} \sum_{n=1}^N x_n \quad (5.17)$$

La desviación estándar σ del SNV se obtiene a partir de la siguiente ecuación.

$$\sigma = \sqrt{\frac{1}{N} \sum_{n=1}^N x_n^2 - A_0^2} \quad (5.18)$$

Donde x_n es la señal del SNV de longitud N de n muestras y A_0 es la media de la muestra.

A fin de observar cómo se distribuye la muestra bajo la hipótesis H_0 , se obtiene la función de densidad de probabilidad.

La función de densidad de probabilidad del SNV la podemos expresar como:

$$f(x | H_0) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-A_0)^2}{2\sigma^2}} \quad (5.19)$$

La P_{fa} es la probabilidad de detectar SV cuando no hay, lo cual corresponde, según la ecuación (5.4), a la probabilidad de detectar un evento no ocurrido (H_0). Esto se expresa en la siguiente ecuación.

$$P_{fa} = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-A_0)^2}{2\sigma^2}} dx \quad (5.20)$$

La P_{fa} se puede expresar en términos de la función de error complementario ($erfc$).

Matemáticamente, la $erfc$ se define a partir de la función de error (erf), que es una función no elemental que se utiliza en el campo de la probabilidad.

La erf de la muestra x se define de la siguiente manera:

$$\operatorname{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt \quad (5.21)$$

La erf no puede ser evaluada en términos de funciones elementales, pero se puede obtener integrando la expansión en serie de Taylor de e^{-x^2} .

$$\operatorname{erf}(x) = \frac{2}{\sqrt{\pi}} \sum_{n=0}^{\infty} \frac{(-1)^n x^{2n+1}}{n!(2n+1)} \quad (5.22)$$

La erfc se define como:

$$\operatorname{erfc}(x) = 1 - \operatorname{erf}(x) \quad (5.23)$$

Modificando los límites de la integral de la ecuación 5.21 se obtiene erfc .

$$\operatorname{erfc}(x) = \frac{2}{\sqrt{\pi}} \int_x^{\infty} e^{-t^2} dt \quad (5.24)$$

Expresamos la P_{fa} en términos de $\operatorname{erfc}(x)$.

$$P_{fa} = \operatorname{erfc}\left(\frac{\lambda - A_0}{\sigma}\right) = \alpha \quad (5.25)$$

Siendo λ el umbral, A_0 la media del SNV, σ la desviación estándar del SNV, y α el parámetro que limita la P_{fa} .

λ se obtiene despejando de (5.25).

$$\lambda = \sigma \cdot \operatorname{erfc}^{-1}(\alpha) + A_0 \quad (5.26)$$

Donde erfc^{-1} corresponde a la función de error complementario inversa de α , que es la constante prefijada.

En Matlab, la $\operatorname{erfc}^{-1}(x)$ se obtiene a partir de la función de error inversa (erf^{-1}), descripta en la siguiente ecuación.

$$\operatorname{erf}^{-1}(x) = \sum_{k=0}^{\infty} \frac{c_k}{2k+1} \left(\frac{\sqrt{\pi}}{2} x\right)^{2k+1} \quad (5.27)$$

Donde $0 \leq k \leq \infty$ y c_k es:

$$c_k = \sum_{m=0}^{k-1} \frac{c_m c_{k-1-m}}{(m+1)(2m+1)} \quad (5.28)$$

Se cumple que cuando $k=0$, entonces $c_k = 1$

Entonces, la $\operatorname{erfc}^{-1}(x)$ de la muestra x , se puede obtener de la siguiente manera:

$$\operatorname{erfc}^{-1}(x) = \sqrt{2} * \operatorname{erf}^{-1}(1 - 2 * \alpha) \quad (5.29)$$

Siendo α la constante prefijada.

Agrupando las ecuaciones (5.26) y (5.29) se obtiene la ecuación del umbral λ .

$$\lambda = \sigma(\sqrt{2} * \operatorname{erf}^{-1}(1 - 2 * \alpha)) + A_0 \quad (5.30)$$

Donde σ es la desviación estándar del SNV, A_0 es la media del SNV y α es la constante prefijada según el criterio de la ecuación (5.16).

5.3 Desarrollo del segmentador de voz

El segmentador de voz desarrollado en este trabajo utiliza el lema de Neyman Pearson para definir un umbral dinámico que se ajusta a los parámetros estadísticos del SNV. Luego se realiza una umbralización que permite identificar SV de SNV.

5.3.1 Esquema del segmentador

La figura 5.1 muestra un esquema básico del segmentador. Cada bloque del diagrama es detallado a continuación.

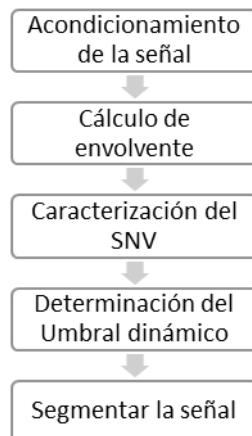


Figura 5.1. Esquema Básico del segmentador

5.3.1.1 Acondicionamiento de la señal

El acondicionamiento de la señal de audio consiste en un re-muestreo inicial a 16KHz, normalización y la posterior limitación en banda frecuencial. La limitación en banda frecuencial se realiza mediante un filtrado pasabanda recursivo Chebyshev tipo 1 con frecuencia de corte inferior en 20Hz y frecuencia de corte superior en 6800 Hz. El filtrado es fijado mediante el criterio de CMU Sphinx [16] y mediante la aplicación de análisis tiempo frecuencia mostrado en la figura 5.2, la cual permite observar que el mayor contenido frecuencial de una muestra de señal de voz tomada aleatoriamente de la Base de datos Berlín, se encuentra por debajo de los 5Khz. (El acondicionamiento de la señal será abordado con mayor detalle en el capítulo 6).

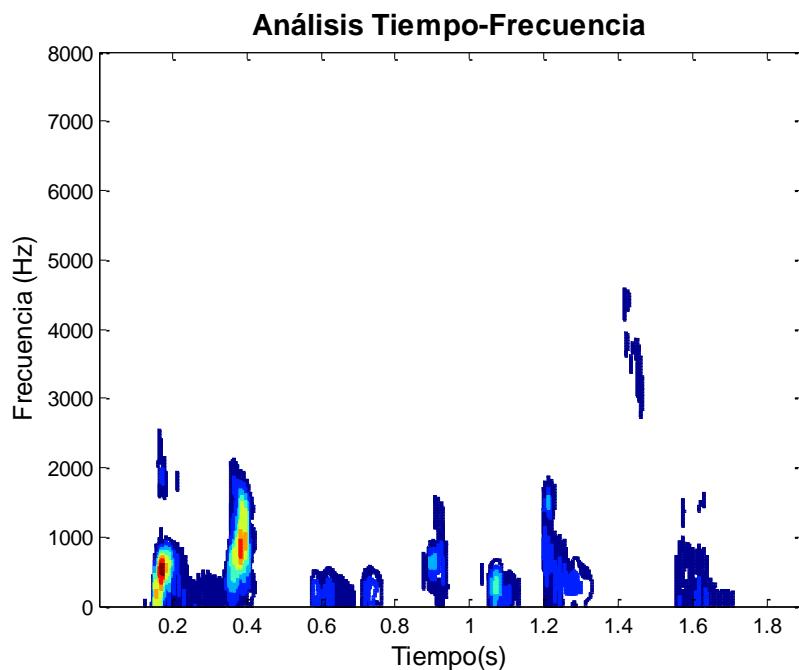


Figura 5.2 Espectrograma

En la figura 5.3 se exhibe el esquema del acondicionamiento de la señal.

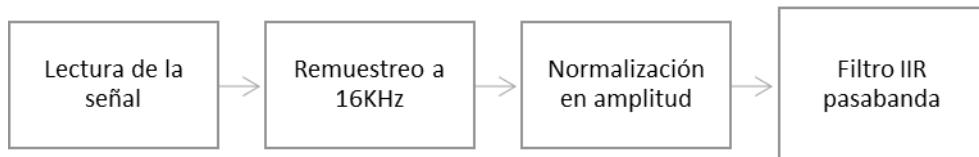


Figura 5.3. Esquema acondicionamiento de la señal

5.3.1.2 Cálculo de envolvente

El procedimiento para obtener la envolvente de la señal consta de la rectificación de la misma y la posterior aplicación de un filtrado IIR pasabajo Chebyshev tipo 1 con frecuencia de corte en 12 Hz. El esquema se presenta en la figura 5.4.



Figura 5.4. Esquema del cálculo de la envolvente

5.3.1.3 Caracterización del SNV

Al SNV se lo puede caracterizar mediante parámetros, tales como A_0 (5.17), σ (5.18), amplitud del SNV (A_{SNV}) y el valor cuadrático medio del voltaje del SNV (P_{SNV}), que representa la potencia del SNV normalizado y se calcula según la siguiente ecuación.

$$P_{SNV} = \sqrt{\frac{1}{T} \int_0^T V_{SNV}^2(t) dt} \quad (5.31)$$

Donde la integral del voltaje del SNV (V_{SNV}) corresponde a la energía de SNV que se estima mediante el RMS y T representa el intervalo de tiempo t del SNV.

5.3.1.4 Determinación del Umbral dinámico

Para realizar el cálculo del umbral dinámico primero se determina el valor del factor α , que es la constante que permite fijar la P_{fa} según la ecuación (5.16). Matemáticamente, α se puede calcular según la siguiente expresión obtenida de [17].

$$\alpha = e^{-\frac{A_{SNV}^2}{2P_{SNV}}} \quad (5.32)$$

Siendo A_{SNV} la amplitud del SNV y P_{SNV} el valor cuadrático medio del voltaje del SNV.

El desarrollo de la ecuación del umbral dinámico consta de dos etapas. En la primera etapa se calcula un parámetro constante (k_{SNV}) que constituye uno de los términos de la ecuación del umbral dinámico y en la segunda etapa se determina el parámetro variable (d_{SNV}) que forma parte del segundo término de la ecuación del umbral.

Etapa 1. Cálculo de k_{SNV}

El parámetro k_{SNV} se obtuvo seleccionando los primeros 130ms de una muestra, tomada en forma aleatoria, de la Base de datos Berlín [18]. Luego, se calcularon los parámetros descriptos en la sección 5.3.1.3 que caracterizan la señal.

k_{SNV} se obtuvo mediante el uso de la ecuación (5.30) que se expresa modificada en la siguiente ecuación.

$$k_{SNV} = \sigma * erfc^{-1}(\alpha) + A_0 \quad (5.33)$$

Donde σ es la desviación estándar de la muestra, A_0 es la media de la muestra y $erfc^{-1}(\alpha)$ es la función de error complementario inversa de α , que es la constante calculada en (5.32).

Esta constante k_{SNV} fue utilizada en primer lugar como el valor del umbralizador que permitía diferenciar los SV de los SNV para la aplicación en las señales de la base de datos Berlín. Sin embargo,

cuando la aplicación se realizaba en señales de otras bases de datos o en señales adquiridas en el laboratorio, la detección no era eficiente, por este motivo se procedió al cálculo del término d_{SNV} , que permitiría al detector ser dependiente de la señal analizada, lo cual le confiere carácter dinámico al umbral. Es importante aclarar que la k_{SNV} es una constante que no es calculada para cada señal analizada si no por única vez en base a los datos obtenidos de una señal muestra de la base de datos usada.

Etapa 2. Cálculo de d_{SNV}

El parámetro d_{SNV} se obtiene seleccionando un intervalo de tiempo de la señal que corresponda a un SNV cuya duración puede variar entre 10-100ms. Por practicidad del algoritmo se escoge el intervalo correspondiente al comienzo de la señal. Posteriormente el algoritmo calcula los parámetros descriptos en la sección 5.3.1.3 del segmento seleccionado y obtiene el d_{SNV} mediante la expresión que se muestra a continuación.

$$d_{SNV} = (\sigma * erfc^{-1}(\alpha) + A_0) * P_{SNV} \quad (5.34)$$

Donde P_{SNV} es el valor cuadrático medio del voltaje del SNV seleccionado de la señal, σ es la desviación estándar de ese intervalo de tiempo, A_0 es la media del segmento elegido y $erfc^{-1}(\alpha)$ es la función de error complementario inversa de α , variable que se calcula a partir de la ecuación (5.32) del SNV escogido de la señal que se pretende umbralizar.

A partir de las ecuaciones (5.33) y (5.34) se obtiene el umbral dinámico, que se expresa en la siguiente ecuación.

$$\text{umbral} = k_{SNV} + d_{SNV} \quad (5.35)$$

Siendo k_{SNV} el término constante obtenido de una señal particular y d_{SNV} el término variable que se calcula de cada señal que se pretende umbralizar.

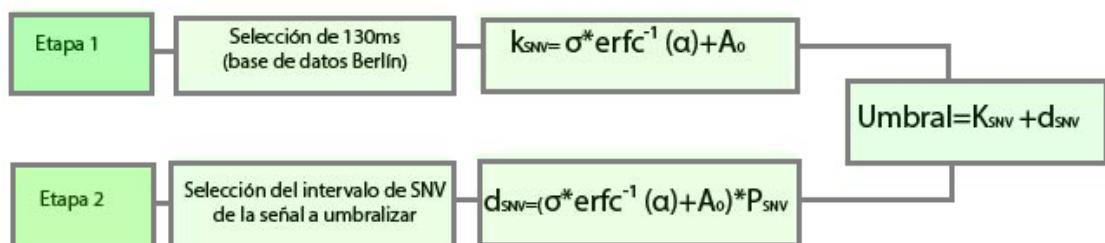


Figura 5.5 Esquema del cálculo del umbral dinámico.

5.3.1.5 Umbralización de la señal

Se desarrolla un algoritmo en el cual el umbral dinámico es aplicado a la envolvente de la señal (descripta en la sección 5.3.1.2) a fin de separar los SNV de los SV.

La salida del algoritmo es una secuencia binaria donde el valor “1” corresponde a un SV y el valor “0” corresponde a un SNV. En la figura 5.6 se exhibe la aplicación del segmentador (la señal de voz se muestra en azul, la envolvente se observa en celeste y la secuencia binaria de umbralización en color rojo).

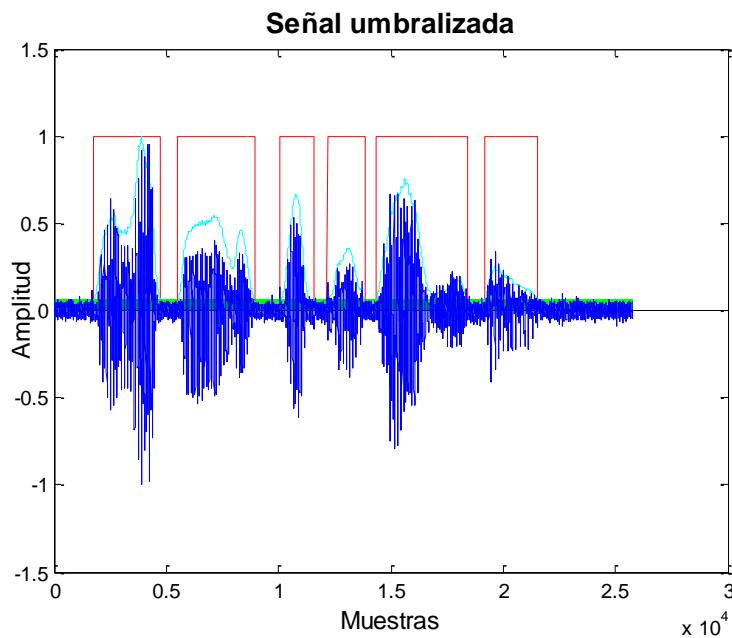


Figura 5.6 Aplicación del segmentador.

5.4 Resultados

A fin de conocer la eficiencia del segmentador se procedió a degradar la señal con ruido aditivo y determinar el error de detección.

Se tomó una señal de la Base de datos Berlín como muestra. Se calculó la SNR mediante la siguiente ecuación.

$$SNR(dB) = 10 \log\left(\frac{P_s}{P_n}\right) \quad (5.36)$$

Donde P_s es la Potencia media de la señal y P_n es la Potencia media del ruido.

La degradación de la señal se alcanzó mediante la incorporación de ruido gaussiano de media cero y varianza conocida a la señal de interés, de tal manera se alcanzó una variación de la SNR desde 8dB a 12.5dB, cabe destacar que la SNR de la señal Berlín sin la adición de ruido blanco es de 12.5dB

5.4.1 Error de detección

Mediante un observador experto se obtuvo el valor de detección del primer SV. Se aplicó el segmentador y se obtuvo el valor de detección del mismo segmento de señal que el observado por el experto.

El error absoluto se calcula sustrayendo del valor detectado por el observador experto (valor_obs), el valor detectado por el segmentador (valor_det).

$$\text{error_abs} = \text{abs}(\text{valor_obs} - \text{valor_det}) \quad (5.37)$$

El error encontrado en número de muestras se convierte a segundos. Este procedimiento se aplica n veces a la misma señal degradada con ruido blanco hasta alcanzar 8.8dB de SNR. Se grafica el error absoluto en función de la relación señal a ruido (figura 5.7). Como se puede observar el error se encuentra en el orden de los milisegundos, inclusive en la peor condición de degradación de la señal.

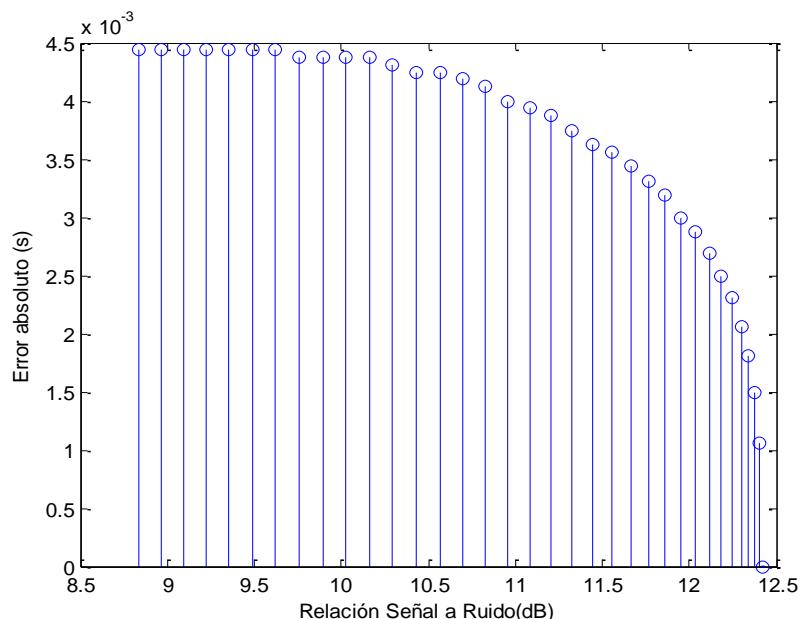


Figura 5.7 Error absoluto en función de la SNR

En la figura 5.8 se observa la señal detectada con 12.42 dB de SNR (a la derecha) y la misma señal detectada con 8.8 dB(a la izquierda).

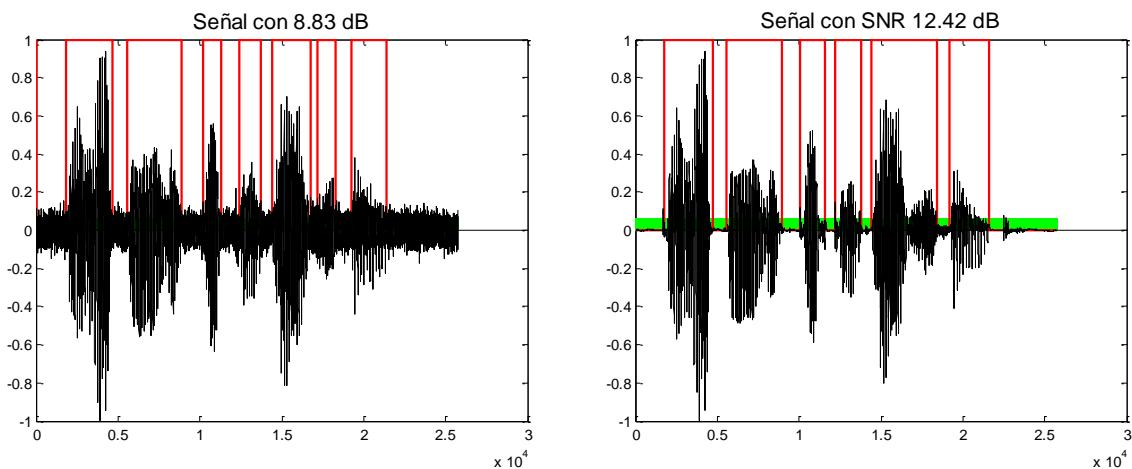


Figura 5.8 Detección de la señal a diferentes SNR

5.4.2 Comparación con otro detector

El Voicebox es un detector comercial que utiliza la entropía de la señal como parámetro que permite detectar los segmentos de actividad vocalica. Su aplicación se muestra en la figura 5.9 aplicado a una señal Berlín con SNR de 12.42dB.

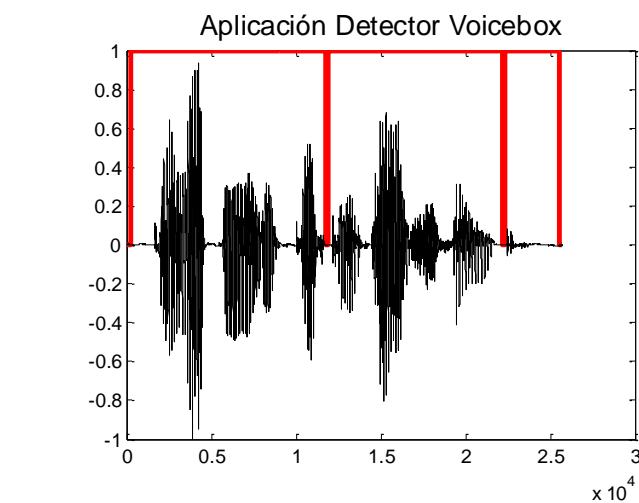


Figura 5.9 Aplicación del detector Voicebox a la señal Berlín.

Posteriormente se obtuvo el error absoluto en la detección con Voicebox y se comparó con la detección del segmentador desarrollado en este trabajo.

En la figura 5.10 se observa a la derecha el segmentador desarrollado en el presente trabajo y a la izquierda el detector comercial. Es notable que el error calculado en Voicebox es varios órdenes superiores al calculado con el segmentador propio.

La VAD con Voicebox no es eficiente, sin embargo su ventaja reside en la robustez ante el ruido, ya que no modifica su detección a pesar de la degradación de la señal.

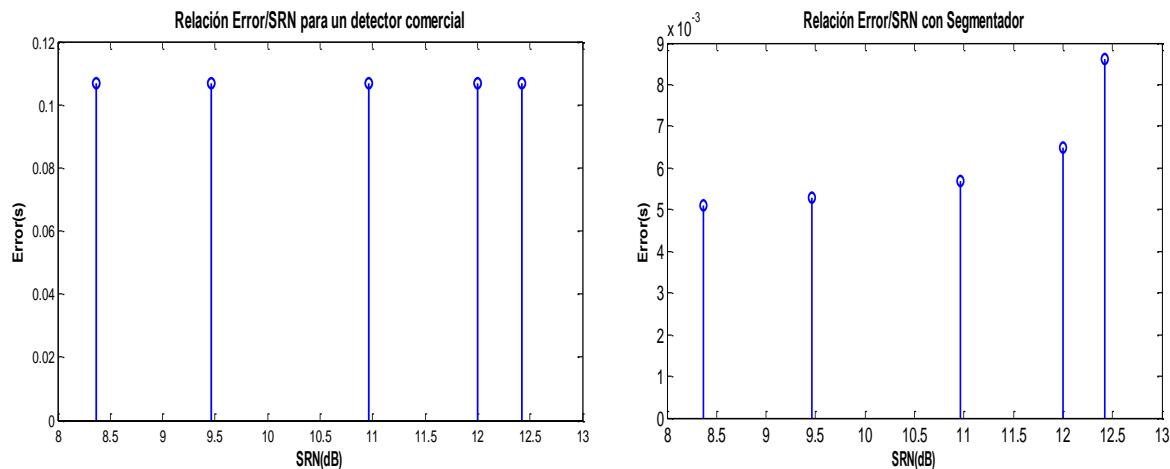


Figura 5.10 Error en función de la SNR

El segmentador propuesto fue validado a partir de la comparación con un detector comercial, el resultado mostró la eficiencia del segmentador desarrollado en el presente trabajo, ante la degradación de la señal con la incorporación de ruido aditivo en comparación con un detector comercial.

El segmentador además fue utilizado para determinar la cantidad de segmentos vocálicos bajo estados emocionales diferentes.

5.5 Conclusión

El segmentador de voz con umbralización dinámica, desarrollado en función de técnicas estadísticas basadas en el Lema de Neyman Pearson fue sometido a distintas pruebas para validar su eficacia. En primer lugar se determinó el error de detección ante señales degradadas con ruido blanco y posteriormente fue comparado con un detector comercial y sometido ambos a la aplicación en señales con diferentes SNR. El resultado mostró la eficiencia del segmentador de voz ante la presencia de ruido aditivo y la superioridad en la diferenciación entre los SV y los SNV en comparación con un detector comercial. El segmentador además fue utilizado para determinar la cantidad de segmentos vocálicos bajo estados emocionales diferentes.

5.6 Referencias

1. Shin, W.; Lee, B.; Lee, Y.; Lee, J. (2000). "Speech/Non-Speech Classification Using Multiple Features For Robust Endpoint Detection", Proc. ICASSP, 1399-1402.
2. Khoa P. (2012). *Noise Robust Voice Activity Detection*. Thesis of Master of Engineering. Nanyang Technological University.
3. Itoh, K.; Mizushima, M. (1997) "Environmental noise reduction based on speech/nonspeech identification for hearing aids," in *Acoustics, Speech, and Signal Processing. ICASSP-97. IEEE International Conference on*, vol. 1, pp. 419–422.
4. Rabiner, L.; Sambur, M. (1975) "An algorithm for determining the endpoints of isolated utterances," *Bell System Tech. Jour.*, vol. 54, no. 2, pp. 297–315.
5. Casals, J.; Puig, P.; Bolaño, R. (2009). *Sencillo detector de actividad de voz basado en Pitch y EMD*. Universidad de Vic. Barcelona.

6. Shen, J.; Hung, J.; Lee, L. (1998). *Robust entropy-based endpoint detection for speech recognition in noisy environments*. In ICSLP (Vol. 98, pp. 232-235).
7. Shafran, I.; Rose, R. (2003). *Robust speech detection and segmentation for real-time ASR applications*. In *Acoustics, Speech, and Signal Processing. Proceedings.(ICASSP'03)*. 2003 IEEE International Conference on (Vol. 1, pp. I-432). IEEE.
8. SKOLNIK, M. I. *Radar Handbook*. Edition ed.: McGraw-Hill, 2008
9. Vidal, J. D. L. C. B., & Fernández, J. R. M. (2014). MATE-CFAR: Ambiente de Pruebas para Detectores CFAR en MATLAB. *Revista Telem@tica*, 13(3), 86-98
10. Mata Moya, D. A. D. L. (2012). *Diseño de detectores robustos en aplicaciones radar*. Tesis doctoral. Universidad de Alcalá.
11. Conte, E., MAIO, A. D., & Galdi, C. (2000). *Signal detection in compound-Gaussian noise: Neyman-Pearson and CFAR detectors*. *Signal Processing, IEEE Transactions on*, 48(2), 419-428.
12. Rangel, E.; Reyes, B.; Pérez, A. (2013). *Lema de Neyman–Pearson para distribuciones de confianza basadas en estadísticas suficientes*. *Lecturas Matemáticas*, 34(2), 205-223.
13. SHEWHART, W. A. *Historia de la Estadística*.
14. Mora Monte, E., Castillo, E., & Puig-Pey Echebeste, J. (1980). *Sobre la obtención de tests de potencia garantizada*. *Qüestiió*. 1980, vol. 4, núm. 2.
15. Kobatake, H., Tawa, K., & Ishida, A. (1989, May). *Speech/nonspeech discrimination for speech recognition system under real life noise environments*. In *Acoustics, Speech, and Signal Processing, 1989. ICASSP-89., 1989 International Conference on* (pp. 365-368). IEEE.
16. Young, S.; Evermann, G.; Gales M. et al. (2005). *The HTK Book*. Copyright 2001-2005 Cambridge University Engineering Departament.
17. HA Larrondo. (2011). *Instrumentación virtual aplicada al estudio de sistemas complejos Capítulo 3: Extractor de Datos de Radar*. Facultad de Ingeniería, Universidad Nacional de Mar del Plata.
18. Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W. F., & Weiss, B. (2005). *A database of German emotional speech*. In *Interspeech* (Vol. 5, pp. 1517-152).

CAPITULO 6

Procesamiento y análisis de base de datos

6.1 Introducción

El procesamiento digital de voz se basa en un conjunto de técnicas que tienen como objetivo encontrar rasgos o características que describan la señal de voz, mediante diferentes procedimientos de parametrización de la voz. La parametrización se combina con la utilización de técnicas discriminativas que permiten seleccionar el conjunto características más eficiente o distintivas.

La señal de entrada suele venir acompañada por efectos perturbadores, los cuales deben ser minimizados. Las técnicas que permiten reducir las componentes indeseables de la señal de entrada antes de la aplicación de la parametrización, forman parte del “Pre-procesamiento”.

Las técnicas de pre-procesamiento permiten acondicionar la señal de voz eliminando componentes que afectan la correcta extracción de características.

Existen varias bases de datos orientadas a la clasificación de estados emotivos. En este trabajo se utilizaron señales de voz de la base de datos de Berlín [1]. La elección de este corpus surge de su amplia trayectoria en el área de procesamiento de emociones y su disponibilidad comercial. Existen otras bases de datos tales como: Enterface [2], Semaine [3], GVEESS [4], IEMOCAP, etc. los primeros tres corpus también fueron utilizados a fin de comparar tasa de error y desempeños de clasificadores respecto al entrenamiento con Berlín.

6.2 Materiales

6.2.1 Base de datos Berlín

Berlín es una base de datos de expresión emocional. Esta base de datos se construyó a partir de la simulación de emociones realizadas por 10 actores (5 mujeres y 5 varones) produciendo 10 expresiones en alemán (5 frases cortas y 5 frases largas). El material de conversación comprende 800 frases. Una emoción simulada involucra una actuación profesional, donde el actor desempeña un papel protagónico que implica un estado emocional en particular.

Las frases de prueba fueron construidas de manera que permitan la eliminación o asimilación de todos los segmentos posibles. Las frases contienen una gran cantidad de contenido vocálico, esto permite el análisis de formantes.

Las adquisiciones se realizaron en la cámara anecoica de la Universidad Técnica de Berlín utilizando un micrófono Sennheiser MKH 40 P 48. Las grabaciones fueron tomadas con una frecuencia de muestreo de 48 kHz y posteriormente remuestreadas a 16 kHz.

Las emociones de la Base de datos Berlín incluyen (términos en alemán en paréntesis):

- Enojo (Ärger)
- Miedo (Angst)
- Neutro (Neutral)
- Tristeza (Trauer)

- Asco (Ekel)
- Felicidad (Freude)
- Aburrimiento (Langeweile)

6.2.2 Base de datos Enterface

La Base de datos Enterface [2] es una base de datos de emociones evocadas que consta de grabaciones audio-visuales en inglés realizadas por 42 sujetos procedentes de 14 nacionalidades diferentes, 81 % de sexo masculino y 19% femenino.

Las emociones de la Base de datos Enterface incluyen:

- Felicidad
- Tristeza
- Sorpresa
- Ira
- Asco
- Miedo

Una emoción evocada implica la inducción a un estado emocional en particular, a partir de imágenes, videos, conversaciones, etc. La persona inducida expresa la emoción correspondiente, en función de su estado emocional particular.

El protocolo de experimentación se desarrolló evocando cada una de las emociones que incluye el corpus. En primer lugar, se le pide al sujeto que escuche un cuento que lo sumerja en el estado emocional adecuado. Una vez que está listo, el sujeto debe pronunciar una de las 5 expresiones propuestas, que constituyen 5 reacciones diferentes para cada emoción en particular.

La base de datos se compone de un total de 1.166 secuencias de vídeo. Los audios son grabados a una frecuencia de muestreo de 48KHz en un formato estéreo de 16bits sin comprimir.

6.2.3 Base de datos Semaine

SEMAINE (Sustained Emotionally coloured Machinehuman Interaction using Nonverbal Expression) [3] es una base de datos audiovisual cuyo enfoque es construir agentes que puedan establecer una conversación emocional con una persona usando el paradigma de Sensitive Artificial Listener (SAL). Las grabaciones son interacciones entre usuarios y operadores. Los operadores simulan un agente SAL, en diferentes configuraciones: Solid SAL (diseñado para que los operadores muestren un adecuado comportamiento no verbal) y SAL semiautomático (diseñado para que las experiencias de los usuarios aproximen la interacción con la máquina).

Contiene un total de 959 conversaciones con una duración de aproximadamente 5 minutos cada una, para un total de 150 participantes. Los audios están muestreados a 48KHz, se utilizan 4 micrófonos en distintas ubicaciones, dos por persona.

Los operadores detectan y expresan las emociones, entonces se definen cuatro personalidades en función de las emociones.

- Spike es enojado. Él intenta hacer enojar al usuario.
- Poppy es feliz. Este operador intenta hacer feliz al usuario.
- Prudence es sensato. Este operador intenta que el usuario sea sensato, prudente.

- Obadiah es depresivo. Este operador intenta deprimir al usuario.

6.2.4 Base de datos GVEESS

La base de datos GVEESS (Genova Vocal Emotion Expression Stimulus Set) [4], contiene una selección de 224 muestras simuladas en alemán, grabadas a una frecuencia de muestreo de 44.1Khz con una duración aproximada de 2-4 seg. por muestra.

Incluye 14 emociones:

- Ansiedad
- Disgusto
- Felicidad
- Ira fuerte
- Interesante
- Ira débil
- Aburrido
- Miedo pánico
- Vergüenza
- Soberbia
- Tristeza
- Desprecio
- Desesperación
- Júbilo

6.3 Métodos

6.3.1 Acondicionamiento de las Bases de Datos

Las bases de datos utilizadas fueron grabadas a diferentes Frecuencias de muestreo. Por este motivo y para estandarizar el procesamiento se procedió a hacer un pre- acondicionamiento de la señal de entrada, el cual se esquematiza en la figura 6.1.



Figura 6.1. Acondicionamiento de la señal

El remuestreo implica submuestrear las señales de entrada a una frecuencia inferior: 16KHz (Frecuencia de muestreo de la Base de datos Berlín) utilizando la función de Matlab® “resample”. De esta manera reducimos muestras redundantes y conservamos las características propias de la señal, ya que se conserva el principio del Teorema de Muestreo, que indica que la frecuencia de muestreo

debe ser al menos el doble de la frecuencia máxima de la señal, siendo ésta en promedio menor a 8 kHz.

La normalización de la señal implica obtener una salida que se encuentre entre 0 y 1 en amplitud, a fin de reducir los errores de baja intensidad producidos cuando la adquisición es de mala calidad. Por ejemplo cuando la distancia del locutor al micrófono es mayor a 30 cm (distancia usada en la adquisición de la base de dato Berlín) la intensidad de la señal adquirida es reducida y puede generar inconvenientes en el procesamiento de la señal. El procedimiento de la normalización consiste en la aplicación de la siguiente operación matemática:

$$s_normalizada = \frac{señal - \min(señal)}{\max(señal) - \min(señal)} \quad (6.1)$$

Donde $\min(señal)$ corresponde al valor mínimo de la señal y $\max(señal)$ corresponde al valor máximo.

Como resultado de la normalización, la señal queda montada sobre una componente de continua, ésta es extraída con el comando detrend de Matlab®, de manera de centrarla respecto del 0 en amplitud. En la figura 6.2 se observa: a la izquierda la señal normalizada y a la derecha, con la eliminación de la componente de continua.

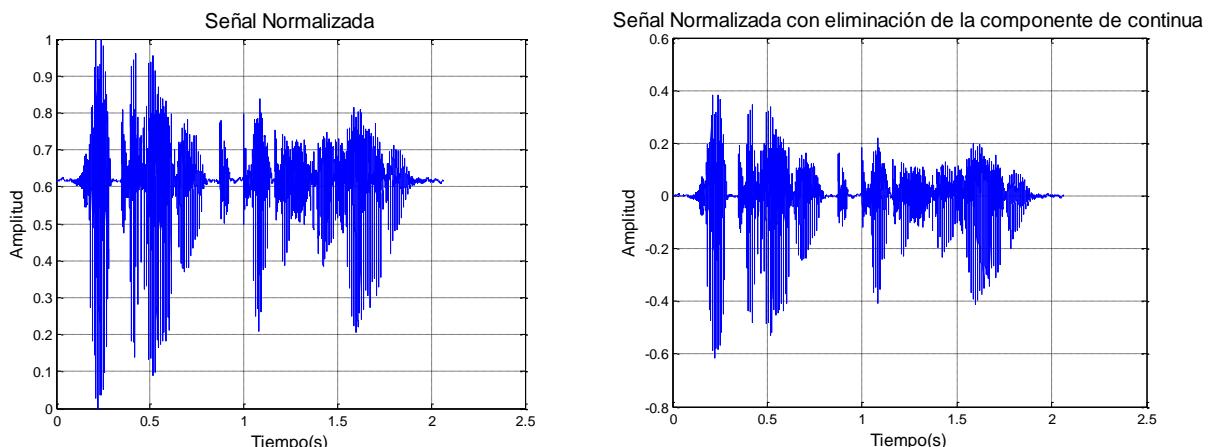


Figura 6.2. Señal normalizada a la izquierda y con eliminación de continua a la derecha

6.3.2 Pre-procesamiento

Posteriormente, la señal es limitada en frecuencia con un filtro pasabanda recursivo Chebyshev tipo 1. A fin de segmentar la señal, extraer los segmentos de voz y eliminar los silencios, se le aplica, a la señal filtrada, un detector de actividad de voz (descripto en el capítulo 5). Posteriormente se procedió a realizar un ventaneo de la señal, estableciendo el límite en 20ms con solapamiento del 50%.

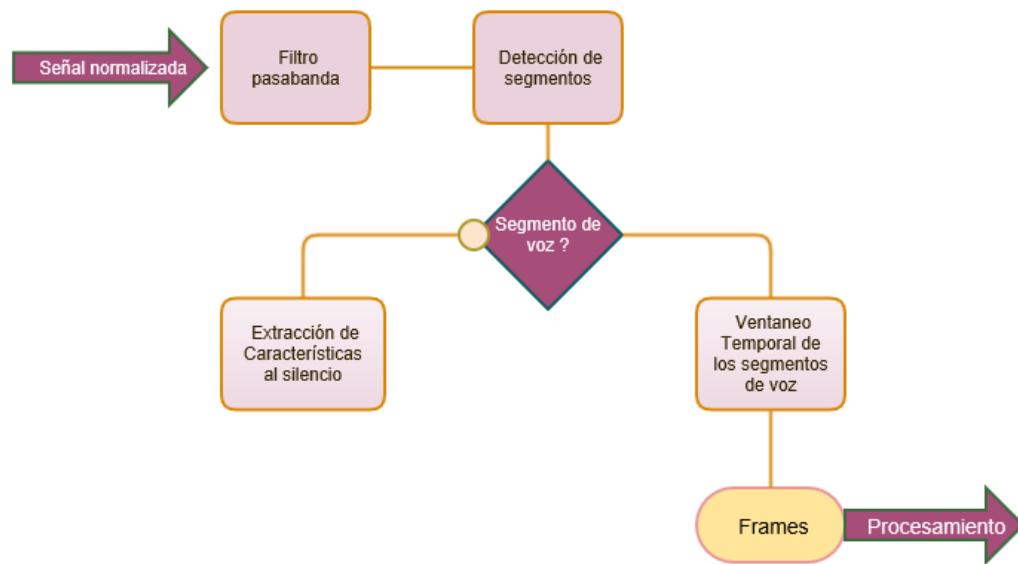


Figura 6.3. Pre-procesamiento.

Como se mencionó, el filtro pasabanda recursivo Chebyshev tipo 1 de orden 2, tiene una frecuencia de corte inferior en 20 Hz y una frecuencia de corte superior en 6.8 KHz. A fin de mantener la linealidad de fase se añadió un filtrado bidireccional, lo cual generó una ligera modificación en las frecuencias de corte. La respuesta frecuencial del filtro se puede observar en la figura 6.4. La gráfica superior describe la respuesta en magnitud del filtro, y la inferior describe la fase. La figura 6.5 corresponde a la ubicación de los ceros y los polos en el plano Z, tal como debe ocurrir, aparecen los ceros ubicados en 1 y -1 del eje real atenuando las componentes de baja y alta frecuencia respectivamente.

La Función de Transferencia en el dominio Z del filtro diseñado se visualiza en la siguiente ecuación.

$$H(z) = \frac{1-z^{-2}}{1-0.21z^{-1}-0.77z^{-2}} \quad (6.2)$$

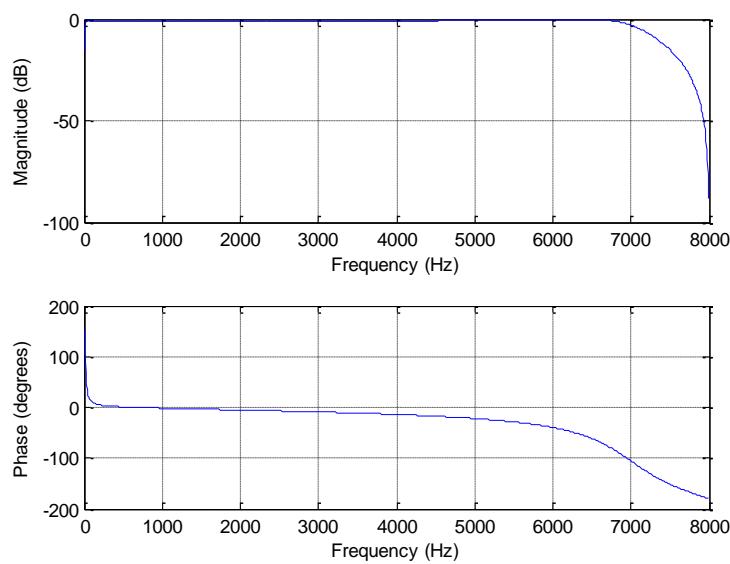


Figura 6.4 Respuesta frecuencial.

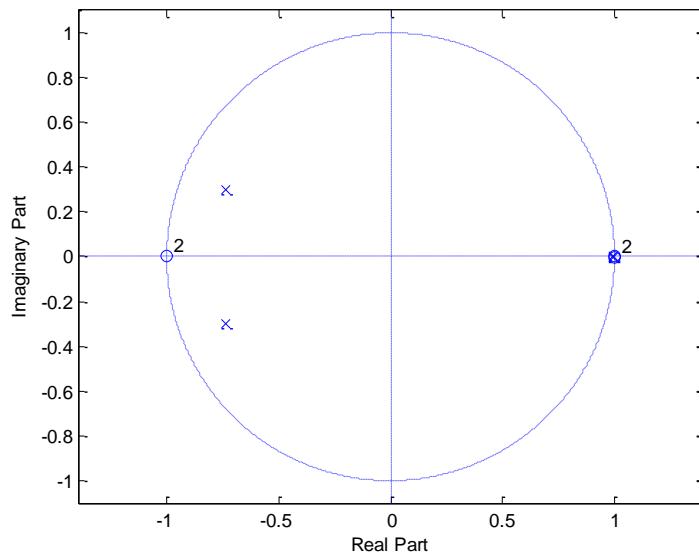


Figura 6.5. Ubicación de polos y ceros en el plano Z.

En la figura 6.6 se observa la señal sin filtrar a la izquierda y a la derecha la señal filtrada.

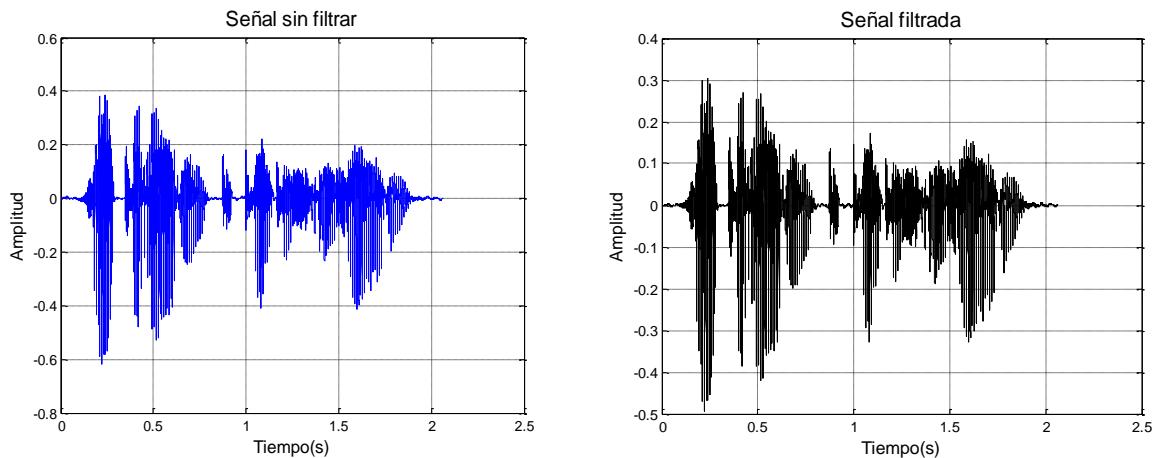


Figura 6.6. Señal sin filtrar a la derecha y filtrada a la izquierda.

A la señal filtrada se le aplicó un detector de segmentos de voz que permite conservar los fragmentos de la señal donde existe una secuencia de voz. Los silencios no fueron descartados por completo. A partir de ellos se extrajo la cantidad total de segmentos de silencio y la duración temporal de cada uno.

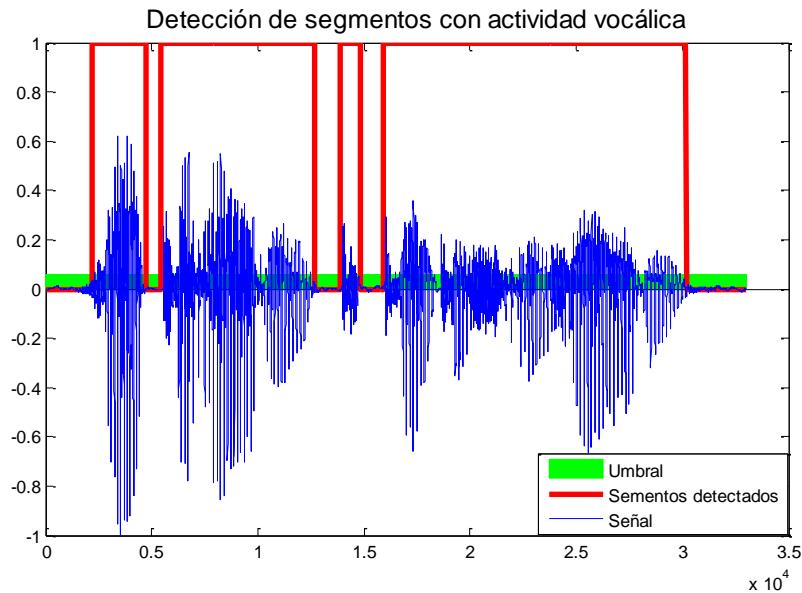


Figura 6.7 Detección de segmentos de actividad de voz

Dada la no estacionariedad de la señal de voz se procedió a realizar un ventaneo temporal que consiste en dividir la secuencia de voz en unidades llamadas cuadros o frames. Se estableció como tamaño de ventana 20ms con solapamiento del 50 % [5], a fin de obtener segmentos de la señal con características cuasiestacionarias.

El principio básico del análisis en ventanas se puede representar mediante la siguiente ecuación [5]:

$$X(n) = \sum_{m=-\infty}^{\infty} T[x(m).w(n-m)] \quad (6.3)$$

Donde $X(n)$ representa el parámetro de análisis o vector de parámetros en tiempo corto. El operador T define la función de análisis y $w(n-m)$ representa la secuencia de ventanas desplazadas en el tiempo. Se utilizó una ventana de Hamming porque ofrece una buena resolución frecuencial.

De esta técnica surgió una nueva característica que permitió la distinción de emociones: el número total de ventanas donde existen segmentos de voz.

6.3.3 Procesamiento de la señal. Extracción de características

El procesamiento de la señal involucra el diseño de algoritmos que permitan extraer características de la voz. Se extrajeron 23 variables, desde enfoques: temporal, prosódico, frecuencial y mixto. En la figura 6.8 se exhiben los parámetros que fueron calculados en esta etapa.

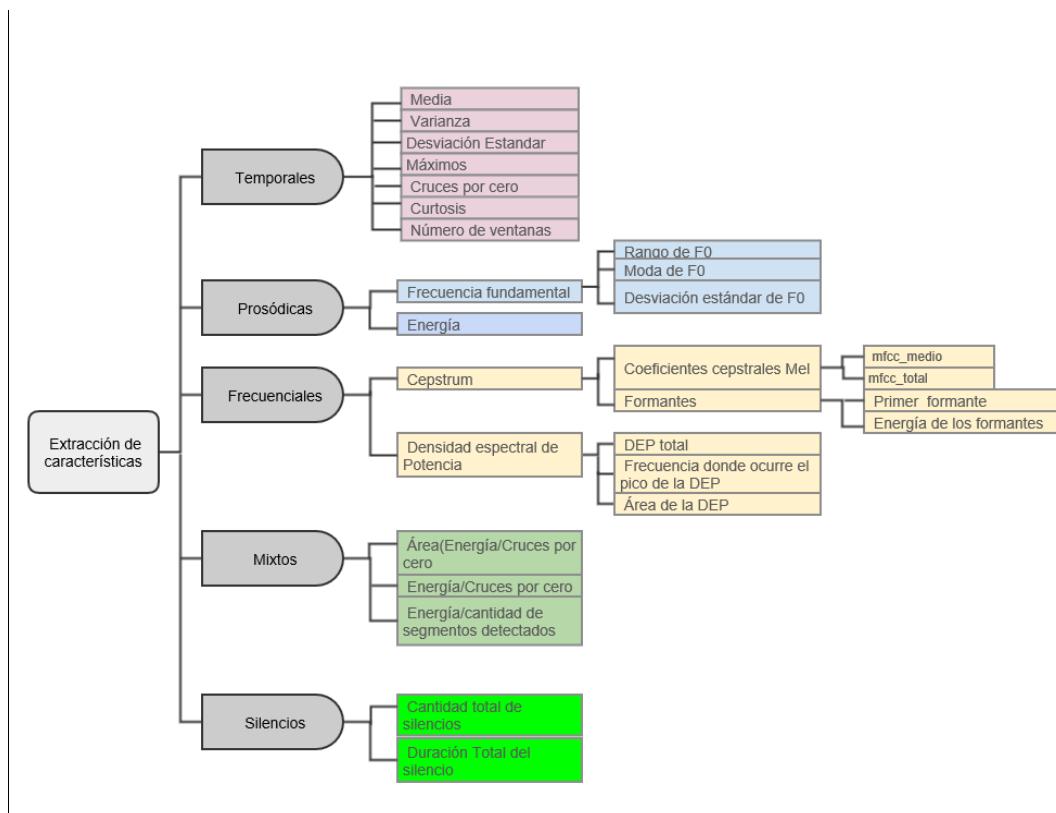


Figura 6.8. Procesamiento de las señales. Parametrización.

6.3.3.1 Características temporales

Los parámetros temporales fueron argumentados en el capítulo 4. Sin embargo, posterior al desarrollo del detector de actividad de voz, se determinó una nueva característica denominada número de ventanas, que se relaciona con la cantidad de segmentos de actividad de voz detectados. Como ya fue mencionado, el tamaño de la ventana es fijo, sin embargo para la misma frase expresada con distintas emociones, la detección de segmentos puede modificar la cantidad de ventanas calculadas, ya que por ejemplo emociones como la tristeza presentan segmentos de actividad vocalica de mayor duración, por lo tanto se calcularan más ventanas. En la siguiente ecuación se describe esta característica.

$$n_win = (l(x(i)) - win) / solap + 1 \quad (6.4)$$

Donde n_win representa el número total de ventanas, $l(x(i))$ es la longitud de la señal, win constituye el tamaño de la ventana (20ms) y el parámetro $solap$ es el solapamiento (50% del tamaño de la ventana).

6.3.3.2 Características prosódicas

De los parámetros prosódicos, se derivaron 3 características de la frecuencia fundamental ($f0$).

- *Rango de la frecuencia fundamental:*

$$Rango_f0 = \max(f0) - \min(f0) \quad (6.5)$$

El rango de la frecuencia fundamental ($Rango_f0$) es la distancia entre el valor máximo ($\max(f0)$) y el mínimo de la frecuencia fundamental ($\min(f0)$). Este parámetro refleja el grado de activación del locutor. Un rango más extenso del normal refleja una excitación emocional.

- *Moda de la frecuencia fundamental:*

La moda de la ($f0$) representa el valor más frecuentemente ocurrido. Este parámetro depende del locutor e indica el valor más frecuente observado a lo largo del discurso.

- *Desviación estándar:*

$$S_f0 = \frac{1}{N-1} \sum_{m=0}^N ([fo(m)]w(n-m)) - \overline{f0} \quad (6.6)$$

Donde S_f0 representa la desviación estándar de la frecuencia fundamental, $\overline{f0}$ representa la media de la frecuencia fundamental, ($fo(m)$) indica la frecuencia fundamental para cada instante m de muestreo y N es el tamaño de la muestra. La Desviación estándar permite determinar el promedio aritmético de fluctuación de la $f0$ respecto a su punto central o media.

6.3.3.3 Características frecuenciales

Los parámetros frecuenciales fueron descriptos en el capítulo 4. Sin embargo, se añadió una característica más que representa la energía de la envolvente del espectro del tracto vocal. El espectro del tracto vocal se consigue a partir del desarrollo de formantes y la envolvente del espectro se obtiene a partir de la aplicación de las técnicas de predicción lineal (LPC). La energía de esta envolvente se calcula a través del RMS, que se usa como estimador de energía, por intervalos. En la siguiente ecuación se describe esta característica.

$$e_envfor = \sum_{m=-\infty}^{\infty} (env[m]w[n-m])^2 \quad (6.7)$$

e_envfor indica la energía en segmentos cortos, $env[m]$ es la envolvente del espectro del tracto vocal y $w(n-m)$ representa el ventaneo temporal.

6.3.3.4 Características mixtas

Los parámetros mixtos se describen a partir de la relación entre las características prosódicas y temporales. Estos parámetros incluyen:

- *Energía/Cruces por cero:*

Este parámetro implica dividir el valor total de la energía de la señal $x(m)$ de longitud N y ventaneo temporal ($w(n-m)$) por el número de cruces por cero (CPZ).

$$E / CPZ = \left(\sum_{m=0}^N (x(m)w(n-m))^2 \right) / CPZ \quad (6.8)$$

- *Área(Energía/Cruces por cero):*

Se calculó el área de la relación entre Energía total de la señal y la cantidad de cruces por cero mediante el método trapezoidal.

- *Energía/cantidad de segmentos detectados:*

Esta característica implica dividir la energía total de la señal $x(m)$ de longitud N y ventaneo temporal $w(n-m)$ en el número de segmentos de voz detectados (sv). Esta relación indica cómo se distribuye la energía en los segmentos de actividad de voz.

$$E / SV = \left(\sum_{m=0}^N (x(m)w(n-m))^2 \right) / sv \quad (6.9)$$

6.3.4 Técnicas de Selección de características

En primera instancia se realizó una discriminación de características a través de la visualización. Se tomaron 4 frases correspondientes a las emociones: felicidad y tristeza. Se planteó el uso de estas emociones ya que forman parte de regiones opuestas en el plano emocional de activación-valencia. Se extrajeron las características que las describen y se procedió al análisis visual. Los parámetros correspondiente a estas emociones fueron comparados en función de la mínima distancia necesaria que permita distinguir entre dos emociones opuestas. La semejanza en los valores obtenidos de las características extraídas para emociones distintas permitió realizar la exclusión de aquellas no relevantes en la categorización.

Los parámetros seleccionados bajo ese procedimiento resultaron ser:

Temporales:

1. Media(Me)
2. Máximos(Ma)
3. Cruces por cero (CPZ)
4. Curtosis(K)
5. Número de ventanas(nV)

Prosódicos:

6. Energía(E)

Frecuenciales:

7. Densidad espectral de Potencia(*DEP*)
8. Área de Densidad Espectral de Potencia(*aDEP*)
9. Frecuencia donde ocurre el pico máximo de la DEP (*fDEP*)
10. Coeficientes cepstrales en frecuencia Mel (*mfcc*)

Mixtos:

11. Energía/Crucos por cero (E/cpz)
12. Área de Energía/Crucos por cero (*aE/cpz*)

La selección definitiva de parámetros tuvo lugar mediante la técnica de análisis discriminante Lineal (LDA).

El Análisis Discriminante equivale a un análisis de regresión donde la variable dependiente es categórica y tiene como categorías la etiqueta de cada uno de los grupos. Las variables independientes son continuas y determinan a qué grupos pertenecen los objetos. Se trata de encontrar relaciones lineales entre las variables continuas que mejor discriminen en los grupos dados a los objetos. Además, se trata de definir una regla de decisión que asigne un objeto nuevo, que no sabemos clasificar previamente, a uno de los grupos prefijados. Es decir, es capaz de identificar características que diferencien a 2 o más grupos. La salida del LDA indica el grupo al que cada valor de la muestra ha sido asignada. Estos datos son analizados a través de una matriz de confusión.

La matriz de confusión es una herramienta de visualización en la cual cada columna de la matriz representa el número de predicciones de cada clase, mientras que cada fila representa a las instancias en la clase real. El error estimado de la clasificación depende de los datos de entrenamiento. El comando de Matlab® *classify* devuelve la tasa de error aparente, esto es el porcentaje de observaciones en el entrenamiento que han sido mal clasificados, ponderados por probabilidades de los grupos.

6.4 Resultados

Se realizaron una serie de experimentos para identificar las variables representativas. De las 66 combinaciones de pares posibles, para las 12 variables (ya nombradas en la sección 6.3.4), se tomaron las 3 que proporcionaron mejores resultados: *fDEP-nV*, *DEP-mfcc*, *Cpz-mfcc*. Al observar que el parámetro *mfcc* se repetía en los grupos de variables que mostraron los mejores resultados, se procedió a evaluar, mediante una experimentación más, su desempeño como una única variable en la clasificación.

A continuación se exhiben los resultados de las combinaciones mencionadas para la diferenciación entre dos emociones: Felicidad y Tristeza.

6.4.1 Experimentación Felicidad-tristeza

Se analizaron combinaciones de variables para la clasificación de dos emociones. Felicidad y tristeza.

6.4.1.1 Prueba 1

Se tomaron las siguientes variables:

- $fDEP$
- nV

El resultado del análisis discriminante, en función de este grupo de variables, es mostrado en la figura 6.9.

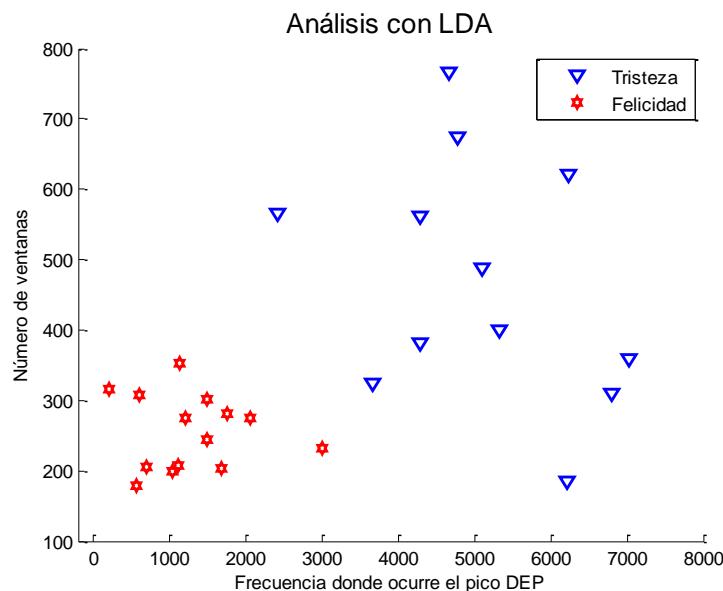


Figura 6.9. Prueba 1

A continuación se exhibe la Matriz de confusión (en porcentajes):

	Tristeza	Felicidad
Tristeza	84.61	15.38
Felicidad	7.69	92.30

Tabla 6.1. Matriz de confusión de Prueba 1

Error de clasificación = 11.53%

En la matriz de confusión de la prueba 1 se observan los porcentajes de correspondencia para las dos emociones (felicidad y tristeza). El porcentaje de detección de tristeza fue del **84.61%** y el de felicidad fue del **92.30%**.

6.4.1.2 Prueba 2

Se tomaron las siguientes variables

- *DEP*
- *mfcc*

En la figura 6.10 se observa el resultado de la experimentación 2.

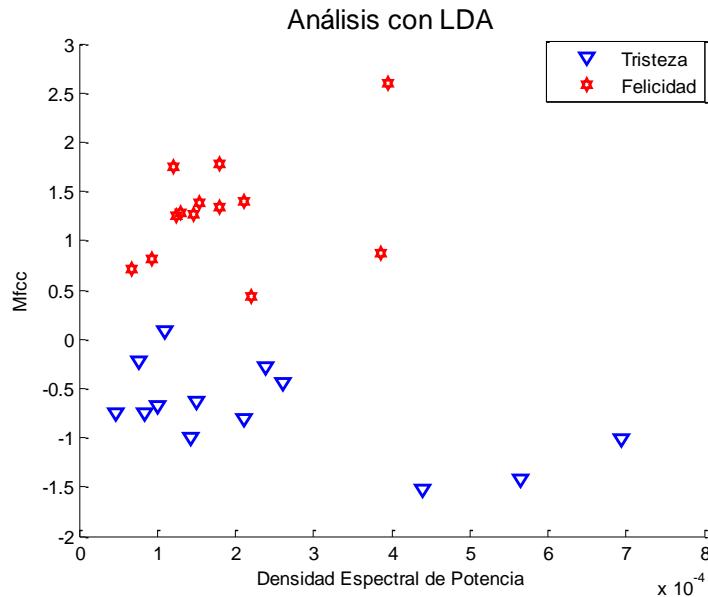


Figura 6.10. Prueba 2

Se expone en la siguiente tabla, la Matriz de confusión (en porcentajes).

	Tristeza	Felicidad
Tristeza	100	0
Felicidad	0	100

Tabla 6.2. Matriz de confusión de Prueba 2

Error de clasificación= 0%

En la matriz de confusión se observa que los porcentajes de detección de tristeza y felicidad fueron del **100%** con un error en la clasificación de 0%.

6.4.1.3 Prueba 3

Se tomaron las siguientes variables:

- *CPZ*
- *mfcc*

El resultado de esta experimentación con LDA es mostrado en la figura 6.11.

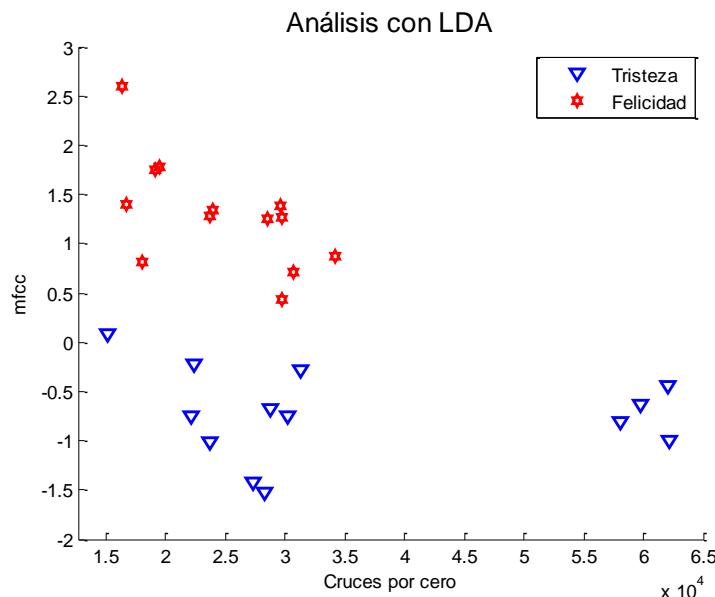


Figura 6.11. Prueba 3

A continuación se presenta la Matriz de confusión (en porcentajes).

		Tristeza	Felicidad
Tristeza	100	0	
Felicidad	0	100	

Tabla 6.3. Matriz de confusión de Prueba 3

Error de clasificación = 0%

El porcentaje de detección para ambas emociones según los datos obtenidos en la matriz de confusión es del **100%**.

6.4.1.4 Prueba 4

Se tomó una sola variable:

- *mfcc*

A continuación se exhiben los resultados de la aplicación de las técnicas de LDA usando una sola variable.

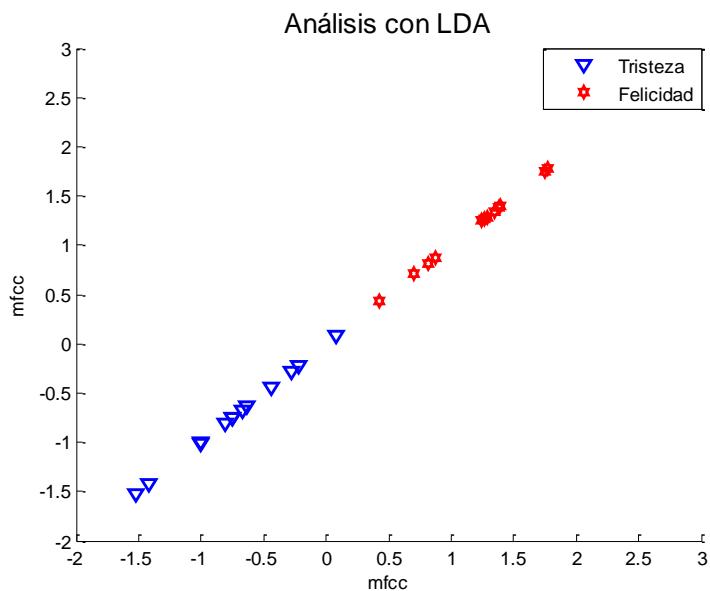


Figura 6.12. Prueba 4

La Matriz de confusión (en porcentajes) utilizando una sola característica se expone en la siguiente tabla.

	Tristeza	Felicidad
Tristeza	100	0
Felicidad	0	100

Tabla 6.4. Matriz de confusión de Prueba 4

Error de clasificación=0%

El porcentaje de detección de tristeza fue del **100%** y el de felicidad fue del **100%**.

En la tabla siguiente se muestra un resumen de los resultados de detección para cada grupo de variables.

Pruebas	Características	Emociones	Porcentaje de emoción detectada	
			Tristeza	Felicidad
1	fDEP-nV	T-F	84.61%	92.30%
2	DEP-mfcc	T-F	100	100
3	Cpz-mfcc	T-F	100	100
4	mfcc	T-F	100	100

Tabla 6.5. Porcentaje de emoción detectada

Se determinaron las características que describen dos emociones: felicidad y tristeza. Posteriormente, se realizó una serie de pruebas variando el número de parámetros obteniéndose de manera cuantitativa las características que permiten diferenciar dos emociones con una tasa de

detección del 100%. Se puede observar en la tabla 6.5 que solo una característica (**mfcc**) es suficiente para realizar la detección de emociones opuestas en el plano de activación-valencia, con una tasa de acierto del 100%.

6.5 Conclusiones

En esta etapa del trabajo, se ejecutaron técnicas de pre-procesamiento tales como filtrado, normalización y segmentación de la señal, que permitieron acondicionar la señal de entrada a fin de reducir las componentes indeseables y facilitar la etapa de procesamiento.

La parametrización de la señal de voz, mediante el desarrollo de algoritmos computacionales, permitió extraer 22 características que describen la señal desde diferentes enfoques: temporal, prosódico, frecuencial y mixto.

A fin de conocer el comportamiento de estas variables en dos estados emocionales, opuestos en valencia y en activación: Felicidad y Tristeza, se ejecutaron técnicas de selección de características. Una primera selección fue realizada mediante la visualización, lo cual permitió la elección de 12 variables. Estas características fueron combinadas en grupos de a dos y analizadas mediante Análisis discriminante. Esta técnica de clasificación lineal permitió hallar las combinaciones que resultaron más eficientes para discernir entre la emoción tristeza y la emoción felicidad con una tasa de acierto del 100%. Además los resultados mostraron que con una sola característica (**mfcc**) se pudo realizar una clasificación, con una tasa de acierto del 100%, de las emociones mencionadas.

6.6 Referencias

1. Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W. F., & Weiss, B. (2005). *A database of German emotional speech*. In *Interspeech* (Vol. 5, pp. 1517-1520).
2. Martin, O., Kotsia, I., Macq, B., & Pitas, I. (2006). *The eINTERFACE'05 audio-visual emotion database*. In *Data Engineering Workshops, 2006. Proceedings. 22nd International Conference on* (pp. 8-8). IEEE.
3. McKeown, G., Valstar, M., Cowie, R., Pantic, M., & Schroder, M. (2012). *The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent*. *Affective Computing, IEEE Transactions on*, 3(1), 5-17.
4. Truong, K., & Van Leeuwen, D. (2007). *An 'open-set'detection evaluation methodology for automatic emotion recognition in speech*. In *Workshop on Paralinguistic Speech-between models and data* (pp. 5-10).
5. Rabiner L. R. y. Schafer R. W. (2007). "Introduction to Digital Speech Processing". *Foundations and Trends® in Signal Processing* Vol. 1, Nos. 1–2.

CAPITULO 7

Determinación de los parámetros invariables al idioma

7.1 Introducción

Una de las bases de datos más exploradas en el campo del reconocimiento de emociones es la Base de datos Berlín, ampliamente utilizada y de difusión gratuita. Sin embargo desde su creación hasta la actualidad se desarrollaron una gran diversidad de corpus en otros idiomas y dialectos. La argumentación de estos centros de investigación es proveer bases de datos con especificidad de idioma [1].

Esto lleva a reflexionar sobre qué tan distinta puede ser una emoción expresada en una lengua o en otra. ¿Hay parámetros en la voz que puedan ser invariantes de una lengua a otra? ¿Qué ocurre si se tiene una base de datos con un determinado idioma y las pruebas experimentales utilizan muestras en otra lengua?, ¿disminuye su rendimiento?, éstas fueron algunas de nuestras inquietudes.

Scherer [2] discrimina a la frecuencia fundamental como un factor de variación entre lenguajes occidentales y las lenguas tonales del Este de Asia y África que pueden afectar el discurso emocional, ya que las lenguas tonales, según el autor, modifican la frecuencia fundamental para transmitir información léxica independiente de la emoción.

En el estudio de la emoción en el habla se supone la hipótesis que la voz sufre cambios acústicos causados por las alteraciones fisiológicas del cuerpo cuando la persona tiene una sensación fuerte y que estos cambios también dependen del idioma. Algunos investigadores indican que los patrones sonoros suprasegmentales [3] son menos variantes al idioma.

En otros trabajos [4] donde se intenta hacer una comparación cruzada entre dos idiomas (holandés y coreano) se crea una base de datos con ambas lenguas. El objetivo es hacer un análisis de rasgos específicos en la expresión emocional tanto en holandés como en coreano a fin de encontrar propiedades invariantes en la expresión de las emociones para ambas culturas.

Nuestra línea de pensamiento es, si los miembros de culturas radicalmente diferentes son capaces de reconocer las emociones expresadas, entonces existe la posibilidad que, para algunas emociones, exista cierta independencia de los factores culturales. En este caso teniendo en cuenta solo el audio, se intentará encontrar características que definan una emoción sin ser afectadas por los diferentes idiomas y culturas. Resulta evidente que si oyentes de una lengua determinada pueden categorizar mejor las expresiones emocionales de personas de la misma lengua o lenguas similares [2] entonces, los sistemas automáticos también mejoraran sus despeños cuando se los entrena con el mismo idioma en el que se va a implementar.

Como ya fue mencionado en el capítulo 6, el corpus utilizado en este trabajo es la Base de datos Berlín (de idioma alemán). Por lo tanto fue indispensable, determinar las características que permitan independizar los algoritmos de este idioma. En función de este objetivo se buscaron los parámetros que redujeran al mínimo la variación entre alemán y el resto de los idiomas, principalmente el español.

7.2 Materiales

En esta etapa del trabajo se utilizó una guía de pronunciación provista por Forvo Media SL® que opera bajo licencia Creative Commons [5]. Esta guía dispone de archivos de audio en diferentes idiomas, muestreados a 44.1Khz.

Se manipuló una misma palabra en castellano, bajo 7 pronunciaciões diferentes (alemán (A), español mexicano (EM), japonés (J), español (E), español argentino (EA), español colombiano (EC), inglés norteamericano (EU) de manera de obtener una gama de posibilidades prosódicas.

Las señales fueron separadas en 2 grupos .Grupo 1: origen español (EM, EA, E, EC). Grupo 2: mixto (A, J, E, EU).

7.3 Procesamiento de la señal

Las señales de esta Base de datos fueron pre-procesadas y parametrizadas según los métodos descriptos en el capítulo 6, donde se determinaron las características representativas de los enfoques temporal, prosódico y frecuencial.

7.3.1 Características temporales y prosódicas

Se determinaron los siguientes parámetros:

- Media (Me)
- Energía (E)
- Funciones de distribución de probabilidad
- Cruces por cero (Cpz)
- Frecuencia fundamental (pitch)

7.3.2 Características frecuenciales

Se tomaron en cuenta las siguientes características:

- Coeficientes cepstrales en frecuencia Mel (mfcc)
- Formantes (Fm)
- Densidad espectral de Potencia (DEP).

7.4 Métodos de análisis

Las variables que surgieron del procesamiento fueron analizadas mediante las siguientes técnicas comparativas:

- Divergencia de Kullback Leibler (DKL)

La divergencia de Kullback Leibler es un indicador de similitud entre dos funciones de distribución y una medida de entropía relativa [6]. Permite mostrar cuánto diverge la distribución de una señal de un determinado acento prosódico respecto de otra. Se describe mediante la siguiente ecuación:

$$dKu = \sum p_i * \log\left(\frac{p_i}{q_i}\right) \quad (7.1)$$

Donde p_i y q_i indican las probabilidades de un evento i de una variable aleatoria discreta en las distribuciones de probabilidad p y q , respectivamente.

- **Técnica de correlación cruzada**

La correlación cruzada entre dos señales es una medida de la similitud de los atributos que la componen. Se describe mediante la siguiente ecuación.

$$r_{dx}(k) = \sum_{n=1}^N d(n)x(n-k) \quad (7.2)$$

Donde $d(n)$ y $x(n)$ son las señales a correlacionar de tamaño N de n muestras y k representa los intervalos de muestreo retrasados de la señal $x(n)$.

- **Distancia euclíadiana**

La distancia es una herramienta matemática que permite estimar el grado de semejanza entre dos elementos, se establece entre los puntos p_i y x_i de las señales p y q , respectivamente.

$$d(p, x) = \sqrt{\sum_{i=1}^N (p_i - x_i)^2} \quad (7.3)$$

7.5 Resultados

A fin de facilitar el análisis de las características extraídas se establecieron las siguientes combinaciones dentro de cada grupo:

- Grupo1 combinado (G1C): (EM-EC), (EM-E), (EM-EA), (EC -EA), (E-A).
- Grupo2 combinado (G2C): (A-EU), (A-J), (EU-J), (EU-E), (J-E).

Posteriormente, a cada combinación, se le aplicaron las técnicas comparativas descriptas en la sección 7.4.

A continuación se mostrarán los resultados para cada enfoque analizado.

7.5.1 Enfoque temporal y prósodico

Los parámetros Me y E fueron evaluados mediante la Técnica de correlación. Es importante aclarar que un valor de salida igual a “1” indica el estado de máxima correlación (máximo grado de similitud) y un valor igual a “0” implica correlación nula (mínima similitud) entre las señales analizadas.

- Correlación de Me (CMe)
- Correlación de E (CE)

La Distancia euclíadiana permitió el análisis comparativo de las variables Cpz y pitch para cada combinación. Un valor de distancia entre dos características igual a “1” implica mínima similitud y un valor igual a “0” indica máxima similitud.

- Distancia de Cpz (DC_{Cpz}).
- Distancia de pitch (D_{pitch}).

De determinó además la DKu de las funciones de probabilidad de cada combinación. Un valor de divergencia entre las distribuciones de dos señales igual a “1” indica mínima similitud y un valor igual a “0” implica máxima similitud entre ellas.

- Divergencia entre funciones de probabilidad (DKu).

Los resultados de la implementación de las técnicas mencionadas se muestran en la siguiente tabla. Están resaltadas aquellas combinaciones donde la técnica empleada indica máxima similitud entre sus elementos.

Combinaciones	DKu	CMe	CE	DCpz	Dpitch
EM-EC	1	0	0.28	0,144	0.36
EM-E	0.66	0,40	0.95	0,089	1
EM-EA	0	0,45	1	1	0.40
EC-E	0.81	0,14	0.59	0	0.59
EC-EA	0.69	0,21	0.70	0,8013	0
E-EA	0.035	1	1	0,8559	0.55
A-EU	0,55	0	0	0,2779	0.43
A-J	0,85	0,23	0,24	0,6389	0.34
A-E	1	0,04	0,07	0,3195	1
EU-J	0,1809	0,86	0,81	0,9584	0
EU-E	0,1131	0,58	0,59	0	0.46
J-E	0	1	1	1	0.56

Tabla 7.1. Análisis de parámetros temporales y prosódicos

La DKu aplicada a las funciones de probabilidad de cada combinación, muestra un valor mínimo (tendiente a cero) para **EM-EA**, **E-EA** y **J-E** lo cual se interpreta como una mínima divergencia entre la representación estadística de estas señales, que en otras palabras implica una máxima similitud entre sus distribuciones de probabilidad.

La correlación cruzada tanto de Me como de E, mostró que las combinaciones **E-EA** y **E-J** exhibieron un alto grado de correlación, que implica máxima similitud entre las señales analizada. La E de la combinación **EM-EA** resultó también altamente correlacionada.

El análisis de DCpZ mostró que los grupos **E-EU** y **E-EC** presentaron distancias mínimas, lo cual indica semejanza entre los elementos de cada pareja.

La evaluación de la Dpitch mostró que las combinaciones **EC-EA** y **EU-J** exhibieron similitud del parámetro analizado para cada combinación. En la figura 7.1 se exponen las distancias entre frecuencias fundamentales de todas las combinaciones analizadas. Los grupos **EC-EA** y **EU-J** presentan un valor nulo de distancia, o sea máxima similitud.

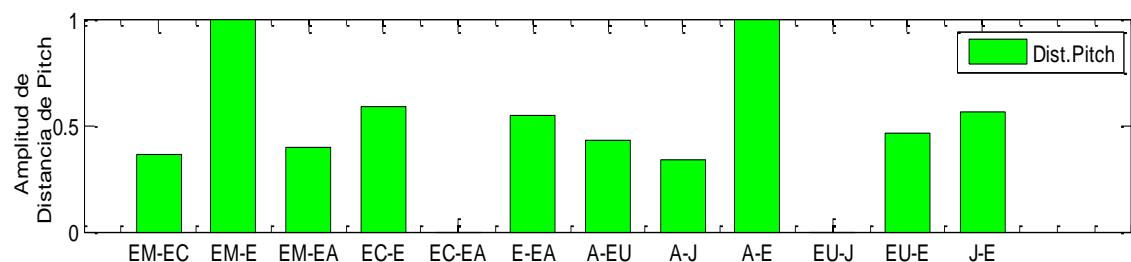


Figura 7.1. Distancia de Frecuencias fundamentales en G1C Y G2C

7.5.2 Características frecuenciales

Los parámetros *Fm*, *DEP* fueron determinados de cada señal de los Grupos 1 y 2, y posteriormente evaluados. Los *mfcc*, obtenidos de cada combinación de señales pertenecientes a un idioma o dialecto, fueron analizados mediante la técnica de correlación.

- Correlación de *mfcc* (*Cmfcc*).

En la siguiente tabla se exhibe el análisis de correlación de los *mfcc* para cada combinación. Aquellos pares resaltados pertenecen a las combinaciones que presentan alta correlación entre sus elementos.

Combinaciones	Cmfcc
EM-EC	0.67
EM-E	0.61
EM-EA	0.69
EC-E	0.93
EC-EA	0.96
E-EA	0.94
A-EU	0.81

A-J	0.74
A-E	0.98
EU-J	0.78
EU-E	0.78
J-E	0.78

Tabla 7.2. Análisis de Correlación de Coeficientes cepstrales en Frecuencia Mel

El análisis de la *Cmfcc* mostró un alto grado de similitud para las siguientes combinaciones: **EC-E**, **EC-EA**, **E-EA** y **A-E**.

La determinación de los primeros 5 *Fm* para cada elemento de los Grupos 1 y 2 se muestra en la figura 7.2.

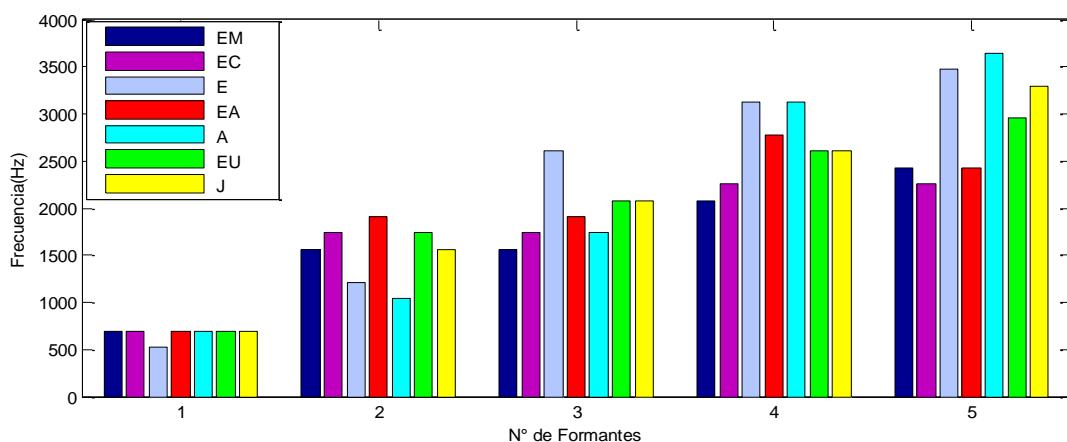


Figura 7.2. Formantes de los Grupos 1 y 2

Es factible observar que para las 7 señales analizadas correspondientes a los Grupos 1 y 2 existe semejanza en el valor de frecuencia del primer formante. Sin embargo ese grado de similitud disminuye en la medida que se analizan formantes de orden superior.

La determinación (DEP) en los Grupos 1 y 2 es mostrada en las figuras 7.3 y 7.4 respectivamente.

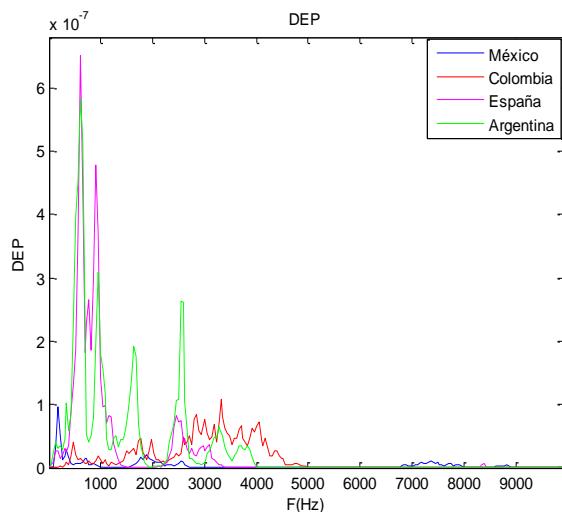


Figura 7.3. Grupo 1

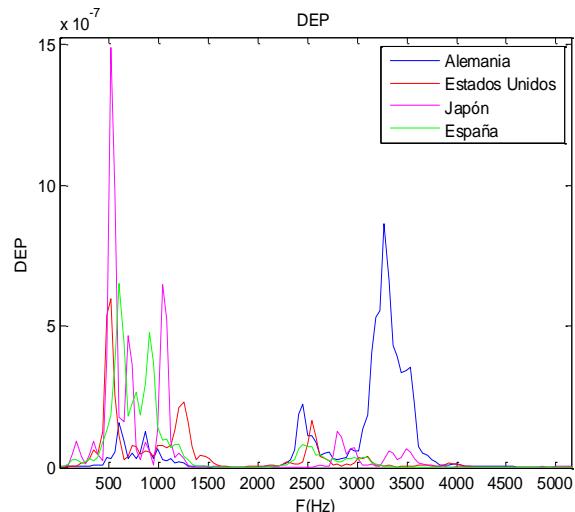


Figura 7.4. Grupo 2

En el Grupo 1 (figura 7.3) se observó alta similitud en la DEP para **E** (magenta) y **EA** (verde) en bajas frecuencias (por debajo de los 2 kHz).

El Grupo 2 (figura 7.4) presenta, para bajas y altas frecuencias, similitud de la DEP entre **EU** (rojo) y **E** (verde). Sin embargo se puede evidenciar que alrededor de los 2.5Khz existe semejanza de las DEP entre **A** (azul), **EU** (rojo) y **E** (verde).

7.6 Conclusiones

El análisis de las características temporales, prosódicas y frecuenciales en señales de audio para diferentes idiomas y dialectos permite mejorar la robustez de los sistemas que realicen procesamiento de señales de voz en diferentes entornos lingüísticos para el reconocimiento de emociones.

La formulación de esta etapa surge de la necesidad de evaluar, a partir de la comparación de características de las señales de voz, la distancia entre idiomas. Esto es debido a que se pretende construir un sistema reconocedor de emociones que no dependa del idioma de la base de datos en la que se probaron los algoritmos ni tampoco del idioma o dialecto del entorno donde se aplicará.

Se utilizó una guía de pronunciación que incluye archivos de audios en diferentes idiomas para encontrar características descriptivas que permitan reducir al mínimo la variación entre el idioma alemán (ya que el sistema de reconocimiento de emociones se realizará en función de este idioma) y el resto de los idiomas, principalmente el español. Los parámetros encontrados que cumplen con este objetivo resultaron ser: los mfcc, cuyo análisis de correlación mostró semejanza entre los idiomas alemán-español y español-español argentino y el primer formante cuyos resultados fueron semejantes para todas las muestras analizadas.

El parámetro mfcc no solo es una característica importante en la detección de emociones (como se demuestra en el capítulo 6) sino que también cumple con ser invariante para los idiomas alemán, español y español argentino.

7.7 Referencias

1. El Ayadi, M., Kamel, M. S., & Karray, F. (2011). *Survey on speech emotion recognition: Features, classification schemes, and databases*. *Pattern Recognition*, 44(3), 572-587.
2. Scherer, K. R., Johnstone, T., & Klasmeyer, G. (2003). *Vocal expression of emotion*. *Handbook of affective sciences*, 433-456.
3. Iriondo, I., Guaus, R., Rodríguez, A., Lázaro, P., Montoya, N., Blanco, J. M. & Longhi, L. (2000). *Validation of an acoustical modelling of emotional expression in Spanish using speech synthesis techniques*. In *ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion*.
4. Goudbeek, M., & Broersma, M. (2010). *The Demo/Kemo corpus: A principled approach to the study of cross-cultural differences in the vocal expression and perception of emotion*. In *7th International Conference on Language Resources and Evaluation (LREC 2010)* (pp. 2211-2215). ELRA.
5. Forvo. Diccionario de Pronunciación <http://es.forvo.com>
6. Carvajal, C. R., Herrero, R. Á., Vázquez, E. F., & Muñiz, A. S. G. *Algunas implicaciones de la teoría matemática de la información al análisis Inpout—output*.

CAPITULO 8

Resultados

8.1 Introducción

Los resultados obtenidos del desarrollo de la tesis serán expuestos en las siguientes secciones de manera de brindar facilidad en la interpretación de los mismos.

- **Sección 1: Resultados y desempeño de los sistemas de clasificación implementados en la Base de datos Berlín.**

En esta sección se mostrarán los resultados de la implementación de técnicas de clasificación lineales y no lineales, para detectar dos y luego cuatro emociones, mediante señales de voz. Para el entrenamiento de los sistemas de clasificación, se utilizaron los parámetros de la señal, descriptos en el capítulo 6. El objetivo es evaluar el desempeño de los diferentes sistemas de clasificación empleados, para detectar emociones mediante señales de voz.

- **Sección 2: Resultados y desempeño de clasificadores no lineales implementados en las bases de datos: Semaine, GVEESS.**

En esta sección se expondrán los resultados de la implementación de dos sistemas de clasificación no lineal, entrenados con bases de datos en inglés (Semaine) y en alemán (GVEESS) para la detección de 4 emociones. Se mostrarán los porcentajes de detección y el desempeño de los clasificadores cuando son aplicados a muestras que corresponden a las mismas bases datos que formaron parte del entrenamiento, o sea no se realiza implementación cruzada de clasificadores.

A fin de evaluar si los resultados de estas dos secciones son satisfactorios se compararán con los obtenidos de dos trabajos de investigación en reconocimiento de emociones utilizando el desempeño humano como detector. En [1] se realizó un experimento en reconocimiento de emociones utilizando un corpus de 1000 instancias, con 5 locutores. En [2] se implementó un programa interactivo que selecciona y reproduce los enunciados de un corpus de emociones en orden aleatorio y permite al individuo clasificar cada enunciado de acuerdo a su contenido emocional. Las emociones implicadas en ambos trabajos fueron: enojo, felicidad, triste, miedo y neutro. Los resultados obtenidos por ambos grupos de investigación dedicados a evaluar el reconocimiento no automático de emociones utilizando como detector el desempeño humano muestran que existe mayor facilidad en la detección de enojo y menor facilidad en la detección de miedo. Estos resultados serán tomados como referencia para el análisis del desempeño en el reconocimiento automático de emociones desarrollado en el presente trabajo.

- **Sección 3: Resultados de la implementación del clasificador entrenado con Base de datos Berlín para reconocimiento de emociones en Bases de datos: Semaine y GVEESS.**

En esta sección se mostrarán los resultados de implementar un clasificador entrenado en idioma alemán (Base de datos Berlín) para la detección de 4 emociones en idiomas inglés (Base de datos Semaine) y en alemán (bases de datos GVEESS), a fin de evaluar el desempeño del sistema de clasificación cuando es aplicado a muestras de otras bases de datos y en otro idioma. O sea, a partir de la implementación cruzada del sistema de clasificación, se determina la variabilidad o no de las características respecto al idioma.

- **Sección 4: Resultados en el plano bidimensional de Russell.**

En esta sección se utilizará el plano de valencia-activación para regionalizar emociones que han formado parte del entrenamiento de un sistema de clasificación y emociones que no han sido entrenadas, a fin de evaluar la posibilidad de ubicar en el plano bidimensional, emociones que pertenezcan a la misma región sin la necesidad de que hayan sido parte del entrenamiento de un sistema de clasificación.

- **Sección 5: Clasificación de Tristeza en 2 idiomas.**

En esta sección se expondrán los resultados de la implementación de tres sistemas de clasificación no lineal que utilizan una única característica descriptiva para detectar la emoción tristeza en idiomas alemán y castellano. Uno de los sistemas fue entrenado en alemán, otro en castellano, y un tercer sistema fue entrenado con ambos idiomas. El objetivo es evaluar el desempeño de los clasificadores, su implementación cruzada y la tasa de detección de la emoción tristeza para ambos idiomas.

8.2 SECCIÓN 1

Resultados y desempeño de los sistemas de clasificación implementados en la Base de datos Berlín

En esta sección se realiza un reporte de los resultados obtenidos mediante dos técnicas de clasificación para 2 y 4 emociones de la Base de Datos Berlín.

8.2.1 Reconocimiento de dos emociones:

El objetivo del sistema es reconocer cuál de las dos emociones ha sido expresada: felicidad o tristeza. Para ello se utilizan técnicas de clasificación que permiten usar características de la señal para identificar a qué clase o grupo pertenece. Se utilizaron técnicas lineales y no lineales para mostrar dos alternativas de clasificación y permitir la comparación entre ambas.

8.2.1.1 Técnica de clasificación Lineal: Análisis discriminante lineal

El análisis discriminante (LDA) permite realizar clasificaciones de objetos. La clasificación de un objeto se basa en una combinación lineal de variables discriminantes, construida de modo que maximice las diferencias entre grupos y minimice las diferencias intragrupo. El comando de Matlab® “classify” permite realizar el entrenamiento de grupos y devuelve la tasa de error aparente, esto es el porcentaje de observaciones en el entrenamiento que han sido mal clasificados, ponderados por probabilidades de los grupos.

Las características del entrenamiento fueron las siguientes:

- Descriptor: Coeficientes cepstrales en frecuencia Mel (*mfcc*)
- Emociones a clasificar: Felicidad-Tristeza
- Etiqueta implementada: Felicidad=1; Tristeza=0.
- Cantidad de muestras entrenadas: 30 muestras por estado emocional. 60 muestras totales.

- Cantidad de muestras analizadas: 68 muestras. 8 corresponden a muestras que no han sido parte del entrenamiento.

En la figura 8.1 se presenta la detección de felicidad y tristeza en 68 muestras utilizando un clasificador no lineal. La etiqueta, que es el estado conocido, es mostrada en violeta. La salida detectada está representada en rojo.

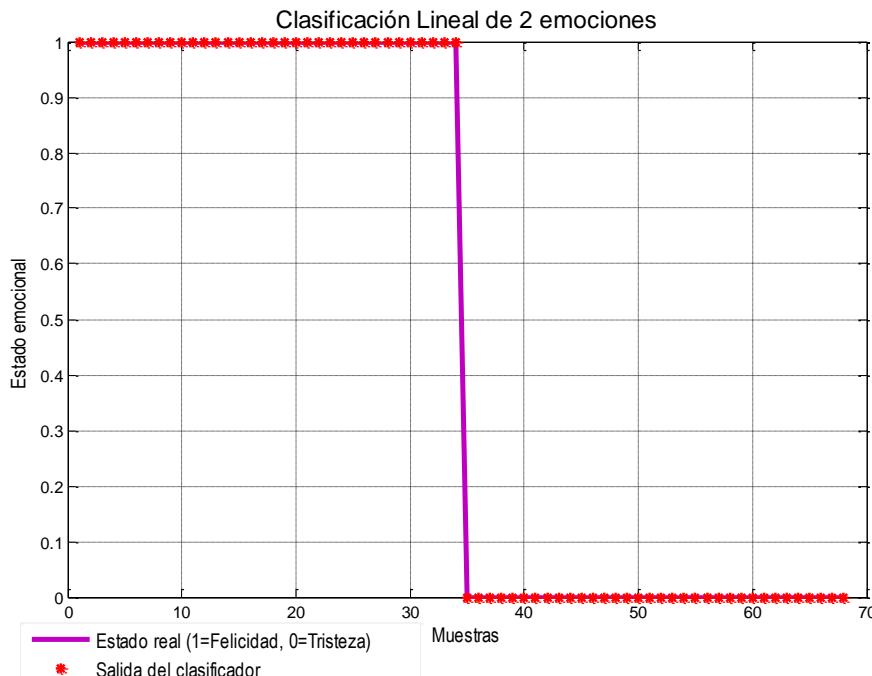


Figura 8.1. Detección de dos emociones mediante LDA

El error aparente que resulta del entrenamiento y de la aplicación en las muestras no entrenadas es del 0%.

Con los resultados obtenidos se realizó una matriz de confusión, siendo cada columna de la matriz, el número de detecciones de cada clase, mientras que cada fila representa a las instancias en la clase real (tabla 8.1).

Emoción	Tristeza	Felicidad	Aciertos
Tristeza	34	0	100 %
Felicidad	0	34	100%

Tabla 8.1. Matriz de confusión para dos emociones mediante LDA

La matriz de confusión obtenida muestra que la detección de emociones: **felicidad y tristeza** en la Base de datos Berlín utilizando como técnica de entrenamiento LDA es del **100%** con *mfcc* como característica de entrada.

8.2.1.2 Técnica de clasificación no lineal: Redes neuronales

Las redes neuronales (RN) son un modelo artificial y simplificado del cerebro humano, capaz de adquirir conocimientos a través de la experiencia. Se los describe como sistemas dinámicos

autoadaptativos. Son adaptables debido a la capacidad de autoajuste de los elementos procesales (neuronas) que componen el sistema. Son dinámicos, pues son capaces de estar constantemente cambiando para adaptarse a las nuevas condiciones.

Las redes neuronales permiten el reconocimiento de patrones, una de las formas de programación usadas para tal fin es la llamada backpropagation. Ese tipo de programación de red implica que el error es propagado hacia atrás a través de la red neuronal. Esto permite que los pesos sobre las conexiones de las neuronas ubicadas en las capas ocultas cambien durante el entrenamiento. El cambio de los pesos en las conexiones de las neuronas influye en la entrada global y en la activación de una neurona. Por lo tanto, es de utilidad considerar las variaciones de la función de activación al modificarse el valor de los pesos.

La RN está constituida por neuronas interconectadas y arregladas en tres capas básicas. Los datos ingresan por medio de la capa de entrada, pasan a través de la capa oculta y salen por la capa de salida. Cabe mencionar que la capa oculta puede estar constituida por varias capas.

En la figura 8.2 se muestra el esquema de RN utilizado.

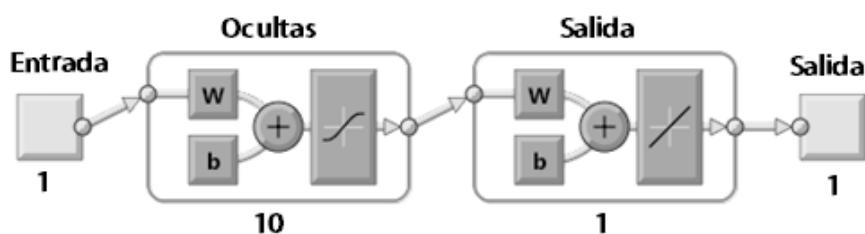


Fig. 8.2 Esquema de la red neuronal utilizada

Donde w representa los pesos sinápticos y b es una constante. El diseño utiliza 10 capas ocultas y 1 capa de salida. Matemáticamente el entrenamiento de la red viene caracterizado por:

$$y_i(t) = f_i(w_{ij} * y_j(t) + b) \quad (8.1)$$

Donde, $y_i(t)$ es la salida del entrenamiento.

La muestra se divide en tres conjuntos: el 70% de entrenamiento, el 15% de validación y el 15% restante de testeo fuera de muestra.

El entrenamiento de la red viene caracterizado por ser un entrenamiento supervisado, lo cual le proporciona a la red el valor de la salida. Partiendo de un conjunto de pesos aleatorio, se inicia el proceso iterativo en el cual se intenta minimizar el error de la estimación. En este caso se ha utilizado el algoritmo Levenberg-Marquardt.

Las características del entrenamiento fueron las siguientes:

- Descriptor: *mfcc*.
- Cantidad de muestras: 30 por estado emocional (60 en total).
- Etiqueta: Felicidad (1)- Tristeza (0).
- Cantidad de neuronas en la capa oculta: 10

Los resultados del entrenamiento son mostrados en la tabla 8.2:

	Entrenamiento	Validación	Prueba
Regresión	0.9482	0.99	0.9882
Error	0.0248	6,06e-05	0.0072

Tabla 8.2. Resultados de entrenamiento mediante RN

El error cuadrático medio de las tres series decrece hasta la iteración 20. Los errores de validación y testeо están muy por debajo del error de entrenamiento.

El valor de R se sitúa por encima del 0.94 para todos los casos. Siendo el valor total de R del 0.96.

El sistema entrenado para dos emociones fue implementado con la Base de datos Berlín utilizando solo las muestras que corresponden a emociones de felicidad y tristeza. El resultado de la detección se muestra en la figura 8.3. El estado conocido o etiqueta se muestra en negro y la salida detectada se representa en rojo.



Figura 8.3 Detección de dos emociones mediante RN.

Este resultado se analizó mediante una matriz de confusión que se muestra en la tabla 8.3

Emoción	Tristeza	Felicidad	Aciertos
Tristeza	29	0	96.6 %
Felicidad	0	30	100%

Tabla 8.3. Matriz de confusión para dos emociones mediante RN

La matriz de confusión obtenida describe que la tasa de acierto es del **100%** para la detección de **felicidad** y del **96.6 %** para la detección de la emoción **tristeza**. El error total de detección para 60 muestras de la Base de datos Berlín es del 1.67%, utilizando los *mfcc* como característica de entrada.

Los resultados de la implementación de las técnicas de clasificación lineal y no lineal, utilizando la misma característica de entrada, para la detección de dos emociones opuestas en el plano de valencia- activación son satisfactorios ya que la tasa de detección para ambos casos supera el 96 %. El resultado en este caso depende de la capacidad de esta característica de describir la señal y no del kernel del clasificador.

8.2.2 Reconocimiento de cuatro emociones:

El sistema de reconocimiento para 4 emociones: Felicidad, Tristeza, Miedo y Enojo fue desarrollado mediante la técnica no lineal de RN. Dada la dificultad en la clasificación de 4 emociones se desarrollaron una serie de pruebas variando la cantidad de características descriptivas de la señal.

En todos los casos, la RN se diseña en base al algoritmo de Levenberg-Marquardt con técnica backpropagation. La muestra se divide en tres conjuntos de forma aleatoria: el 70% de entrenamiento, el 15% de validación y el 15% restante de testeo fuera de muestra. Se utilizan 120 instancias que incluyen 30 muestras por emoción. La salida del clasificador depende de las etiquetas seleccionadas.

Estas etiquetas se designan de la siguiente manera:

- Enojo= 3
- Felicidad=2
- Miedo=1
- Tristeza=0

8.2.2.1 Prueba 1. Uso de 11 características

En la prueba 1 se desarrolló el sistema de clasificación incluyendo 11 características de entrada (tabla 8.4) (descriptas en el capítulo 6).

Características	
nV	
CPZ	
E/cpz	
aE/cpz	
K	
fDEP	
DEP;	
aDEP	
Mfcc_media	
Mfcc_total	
Ma	

Tabla 8.4. Características utilizadas en la clasificación de la prueba 1

La red usa 11 capas de entrada, 25 capas ocultas y 1 capa de salida. El esquema de la red se presenta en la figura 8.4.

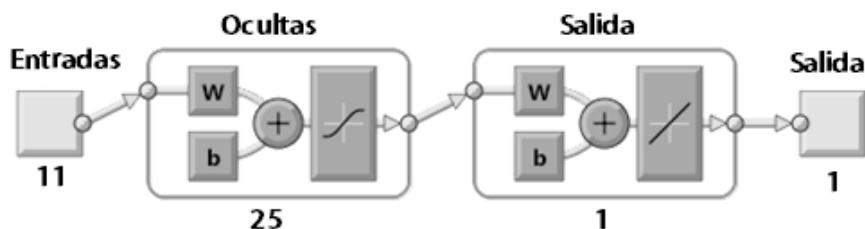


Figura 8.4. Esquema de la red neuronal .Prueba 1

Los resultados del entrenamiento son mostrados en la tabla 8.5:

	Entrenamiento	Validación	Prueba
Regresión	0.9569	0.9352	0.8136
Error	0.5042	0.2045	0.1223

Tabla 8.5. Resultados del entrenamiento para 4 emociones. Prueba 1

El valor de R se sitúa por encima del 0.9 para entrenamiento y validación, siendo 0.81 el obtenido en la prueba de testeо. El valor total de R resulta de 0.92, este resultado indica el buen aprendizaje de la red.

El sistema entrenado con cuatro emociones fue implementado en 120 muestras de la Base de datos Berlín que incluyen sólo las emociones felicidad, enojo, miedo y tristeza (el sistema de clasificación no fue aplicado en la base de datos completa que incluye 6 emociones en total).

El resultado de la detección se muestra en la figura 8.5. El resultado conocido (etiqueta) se muestra en negro y la salida detectada se representa en rojo.

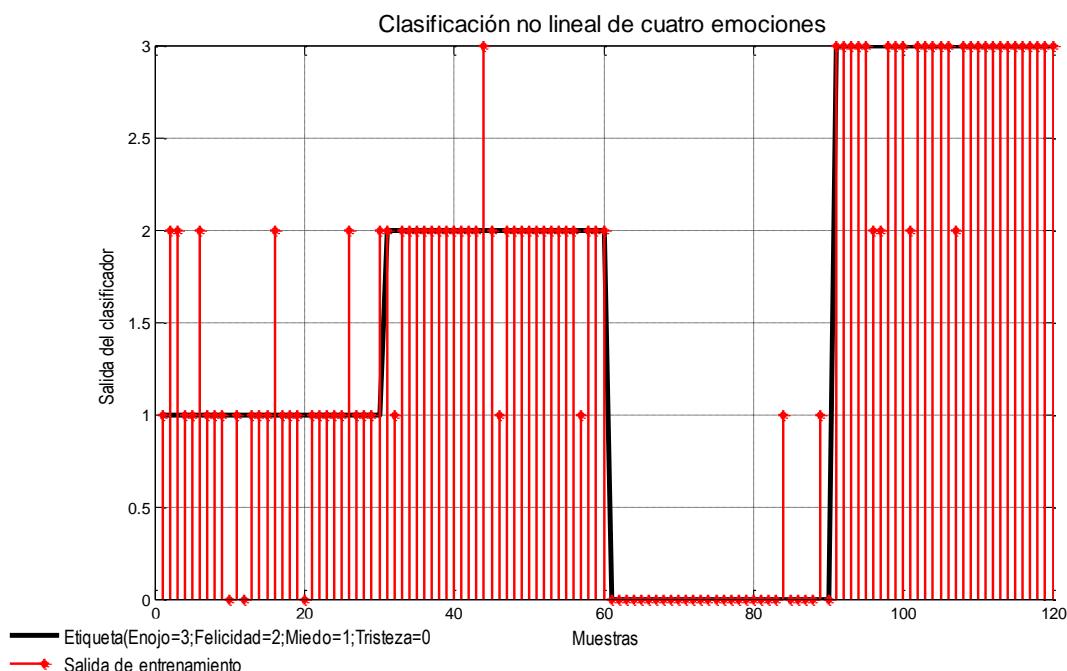


Figura 8.5 Detección de cuatro emociones. Prueba 1

Este resultado se analizó mediante una matriz de confusión que se muestra en la tabla 8.6

Categoría	Miedo	Felicidad	Enojo	Tristeza	Aciertos
Miedo	21	6	0	3	70%
Felicidad	3	26	1	0	86.66%
Enojo	0	4	26	0	86.66%
Tristeza	2	0	0	28	93.34%

Tabla 8.6. Matriz de confusión para 4 emociones. Prueba 1.

La matriz de confusión obtenida muestra a la **tristeza** como la emoción con mayor tasa de acierto (**93.34 %**) y al miedo como la emoción con menos porcentaje de detección (70%). El error de detección total es de 15.83 %.

8.2.2.2 Prueba 2. Uso de 12 características

La experimentación 2 incluye 12 características que describen la señal de voz. Estas variables se detallaron en el capítulo 6. En la siguiente tabla se indican los parámetros que se mantuvieron constantes respecto a la prueba 1 y aquellos que fueron agregados en esta experimentación. La

adición de estas características se deriva del notable uso de las mismas en la bibliografía [3] [4] [5] [6] [7] [8], ya que permiten una buena diferenciación en los niveles de excitación.

Características que se preservaron de la prueba 1	Características agregadas
CPZ	E
E/Cpz	Me
DEP	Rango_fo
aDEP	E/cant de segmentos detectados
mfcc_media	Duración total del silencio
mfcc_total	
Ma	

Tabla 8.7 Características utilizadas en la clasificación de la prueba 2

La RN utiliza 12 capas de entrada, 25 capas ocultas y 1 capa de salida. El esquema de la red se presenta en la figura 8.6.

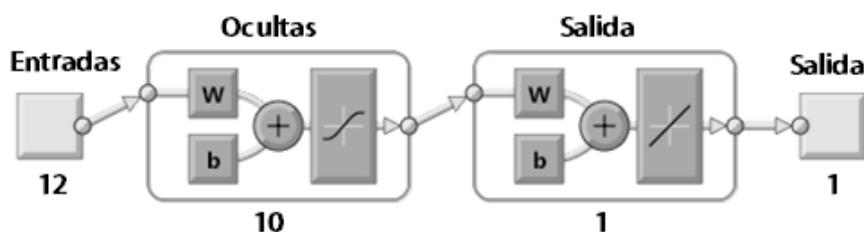


Figura 8.6. Esquema de la red neuronal .Prueba 2

Los resultados del entrenamiento son mostrados en la tabla 8.8:

	Entrenamiento	Validación	Prueba
Regresión	0.9748	0.8142	0.9172
Error	0.0629	0.4128	0.2985

Tabla 8.8. Resultados del entrenamiento para 4 emociones. Prueba 2

El valor de R se sitúa por encima del 0.9 para entrenamiento y prueba, siendo 0.81 el obtenido en la validación. El valor total de R resulta de 0.93, superior al obtenido en la prueba 1, esto se debe a que la red tuvo un mejor aprendizaje con estas características.

El sistema de 12 características fue implementado en 120 muestras de la Base de datos Berlín que incluyen sólo las emociones felicidad, enojo, miedo y tristeza (son las mismas muestras que se utilizaron en la prueba 1).

El resultado de la detección se expone en la figura 8.7. El resultado conocido (etiqueta) se muestra en negro y la salida detectada se presenta en rojo.

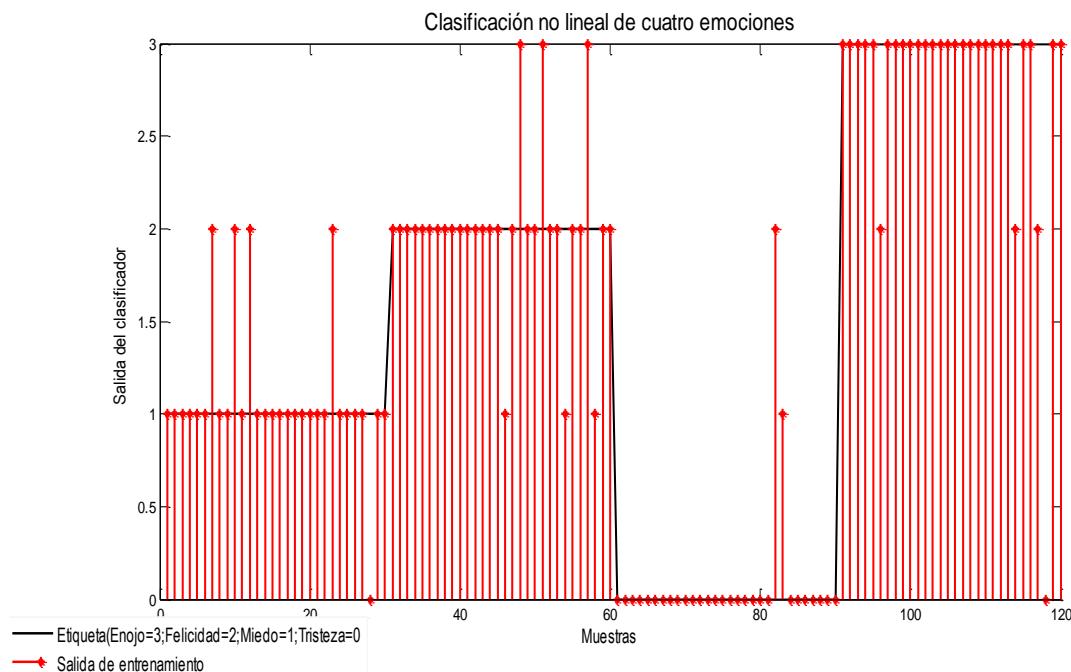


Figura 8.7 Detección de cuatro emociones. Prueba 2

Posteriormente, se analizaron estos resultados mediante una matriz de confusión que se muestra en la tabla 8.9

Categoría	Miedo	Felicidad	Eñojo	Tristeza	Aciertos
Miedo	25	4	0	1	83.33%
Felicidad	3	24	3	0	80%
Eñojo	0	3	26	1	86.66%
Tristeza	1	1	0	28	93.34

Tabla 8.9. Matriz de confusión para 4 emociones. Prueba 2

El análisis de los resultados muestra a la **tristeza** como la emoción mejor detectada (**93.34%**) y la felicidad como la emoción con menor tasa de aciertos (80%). El error de detección total es del 14.16 %, inferior al obtenido con el sistema de clasificación de 11 características.

8.2.2.3 Comparación entre Sistema de Reconocimiento Automático y no Automático.

A continuación, se expone una tabla comparativa (tabla 8.10) entre el reconocimiento no automático de emociones (desempeño humano) y los dos resultados del reconocimiento automático: prueba 1 (11 características) y prueba 2 (12 características).

Emoción	Reconocimiento No automático		Reconocimiento Automático	
	% Aciertos [1]	%Aciertos [2]	%Acierto Prueba 1	%Acierto prueba 2
Felicidad	88%	61.4 %	86.66%	80%
Tristeza	80%	68.3 %	93.34	93.34%
Enojo	96%	72.5 %	86.66%	86.66%
Miedo	64%	49.3 %	70%	83.33%

Tabla 8.10. Comparación entre el desempeño humano y el reconocimiento automático

En ambos sistemas automáticos de reconocimiento de emociones, el de 11 y 12 características, la **detección de las 4 emociones clasificadas es 15% superior** a la obtenida mediante el reconocimiento no automático de emociones citadas en [2].

La detección de la **tristeza** del sistema automático refiere una **mejora de 13%** en relación con el mejor resultado del desempeño humano que corresponde al trabajo de Dellaert et al. [1].

El miedo es la emoción con menor tasa de detección del sistema no automático (49 % [2], 64% [1]), por lo tanto, la emoción más difícil de clasificar por el hombre. Si sólo hacemos la comparación con el mejor resultado del desempeño humano que corresponde al trabajo [1], observamos que ambos sistemas desarrollados para reconocimiento automático permiten una **mejora del 6% y del 19%** en la detección según se utilicen 11 o 12 características, respectivamente.

8.3 SECCIÓN 2

Resultados y desempeño de clasificadores no lineales implementados en las bases de datos: Semaine (inglés), GVEESS (alemán)

En esta sección se diseñaron dos sistemas de reconocimiento (uno para cada corpus) de 4 emociones: Felicidad, Tristeza, Miedo y Enojo, implementando la técnica no lineal de RN con algoritmo de Levenberg-Marquardt con técnica backpropagation. La red usa 12 capas de entrada, 25 capas ocultas y 1 capa de salida. Para todos los casos se dividió la muestra en tres conjuntos de forma aleatoria:

- Conjunto 1: 70% de entrenamiento
- Conjunto 2: el 15% de validación
- Conjunto 3: 15% de testeо

Dados los buenos resultados del sistema de clasificación de 12 características desarrollado en la sección anterior, se decidió implementar estas características en los sistemas de clasificación entrenados con las Bases de datos Semaine y GVEESS y cuyos resultados son expuestos en esta sección.

El sistema de etiquetas es el mismo al usado en la sección anterior.

8.3.1 Implementación de técnicas de clasificación en Base de datos Semaine (inglés)

El sistema de clasificación utiliza 56 instancias de la Base de datos Semaine. Este corpus no incluye la emoción miedo, por ello sólo se analizarán tres emociones, bajo las siguientes etiquetas:

- Enojo =etiqueta 3
- Felicidad= etiqueta 2
- Tristeza=etiqueta 0

Los resultados del entrenamiento utilizando la Base de datos Semaine se presentan en la tabla 8.11.

	Entrenamiento	Validación	Prueba
Regresión	0.9769	0.7361	0.80
Error	0.073	1.4926	0.7819

Tabla 8.11. Resultados del entrenamiento para 3 emociones. Semaine

El valor de R se sitúa en 0.97 para entrenamiento, 0.8 para el testeо y 0.73 para validación. El valor total de R resulta de 0.88, inferior al obtenido en la sección 1.

El sistema de clasificación desarrollado fue implementado en 56 muestras, de la Base de datos Semaine, que incluyen sólo las emociones felicidad, enojo y tristeza (no fue implementado en la totalidad de la base de datos ya que ésta contiene emociones que no han sido tomadas en cuenta en el presente desarrollo).

Los resultados de la implementación fueron analizados mediante una matriz de confusión que se muestra en la tabla 8.12

Categoría	Felicidad	Eñojo	Tristeza	Aciertos
Felicidad	13	4	1	72.22%
Eñojo	2	17	1	85%
Tristeza	0	0	18	100%

Tabla 8.12 Matriz de confusión para 3 emociones

La matriz de confusión muestra que la emoción **mejor detectada fue la tristeza (100%)** y la emoción con menor tasa de detección fue la felicidad (72.2%). El error de detección total es del 14.28%, semejante al obtenido con el sistema de clasificación de 12 características entrenado con la base de datos Berlín.

8.3.2 Implementación de técnicas de clasificación en Base de datos GVEESS (alemán)

Este sistema de clasificación usa 63 instancias de la Base de datos GVEESS para el entrenamiento de 4 emociones: enojo, felicidad, miedo, tristeza.

Los resultados del entrenamiento de la RN se presentan en la tabla 8.13.

	Entrenamiento	Validación	Prueba
Regresión	0.9395	0.7169	0.7063
Error	0.1442	0.5863	0.7490

Tabla 8.13. Resultados del entrenamiento para 4 emociones. GVEESS

El valor de R se encuentra en 0.93 para entrenamiento y 0.7 para el testeо y validación. El valor total de R resulta de 0.88, similar al obtenido en el entrenamiento de la base de datos Semaine.

El clasificador entrenado fue implementado en 63 muestras, de la Base de datos GVEESS, correspondientes a las emociones felicidad, enojo, miedo y tristeza (no fue implementado en el corpus completo que incluye 14 emociones). Luego, se determinó el porcentaje de aciertos para cada emoción mediante el uso de la siguiente matriz de confusión (tabla 8.14).

Categoría	Miedo	Felicidad	Enero	Tristeza	Aciertos
Miedo	10	3	0	3	62.52%
Felicidad	3	13	0	0	81.25%
Enero	2	2	12	0	75%
Tristeza	2	0	0	13	86.66%

Tabla 8.14 Matriz de confusión para 4 emociones. GVEESS

Los resultados obtenidos muestran que la **emoción tristeza presenta la mejor tasa de detección** para este corpus (**86.66%**).

8.3.3 Comparación entre Sistema de Reconocimiento Automático y no Automático

En esta subsección se realiza una comparación entre el reconocimiento no automático de emociones que corresponde al desempeño humano y el reconocimiento automático entrenado en bases de datos de diferentes idiomas. Los resultados de esta comparación se presentan en la tabla 8.15.

Emoción	Reconocimiento No automático		Reconocimiento Automático	
	% Aciertos [1]	%Aciertos [2]	%Aciertos Semaine (inglés)	%Aciertos GVEESS (alemán)
Felicidad	88%	61.4 %	72.22 %	81.25%
Tristeza	80%	68.3 %	100 %	86.66%
Enero	96%	72.5 %	85 %	75%
Miedo	64%	49.3 %	--	62.52%

Tabla 8.15. Comparación entre el desempeño humano y el reconocimiento automático

El análisis comparativo de los dos sistemas automáticos en relación al desempeño humano evidencia que la detección de las 4 emociones mediante los clasificadores entrenados en Semaine y en GVEESS **supera en más del 10%** a la detección mediante el desempeño humano expuesto en el trabajo [2].

En relación al trabajo [1], la **detección de tristeza** (resaltada en rosado) es la única emoción cuya tasa de acierto es **superada en un 20% y 6%** por el sistema automático de detección entrenado con Semaine y GVEESS, respectivamente. La detección del resto de las emociones es mejor resuelta por el sistema no automático.

8.4 SECCIÓN 3

Resultados de la implementación del clasificador entrenado con Base de datos Berlín en el reconocimiento de emociones de las Bases de datos: Semaine (inglés) y GVEESS (alemán)

En esta sección se implementó el sistema de clasificación de 12 características entrenado con la Base de datos Berlín, en los corpus Semaine y GVEESS. El objetivo es exhibir el desempeño del clasificador entrenado en alemán aplicado a bases de datos en inglés y alemán, o sea determinar el desempeño cruzado del sistema de clasificación. Es importante aclarar que debido a que la emoción miedo no está incluida en Semaine, solo se determinan los porcentajes de acierto para las emociones: felicidad, enojo y tristeza.

Los resultados de la aplicación del clasificador entrenado con la base de datos Berlín en 56 muestras de la Bases de datos Semaine y en 63 muestras de la base de datos GVEESS se exponen en la tabla 8.16.

Categoría	% de Aciertos	
	Semaine	GVEESS
Miedo	--	37.5%
Felicidad	16.66%	12.5%
Enojo	65%	0%
Tristeza	27.77%	53.33%

Tabla 8.16. Porcentajes de aciertos obtenidos de la implementación del clasificador Berlín en las Bases de datos Semaine y GVEES.

Los resultados de la aplicación del clasificador Berlín en las Bases de datos Semaine y GVEESS no son satisfactorios, demostrando la dependencia al idioma de las características implementadas y del esquema de clasificación. Es muy frecuente que los sistemas de reconocimiento automático entrenados en un idioma determinado bajen su rendimiento cuando son aplicados en muestras de otros idiomas.

8.5 SECCIÓN 4

Resultados en el plano bidimensional de Russell

El diagrama emocional de Russell [9] tal como se detalla en el capítulo 2 regionaliza las emociones en un espacio bipolar de dos dimensiones. La dimensión horizontal, también llamada valencia, corresponde al carácter positivo o negativo de las emociones (emociones de placer para el eje positivo de la abscisa y emociones de desagrado para el eje negativo) y la dimensión vertical corresponde a la activación de las emociones (emociones de excitación para el eje positivo de la ordenada y emociones de relajación para el eje negativo) (figura 8.8).

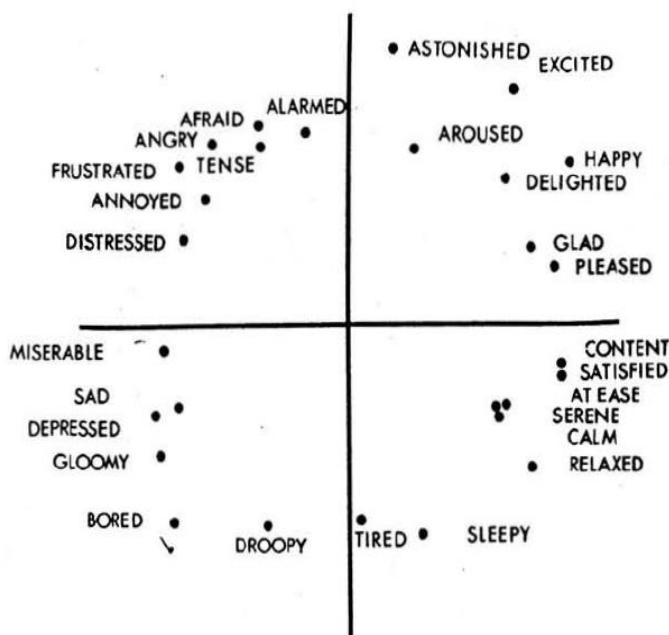


Figura 8.8. Modelo Circumplejo de Russell, J.A. (1980)

Teniendo en cuenta este diagrama y la posición angular donde se ubican las emociones según el modelo Circumplejo de escalamiento unidimensional, se localizaron en el plano las emociones detectadas por el sistema de clasificación.

Se tomaron como referencia las emociones felicidad, tristeza, enojo y miedo. Las posiciones angulares de estas emociones de referencia, según el Escalamiento Unidimensional, se detallan en la tabla 8.17.

Escalamiento Unidimensional

Emoción	Ángulo correspondiente
Felicidad	34°
Enojo	130°
Miedo	117°
Tristeza	203°

Tabla 8.17. Posiciones angulares de las emociones de referencia

Se tuvieron en cuenta sólo los ángulos de las emociones de referencia, por ello éstas se ubicarán en el círculo unitario del diagrama polar (figura 8.9).

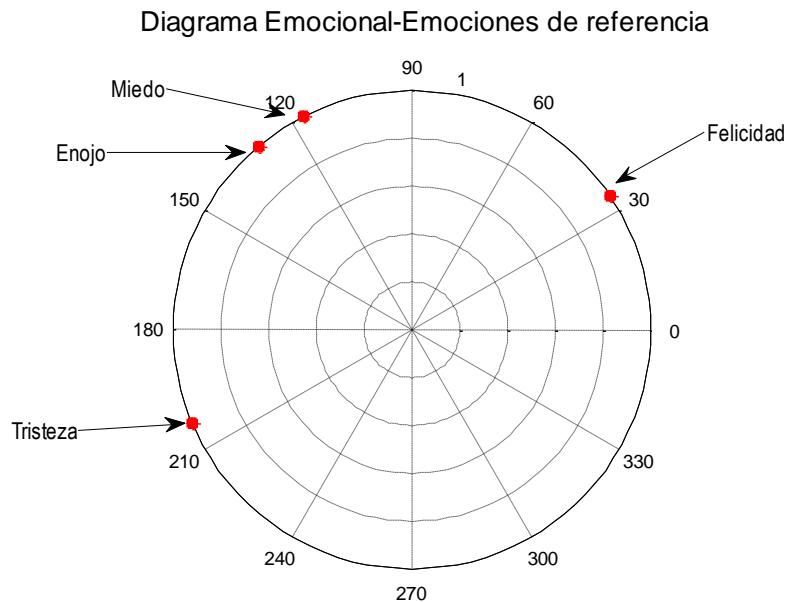


Figura 8.9. Diagrama emocional con las emociones de referencia

8.5.1 Regionalización de emociones entrenadas

Se utilizaron las Bases de datos: Berlín y GVEESS completas (o sea, con la totalidad de emociones que incluyen, y que son descriptas en el capítulo 6).

El clasificador de 11 características entrenado con sólo cuatro emociones de las 6 que forman parte de la Base de datos Berlín (detallado en la [sección 8.2.2.1](#)) se aplicó a este corpus completo (que incluye las 4 emociones que han sido entrenadas por el clasificador (felicidad, miedo, enojo y tristeza) y 2 emociones que no han formado parte del entrenamiento (asco y aburrimiento)).

Por otro lado, el clasificador de 12 características entrenado con la Base de datos GVEESS (detallado en la [sección 8.3.2](#)), fue aplicado a la base completa (que incluye 4 emociones que han sido entrenadas por el clasificador (felicidad, miedo, enojo y tristeza) y 10 emociones que no han formado parte del entrenamiento).

En primer lugar, sólo se graficaron en el mapa, las emociones que formaron parte del entrenamiento del clasificador. Ambas bases de datos son alemanas, de manera que la variable idioma se mantuvo constante. Se utilizó el clasificador de 12 características para GVEESS porque dio mejores resultados que el de 11 características.

Las emociones entrenadas que corresponden a: felicidad, tristeza, miedo y enojo, se ubicaron en el diagrama aplicando una correspondencia entre la salida del clasificador y las posiciones angulares de las emociones, proporcionadas por la experimentación de Russell. La salida del clasificador depende de las etiquetas seleccionadas. Las etiquetas se mantienen como se mencionó en la [sección 8.2.2](#).

Se estableció una correspondencia directa entre la etiqueta, que indica una emoción entrenada, y las posiciones angulares del plano de Russell. Por ejemplo, la emoción felicidad que corresponde a la etiqueta 2, se designa en el plano activación-valencia en la posición angular de 34°. Entonces, todas las señales del corpus ingresadas a este algoritmo que sean interpretadas por el clasificador como felicidad se ubicarán a 34° en el diagrama polar.

En las siguientes gráficas, se muestran las emociones: enojo (negro) y felicidad (magenta). Las emociones de referencia se exhiben en rojo. Se muestran los resultados en gráficas individuales para una mejor visualización.

En la figura 8.10 se ubican dos de las cuatro emociones que formaron parte del entrenamiento con Berlín y en la figura 8.11 dos de las cuatro emociones entrenadas con GVEESS. Cada punto en el diagrama se corresponde a una salida del clasificador de una emoción entrenada, en función de su etiqueta. Por ejemplo, en el caso de la figura 8.10, todos los puntos de color magenta corresponden a la emoción felicidad (para este caso, 30 muestras de felicidad que pertenecen a la base de datos Berlín), la salida del clasificador debería ubicar estos puntos a 34° en el plano de emociones, siempre que su tasa de detección sea del 100%. Sin embargo como existe error en la detección, algunos puntos de la emoción felicidad aparecen ubicados en el plano a 130° , posición angular del enojo.

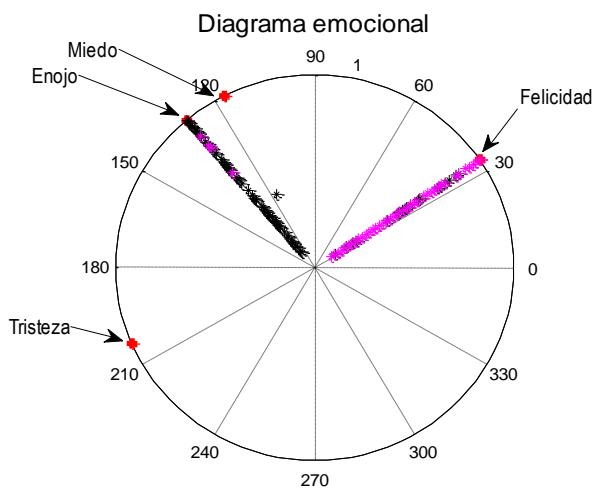


Figura 8.10 Plano emociones Berlín: enojo (negro), felicidad (magenta)

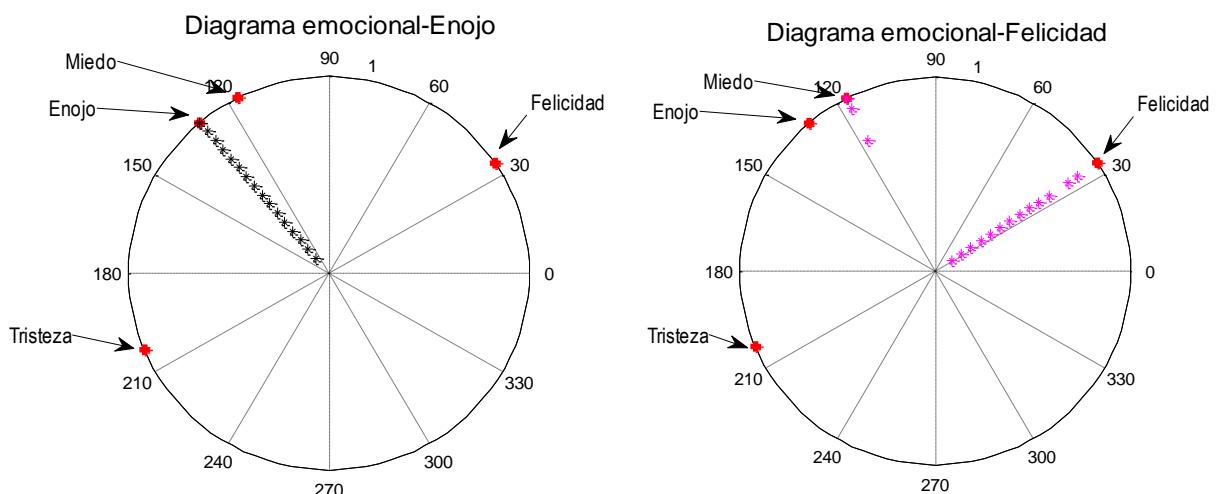


Figura 8.11 Plano emociones GVEESS: enojo (izquierda), felicidad (derecha).

Las figuras siguientes corresponden a las emociones miedo (verde) y tristeza (azul). Se muestran los resultados en gráficas individuales para una mejor visualización. En la figura 8.12 son expuestas dos las cuatro emociones entrenadas con Berlín y en la figura 8.13 se ubican dos de las cuatro emociones entrenadas con GVEESS.

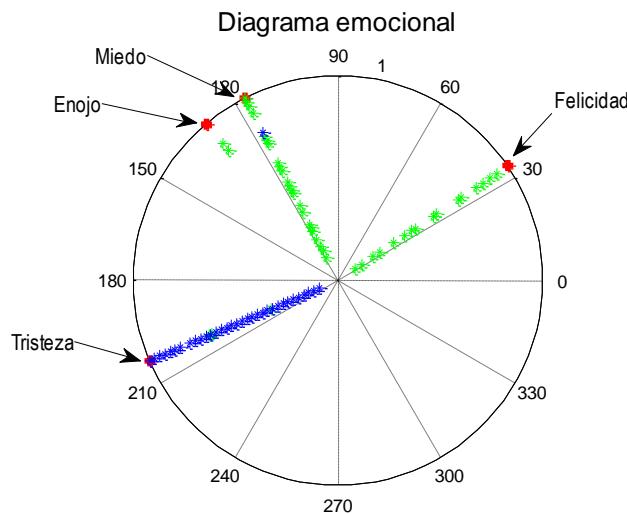


Figura 8.12 Plano emociones Berlín: miedo (verde), tristeza (azul)

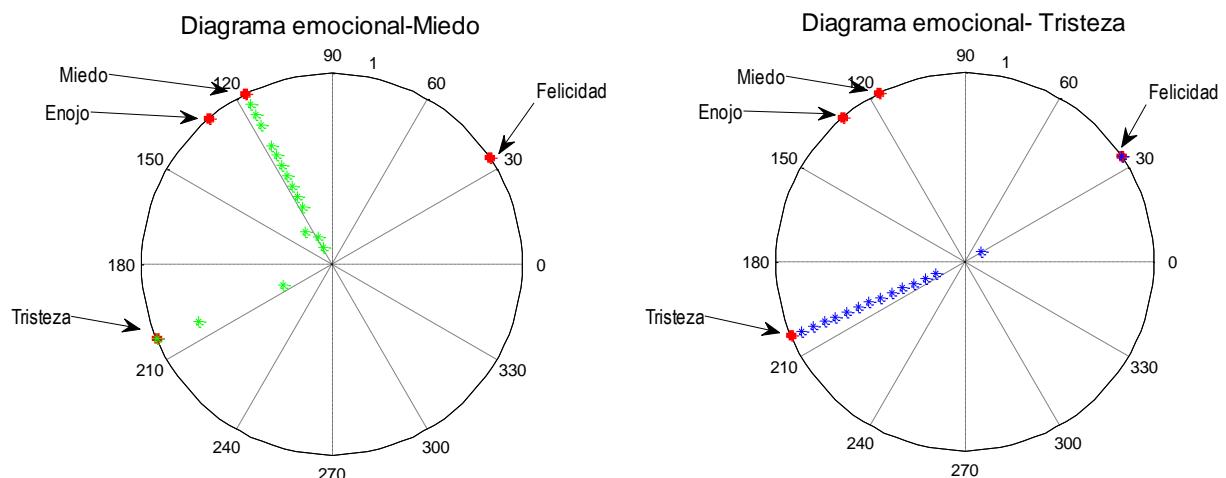


Figura 8.13. Plano emociones GVEESS: miedo (izquierda), tristeza (derecha).

Es importante aclarar que la detección y posterior regionalización de estas cuatro emociones se realiza aplicando los clasificadores Berlín y GVEESS entrenados con 4 emociones a las bases de datos completas (6 emociones en Berlín y 14 emociones en GVEESS), por lo tanto la probabilidad de que una emoción sea mal clasificada se incrementa.

La tasa de detección por emoción para las dos bases de datos se muestra en la tabla 8.18.

Emoción	Porcentaje de detección	
	B.D.Berlín	B.D GVEESS
Felicidad	93.44 %	81.25 %
Enojo	86.67%	100 %
Miedo	70 %	75 %
Tristeza	96.67%	86.66 %

Tabla 8.18. Porcentaje de detección aplicado a los corpus Berlín y GVEESS completos

La emoción mejor detectada por el clasificador Berlín de 11 características resultó ser la tristeza, con un porcentaje de acierto del **96.67 %**. Los resultados del clasificador entrenado con la base de datos GVEESS, mostraron que la emoción con mayor tasa de acierto, fue el enojo con un porcentaje de detección del **100%**, siendo la tristeza la segunda emoción mejor detectada por este clasificador. Para ambos clasificadores, el miedo fue la emoción peor detectada con porcentajes de detección similares (70-75%).

En una segunda etapa, se eligieron dos emociones que están incluidas en ambas bases de datos: Berlín y GVEESS, pero que no formaron parte del entrenamiento. Se seleccionaron las siguientes emociones no entrenadas:

- Aburrimiento
- Asco

8.5.2 Regionalización de emociones no entrenadas

La regionalización de las emociones no entrenadas se desarrolló haciendo uso de los clasificadores: Berlín y GVEESS, que permiten la clasificación de solo 4 emociones, como ya fue mencionado. Estos clasificadores aplicados al resto de las muestras que forman parte de las bases de datos , y que no formaron parte del entrenamiento, generan salidas que según las etiquetas ya mencionadas, pueden indicar solo una de las 4 emociones entrenadas.

Luego, cada una de las salidas clasificadas es graficada en el plano activación-valencia bajo el mismo procedimiento que las emociones entrenadas.

Por ejemplo, se toma una muestra del corpus del cual se conoce su pertenencia a las emociones de aburrimiento. A esta muestra, se le aplica el clasificador Berlín, y su salida indica que esa instancia pertenece a la emoción Tristeza (o sea su salida corresponde a 0). Entonces, se realiza la correspondencia de Tristeza con la posición angular asociada (203°), por lo tanto se genera una marca en esa posición angular de la gráfica polar. Cada instancia clasificada es graficada como un punto en el diagrama de Russell y ubicada en la región de una de las 4 emociones de referencia.

8.5.2.1 Aburrimiento

La emoción aburrimiento (bored en inglés) según el escalamiento Unidimensional del diagrama de Russell está ubicada en el tercer cuadrante a 224° (figura 8.14). Esta emoción (marcada en rojo)

comparte región con las emociones tristeza, depresión y decaimiento. Se ubica en esta región ya que presenta valencia negativa y baja activación.

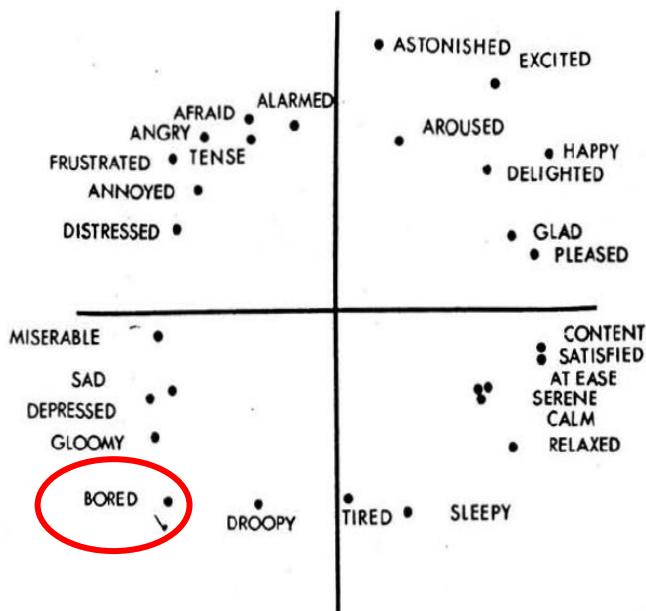


Figura 8.14 Plano de Russell. Aburrimiento

Las muestras correspondientes a la emoción aburrimiento fueron ingresadas a los clasificadores: Berlín y GVEESS. Los resultados se muestran, respectivamente, en las figuras 8.15 y 8.16.

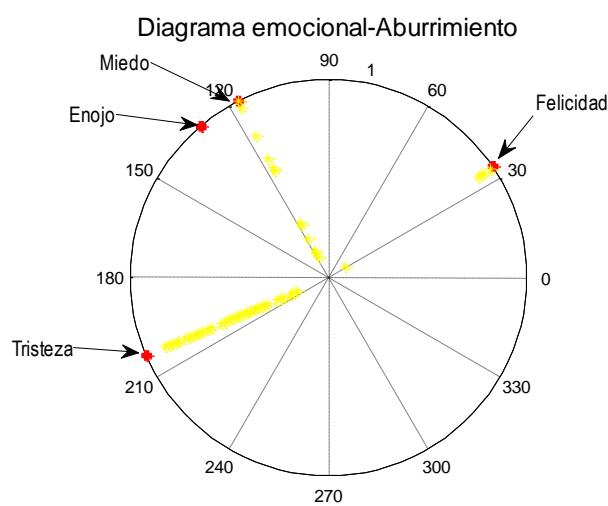


Figura 8.15. Plano emociones Berlín.
Aburrimiento

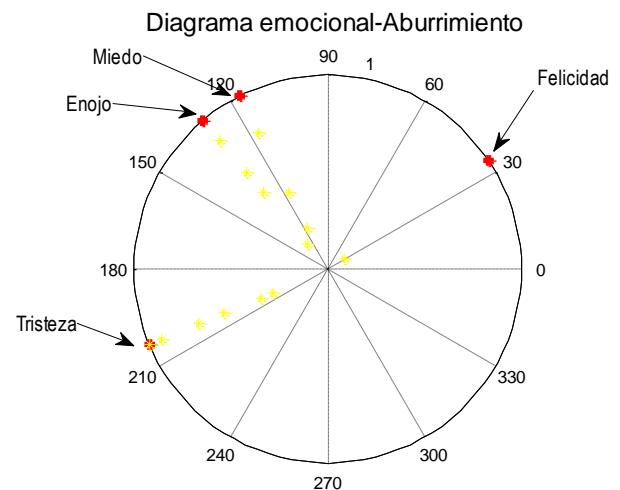


Figura 8.16. Plano emociones GVEESS
Aburrimiento

En la tabla 8.19 se muestran los porcentajes de detección de cada emoción detectada, utilizando ambos clasificadores.

Emoción	Porcentaje de detección- Aburrimiento	
	B.D. Berlín	B.D GVEESS
Felicidad	9.80 %	7.14 %
Enojo	0 %	28.57 %
Miedo	23.52 %	21.42 %
Tristeza	66.66%	42.85 %

Tabla 8.19. Porcentaje de detección de emociones. Aburrimiento

Como ya fue mencionado, la emoción que debería tener una mayor tasa de detección es la tristeza ya que comparte región con el aburrimiento, emoción que se pretende analizar. Evaluando los resultados de ambos clasificadores, se puede observar que la emoción con mayor tasa de detección cuando se analiza aburrimiento, es la tristeza, lo cual indica la buena regionalización de los clasificadores para esta emoción no entrenada, siendo superior el desempeño del Clasificador Berlín (66.66%).

8.5.2.2 Asco

El estado emocional asco, según el escalamiento Unidimensional de Russell, se encuentra incluido en el grupo de emociones “Distressed”. Este grupo que indican angustia se ubica en el segundo cuadrante, a 154° en el plano polar (figura 8.17, marcado en celeste), ya que presentan valencia negativa y alta activación. Este conjunto comparte región con las emociones: miedo, enojo y frustración entre otras.

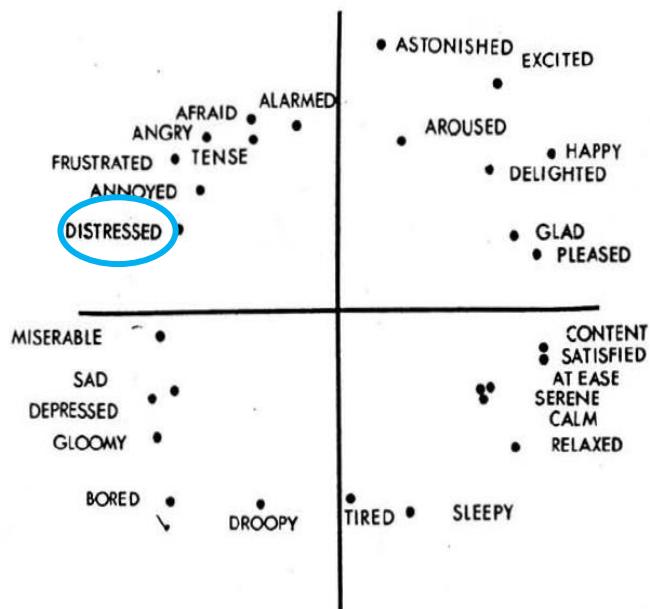


Figura 8.17 Plano de Russell. Asco

Los resultados de implementar los clasificadores: Berlín y GVEESS en las instancias correspondientes a asco se presentan en las siguientes figuras 8.18 y 8.19, respectivamente.

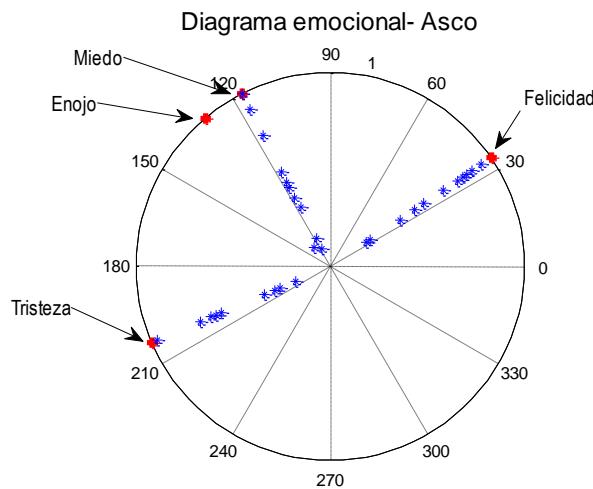


Figura 8.18 Plano emociones Berlín.
Asco

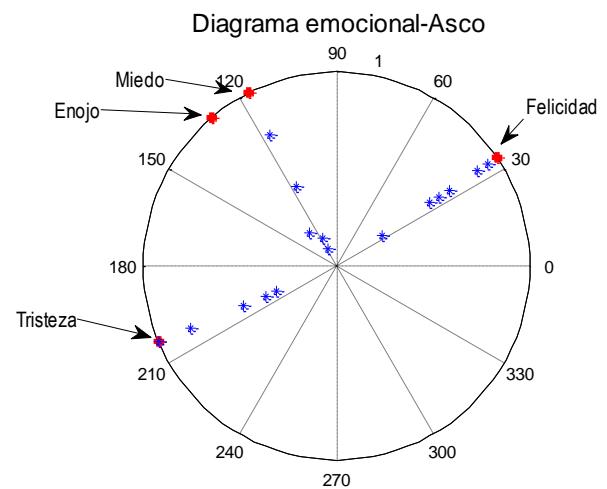


Figura 8.19 Plano emociones GVEESS.
Asco

En la tabla 8.20 se muestran los porcentajes de detección para ambos clasificadores. Se realizó la unificación de la detección de las emociones: enojo y miedo. Esta unificación es re-definida en la tabla como “emoción del segundo cuadrante” y fue elaborada para simplificar el análisis de la regionalización del estado emocional asco. En otras palabras, la detección de la emoción asco como enojo o miedo, representaría una buena regionalización del clasificador ya que ambas emociones se ubican en el mismo cuadrante.

Porcentaje de detección- Asco		
Emoción	B.D. Berlín	B.D GVEESS
Felicidad	35.48 %	37.5 %
“Emociones del segundo cuadrante”	35.48%	37.5 %
Tristeza	29.03 %	25 %

Tabla 8.20. Porcentaje de detección de emociones. Asco

El análisis de los resultados reflejados en la tabla 8.20 permite visualizar el conflicto en la regionalización de esta emoción y la confusión con la emoción felicidad del primer cuadrante. Características dependientes de la energía de la señal son determinantes en la diferenciación de los diferentes grados de activación. Emociones que presentan una alta activación (o en otras palabras una alta excitación) tales como la felicidad y el miedo, tienen características de energía elevadas respecto a las emociones con baja activación tales como la tristeza. En ello reside la dificultad en la regionalización.

En general, las emociones entrenadas, mejor detectadas, fueron: **tristeza y enojo**, con una tasa de detección del **96.67%** y **100%**, respectivamente. La emoción no entrenada que mejor fue regionalizada fue el **aburrimiento** con una tasa de acierto del **66.66%** para la Base de datos Berlín.

Las emociones de tristeza y aburrimiento resultan importantes si se pretende analizar patologías del trastorno del espectro emotivo. Son estados emocionales que siendo pasajeros no generan mayor alteración, sin embargo cuando se hacen permanente pueden ocasionar grandes alteraciones psicofísicas.

8.6 SECCIÓN 5

Clasificación de tristeza en dos bases de datos. Validación cruzada

La tristeza es uno de los estados emocionales que más se asocia a patologías del tipo depresivo. Cuando esta emoción surge en situaciones particulares y pasajeras, no es considerada patológica. Sin embargo el incremento en la ocurrencia y en la durabilidad de este tipo de emoción puede propiciar la persistencia de la misma, favoreciendo la aparición del estado depresivo y generando una alteración directa al sistema inmunológico (ver en capítulo 2).

En esta sección se utilizarán corpus en dos idiomas y se detectará la emoción tristeza.

Las bases de datos utilizadas se encuentran en idiomas: alemán (Base de datos Berlín, descripta en el capítulo 6) y castellano (Base de datos EAFIT).

La base de datos EAFIT es un corpus que fue desarrollado en la Universidad de EAFIT Medellín (Colombia) validada por profesionales del área de la psicología. Son 101 instancias de canal telefónico grabadas en castellano a 8kHz y re-muestreadas a 44.1kHz, incluyen las emociones:

- Felicidad
- Ira
- Tristeza
- Miedo
- Frustración
- Sorpresa
- Tranquilidad
- Neutro

El procesamiento de la señal se realiza según la descripción del capítulo 6. Se **utiliza una única característica descriptiva: los Coeficientes Cepstrales en frecuencia Mel (mfcc)**. El motivo de usar este parámetro es fundamentado en el capítulo 6 y en el capítulo 7.

Se usaron tres sistemas de clasificación. El primero es entrenado en alemán y aplicado en castellano. El segundo es entrenado en castellano y aplicado en alemán. Por último, el tercer sistema mixto fue entrenado con ambos idiomas y aplicado a los dos corpus.

Los sistemas de clasificación utilizados se desarrollaron mediante la técnica no lineal de RN en base al algoritmo de Levenberg-Marquardt con técnica backpropagation, usando 1 capa de entrada, 25 capas ocultas y 1 capa de salida. En todos los casos la muestra se dividió en tres grupos de forma aleatoria:

- Entrenamiento :70%
- Validación: 15%

- Testeo:15%

La evaluación del desempeño de los clasificadores se realizará mediante la determinación de la tasa de aciertos y el error falso positivo. La tasa de aciertos implica determinar la cantidad de muestras detectadas correctamente y el error falso positivo se refiere a la cantidad de muestras que han sido detectadas como verdaderas siendo que no han sucedido.

8.6.1 Entrenamiento en alemán

El entrenamiento de la RN utiliza 120 instancias de la Base de datos Berlín, que incluyen emociones: felicidad, miedo, enojo y tristeza. Como se pretende sólo entrenar tristeza, las emociones que no corresponden a este estado emocional se entranan bajo la denominación: "resto de emociones".

La salida del clasificador depende de las etiquetas seleccionadas. Estas etiquetas se designan de la siguiente manera (tabla 8.21):

Emoción	Etiqueta
Tristeza	1
Resto de emociones	0

Tabla 8.21. Etiquetas correspondientes a las emociones entrenadas.

Los resultados del entrenamiento son mostrados en la tabla 8.22:

	Entrenamiento	Validación	Prueba
Regresión	0.7917	0.7447	0.8371
Error	0.0761	0.0826	0.0527

Tabla 8.22. Resultados del entrenamiento para detección de Tristeza (entrenamiento en alemán).

Los errores de entrenamiento, validación y testeo se encuentran en el mismo orden, por debajo de los clasificadores antes analizados. El valor de R se sitúa por encima del 0.74 para entrenamiento y validación y de 0.83 para el test de prueba. Siendo el valor total de R de 0.78.

El sistema clasificador entrenado con sólo una característica fue implementado en 120 muestras de la Base de datos en alemán y en 101 muestras de la Base de datos en español. Estos resultados se exhiben en la tabla 8.23.

Detección: Tristeza			
Entrenamiento en	Aplicado en	Porcentaje de aciertos	Error falso positivo
Alemán	Alemán	83.33 %	7.77
Alemán	Castellano	40%	8.33%

Tabla 8.23 Detección de Tristeza en Alemán y Castellano con un clasificador Alemán

El análisis de los resultados permite observar que el porcentaje de detección de Tristeza utilizando un reconocedor entrenado con muestras en alemán con una sola característica descriptiva es óptimo (83,33%), cuando se aplica en instancias que presentan el mismo idioma con que fue realizado el

entrenamiento. Sin embargo, el desempeño del clasificador decae cuando es aplicado a muestras en otro idioma (en este caso, castellano).

8.6.2 Entrenamiento en castellano

Se utilizaron 101 instancias de la Base de datos EAFIT en castellano para realizar el entrenamiento del sistema de clasificación.

Los resultados del entrenamiento se exponen en la tabla 8.24.

	Entrenamiento	Validación	Prueba
Regresión	0.7077	0.6261	0.1547
Error	0.0203	0.0403	0.1579

Tabla 8.24. Resultados del entrenamiento para detección de Tristeza (entrenamiento en castellano).

El valor de R para el entrenamiento se encuentra en 0.70, siendo inferior para el caso de validación y testeo. Los errores de entrenamiento y validación se sitúan en valores óptimos, 0.02 y 0.04, respectivamente. El valor total de R es de 0.48. Este valor no satisfactorio, puede deberse a que esta base de datos contiene emociones, como la tranquilidad, con niveles de excitación, muy semejantes a la tristeza.

El sistema de clasificación entrenado con muestras en castellano fue implementado en 120 muestras de la Base de datos en alemán y en 101 muestras de la Base de datos en castellano. Estos resultados se muestran en la tabla 8.25.

Detección: Tristeza			
Entrenamiento en	Aplicado en	Porcentaje de aciertos	Error falso positivo
Castellano	Castellano	60%	8.33%
Castellano	Alemán	30%	10%

Tabla 8.25 Detección de Tristeza en Castellano y Alemán, con un clasificador español.

Los resultados exhibidos en la tabla 8.25 muestran una tasa de detección superior cuando se implementa este sistema de reconocimiento en muestras que presentan el mismo idioma con que fue realizado el entrenamiento. En este caso, la tasa de aciertos es del **60%** cuando el clasificador entrenado con muestras en español es implementado en muestras del mismo idioma, éste es un porcentaje de aciertos que no alcanza nuestras expectativas, sin embargo es superior al resultado que se obtiene de la implementación de la clasificación cruzada, en la que se puede observar que la tasa de detección cae notablemente.

El análisis de los resultados de las tablas 8.23 y 8.25 refiere que los sistemas de clasificación para la detección de la emoción tristeza tienen buen desempeño cuando se los aplica a muestras de señales de voz que pertenecen al mismo idioma con que fue entrenado el clasificador. Sin embargo cuando se realiza la detección cruzada, el desempeño del clasificador disminuye notablemente. A pesar de ello, es importante remarcar que los sistemas de clasificación implementados en esta sección, utilizan una sola característica (mfcc) para la detección de la emoción tristeza.

8.6.3 Entrenamiento mixto

Por último, se realizó el entrenamiento de la RN con una base de datos mixta de 221 muestras que contiene instancias en: alemán (B.D. Berlín) y castellano (B.D. EAFIT)

Los resultados del entrenamiento utilizando la base de datos mixta se presentan en la tabla 8.26.

	Entrenamiento	Validación	Prueba
Regresión	0.7179	0.8003	0.7022
Error	0.0756	0.0321	0.0374

Tabla 8.26. Resultados del entrenamiento para detección de Tristeza (entrenamiento mixto).

El valor de R se sitúa por encima del 0.7 para entrenamiento y prueba, siendo 0.80 el obtenido en la prueba de validación. El valor total de R resulta de 0.75. Los errores de entrenamiento, validación y testeo se encuentran en valores óptimos para el sistema de clasificación.

El clasificador bilingüe fue implementado en 120 muestras de la Base de datos Berlín y en 101 muestras de la base de datos EAFIT. Los resultados del porcentaje de aciertos y del porcentaje de error falso positivo se detallan en la tabla 8.27.

Detección: Tristeza			
Entrenamiento en	Aplicado en	Porcentaje de aciertos	Error falso positivo
Mixto	Alemán	83.33%	6.66%
Mixto	Castellano	60%	6.25%

Tabla 8.27. Detección de Tristeza en Alemán y Castellano con un clasificador bilingüe.

El análisis de los resultados muestra que utilizando un sistema de clasificación, de un único descriptor, entrenado en dos idiomas (alemán y castellano) para la detección de Tristeza, se consiguen mejores porcentajes de detección de la emoción en cuestión. Por lo tanto, la clasificación bilingüe de una emoción tal como la tristeza, descripta con una sola característica es factible con la premisa de que el sistema de clasificación tenga en cuenta el idioma donde va a ser aplicado el detector.

8.7 Conclusiones

En este capítulo se presentaron los resultados del desarrollo de la tesis.

En la primera sección, se procedió a la detección de **dos emociones** opuestas en el plano de valencia-activación (emociones de tristeza y felicidad) utilizando una sola característica de entrada (*mfcc*) e implementando técnicas de clasificación lineal y no lineal para el entrenamiento de señales de voz de la Base de datos Berlín, con resultados de detección superior al **96%**. Luego, se desarrollaron dos sistemas de clasificación, uno con 11 y otro con 12 características, utilizando para el entrenamiento la Base de datos Berlín, a fin de detectar **4 emociones** (felicidad, enojo, miedo y tristeza). El valor total de R obtenido para ambos casos superó el 0.9, lo cual indica el buen aprendizaje de la RN con las características propuestas. La tasa de detección de ambos sistemas de reconocimiento automático de emociones fue comparada con las obtenidas por sistemas de

reconocimiento No automáticos de emociones que utilizan como detector el desempeño humano. Los resultados mostraron que ambos sistemas automáticos (el de 11 y 12 características) presentaron una tasa de detección, para las 4 emociones, **15% superior** que la obtenida por el reconocimiento no automático de emociones del trabajo [2], siendo la detección automática de la tristeza **13 % superior** al obtenido por el desempeño humano del trabajo [1].

En la segunda sección, se diseñaron dos sistemas de clasificación no lineal implementando las Bases de datos Semaine (en idioma inglés) y GVEESS (en idioma alemán) con el mismo esquema de 12 características implementado en Berlín. Se realizó la detección de **4 emociones** (felicidad, enojo, miedo y tristeza). El valor total de R para ambos casos resultó de 0.88, inferior que el obtenido en la primera sección. El análisis comparativo de los dos sistemas automáticos, en relación con el desempeño humano, evidenció que la detección de las 4 emociones mediante los clasificadores entrenados en Semaine y en GVEESS, **superan en más del 10%** a las detectadas mediante el desempeño humano expuesto en el trabajo [2], mientras que la detección automática de tristeza es la única emoción cuya tasa de acierto **supera en un 20%(Semaine) y 6% (GVEESS)** al sistema no automático de detección.

En la tercera sección, se implementó el clasificador entrenado con la Base de datos Berlín (desarrollado en la primera sección) en las Bases de datos Semaine y GVEESS con resultados no satisfactorios debidos a la dependencia al idioma de las características empleadas y del esquema de clasificación.

En la cuarta sección, se utilizó el modelo circumplejo de emociones de Russell para regionalizar las **4 emociones** (felicidad, enojo, miedo y tristeza) entrenadas con las bases de datos Berlín y GVEESS y las **2 emociones (aburrimiento y asco) que no formaron parte del entrenamiento**. Ambos sistemas de clasificación fueron aplicados a las bases de datos completas (6 emociones de la base Berlín) y (14 emociones de la base de datos GVEESS) y posteriormente ubicadas en el diagrama valencia-activación, considerando las posiciones angulares de la experimentación de escalamiento unidimensional de Russell. Los resultados mostraron que las emociones entrenadas, mejor detectadas, fueron: tristeza y enojo, con una tasa de detección del **96.67% y 100%**, respectivamente y la emoción no entrenada que mejor fue regionalizada fue el aburrimiento con una tasa de acierto del **66.66%** para la Base de datos Berlín.

Por último, en la quinta sección, se diseñaron tres sistemas de clasificación utilizando una única característica de entrada (*mfcc*) para la detección de tristeza. El primer sistema fue entrenado con la base de datos Berlín (en idioma alemán) y aplicado a muestras en alemán (Berlín) y en castellano (EAFIT). El segundo sistema fue entrenado con la base de datos EAFIT (en idioma castellano) y aplicado a muestras en castellano (EAFIT) y en alemán (Berlín). El tercer sistema fue entrenado con ambas bases de datos y aplicado a ambos corpus por separado. El análisis de los resultados mostró que la tasa de detección de tristeza es superior cuando el sistema de clasificación usado se entrena con el mismo idioma donde va a ser implementado, decayendo el porcentaje de aciertos cuando se intenta la implementación cruzada. Utilizando una única característica como descriptora de la señal de voz es factible la detección de tristeza con la premisa que el sistema de clasificación tenga en cuenta el idioma donde va a ser aplicado el detector.

8.8 Referencias

1. Dellaert, F., Polzin, T., & Waibel, A. (1996, October). *Recognizing emotion in speech*. In *Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on* (Vol. 3, pp. 1970-1973). IEEE.
2. Petrushin, V. (1999, November). *Emotion in speech: Recognition and application to call centers*. In *Proceedings of Artificial Neural Networks in Engineering* (pp. 7-10).
3. Bänziger, T., & Scherer, K. R. (2005). *The role of intonation in emotional expressions*. *Speech communication*, 46(3), 252-267.
4. Ververidis, D., & Kotropoulos, C. (2006). *Emotional speech recognition: Resources, features, and methods*. *Speech communication*, 48(9), 1162-1181.
5. de Diego, I. M., Serrano, Á., Conde, C., & Cabello, E. (2006). *Técnicas de reconocimiento automático de emociones*. *Teoría de la Educación: Educación y Cultura en la Sociedad de la Información*, 7(2), 7.
6. Luengo, I., Navas, E., Hernández, I., & Sánchez, J. (2005). *Reconocimiento automático de emociones utilizando parámetros prosódicos*. *Procesamiento del lenguaje natural*, 35, 13-20.
7. Resa, C. O. (2009). *Detección de emociones en voz espontánea*.
8. El Ayadi, M., Kamel, M. S., & Karray, F. (2011). *Survey on speech emotion recognition: Features, classification schemes, and databases*. *Pattern Recognition*, 44(3), 572-587.
9. Russell, J. A. (1980). *A circumplex model of affect*. *Journal of personality and social psychology*, 39(6), 1161.

CAPITULO 9

Conclusiones

9.1 Introducción

La nueva perspectiva de las emociones, desde el área de las neurociencias, la definen como procesos influenciados por el pasado evolutivo y personal que se ejecutan a partir de la interconexión de los elementos del sistema nervioso y que desencadenan un conjunto de cambios fisiológicos (viscerales, del sistema autónomo, de los gestos, del tono de voz, etc.) y del comportamiento (memoria, toma de decisiones). A partir de diversos estudios se ha propuesto que existen diferentes tipos de emociones entre las cuales se destacan las emociones básicas, que son consideradas innatas y están presentes en todas las culturas (ira, miedo, alegría, tristeza, sorpresa y asco).

Emociones perturbadoras pueden resultar nocivas para la salud. Investigaciones llevadas a cabo en miles de personas confirman que individuos que sufren ansiedad crónica, largos episodios de tristeza y pesimismo o irritación constante son propensos a contraer enfermedades. Las emociones negativas como el enojo, la tristeza, el miedo, la frustración, la ansiedad, la depresión, etc. son un factor de riesgo para el desarrollo de cualquier enfermedad y constituyen una amenaza para la salud.

Los trastornos emotivos que incluyen a la depresión, el trastorno bipolar, el aspecto emotivo del trastorno del espectro autista, son manifestaciones clínicas de alteraciones en el neurodesarrollo, esto indica que el proceso que las determina ha ocurrido mucho antes de que se manifiesten los primeros síntomas. Estos trastornos comienzan en una etapa muy temprana de la vida. Numerosos estudios han detectado que el 13% de los chicos entre 8 y 15 años tiene alguna forma de trastorno mental y menos de la mitad recibe tratamiento.

El trastorno depresivo es un desorden emotivo que genera una serie de síntomas tales como irritabilidad (ira) o ánimo depresivo (tristeza), disminución del interés por actividades diarias, variaciones significativa en el peso, insomnio o hipersomnia, agitación o retardo psicomotor. Actualmente una de cada seis personas padece depresión clínica al menos una vez en su vida y un 7% de la población sufre esta enfermedad al cabo del año. Se calcula que afecta a unos 350 millones de personas en todo el mundo, constituyéndose en una de las principales causas de discapacidad.

La detección temprana se ha convertido en el objetivo primario del trabajo en salud mental. A pesar de los avances en la neurociencia, los diagnósticos en psiquiatría siguen siendo subjetivos, llevados a cabo a partir de conversaciones con el paciente y con su familia sobre sus síntomas y llenando formularios de actividades rutinarias.

Dado que los trastornos emotivos son alteraciones cerebrales que desencadenan estados fisiológicos particulares al tipo de emoción que está involucrada en el trastorno, se puede esperar que indicadores biológicos no invasivos tales como la prosodia de la voz podrían ser detectados antes de la aparición de los primeros síntomas de la enfermedad, lo cual mejoraría el pronóstico.

El análisis de características que permiten describir una señal de voz proporciona una evaluación del contenido frecuencial, temporal y prosódico de la señal dejando de lado el contenido lingüístico. O sea, se valora “cómo se dice” y no “qué es lo que se dice”. A partir de estos parámetros descriptivos de la señal de voz se puede establecer una correspondencia con un estado emotivo particular.

En este trabajo de tesis se desarrollaron algoritmos que permitieron parametrizar la señal de voz y estimar un estado emocional mediante un sistema de reconocimiento de patrones basado en aprendizaje supervisado.

La detección automática de emociones puede integrarse a los sistemas de diagnóstico de patologías del espectro emotivo como una herramienta objetiva destinada a mejorar las técnicas de detección de desórdenes emotivos.

9.2 Conclusiones

- **El sistema de reconocimiento de emociones desarrollado en el presente trabajo se abordó desde un enfoque multidisciplinario.** El análisis de la señal de voz fue planteado desde la generación fisiológica de la misma, y luego desarrollado desde un punto de vista físico-matemático, a fin de determinar variables descriptivas relevantes. Por otro lado, desde un enfoque neurocientífico se realizó una investigación bibliográfica exhaustiva de la teoría de las emociones, de los procesos anatómicos donde se conciben y de los estados fisiológicos desencadenados por las mismas. Se priorizó la búsqueda bibliográfica de indicadores fisiológicos en la voz, representativos de un estado emocional en particular.
Como herramienta de desarrollo para la ejecución de este trabajo se utilizó el programa Matlab® y las bases de datos: Berlín (de dominio público), Semaine, GVEESS y EAFIT.
- **Se desarrollaron algoritmos computacionales a fin de extraer 23 características descriptoras de la señal de voz** desde enfoques: temporal, prosódico y frecuencial. Posteriormente se evaluaron estas características mediante la técnica de análisis discriminante a fin de diferenciar los parámetros que mejor representaban los estados emotivos propuestos. En primer lugar se propuso el reconocimiento de dos emociones (Felicidad y Tristeza), en una etapa posterior se realizó la clasificación de 4 estados emocionales (Felicidad, Enojo, Miedo y Tristeza) y en una tercera etapa se evaluó sólo el estado emocional Tristeza.
- **Con el propósito de mejorar la parametrización de la señal se realizó un pre-procesado de la misma teniendo relevancia en esta etapa el Sistema Segmentador de la Señal de Voz que fue un aporte importante del presente desarrollo.** El segmentador de señales de voz permite diferenciar segmentos de voz y de no voz, lo cual mejora el tiempo de cómputo, permite eliminar el ruido propio de los segmentos de silencio, disminuir la tasa de error en características tales como cruces por cero, y por otro lado permitió extraer nuevas características, por ejemplo la duración del silencio. Este sistema segmentador de voz fue desarrollado desde la perspectiva de la estadística utilizando el lema de Neyman-Pearson y luego fue comparado con un detector comercial. Los resultados de la comparación mostraron que el error absoluto en la detección de segmentos de voz mediante la aplicación del segmentador desarrollado en el presente trabajo era varios órdenes inferiores al obtenido utilizando el detector comercial.
- **Se utilizaron tres esquemas de características descriptivas de 1, de 11 y de 12 características.** Se eligieron dos sistemas de clasificación (con técnicas lineal y no lineal) para el primer esquema, el cual fue implementado en la Base de datos Berlín para la detección de dos emociones opuestas en el plan valencia-activación (Felicidad-Tristeza). La comparación entre ambas técnicas de clasificación mostró la semejanza en las tasas de detección con resultados superior al 96%.

- Luego se utilizaron los esquemas de extracción de 11 y 12 características en la Base de datos Berlín con un sistema de clasificación no lineal de redes neuronales para la detección de 4 emociones (Felicidad, Miedo, Enojo y Tristeza). El desempeño de los clasificadores fue comparado con sistemas de reconocimiento no automático de emociones que utiliza el desempeño humano como detector. Los resultados fueron satisfactorios ya que superaron en 15% los porcentajes de detección tomados como referencia en la comparación.
- Se implementó un sistema de clasificación no lineal, utilizando el esquema de 12 características, entrenado con las Bases de datos Semaine y GVEESS a fin de evaluar el comportamiento de las características elegidas bajo señales de voz en otros idiomas. Los resultados mostraron el buen desempeño del esquema de características empleado cuando se entrena el sistema de clasificación con muestras en el mismo idioma donde va a ser implementado.
- La teoría de regionalización de emociones de Russell fue utilizada para presentar los resultados del sistema de clasificación desarrollado con los esquemas de 11 y 12 características, aplicado a las bases de datos Berlín y GVEESS completas con el objetivo de determinar los porcentajes de detección de emociones cuando las muestras incluidas en la aplicación no correspondían a emociones antes entrenadas por el sistema de clasificación. La ubicación de las emociones en el plano valencia-activación proporcionó información acerca de la similitud de algunos estados emocionales y permitió comprender por qué ciertas emociones son difíciles de detectar. El sistema de clasificación implementado mostró un buen desempeño en la regionalización de las emociones del tercer cuadrante que corresponden al grupo de las emociones negativas con bajo nivel de excitación (tristeza, aburrimiento).
- Por último, usando una única característica, se diseñó un sistema de clasificación no lineal a fin de detectar la emoción tristeza en muestras de idiomas alemán y español. El análisis de los resultados mostró que la tasa de detección de tristeza es superior cuando el sistema de clasificación se entrena con el mismo idioma donde va a ser implementado, decayendo el porcentaje de aciertos cuando se intenta la implementación cruzada.

Es importante destacar que los aportes relevantes del presente trabajo fueron el **desarrollo del sistema segmentador de señales de voz** que mostró eficacia en su función y permitió determinar nuevas características descriptivas de la señal de voz y por otro lado, la **determinación y desarrollo de parámetros que mejor representarían las señales de voz** ante diferentes estados emotivos, priorizando la emoción tristeza. Esta emoción forma parte fundamental de patologías del espectro emotivo, por ejemplo de la depresión, por ello es clave la correcta detección de la misma. El uso de características que resalten las diferencias en los niveles de activación define los estados emotivos de alta excitación tales como enojo o felicidad y estados de relajación como la tristeza o aburrimiento.

Las **características propuestas resultaron eficientes para la detección de tristeza**, esto es concluído dado el buen desempeño de los clasificadores implementados para su detección. Sin embargo se debe tener en cuenta que los sistemas de clasificación deben entrenarse con muestras que presenten el mismo idioma donde va a ser aplicado el detector, debido a que estos parámetros no tienen la capacidad de ser invariantes a los idiomas.

9.3 Líneas futuras

El área de procesamiento de señales de voz para la detección de estados emotivos implicados en trastornos emotivos es muy incipiente. Los trabajos publicados donde se intenta establecer una correlación entre estados emocionales patológicos y prosodia de la voz son muy pocos. Este proyecto fue desarrollado con el fin de enfocar estos conocimientos y desarrollos en los sistemas de reconocimiento de emociones en el área de la biomedicina, a fin de permitir la colaboración de la ingeniería en el área de salud. Para ello, resulta fundamental la experimentación en casos patológicos, sería apropiado utilizar bases de datos con especificidad en alguna patología del espectro emotivo, por ejemplo en la depresión. Con esto se ofrecería una mayor caracterización del estado emotivo y permitiría una mejor evaluación del desempeño del sistema de clasificación.

Es importante considerar que la aplicación de este sistema de reconocimiento de emociones se debe realizar en conjunto con personal médico que permita corroborar el correcto funcionamiento del sistema de clasificación a fin de ser una herramienta confiable para su uso. Por ello, sería interesante implementar el sistema de reconocimiento de emociones en pacientes que estén diagnosticados con patologías del espectro emotivo y generar una base de datos a fin de poder hacer las pruebas necesarias para mejorar la detección.