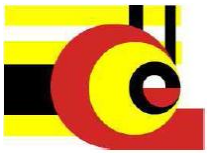


INSTITUTO POLITÉCNICO NACIONAL

CENTRO DE INVESTIGACIÓN EN COMPUTACIÓN

MODELO DE PROCESAMIENTO DE VOZ PARA LA CLASIFICACIÓN DE ESTADOS

T E S I S

QUE PARA OBTENER EL GRADO DE
DOCTOR EN CIENCIAS DE LA COMPUTACIÓN

PRESENTA:

JOSÉ FRANCISCO SOLÍS VILLARREAL

ASESORES DE TESIS:

DR. SERGIO SUÁREZ GUERRA
DR. CORNELIO YÁÑEZ MÁRQUEZ

México D.F. Junio del 2011

SIP-14 BIS

INSTITUTO POLITÉCNICO NACIONAL

SECRETARÍA DE INVESTIGACIÓN Y POSGRADO

ACTA DE REVISIÓN DE TESIS

En la Ciudad de México, D.F. siendo las 10:00 horas del día 14 del mes de Junio de 2011 se reunieron los miembros de la Comisión Revisora de la Tesis, designada por el Colegio de Profesores de Estudios de Posgrado e Investigación del:

Centro de Investigación en Computación

para examinar la tesis titulada:

"UN MODELO DE PROCESAMIENTO DE VOZ PARA CLASIFICACIÓN DE ESTADOS"

Presentada por el alumno:

SOLÍS	**VILLARREAL**	**JOSÉ FRANCISCO**
Apellido paterno	Apellido materno	Nombre(s)

Con registro:

B	0	7	1	2	4	0

aspirante de: **DOCTORADO EN CIENCIAS DE LA COMPUTACIÓN**

Después de intercambiar opiniones los miembros de la Comisión manifestaron ***APROBAR LA TESIS***, en virtud de que satisface los requisitos señalados por las disposiciones reglamentarias vigentes.

LA COMISIÓN REVISORA

Directores de tesis

Dr. Sergio Suárez Guerra

Dr. Cornelio Yáñez Márquez

Dr. Jesús Guillermo Figueroa Nazuno

Dr. Oleksiy Pogrebnyak

Dr. Héctor Manuel Pérez Meana

PRESIDENTE DEL COLEGIO DE PROFESORES

Dr. Luis Alfonso Villa Vargas

INSTITUTO POLITÉCNICO NACIONAL
CENTRO DE INVESTIGACIÓN EN COMPUTACIÓN
DIRECCIÓN

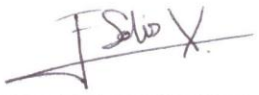
INSTITUTO POLITÉCNICO NACIONAL

SECRETARÍA DE INVESTIGACIÓN Y POSGRADO

CARTA DE CESIÓN DE DERECHOS

En la Ciudad de **México, D. F.** el día **23** del mes de **Junio** del año **2001**, el que suscribe **José Francisco Solís Villarreal** alumno del Programa de **Doctorado en Ciencias de la Computación** con número de registro **B071240**, adscrito al **Centro de Investigación en Computación**, manifiesta que es autor intelectual del presente trabajo de Tesis bajo la dirección del **Dr. Sergio Suárez Guerra** y **Dr. Cornelio Yáñez Márquez** y cede los derechos del trabajo intitulado **Modelo de Procesamiento de Voz para el Procesamiento de Estados**, al Instituto Politécnico Nacional para su difusión, con fines académicos y de investigación.

Los usuarios de la información no deben reproducir el contenido textual, gráficas o datos del trabajo sin el permiso expreso del autor y/o directores del trabajo. Este puede ser obtenido escribiendo a la siguiente dirección tlilectic.mixtzin@gmail.com. Si el permiso se otorga, el usuario deberá dar el agradecimiento correspondiente y citar la fuente del mismo.



José Francisco Solís Villarreal.

Resumen

En esta tesis se reportan los avances obtenidos en la clasificación de emociones a partir de señales de voz, los objetivos que se plantean para la conclusión del mismo y una descripción de los trabajos relacionados encontrados hasta el momento.

El reconocimiento de emociones es un problema que ha sido abordado de diferentes maneras, teniendo en cuenta diversas formas de obtener los datos, como son la lectura de señales biométricas (presión arterial, pulso, entre las más importantes), detección de secuencias de movimientos del cuerpo humano al andar, el seguimiento de la expresión facial, la señal de voz, entre otras.

Hay trabajos que abordan el tema analizando diferentes tipos de señales y otros que tratan con un solo tipo de señal; para los fines del presente proyecto, se trabaja con la señal de voz.

Las técnicas utilizadas hasta el presente para el reconocimiento y clasificación de emociones son redes neuronales artificiales, mixturas Gaussianas, Modelos Ocultos de Markov, principalmente.

En este trabajo de tesis se hace uso de técnicas de soporte vectorial y memorias asociativas, la combinación de ambas dan como resultados una efectividad cercana al 99 % de acierto en la clasificación de emociones.

Se utiliza la base de datos de Berlín [2], la cual es gratuita y con la cual muchos investigadores han llevado a cabo sus trabajos y donde los resultados reportados no han sobrepasado al 82 %.

Abstract

In this work, in general terms, we report the progress made in the recognition and classification of emotions from voice signals, the objectives proposed for the conclusion and a description of relevant work found so far.

Emotion recognition is a problem that has been treated in different manners, taking into account various ways of obtaining data, such as reading biometric signals (blood pressure, pulse, among the most important), detection of sequences of movements of the human body walking, monitoring of facial expression, voice signal, among others.

There are works dealing with the issue by analyzing different types of signals and others who deal with one type of signal, for the purposes of this project, working with the voice signal.

Actually, the techniques used for the recognition and classification of emotions are neural networks, Gaussian mixtures, Hidden Markov Models, mainly.

In this thesis makes use of support vector techniques and associative memories, the combination of both result in an overall benefit of 99 % accuracy in the classification of emotions.

It uses the database Berlin [2], which is free and with which many researchers have carried out their work and where the reported results have not exceeded 82 %.

Agradecimientos

Este trabajo no habría sido posible sin el apoyo y estímulo de mis asesores Dr. Segio Suárez Guerra y Dr. Cornelio Yáñez Márquez. Les estaré siempre agradecido por su calidad como científicos, su devoción y nobleza extraordinarias para apoyarme de forma incondicional.

A mi sinodales Dr. Jesús Guillermo Figueroa Nazuno, Dr. Oleksiy Pogrebnyak, Dr Héctor Manuel Pérez Meana y Dr. José Luis Oropeza Rodríguez por sus diversas e invaluables contribuciones vertidas en este trabajo, sus oportunos comentarios ayudaron a mejorar esta tesis.

Al Dr. Mario Aldape Pérez, gracias por el apoyo recibido para la elaboración del trabajo.

Al Dr. Itzamá López Yáñez, gracias por el apoyo brindado en esta investigación.

A todos los miembros del Grupo Alfa-Beta, por todos sus aportes e ideas que fueron surgiendo en los seminarios de investigación.

Al Centro de Investigación en Computación (CIC) y al Instituto Politécnico Nacional (IPN), que les debo la oportunidad de poder acceder a una formación de esta máxima case de estudios, estaré siempre pendiente de poner en alto su nombre.

Al CONACyT, gracias por todo el soporte económico durante el desarrollo del presente trabajo.

A mi familia y allegados por su incalculable apoyo.

Índice general

Índice de figuras

Índice de tablas

Glosario

1. SVM - Máquinas de Soporte Vectorial.
2. GMM - Mixturas Gaussianas.
3. F0 - Frecuencia Fundamental.
4. MLB - Verosimilitud Máxima de Bayes.
5. KR - Kernel de Regresión.
6. KNN - K - Vecinos Próximos.
7. FS - Selección de Parámetros.
8. PFS - Selección de Rasgos por el Método de los Primeros más Significativos.
9. FS - Selección de Rasgos Hacia Adelante.
10. HMM - Modelos Ocultos de Markov.
11. GMD - Densidad de Mixturas Gaussianas
12. DSE - Parámetros extraídos Específicamente de la Base de Datos Emocional.
13. AHL - Parámetros de Todo Alto Nivel.
14. MLP - Perceptrón Multicapa.
15. SFS - Selección Secuencial hacia Adelante.
16. LDA - Análisis Discriminante Lineal.

17. LBG - Linde-Buzo-Gray.
18. LOO - Leave One Out.
19. FSS - Feature Subset Selection.
20. EDA - Algoritmo de Estimación de la Distribución.
21. ARFF - Atribute-Relation File Format.
22. MFCC - Mel Frequency Cepstral Coefficients.

Capítulo 1

Introducción

1.1. Antecedentes

En el presente trabajo se parte del uso de las computadoras como medio de comunicación entre los humanos, para lo cual se ha trabajado durante mucho tiempo en reconocimiento, síntesis y traducción por mensajes hablados. Pero eso no es lo único que se puede extraer de la voz, la cual es una señal con información inteligente; también la voz refleja el estado de ánimo del que habla, o si está diciendo una verdad o una mentira. Esta otra parte de la información está oculta; es decir, es muy diferente del mensaje inteligente que la comunicación oral quiere expresar.

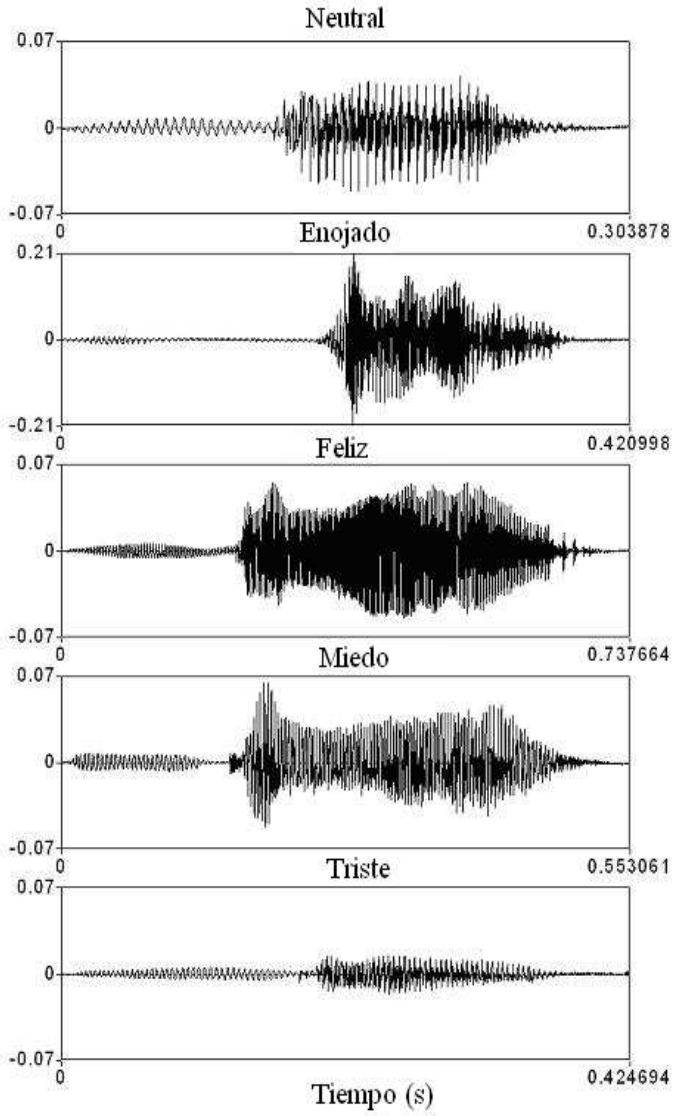


Figura 1.1: Palabra "da" en serbio, se traduce como "si" en castellano.

Según Paul Ekman [15] las emociones pueden ser vistas por su valor adaptativo con las tareas fundamentales de la vida. Cada emoción tine características únicas y otras que son comunes que son producto de nuestra evolución y se distinguen así de otros fenómenos emotivos. Para este autor, basa la existencia de las emociones básicas: enojado, miedo, tristeza, alegría, disgustado y sorpresa; en su origen evolutivo. Por lo que en la mayor parte de bases de datos para el reconocimiento de voz emotiva

consideran dichas emociones, es decir, son las más comunes usadas por el hombre.

En [23] se reportan impresiones de apreciación de las 5 emociones más comunes usadas en las bases de datos orientadas a la clasificación de voz emotiva. Primeramente tenemos a la voz neutral, que se puede percibir de una forma uniforme, calmada, con un tono más o menos idéntico, sin alteraciones o interrupciones; posteriormente, la emoción de enojado se puede apreciar una voz determinante, fuerte, irritable, agresiva, severa.

Para el estado de felicidad, se le puede considerar como una voz cantada, llena de alegría, de alguna forma como si el locutor tuviera una sonrisa en la cara; la forma de expresarse con la emoción del miedo denota una voz cambiante, interrumpida, un tono casi chillón, voz ansiosa, con susurros. Por último el estado emocional de tristeza puede ser percibido como monótono, depresivo, lento, melancólico, lento.

En la Figura 1.1 se puede percibir, en las gráficas, las señales de voz que contienen o que se expresan en la palabra en serbio "da", que en castellano se puede traducir como "si"; dichas señales fueron expresadas con 5 diferentes emociones y cabe hacer notar las diferencias en duraciones de tiempo así como las diferencias en amplitud. [23]

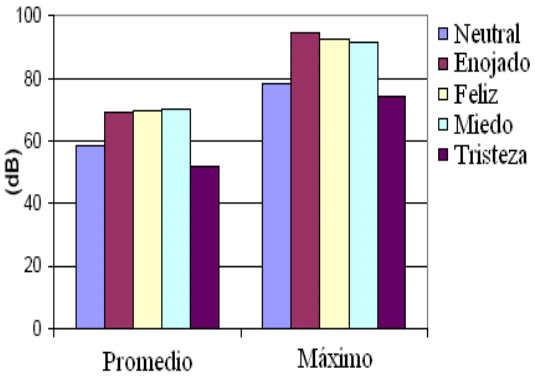


Figura 1.2: Valor promedio y máximo de la energía para cada emoción.

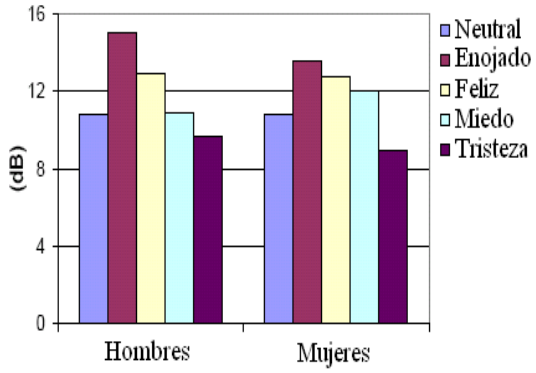


Figura 1.3: Desviación estándar de la energía por emoción y por género.

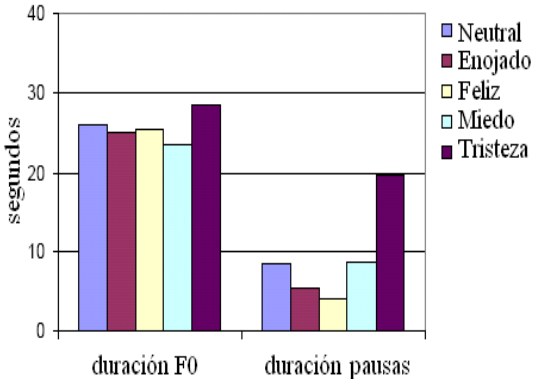


Figura 1.4: Promedio de las duraciones de señal de voz hablada y pausas por emoción.

A continuación se muestran algunas gráficas [23] que muestran diferencias entre 5 emociones básicas (neutral, enojado, felicidad, miedo y tristeza). En la Figura 1.2, podemos apreciar el promedio y el máximo valor de energía para las 5 emociones en una escala de 0 a 100 decibeles; también se muestran los diferentes valores de la desviación estándar de la energía para cada emoción y para cada género (ver Figura 1.3). Por último se presenta la Figura 1.4, donde se puede apreciar que para la emoción de la tristeza, se tiene una diferencia mayor con respecto a las otras 4 emociones, al menos para la duración de las pausas.

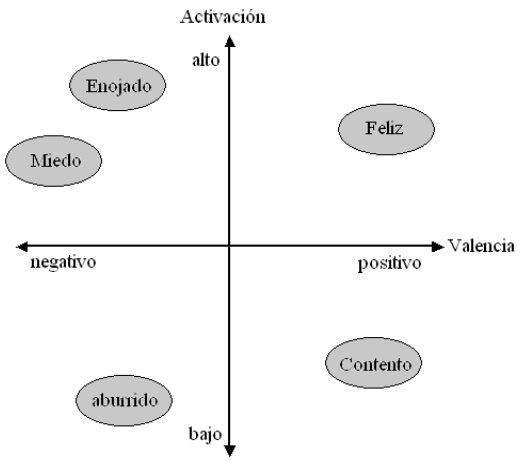


Figura 1.5: Emociones básicas espaciadas en 2 dimensiones por los ejes de valencia y actividad.

Para tener una idea un poco más ilustrativa de cómo se podrían clasificar las emociones, puede verse en [47] una distribución de emociones básicas clasificadas por medio del eje de actividad y el eje de la valencia; por ejemplo, puede verse en la Figura 1.5 que se puede distinguir la emoción de estar feliz con respecto a estar enojado por medio de la valencia (emociones positivas y negativas) y para diferenciar al estado enojado del aburrido, se puede llevar a cabo por el eje de la actividad (enojado es un estado más activo que aburrido).

El reconocimiento automático de emociones a partir de la voz es un área de investigación relativamente nueva [42]; sin embargo, se tienen trabajos desde el año 1996 [14, 28, 35] y hay otros trabajos como el [50], donde se reporta un estudio experimental en el que trabaja con 4 emociones: enojado, feliz, triste y neutral, utilizando un corpus de 721 instancias cortas.

Posteriormente, en [41] 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 del placer, con el fin de mejorar la tasa de reconocimiento. Para [19], el problema de reconocimiento de emociones se lleva a una frontera más lejos al realizar un experimento donde se intenta realizar clasificación de estados emotivos dentro de un contexto multi-lenguaje. El experimento se llevó a cabo usando bases de datos en inglés, eslovenio, castellano y francés.

En 2004, por parte de [44], se reporta una clasificación de emociones usando la base de datos de voz emocional danesa, donde se extrajeron 87 rasgos y se usó un criterio de Selección Secuencial hacia Adelante. En el siguiente año, [45] presenta un artículo donde se realiza una minería de datos sobre 1000 rasgos extraídos del pitch, energía y MFCC's, usando las bases de datos de Berlín [2] y la del mago de Oz.

En el mismo año, [34] se hace un reporte de detección de emociones usando una base de datos en lenguaje Mandarín, alcanzando una precisión de 88.7 %, usando Análisis Discriminante Lineal, K-vecinos y Modelos Ocultos de Markov. En el trabajo reportado por [31], se hace uso de una base de datos en euskara, la cual contiene 6 emociones (alegría, asco, ira, miedo, sorpresa y tristeza), dicho corpus contiene 582 instancias; para llevar a cabo la clasificación fueron utilizadas Máquinas de Soporte Vectorial (SVM) y Mixturas Gaussianas (GMM); se reportan resultados de 98.4 y 92.3 %.

Para [46], el problema de reconocimiento de emociones se mejora con una pre-clasificación del género, con un 2 al 4 %; para este caso se trabaja con 2 bases de datos. En el 2007, en [42] se presenta un trabajo donde se pretende detectar nuevas emociones, además de las emociones de entrenamiento, propias de la base de datos.

Basándose en una selección de rasgos a partir de algoritmos evolutivos, [7] usa una base de datos bilingüe, con los lenguajes castellano y vasco. Usando técnicas basadas en computación evolutiva, se seleccionan grupos de rasgos para optimizar el reconocimiento automático de emociones. Por otro lado, en [13] se presenta una aproximación usando un clasificador basado en árboles de decisión binarios, en donde se usan 2 bases de datos en alemán y en polaco, alcanzando 72 % de reconocimiento.

Por medio de una clasificación jerárquica [48], mediante 68 parámetros extraídos a la base de datos de Berlín, se alcanzó un reconocimiento del 79.47 % donde se realizó una pre-clasificación del género.

En el 2010, [51] usa un modelo basado en el algoritmo k-vecinos próximos que toma en cuenta la estimación del costo del error, ese trabajo reporta un desempeño aproximado del 82 % de clasificación de emociones usando las 7 emociones de la base de datos de Berlín [2]. La experimentación se llevó a cabo por medio de una validación cruzada de 2 pruebas en la que se entrena con el 50 % de la base de datos y se prueba con el otro 50 %, se reordena de forma aleatoria en cada prueba, estimando

el resultado final como el promedio del desempeño de las 2 pruebas.

En marzo del presente año (2011), [16] se reporta una extendida revisión de las bases de datos orientadas al reconocimiento automático de emociones; los resultados más altos alcanzados en la clasificación de emociones gira alrededor del 80 %. Una gran limitante que hay para trabajar en este campo es la poca disponibilidad de los corpus que se utiliza para trabajar en esta área.

1.2. Hipótesis

La voz es una señal que lleva información dinámica; es decir, la secuencia en el tiempo representa qué se dice y cómo se dice (emoción). Si se hace una representación de determinados parámetros de la voz de manera bidimensional y se aplican las técnicas que se utilizan en el procesamiento de imágenes de clasificación, es posible obtener un clasificador de emociones a partir de este tipo de representación.

1.3. Objetivo

Objetivo principal. Obtener el conjunto de parámetros de la señal de voz que permitan caracterizar eficientemente la información, para poder hacer la clasificación de los estados de ánimo y su uso mediante un modelo asociativo, para incrementar los resultados que se han obtenido a la fecha.

1.3.1. Objetivos específicos

-Disponer de un corpus de voces para la clasificación de estados.

-Parametrizar la base de datos.

-Realizar selección de rasgos más representativos.

-Diseño de experimentos para la clasificación de estados de ánimo mediante el uso de parámetros más representativos y técnicas utilizadas en procesamiento de voz.

-Diseño de experimentos para la clasificación de estados de ánimo a partir de representaciones bidimensionales de los parámetros utilizados en procesamiento de voz.

-Proponer un nuevo modelo que permita mejorar la clasificación de la emoción hablada, de la presentada hasta la actualidad mediante un modelo asociativo.

1.4. Contribuciones

Extracción y selección de parámetros (rasgos), de la señal de voz para la clasificación de estados emocionales.

A partir de la señal de voz, hacer representaciones de sus parámetros en forma de representaciones bidimensionales (no imágenes) y utilizar estas representaciones como elementos para clasificar estados emocionales.

Aplicación de los modelos asociativos basados en el uso de técnicas de soporte vectorial con los operadores alfa-beta para la clasificación de estados emocionales.

1.5. Justificación

Hay múltiples razones por las que el hacer reconocimiento de emociones es un problema difícil [9]. En la última década, por ejemplo, no se ha tenido un gran progreso, como lo han tenido otros campos en el área de procesamiento de voz; de hecho, se ha alcanzado solo un 50 o 60 % de precisión en la clasificación. Esto es porque la mayor parte de la investigación en este campo se ha enfocado más a la síntesis de voz emocional, que al reconocimiento automático de emociones [36]. Con los enfoques que se han usado hasta ahora, la cota máxima de precisión en la clasificación de algunas emociones gira alrededor del 80 % dentro de un área particular, tomando en cuenta muchas consideraciones, como la creación de la base de datos [40] y el género [47]; es decir, cuestiones como el número de emociones a clasificar, dependiente o independiente del idioma, género o locutor.

La principal motivación para elaborar el presente trabajo es el desarrollo de un nuevo enfoque dentro del campo del reconocimiento de emociones a partir de una señal de voz con la finalidad de lograr una mejora en la clasificación, esto implica una selección rigurosa de parámetros acorde a su aportación en la clasificación y el uso de modelos asociativos.

Por otro lado, en esta tesis se experimentó con otros enfoques que aún no han

sido explorados y/o reportados en la literatura relacionada con el reconocimiento de emociones, como lo son el uso de las representaciones de los parámetros de procesamiento de voz como representaciones bidimensionales para la clasificación de las emociones.

1.6. Organización del documento

En este Capítulo se han presentado: los antecedentes, la hipótesis, el objetivo, los objetivos específicos, las contribuciones de este trabajo de tesis y su justificación. El resto del documento está organizado de la siguiente manera:

En el Capítulo 2 se presenta el estado del arte en el campo del reconocimiento automático de emociones a partir de señales de voz. A su vez, el Capítulo 3 se describen los materiales y métodos usados para el desarrollo de este trabajo, como son las máquinas asociativas Alfa Beta con Soporte Vectorial, la base de datos que usamos para hacer el reconocimiento de emociones así como el software utilizado durante el desarrollo. En el Capítulo 4 se presenta el Modelo propuesto, dentro de este capítulo tenemos el aporte más importante de la tesis, donde se aborda el problema de reconocimiento de emociones con un enfoque nuevo que no ha sido reportado antes en la literatura. El Capítulo 5 presenta los resultados experimentales del nuevo modelo con la base de datos y en el Capítulo 6 se comparan los resultados obtenidos con los reportados en la literatura, a su vez, se presentan las presentaciones y publicaciones derivadas de este trabajo de tesis. Finalmente, se incluyen las referencias bibliográficas y los apéndices.

Capítulo 2

Estado del Arte

Cronología del reconocimiento de emociones

En 1996 [14] se realiza un trabajo de reconocimiento de emociones utilizando un corpus de 1000 instancias, con 5 locutores, 50 sentencias cortas grabadas con las emociones de enojado, felicidad, triste, miedo y normal; se tomaron 250 instancias como entrenamiento y la frecuencia de muestreo fue de 16kHz.

Tabla 2.1: Matriz de confusión del desempeño humano. [14]

Categoría	Feliz	Triste	Enojado	Miedo	Error
Feliz	44	2	2	2	3 %
Triste	1	40	3	6	5 %
Enojado	2	0	48	0	1 %
Miedo	8	7	3	32	9 %
					18 %

En la Tabla 2.1 se puede observar la matriz de confusión generada por una persona al etiquetar algunas grabaciones de la base de datos antes mencionada; dicha matriz será usada para poder evaluar los resultados del reconocimiento automático.

Toda la extracción de parámetros se hizo a partir del pitch (F0), extrayendo así un total de 70 rasgos agrupados de la siguiente forma:

-Mediciones estadísticas relacionadas con el ritmo: razón del habla, promedio de longitud entre regiones habladas, número de las curvas positivas entre las negativas, curva máxima, entre otras.

-Relación estadística del pitch suavizado: mínimo, máximo, promedio y desviación estándar.

-Relación estadística de la gradiente del pitch suavizado: mínimo, máximo, mediana y desviación estándar.

-Aproximación estadística de las partes individuales habladas: promedio mínimo, promedio máximo.

-Valores estadísticos de las curvas individuales: promedio positivo, promedio negativo.

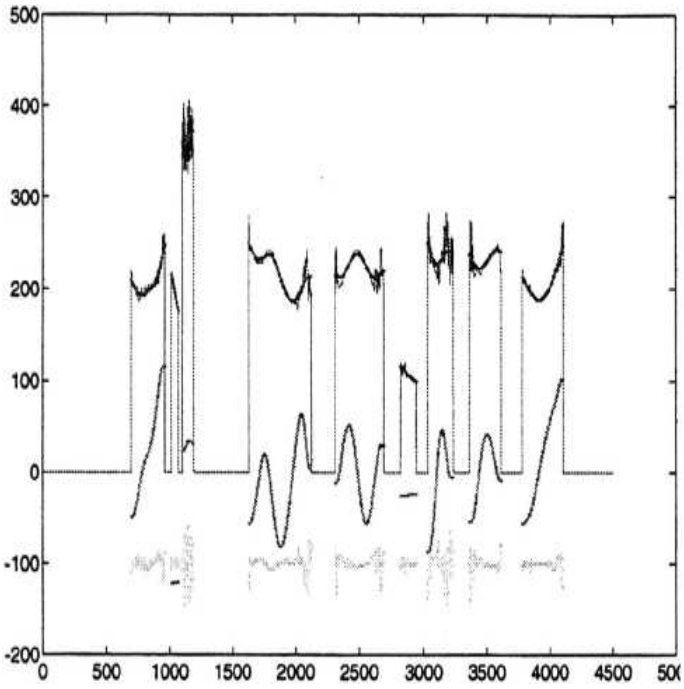


Figura 2.1: Aproximación suavizada del contorno del pitch. [14]

Los rasgos son agrupados en 2 grupos, los parámetros "A", que consisten en 7 estadísticas globales de la señal de pitch, que son el promedio, desviación estándar, mínimo, máximo y rango del pitch, y medidas de las curvas y la razón del habla. Los parámetros "B" es un conjunto más grande de rasgos, que fueron extraídos a partir

del suavizado de la señal del pitch Figura 2.1

Tabla 2.2: Modelos de clasificación clásicos con los 2 grupos de rasgos. [14]

Método	p	Error(A)	p	Error(B)
MlB	-	41.5 %		44 %
KR	kw=1.2	37 %	kw=1.1	35 %
KNN	k=19	36 %	k=11	32 %

Se usaron 3 métodos de clasificación, el de Verosimilitud Máxima de Bayes (MLB), Kernel de Regresión (KR) y K-vecinos próximos (KNN). En la Tabla 2.2 se puede observar que el método KNN con $k = 11$ usando el conjunto B de rasgos, tiene el menor error.

Para disminuir el error fueron elegidas 2 estrategias de selección de parámetros (Feature Selection FS). La primera FS consiste en hacer una graduación de parámetros, de tal forma que queden ordenados en relación a su aportación individual a la clasificación global y posteriormente se hace una selección de parámetros hacia adelante respetando ese orden, mientras que la segunda FS consiste en hacer la búsqueda del conjunto de parámetros midiendo las combinaciones entre los rasgos de 1 en 1, empezando por un rasgo seleccionado al azar. Los resultados de ambas estrategias pueden verse en la Tabla 2.3.

Tabla 2.3: Resultados de 2 selecciones de rasgos, los primeros más significativos (PFS) y selección de rasgos hacia adelante (FS). [14]

Método	Error (A)	Error (B)
PFS	36 % (4)	28 % (8)
FS	34.5 % (4)	28.5 % (5)

Posteriormente, en [35] es reportado el uso de una base de datos generada a partir de 50 sentencias que van desde 2 a 12 palabras y como locutores se solicitó a 5 estudiantes de drama que pronunciaran las sentencias con la etiqueta de la emoción correspondiente. Las emociones manejadas fueron feliz, triste, enojado, miedo y neutral. De tal forma que se tienen un máximo de 250 sentencias por estudiante.

Algunas personas hicieron el reconocimiento de las cuatro emociones (feliz, triste, enojado y miedo) en un orden aleatorio, obteniendo así un desempeño del 70 %. En este trabajo se utilizaron Modelos Ocultos de Markov (HMM), utilizando parámetros prosódicos y la validación de los resultados fue hecha mediante el método "Hold out" usando un 70 % de entrenamiento y el resto para prueba, resultados que se muestran en la Tabla 2.4.

Tabla 2.4: Desempeño por emoción, usando parámetros prosódicos con Modelos Ocultos de Marcov. [35]

Emoción	Feliz	Miedo	Enojado	Triste
Precisión	93.8 %	60.0 %	77.9 %	59.6 %

En otro artículo [28] la base de datos se colectó mediante 5 estudiantes voluntarios sin entrenamiento (3 hombres y 2 mujeres), cada locutor grabó 20 sentencias para cada emoción (neutral, feliz, enojado, miedo, sorprendido y triste), de las cuales 15 fueron usadas para el entrenamiento y 5 como conjunto de prueba.

El análisis de los parámetros se hizo mediante 2 aproximaciones, las de corto plazo: primeras 4 formantes, primeros anchos de banda de los 4 formantes, pitch, energía en escala logarítmica y los coeficientes de autocorrelación de primer orden normalizados. Para las de largo plazo, se calcula para cada rasgo de corto plazo, los siguientes valores: promedio del parámetro sobre toda la sentencia, promedio de la primera y de la segunda parte de la sentencia y promedio de cada tercio de la sentencia.

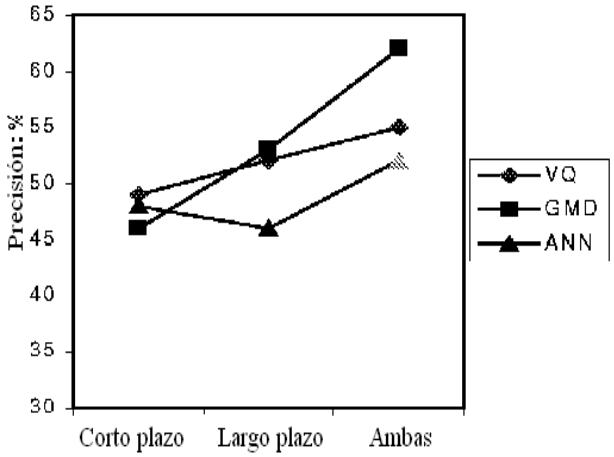


Figura 2.2: Comparación de clasificación usando los 3 modelos. [28]

En la Figura 2.2, podemos observar que el modelo de Densidad de Mixturas Gaussianas (GMD), tiene el mejor desempeño usando la combinación de los parámetros de corto y largo plazo, aunque utilizando únicamente las variables de corto plazo, mientras que GMD es el modelo que tiene el desempeño más bajo. En las Tablas 2.5 y 2.6, se puede observar el desempeño del uso de rasgos de corto plazo y largo plazo respectivamente.

Tabla 2.5: Resultados usando parámetros de corto plazo con GMD. [28]

	Neutral	Feliz	Enojado	Miedo	Sorpresa	Triste
Neutral	0.4	0.1	0	0.2	0	0.3
Feliz	0.005	0.6	0.1	0	0.25	0
Enojado	0.05	0.2	0.45	0.1	0.2	0
Miedo	0.25	0.05	0	0.4	0.1	0.2
Sorpresa	0	0.35	0.3	0.1	0.25	0
Triste	0.1	0	0.1	0.15	0	0.65

En este artículo, se usaron 3 modelos, el vector de cuantificación, redes neuronales artificiales y el modelo de Densidad de Mixturas Gaussianas (GMD). Este último modelo es el que presentó mejores resultados, ver Tabla 2.7, combinando parámetros tanto de corto como de largo plazo.

Tabla 2.6: Resultados usando parámetros de largo plazo con GMD. [28]

	Neutral	Feliz	Enojado	Miedo	Sorpresa	Triste
Neutral	0.4	0.1	0	0.25	0	0.25
Feliz	0	0.6	0	0.05	0.35	0
Enojado	0.05	0.1	0.5	0.05	0.1	0.2
Miedo	0.25	0.05	0	0.35	0.05	0.3
Sorpresa	0	0.3	0.05	0.25	0.4	0
Triste	0.05	0.05	0	0	0	0.9

Tabla 2.7: Resultados usando parámetros de corto y largo plazo con GMD. [28]

	Neutral	Feliz	Enojado	Miedo	Sorpresa	Triste
Neutral	0.45	0.1	0	0.2	0	0.25
Feliz	0.05	0.85	0	0	0.1	0
Enojado	0.05	0.1	0.5	0.05	0.1	0.2
Miedo	0.2	0.05	0	0.45	0.05	0.25
Sorpresa	0	0.25	0.05	0.15	0.55	0
Triste	0.05	0.05	0	0	0	0.9

Otra aproximación [50] se utiliza una base de datos extraída de películas o televisión, la cual es un conjunto de 721 sentencias cortas (ver Tabla 2.8) que con tienen 4 emociones (enojado, feliz, triste y neutral)

Tabla 2.8: Todos los archivos tienen una frecuencia de muestreo de 16kHtz. [50]

Enojado	Feliz	Neutral	Tristeza
215	136	242	128

Fueron extraídos un total de 60 variables, agrupadas de la siguiente forma:

-Rasgos estadísticos relacionados con el ritmo: velocidad del habla, promedio de la longitud entre regiones vocalizadas, entre otras.

-Relaciones estadísticas de la señal del pitch suavizado: mínimo, máximo, mediana y desviación estándar.

-Variables estadísticas del gradiente del pitch suavizado: mínimo, máximo, mediana y desviación estándar.

-Estadísticas sobre las partes individuales vocalizadas: promedio del mínimo y promedio del máximo.

-Estadísticas sobre partes individuales de las curvas: Promedio positivo y promedio negativo.

Todos los parámetros fueron calculados solo en las regiones válidas, las cuales empiezan en el primer valor no cero del pitch (F0) y terminan en el último valor no cero del pitch.

Para la clasificación fueron usados 3 modelos, redes neuronales artificiales, k-vecinos próximos y máquinas de soporte vectorial (SVM). En el caso de las SVM's, se utilizó una SVM por emoción, y para la validación de resultados fueron utilizadas 100 sentencias de cada emoción para el entrenamiento y el resto para prueba. Los resultados pueden verse en la Tabla 2.9

Tabla 2.9: Comparación del desempeño de los 3 modelos. [50]

Método	Enojado	Feliz	Neutral	Triste
NN	40 %	27.78 %	62.68 %	35.71 %
KNN	42.86 %	39.28 %	89.29 %	32.14 %
SVM	77.16 %	65.64 %	83.73 %	70.59 %

En [41] se hace uso de una base de datos en alemán, que contiene 40 comandos con las emociones de enojado, feliz, triste, aburrido y neutral, los cuales fueron grabados por 14 locutores no-actores (7 hombres y 7 mujeres) con un total de 2800 instancias. Para la evaluación de resultados se usó una validación cruzada dejando un locutor afuera para los resultados independientes del locutor, mientras que para los resultados dependientes del locutor se tomó un 80 % de la base de entrenamiento y el resto de prueba.

Los parámetros fueron extraídos en 2 grupos principales, los prosódicos y los de calidad. Para los primeros, se obtuvo un conjunto de 37 rasgos como los siguientes:

-Logaritmo de F0: máximo, mínimo, posición máxima, posición mínima, promedio, desviación estándar, coeficientes de regresión, F0 para el primer y último

segmentos de señal vocalizada.

-Energía: máximo, posición máxima, posición mínima, promedio, coeficientes de regresión y error cuadrático promedio para los coeficientes de regresión.

-Aspectos de duración: número de regiones vocalizadas y no vocalizadas, número de segmentos vocalizados y no vocalizados, mayor región vocalizada y no vocalizada, razón del número de segmentos vocalizados entre los no vocalizados, razón del número de regiones vocalizadas entre las no vocalizadas, razón del número de segmentos vocalizados entre el total y razón del número de regiones vocalizadas entre el total.

Los rasgos de calidad, describen las 3 primeras formantes, sus anchos de banda, distribución espectral de la energía, razón entre la energía vocalizada entre la no vocalizada y flujo glotal. Estos parámetros fueron extraídos usando un software de análisis fonético PRAAT.

Tabla 2.10: Clasificación dependiente del locutor. [41]

	Alto	Neutral	Bajo
Alto	82.1 %	17.9 %	0 %
Neutral	10.3 %	82.8 %	6.9 %
Bajo	0 %	13 %	87 %

Tabla 2.11: Clasificación dependiente del locutor. [41]

	Feliz	Enojado		Aburrido	Triste
Feliz	75 %	25 %	Aburrido	76 %	24 %
Enojado	28 %	72 %	Triste	44 %	56 %

Tabla 2.12: Clasificación independiente del locutor. [41]

	Alto	Neutral	Bajo
Alto	68.1 %	17.3 %	14.5 %
Neutral	14.4 %	3.7 %	81.8 %
Bajo	14.4 %	3.7 %	81.8 %

Para la selección de rasgos se usaron modelos de regresión lineal. La clasificación fue llevada a cabo por modelos de redes neuronales artificiales, y se observa que para una clasificación dependiente del locutor, se tiene un reconocimiento del 83.7 % para clasificar los estados neutro, alto y bajo (ver Tabla 2.10). La clasificación de los estados feliz-enojado se alcanzó un reconocimiento del 73.5 % y para los estados aburrido-triste fue del 66 % (ver Tabla 2.11); por último, en la clasificación independiente del locutor, se logró un 77 % para los estados alto, neutral y bajo (ver Tabla 2.12).

En el artículo [19] se presenta un análisis de reconocimiento de emociones dentro de un contexto multi-lenguaje, con bases de datos en idioma inglés, esloveno, castellano y francés. Dichas bases de datos incluyen varios estilos neutrales, y 6 emociones: disgustado, sorprendido, alegre, miedo, enojado y triste. La base de datos en inglés fue elaborada mediante 2 locutores varones adultos y una mujer, las demás bases, utilizaron un varón y una mujer únicamente.

Para la base de datos en idioma inglés, se grabaron 186 sentencias, en esloveno 190, castellano 184 y francés 175. Las sentencias de los corpus contienen palabras aisladas, oraciones cortas, medias y largas, las cortas están conformadas de 5 a 8 palabras, las medias de 9 a 13 y las largas de 14 a 18. Dichas frases fueron expresadas en forma interrogativa y afirmativa.

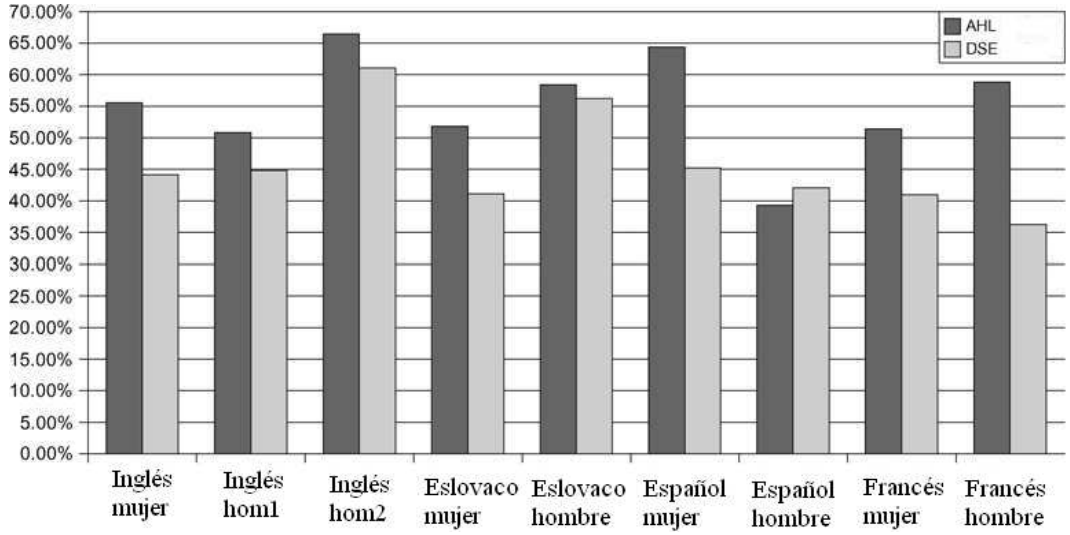


Figura 2.3: Clasificación dependiente del locutor. [19]

Los parámetros fueron extraídos en 2 grupos, los de bajo nivel: calculados a partir del pitch, gradiente del pitch, energía, gradiente de la energía y duración de los segmentos vocalizados, los de alto nivel: son representaciones estadísticas de los rasgos de bajo nivel. A partir de estos datos se establecieron 2 tipos de datos, los Específicamente tomados de la Base de Datos Emocional (DSE por sus siglas en inglés) son 14 parámetros extraídos exclusivamente a partir del pitch (F0), gradiente del pitch y duración, por otro lado los de Todo Alto Nivel (AHL por su siglas en inglés) suman 26 parámetros. El modelo usado para el reconocimiento de emociones es el de redes neuronales artificiales.

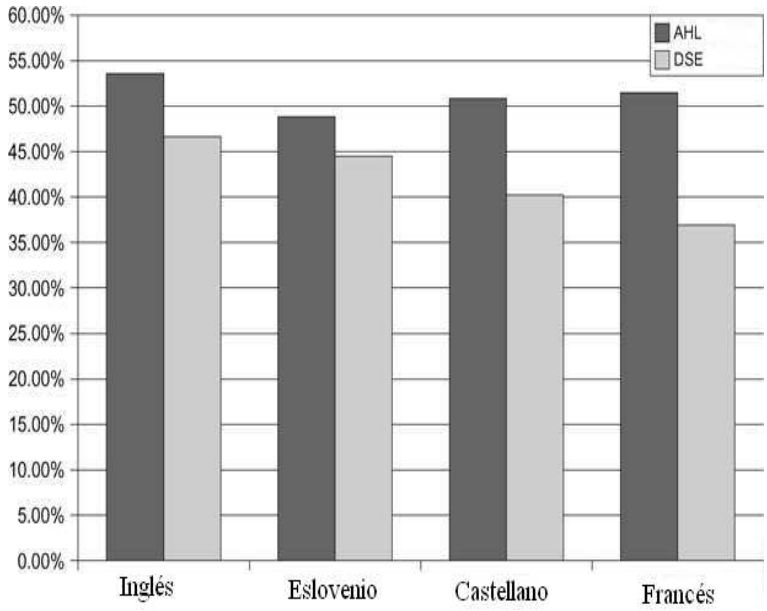


Figura 2.4: Clasificación mono-lenguaje usando parámetros DSE y AHL. [19]

Se generaron 4 topologías diferentes de perceptrón multi-capa (MLP), debido a los distintos tamaños de las entradas y salidas. Las 4 topologías tienen 26 neuronas en la capa oculta, en la capa de salida; la primera y la segunda tienen 8 neuronas, la tercera y cuarta tienen 7; la primera y la tercera tienen 26 neuronas en la capa de entrada mientras que la segunda y la cuarta tienen 14. Todas las neuronas manejan la función de tangente hiperbólica como función de activación.

En la Figura 2.3 se puede observar que para el segundo locutor masculino, se logró el mejor reconocimiento de emociones. En idioma inglés, se encontró una mayor precisión en el reconocimiento de emociones (ver Figura 2.4). Para el reconocimiento multi-lenguaje, la emoción de tristeza es la que mejor se identifica (ver Figura 2.5).

En [44], se trabajó con la base de datos emocional en Danés, dicho corpus, consta de 500 registros (sin silencios), generados a partir de 4 actores profesionales (2 hombres y 2 mujeres), expresando 5 estados emocionales: enojado, feliz, neutral, tristeza y sorprendido.

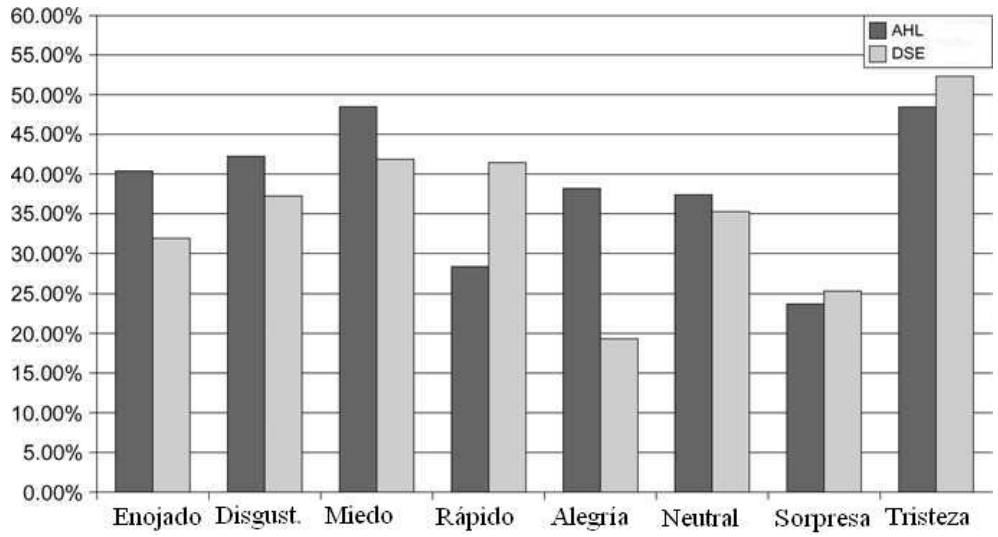


Figura 2.5: Clasificación multi-lenguaje de emociones con parámetros DSE y AHL. [19]

Se extrajeron 87 parámetros basados en el pitch y la energía, de los cuales, mediante un método de Selección Secuencial hacia Adelante (SFS) es encontrado un subconjunto de 5 rasgos para realizar la clasificación. El reconocimiento fue realizado por el método de Naive Bayes, puede verse en la Tabla 2.13 los resultados alcanzados mediante este enfoque.

Tabla 2.13: Matriz de confusión del clasificador Naive Bayes y porcentajes de reconocimiento obtenidos por personas. [44]

Matriz de confusión para clasificación bayesiana

	Neutral	Sorpresa	Feliz	Tristeza	Enojado
Neutral	56	13	3	25	3
Sorpresa	6	65	5	9	15
Feliz	9	24	39	14	14
Tristeza	17	6	1	72	4
Enojado	14	14	20	12	40

Clasificación alcanzada por humanos

	Neutral	Sorpresa	Feliz	Tristeza	Enojado
Neutral	60.8	2.6	0.1	31.7	4.8
Sorpresa	10	59.1	28.7	1.0	1.3
Feliz	8.3	29.8	56.4	1.7	3.8
Tristeza	12.6	1.8	0.1	85.2	0.3
Enojado	10.2	8.5	4.5	1.7	75.1

En el trabajo [34], se trabajó con 2 bases de datos en mandarín, la primera consta de 538 registros grabados por 12 locutores nativos y la otra contiene 503 sentencias hechas por 2 actores profesionales. Los parámetros fueron extraídos a partir de 16 coeficientes LPC's, 12 LPCC, 16 LFPC, 16 PLP, 20 MFCC's y el jitter. Los modelos usados para este artículo fueron Análisis Discriminante Lineal (LDA), k-vecinos (K-NN) y Modelos Ocultos de Markov (HMM). Se obtuvo un reconocimiento del 88.3 % para la primera base de datos y 88.7 % para la segunda.

En la Tabla 2.14, puede observarse la distribución de registros según su base de datos, emoción y género.

La Figura 2.6 muestra un diagrama de bloques de como es la extracción de rasgos, el vector Y_1 es generado a partir de la obtención de varios centroides por medio del algoritmo Linde-Buzo-Gray (LBG) [34], el segundo vector (Y_2) es obtenido con el promedio.

La Figura 2.6 muestra un diagrama de bloques del proceso de extracción de parámetros. En el preprocesamiento primero se localizan los puntos finales, seguidamente se pasa la señal de voz por un filtro pasa altos, para enfatizar las componentes de alta frecuencia, después la señal es particionada en ventanas de 256 muestras;

Tabla 2.14: Número de registros por género, por emoción y por base de datos. [34]

Sentecias del Corpus I			
	Mujer	Hombre	Total
Enojado	75	76	151
Aburrido	37	46	83
Feliz	56	40	96
Neutral	58	58	116
Tristeza	54	58	112
Total	280	278	558
Sentencias del Corpus II			
	Mujer	Hombre	Total
Enojado	36	72	108
Aburrido	72	72	144
Feliz	36	36	72
Neutral	36	36	72
Tristeza	72	35	107
Total	252	251	503

posteriormente se aplica una ventana de Hamming a cada ventana individualmente para minimizar las discontinuidades de la señal.

Con el fin de encontrar una combinación adecuada de parámetros extraídos, se utilizó el método de selección por regresión para determinar los rasgos más benéficos de entre más de 200 parámetros de voz. Diez candidatos fueron seleccionados: LPC, LPCC, MFCC, Delta MFC, Delta-Delta MFCC, PLP, RastaPLP, LFPC, jitter y shimmer. Como método de validación de resultados, se usó el esquema Leave-One-Out (LOO), los porcentajes de reconocimiento para cada emoción con cada algoritmo (LDA, K-NN y HMM's) se pueden observar en la Tabla 2.15

En [31] se utilizó una base de datos en lengua euskara, que contiene 6 emociones (alegría, asco, ira, miedo, sorpresa y tristeza), se utilizó una actriz profesional para hacer las grabaciones. En total son 97 grabaciones por emoción. Los modelos usados para clasificación son máquinas de soporte vectorial (SVM) y Mixturas Gaussianas (GMM). Los resultados están reportados con validación cruzada.

Fueron extraídos un total de 86 parámetros prosódicos, de ése conjunto se obtuvieron 6 parámetros usando Máquinas de Soporte Vectorial, usando un método

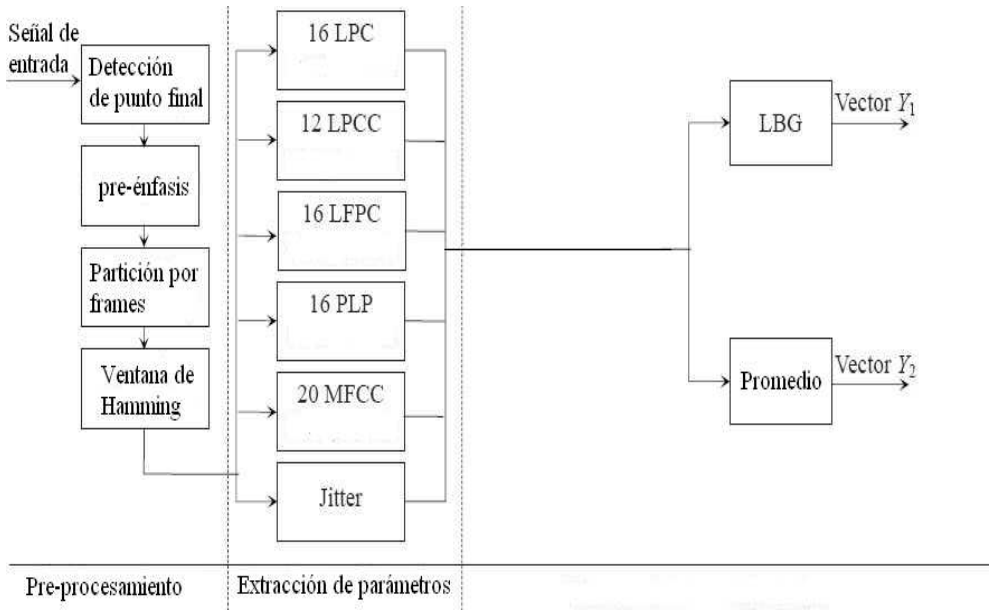


Figura 2.6: Diagrama de la extracción de parámetros. [34]

llamado Jack-knife [31], con los 6 parámetros prosódicos y SVM's, los resultados pueden ser vistos en la Tabla 2.16; usando los mismos 6 parámetros con GMM, se obtuvo la matriz de confusión mostrada en la Tabla 2.17, finalmente la matriz de confusión de la Tabla 2.18 muestra los resultados obtenidos al emplear todos los parámetros con Mixturas Gaussianas.

Para [46], la clasificación de emociones se mejora al usar un detector de género previo a la clasificación de estados emocionales (ver Figura 2.7). En este trabajo se usaron 2 bases de datos, la base de datos de Berlín y la base de datos "SmartKom mobile", ambas están en Alemán, la primera ya ha sido descrita anteriormente y fue usada dejando 1 hombre y 1 mujer para pruebas y los otros locutores para entrenamiento, la segunda fue usada con 56 locutores (24 hombres y 32 mujeres) como entrenamiento y 14 (7 hombres y 7 mujeres) para pruebas; aunque esta base de datos consta de 12 emociones, fueron utilizadas las emociones neutral, alegría, impotencia y enojo.

Tabla 2.15: Resultados del reconocimiento de las 5 emociones con varios clasificadores. [34]

Resultados experimentales de 5 emociones para el corpus I						
Precisión (%)	LDA Y1	LDA Y2	K-NN Y1	K-NN Y2	HMMs Y1	HMMs Y2
Enojado	81.5	80.4	82.3	84.8	86.4	86.7
Aburrido	80,3	79.8	84.9	82.3	89.1	88.4
Feliz	76.5	72.3	79.5	82.1	82.3	83.6
Neutral	78.4	80.4	80.4	81.2	84.5	90.5
Tristeza	82.5	81.3	91.2	89.1	92.4	92.3
Promedio	79.8	78.8	83.6	83.9	86.9	
Resultados experimentales de 5 emociones para el corpus II						
Precisión (%)	LDA Y1	LDA Y2	K-NN Y1	K-NN Y2	HMMs Y1	HMMs Y2
Enojado	82.4	76.2	83.2	84.5	90.2	91.4
Aburrido	78.9	80.2	81.5	80.9	84.3	86.7
Feliz	81.4	77.8	86.4	82.5	87.5	88.1
Neutral	76.5	79.8	84.1	83.2	90.3	86.0
Tristeza	80.3	76.5	86.0	87.5	89.5	91.5
Promedio	79.9	78.1	84.2	83.7	88.3	88.7

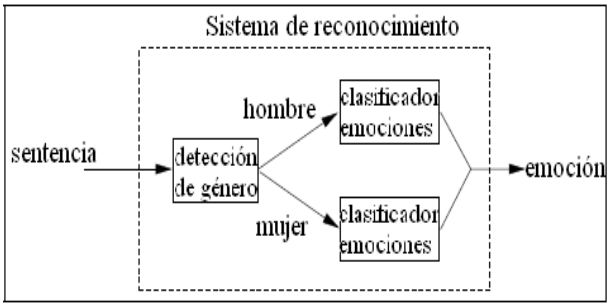


Figura 2.7: Detección de género previa a la clasificación de emociones. [46]

Se extrajeron un total de 1289 parámetros, de éstos, 20 fueron usados para la detección de género para la base de datos de Berlín y 12 para la SmartKom (ver Tabla 2.19), no se reporta cómo se extraen los parámetros.

Tabla 2.16: Matriz de confusión usando 6 parámetros prosódicos y SVM. [31]

	Ira	Miedo	Sorpreasa	Asco	Alegría	Tristeza	Neutral
Ira	92	-	-	1	2	-	-
Miedo	-	94	9	-	-	-	-
Sorpresa	-	3	88	-	-	-	-
Asco	-	-	-	80	-	4	3
Alegría	2	-	-	-	88	-	1
Tristeza	2	-	-	10	-	93	1
Neutral	1	-	-	6	7	-	92
Eficiencia (%)	94.9	96.9	90.7	82.5	90.7	95.9	94.9

Tabla 2.17: Matriz de confusión usando 6 parámetros prosódicos y GMM. [31]

	Ira	Miedo	Sorpreasa	Asco	Alegría	Tristeza	Neutral
Ira	89	2	4	-	4	-	-
Miedo	-	90	8	-	-	-	-
Sorpresa	1	5	83	-	-	-	-
Asco	2	-	-	73	-	14	1
Alegría	4	-	-	-	82	-	8
Tristeza	-	-	-	14	-	83	1
Neutral	1	-	-	10	11	-	87
Eficiencia (%)	91.8	92.8	87.4	75.3	84.5	85.6	89.7

En la Tabla 2.20 se muestra una comparativa de la detección de género, una a partir del pitch exclusivamente, y la otra clasificación fue usando los parámetros de la Tabla 2.19. Los resultados de la clasificación global se muestran en la Tabla 2.21

En el trabajo [7], se reporta el uso de la base de datos bilingüe "RekEmozio", que contiene registros en idioma español y vasco. Los parámetros extraídos están basados en la frecuencia fundamental (F0), energía, distribución espectral de la energía, sonoridad, formantes y sus bandas de frecuencia, jitter, shimmer y velocidad del habla.

Tabla 2.18: Matriz de confusión usando los 86 parámetros prosódicos y GMM. [31]

	Ira	Miedo	Sorpreasa	Asco	Alegría	Tristeza	Neutral
Ira	88	3	4	1	4	-	-
Miedo	1	89	13	-	-	1	-
Sorpresa	1	5	78	-	-	-	-
Asco	3	-	-	76	1	7	2
Alegría	4	-	-	3	68	-	8
Tristeza	-	-	-	7	-	89	1
Neutral	-	-	-	10	24	-	86
Eficiencia (%)	90.7	91.8	82.1	78.4	70.1	91.8	88.7

Tabla 2.19: Parámetros relevantes para la detección del género. [46]

Rasgos	Berlin	SmartKom
Pitch	1	2
Energía	2	3
MFCC	17	7
Σ	20	12

La selección de parámetros fue llevada a cabo mediante (Feature Subset Selection - FSS) con la estimación de distribución de algoritmos (Estimation of Distribution Algorithms - EDA). Los modelos usados para la clasificación fueron: árboles de decisión, aprendizaje basado en instancias, árboles C4.5, Naive Bayes y el árbol Naive Bayes de aprendizaje. Los resultados pueden verse en la Figura 2.8, donde se muestran los resultados de clasificación usando todos los parámetros y el resultado de usar los parámetros encontrados en la selección de rasgos. Los resultados se pueden ver independientes para cada base de datos y para cada algoritmo de clasificación.

Tabla 2.20: Parámetros relevantes para la detección del género. [46]

	F0 promedio	Conjunto optimizado
Berlín	69.37 %	90.26 %
SmartKom	87.56 %	91.85 %

Tabla 2.21: Resultados de los tipos de clasificación con detección de género y sin detección de género. [46]

		Berlín	SmartKom
Sin información del género		81.14 %	75.11 %
Con información correcta del género	mujer	84.62 %	78.99 %
	hombre	87.92 %	75.36 %
	combinado	86.00 %	76.74 %
Con información del género reconocida	mujer	84.93 %	81.38 %
	hombre	80.09 %	75.84 %
	combinado	82.76 %	78.22 %

La validación de la medida del error fue k-fold cross validation con $k = 10$.

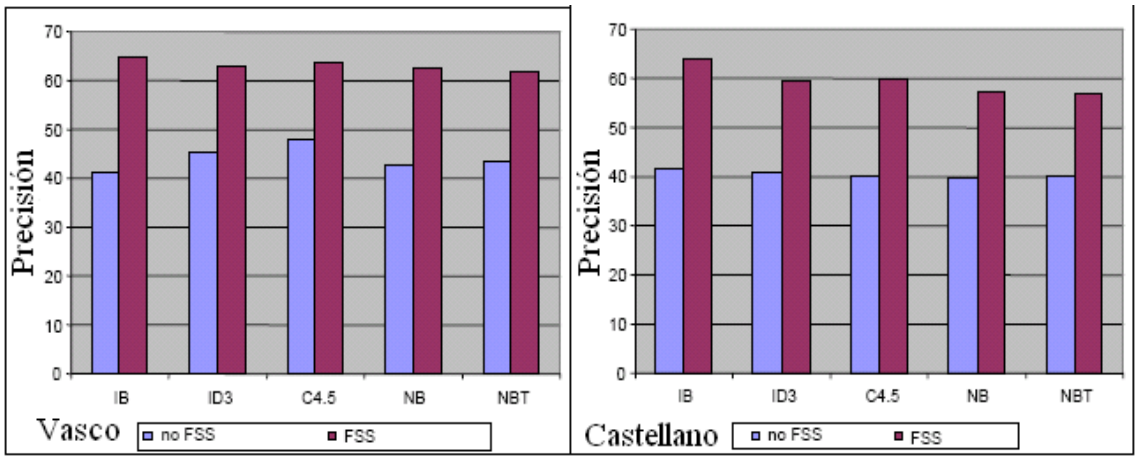


Figura 2.8: Mejora para ambas bases de datos. [7]

En el trabajo publicado por [13], se ocuparon las bases de datos de Berlín y la de Polonia, reportando un reconocimiento del 72 % para la clasificación independiente del locutor. Fueron extraídos 102 parámetros basados en 3 grupos principales: la frecuencia fundamental, la energía y parámetros temporales como las pausas.

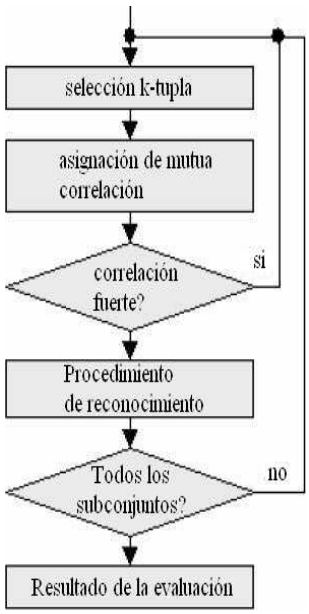


Figura 2.9: Diagrama de bloques de la selección de rasgos. [13]

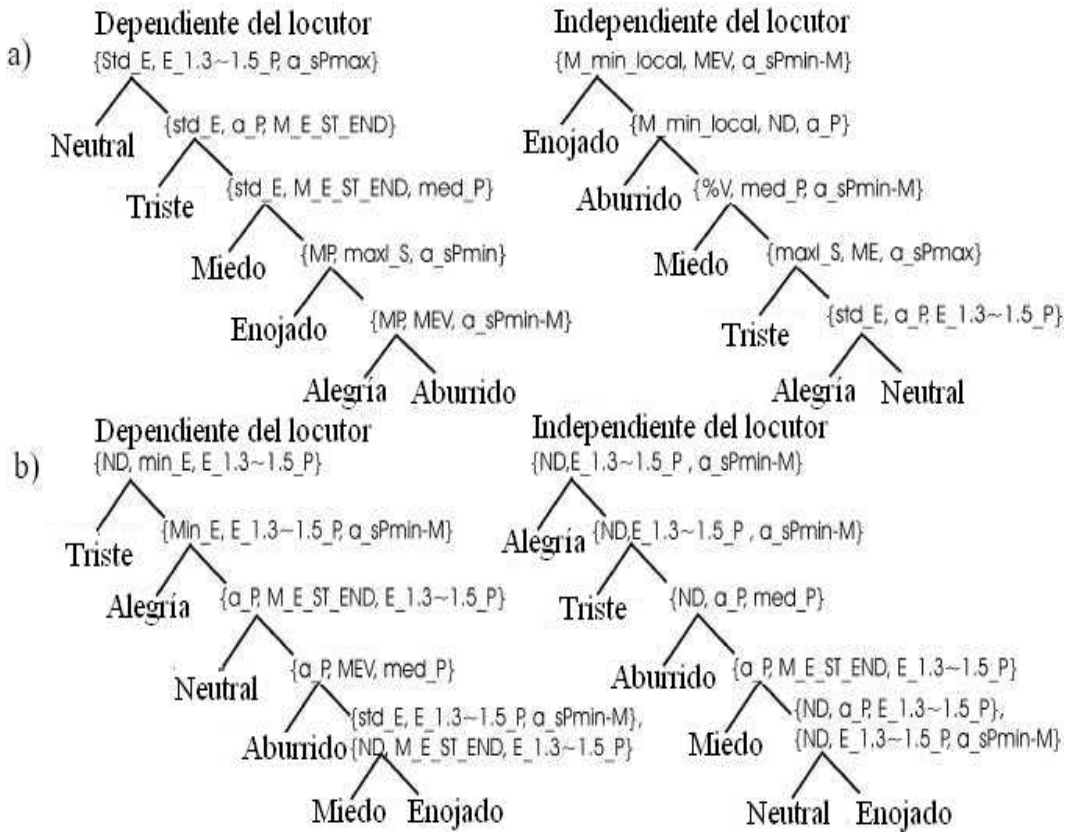


Figura 2.10: Modelos encontrados para cada base de datos (a) es para la base de datos de Berlín, (b) para el corpus en polaco. [13]

La selección de rasgos por medio de estos árboles de decisión binaria, fue usando subconjuntos de 3 rasgos, 1 por cada grupo (F0, energía y duraciones), se evalúa la correlación del subconjunto, si la correlación rebasa cierto umbral, el conjunto es desechado y se selecciona otro. Esto puede ser visto en la Figura 2.9.

Para la base de datos de Berlín, solo se usaron 6 emociones, la emoción de disgustado no fue ocupada, para poder hacer comparaciones entre las 2 bases de datos. Los resultados del reconocimiento pueden ser vistos en la Tabla 2.22

Tabla 2.22: Resultados de clasificación para ambas bases de datos. [13]

Base de datos	Mejor resultado	
	Dependiente del locutor	Independiente del locutor
Polaco	76.30 %	64.18 %
Alemán	74.39 %	72.04 %

En [48], se experimenta con la base de datos de Berlín, se alcanza un desempeño del 76.22 % y del 79.47 % cuando se hace una clasificación previa del género. En este trabajo no se trabajó con la emoción de disgustado, por lo que la clasificación corresponde a 6 estados emocionales (enojado, aburrido, miedo, alegría, neutral y tristeza). La validación de los resultados fue hecha mediante 10 corridas donde se tomó el 50 % de la base de datos para entrenamiento y el otro 50 % de prueba, para cada corrida las sentencias fueron tomadas aleatoriamente.

Fueron extraídos 68 parámetros basados en la frecuencia fundamental, los 3 primeros formantes, energía, duración, y 2 grupos más de parámetros unos basados en el análisis armónico de la señal (extraídos a partir del contorno de la frecuencia fundamental, pasado por un banco de filtros, se obtiene su envolvente y se saca la FFT de dicha envolvente) (ver Figura 2.11); el último grupo de parámetros son extraídos a partir de una ley propuesta empíricamente [48], de la cual se extraen los rasgos "Zipf".

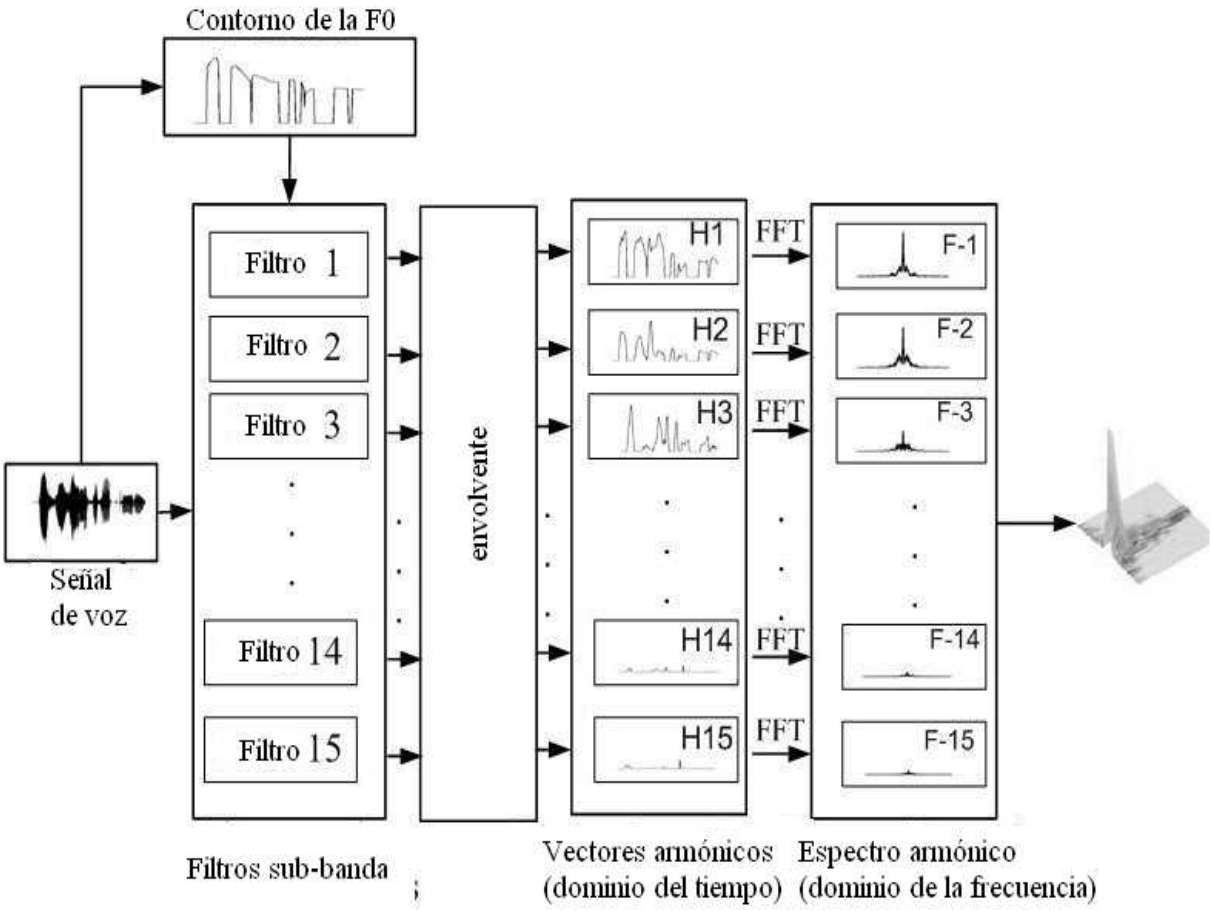


Figura 2.11: Análisis armónico de la señal. [48]

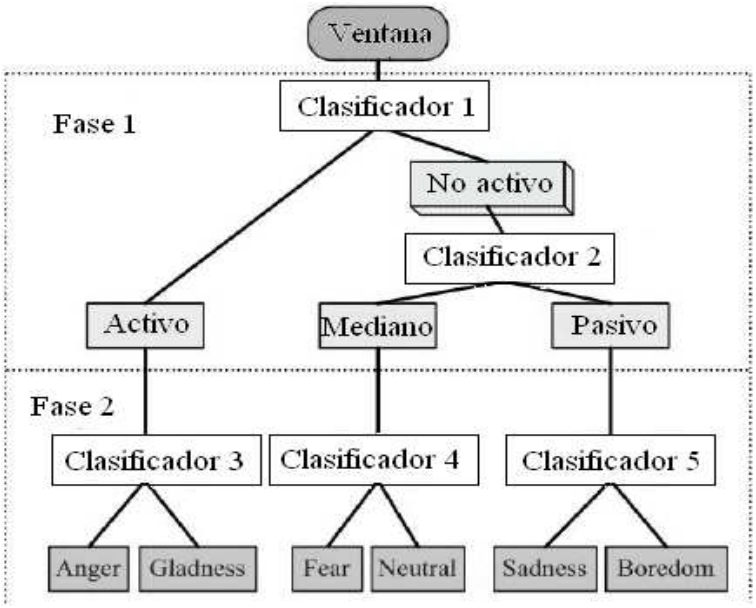


Figura 2.12: Clasificación de emociones mediante 2 etapas. [48]

Se utilizaron redes neuronales artificiales (Perceptrón Multicapa) con 2 capas ocultas, la función de transferencia es la función logarítmica sigmoidal. En la capa de salida solo hay una neurona que separa 2 clases con un umbral de 0.5. En la Figura 2.12 se puede observar como es el esquema de clasificación independiente del locutor.

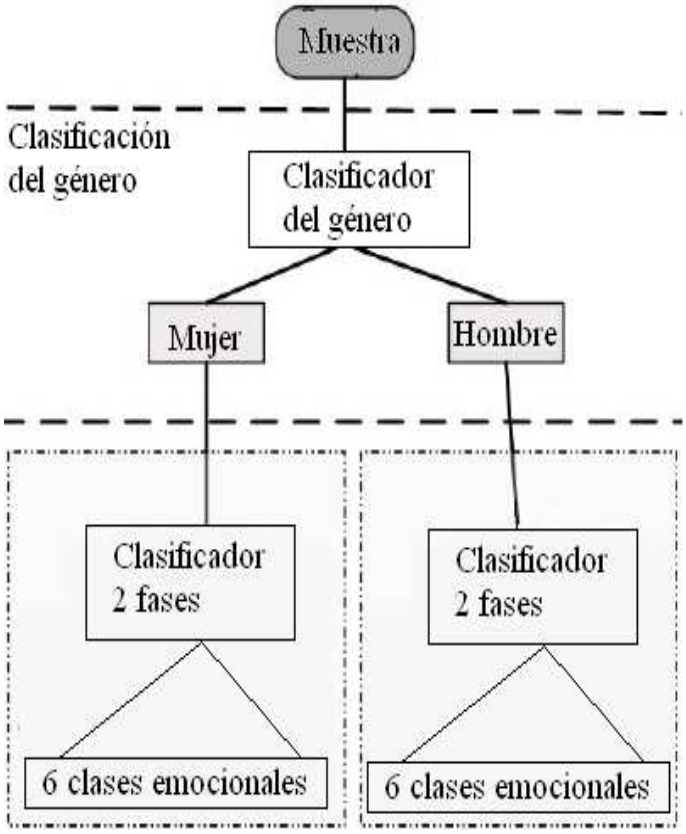


Figura 2.13: Clasificación jerárquica con información del género. [48]

En la Figura 2.13 se ilustra cómo se realiza la clasificación dependiente del locutor. Los resultados obtenidos mediante esta estrategia, se observan en la Tabla 2.23

Tabla 2.23: Porcentajes de clasificación de emociones para cada caso. [48]

	Hombre	Mujer	Promedio	Mezclado	Mezclado con clasificador del género
Global	81.56	76.76	78.86	75.12	76.95
2 fases	84.17	79.88	81.75	76.22	-
3 fases	-	-	-	-	79.47

En el trabajo [51] se reporta el uso de un modelo basado en el k-vecinos próximos con una estimación de costo del error en dicho trabajo usa las 7 emociones de la base de datos de Berlín [2], y se reporta un resultado del 82.44 % de reconocimiento, obtenido con una validación cruzada de 2-folds (50 % de los datos son usados para entrenamiento y 50 % para prueba.

En [16], se muestra una amplia información de las diversas estrategias que se han tomado para abordar el problema de reconocimiento de emociones; dicha tarea es muy desafiante debido a 3 razones principales; la primera es que no se sabe cuáles son los parámetros más potentes para clasificar emociones, la segunda es que la expresión de emociones depende de las raíces culturales del locutor, de su idioma, además de que hay sentencias en las que se pueden percibir más de una emoción y es muy difícil tratar de encontrar los límites de frontera entre las emociones. Finalmente hay emociones como tristeza que pueden tardar unas pocas horas, unos días o más aún meses, mientras que hay otras que cuando se manifiestan solo duran unos minutos como la ira.

Por otro lado, la mayor parte de bases de datos de voz emotiva no están disponibles al público, como se muestra en la Tabla 2.24. En dicha Tabla se muestran las bases de datos de voz emotiva más usadas en la investigación de reconocimiento automático de emociones a partir de la voz, el principal problema de la mayor parte de estas bases de datos es que no simulan lo suficientemente clara y natural las emociones, por lo que el reconocimiento llevado a cabo por personas está por debajo del 80 % de reconocimiento.

Un sistema automático de clasificación de emociones a partir de la voz consta de 2 etapas, donde la primera es llevar a cabo un proceso para extraer los parámetros apropiados de los datos disponibles (en este caso la señal de voz), y la segunda etapa es la selección del clasificador, es importante mencionar que mucho del trabajo publicado recientemente se enfoca más a esta segunda etapa. Otro problema que se tiene en relación a los clasificadores es que no se ha encontrado uno que sea el más apropiado para resolver este problema del reconocimiento de emociones y es un hecho que cada clasificador tiene sus ventajas y limitaciones. [16]

La mayor parte de los trabajos reportan clasificación de emociones usando 4, 5 o 6 de ellas, específicamente, cuando se reporta el uso de la base de datos de Berlín,

Tabla 2.24: Desempeño por emoción, usando parámetros prosódicos con Modelos Ocultos de Marcov. [16]

Corpus	Tamaño	Emociones
LDC Voz Prosódica Emocional y Transcripciones[6]	7 actores x 15 emociones x 10 sentencias	Neutral, pánico, ansiedad, enojado fuerte, enojado calmado, desdesperación, tristeza, júbilo, alegría, interesado, aburrido, vergüenza, orgullo y desprecio
Base de Datos emocional de Berlín[2]	535 instancias (10 actores x 7 emociones x 10 sentencias)	Enojado, alegría, tristeza, miedo, disgustado, aburrido y neutral
Base de Datos emocional Danesa[5]	4 actores x 5 emociones (2 palabras + 9 oraciones + 2 pasajes)	Enojado, alegría, tristeza, sorpresa y neutral
Natural[32]	388 instancias, 11 locutores, 2 emociones	Enojado y neutral
ESMBS[33]	720 sentencias, 12 locutores, 6 emociones	Enojado, alegría, tristeza, disgustado, miedo y sorpresa
INTERFACE[20]	Inglés (186 sentencias), eslovaco (190 sentencias), español (184 sentencias) y francés (175 sentencias)	Enojado, disgustado, miedo, alegría, sorpresa, tristeza, neutral lento y neutral rápido
KISMET[11]	1002 instancias, 3 locutoras y 5 emociones	Aprobación, atención, prohibición, calmante y neutral
BabyEars[39]	509 instancias, 12 actores (6 hombres y 6 mujeres), 3 emociones	Aprobación, atención y prohibición
MPEG-4[38]	2440 Instancias, 35 locutores	Alegría, enojado, disgustado, miedo, tristeza, sorpresa, neutral
Universidad de Beihang[17]	7 actores x 5 emociones x 20 sentencias	Enojado, alegría, tristeza, disgustado y sorpresa
FERMUS III[37]	2829 instancias, 7 emociones y 13 actores	Enojado, disgustado, alegría, neutral, tristeza y sorprendido
KES[24]	5400 instancias, 10 actores	Neutral, alegría, tristeza y enojado
CLDC[52]	1200 instancias, 4 actores	Alegría, enojado, sorpresa, miedo, neutral y tristeza
Hao Hu[22]	8 actores x 5 emociones x 40 instancias	Enojado, miedo, alegría, tristeza, neutral
Amir[8]	60 actores Hebreos y 1 Ruso	Enojado, disgustado, miedo, alegría, neutral y tristeza
Pereira[21]	2 actores x 5 emociones x 8 instancias	Enojado Fuerte, enojado calmado, alegría, neutral, tristeza.

Tabla 2.25: Desempeño de los clasificadores mas usados en el reconocimiento de emociones. [16]

Clasificador	HMM	GMM	ANN	SVM
Precisión promedio	75.5-78.5 %	74.83-81.94 % /63-70 %	51.19-52.82 %	75.45-81.29 %

por lo general no se toma en cuenta la emoción de disgustado. Por otro lado, cuando la validación de resultados se hace por medio de una validación cruzada de k-folds, se sugiere que k tenga un valor entre 10 a 20 [25]. No se ha reportado el uso de varios modelos de clasificación entre ellos, los modelos asociativos.

Como se pudo ver en [16], en lo que respecta al campo de reconocimiento de emociones a partir de la voz, aún no se han identificado ni los parámetros clave ni el modelo más significativo para poder encontrar un solución o un marco de trabajo óptimo.

El aporte más significativo de esta tesis consiste en una representación bidimensional de la energía, mientras que el modelo para clasificar asignado a esta tarea son las máquinas asociativas Alfa-Beta con Soporte Vectorial. Es preciso hacer notar que los modelos basados en memorias asociativas no han sido reportadas en el área del reconocimiento y clasificación de emociones en la literatura hasta el momento, no obstante que han demostrado ofrecer buenos resultados cuando se entrenan con datos similares a los de la representación de la energía en un arreglo bidimensional con datos binarios. [29]

Capítulo 3

Materiales y Métodos

3.1. Alfa-Beta con soporte vectorial

A continuación se presenta un ejemplo que ayuda a describir el modelo de las máquinas asociativas Alfa-Beta con soporte vectorial [29]. Dicho modelo consiste originalmente en el aprovechamiento de la información repetida entre los patrones y esta información da lugar al vector soporte.

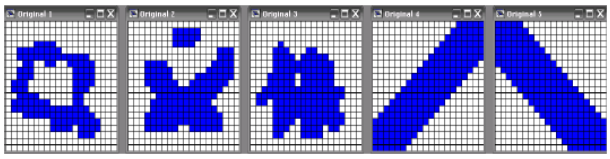


Figura 3.1: Conjunto fundamental. [29]

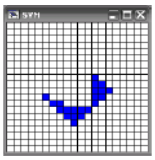


Figura 3.2: Patrón con la información repetida. [29]

Primeramente, en la Figura 3.1 se pueden observar los patrones del conjunto fundamental; posteriormente se obtiene el vector soporte, el cual consiste en encontrar la información que se repite en los patrones, como puede verse en la Figura 3.2.

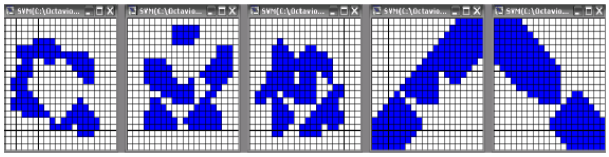


Figura 3.3: Conjunto fundamental con la información del vector soporte eliminada. [29]

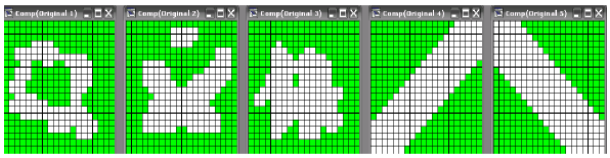


Figura 3.4: Conjunto fundamental negado. [29]

A partir de los patrones originales y del vector soporte, la información contenida en el vector soporte es eliminada de los patrones originales, dando lugar a los patrones de la Figura 3.3. El siguiente paso es negar los patrones del conjunto fundamental para realizar el mismo proceso con la información ausente, ver Figura 3.4.

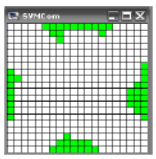


Figura 3.5: Vector soporte del conjunto fundamental negado. [29]

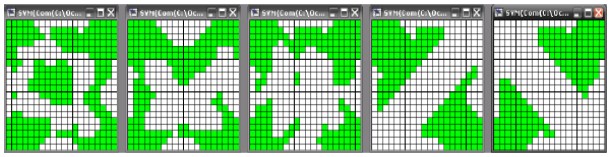


Figura 3.6: Conjunto fundamental negado sin la información del vector soporte. [29]

De la misma forma que se trabajó con los patrones originales, con los negados se obtiene el patrón con la información repetida (Figura 3.5), y dicha información es eliminada del conjunto fundamental negado, ver Figura 3.6.

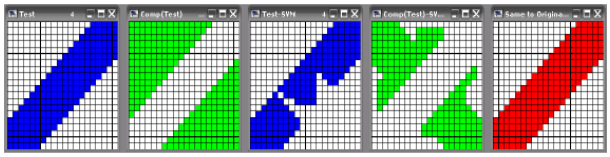


Figura 3.7: Recuperación de uno de los patrones del conjunto fundamental. [29]

Hasta aquí fue la fase de aprendizaje. En la Figura 3.7, se pueden observar los pasos del proceso de recuperación: del patrón original, se obtiene el patrón negado, posteriormente al patrón original se le elimina la información del vector soporte del conjunto fundamental y al patrón original negado se le elimina la información del vector soporte extraído de la información repetida del conjunto fundamental negado. Si estos 2 patrones se parecen de igual forma a sus contrapartes del conjunto fundamental y conjunto fundamental negado con información eliminada, entonces se va a elegir al patrón del conjunto fundamental con información eliminada que más se parezca; en caso contrario se elige al patrón que más parecido tenga.

El modelo de clasificación usado en esta tesis es la Memoria Asociativa Alfa-Beta SVM [29]. Los conceptos básicos concernientes a las memorias asociativas han sido reportados desde [18, 26, 27]; sin embargo, en esta tesis, lo referente a las Memorias Asociativas, se usa la notación y conceptos introducidos en [49]. Una Memoria Asociativa M relaciona patrones como: $x \rightarrow M \rightarrow y$ con x como patrón de entrada y y como patrón de salida. Por cada x se forma una asociación con una correspondiente y. La asociación correspondiente está dada por $\left(x^k, y^k\right)$, donde k es un entero positivo. La Memoria Asociativa M está representada por una matriz cuya $ij - e'sima$ componente es m_{ij}.

M es generada del conjunto fundamental, representada como: $\{(x^\mu, y^\mu) \mid \mu = 1, 2, ..., p\}$ es la cardinalidad del conjunto. Si $x^\mu = y^\mu, \forall \mu \in \{1, 2, .., p\}$, M es autoasociativa, de otro modo es heteroasociativa. La versión distorsionada del patrón x^k a ser recuperado, está denotado como $\tilde{x}^k$. Cuando se presenta una versión distorsionada de $x^{\tilde{\omega}}$ con $\tilde{\omega} = \{1, 2, ..., p\}$ a una Memoria Asociativa M, y sucede que la salida correspondiente exactamente a su patrón de salida $y^{\tilde{\omega}}$, se dice que la recuperación es correcta.

Las Memorias Asociativas Alfa-Beta operan en dos modos. El operador α es usado en la fase de aprendizaje y el operador β es útil para la fase de recuperación. α y β

son dos operadores binarios especialmente diseñados para Memorias Alfa-Beta. Sean los conjuntos $A = \{0, 1\}$ y $B = \{00, 01, 10\}$, entonces α y β están definidos por los operadores 3.1 y 3.2

$$\begin{array}{c} \alpha : A \times A \to B \\ \begin{array}{cc|c} x & y & \alpha(x,y) \\ \hline 0 & 0 & 01 \\ 0 & 1 & 00 \\ 1 & 0 & 10 \\ 1 & 1 & 01 \end{array} \end{array} \tag{3.1}$$

$$\begin{array}{c} \beta : B \times A \to A \\ \begin{array}{cc|c} x & y & \beta(x,y) \\ \hline 00 & 0 & 0 \\ 00 & 1 & 0 \\ 01 & 0 & 0 \\ 01 & 1 & 1 \\ 10 & 0 & 1 \\ 10 & 1 & 1 \end{array} \end{array} \tag{3.2}$$

Los conjuntos A y B, los operadores α y β, $\wedge$ (mínimo) y $\vee$ (máximo) forman el sistema algebraico $A, B, \alpha, \beta, \wedge, \vee$ que es la base matemática para las Memorias Asociativas Alfa-Beta.

Todos los conceptos básicos descritos anteriormente [49], son necesarios para describir el algoritmo principal de Alfa-Beta SVM [29, 30]. Se tiene un problema de reconocimiento de patrones, donde el conjunto fundamental se describe como $\{(x^\mu, y^\mu) \,|\mu = 1, 2, ..., p\}$, con $x^\mu \in A^n \forall \mu \in \{1, 2, ..., p\}$, y $n, p \in \mathrm{Z}^+$ y $A = \{0, 1\}$. El algoritmo de Alfa-Beta SVM tiene dos fases:

Fase de aprendizaje:

1. A partir del conjunto fundamental, se calcula el vector soporte S.

2. Para cada $\mu \in \{1, 2, ..., p\}$, obtener $x^\mu|_S$. A partir de los resultados se forma el conjunto fundamental restringido $\{(x^\mu|_S, x^\mu|_S) \,|\mu = 1, 2, ..., p\}$.

3. Para cada $\mu \in \{1, 2, ..., p\}$, obtener $\overline{x^\mu}$, el vector negado de x^μ. Con los p vectores negados, se forma el conjunto fundamental negado $\{(\overline{x^\mu}, \overline{x^\mu}) \,|\mu = 1, 2, ..., p\}$.

4. A partir del conjunto fundamental negado, se calcula e vector soporte $\widehat{S}$.

5. Para cada $\mu \in 1, 2, ..., p$, se obtiene $\overline{x^\mu}|_{\widehat{S}}$. A partir de estos resultados, se forma el conjunto fundamental negado restringido $\left\{\left(\overline{x^\mu}|_{\widehat{S}}, \overline{x^\mu}|_{\widehat{S}}\right) | \mu = 1, 2, ..., p\right\}$

Fase de Recuperación:

Siendo $\tilde{x} \in A^n$ cuyo patrón asociado x^μ es previamente desconocido, es el siguiente:

1. Obtener la restricción $\bar{x}|_S$.

2. Por cada $\mu \in \{1, 2, ..., p\}$, se obtiene $\tau\left(\bar{x}|_S, x^\mu|_S\right)$.

3. Por cada $\mu \in \{1, 2, ..., p\}$, se obtiene $\tau\left(x^\mu|_S, \bar{x}|_S\right)$.

4. Por cada $\mu \in \{1, 2, ..., p\}$, se obtiene $\theta\left(\bar{x}|_S, x^\mu|_S\right)$.

5. Encontrar $\psi \in \{1, 2, ..., p\}$ tal que $\theta\left(\bar{x}|_S, x^\psi|_S\right) = \bigwedge_{\mu=1}^{p} \theta\left(\bar{x}|_S, x^\mu|_S\right)$.

6. Obtener $\overline{\overline{x}}$, el vector negado de $\bar{x}$.

7. Obtener la Restricción $\overline{\overline{x}}|_{\widehat{S}}$.

8. Por cada $\mu \in \{1, 2, ..., p\}$, calcular $\tau\left(\overline{\overline{x}}|_{\widehat{S}}, \overline{x^\mu}|_{\widehat{S}}\right)$.

9. Por cada $\mu \in \{1, 2, ..., p\}$, calcular $\tau\left(\overline{x^\mu}|_{\widehat{S}}, \overline{\overline{x}}|_{\widehat{S}}\right)$.

10. Por cada $\mu \in \{1, 2, ..., p\}$, calcular $\theta\left(\overline{\overline{x}}|_{\widehat{S}}, \overline{x^\mu}|_{\widehat{S}}\right)$.

11. Encontrar $\varphi\{1, 2, ..., p\}$ tal que $\theta\left(\overline{\overline{x}}|_{\widehat{S}}, \overline{x^\varphi}|_{\widehat{S}}\right) = \bigwedge_{\mu=1}^{p} \theta\left(\overline{\overline{x}}|_{\widehat{S}}, \overline{x^\mu}|_{\widehat{S}}\right)$.

12. Si $\theta\left(\overline{x}|_S, x^\psi|_S\right) \leq \theta\left(\overline{\overline{x}}|_{\widehat{S}}, \overline{x^\varphi}|_{\widehat{S}}\right)$, se realiza la asignación $\omega = \psi$; de otro modo se realiza la asignación $\omega = \varphi$.

13. Se obtiene $\left(x^\omega|_S\right)|^S$.

Para un análisis más detallado del proceso de este modelo, ver el Apéndice A.

3.2. Base de datos

Al empezar a trabajar con reconocimiento de emociones, se tiene que trabajar con una base de datos, preferentemente orientada a la clasificación de estados emotivos. Hay varias bases de datos que fueron diseñadas para estos propósitos [43] y las emociones más comunes empleadas en estos corpus de voces y en orden de mayor a menor frecuencia se tiene:

-Enojado.

-Tristeza.

-Felicidad.

-Miedo.

-Disgustado.

-Alegría.

-Sorprendido.

-Aburrido, etc.

Aunque existen varias bases de datos orientadas al reconocimiento de emociones, en el presente proyecto se va a trabajar con la base de datos de Berlín [12] por su disponibilidad [2]. Esta base de datos cuenta con 7 emociones, 10 actores profesionales (5 hombres y 5 mujeres) que expresan 10 diferentes oraciones en idioma Alemán. Este corpus fue grabado mediante una frecuencia de muestreo de 16,000 Hz, con 16 bits de precisión en formato .wav.

Las oraciones que se usaron para la elaboración de la base de datos utilizada, son las siguientes:

1) Der Lappen liegt auf dem Eisschrank (The tablecloth is laying on the fridge) (El mantel está colocado sobre la nevera).

2) Das will sie am Mittwoch abgeben (She will hand it in on Wednesday) (Ella se encargará el miércoles).

3) Heute abend könnte ich es ihm sagen (Tonight I could tell him) (Esta noche podría decirle).

4) Das schwarze Stück Papier befindet sich da oben neben dem Holzstück (The black sheet of paper is located up there besides the piece of timber) (La hoja de papel negro se encuentra allá arriba, además de la pieza de madera).

5) In sieben Stunden wird es soweit sein (In seven hours it will be) (Ocurrirá en siete horas).

6) Was sind denn das für Tüten, die da unter dem Tisch stehen? (What about the bags standing there under the table?) (¿Qué pasa con las bolsas que están ahí debajo de la mesa?).

7) Sie haben es gerade hochgetragen und jetzt gehen sie wieder runter (They just carried it upstairs and now they are going down again) (Simplemente lo llevaron escaleras arriba y ahora lo devuelven abajo de nuevo).

8) An den Wochenenden bin ich jetzt immer nach Hause gefahren und habe Agnes besucht (Currently at the weekends I always went home and saw Agnes) (En la actualidad los fines de semana siempre fui a casa y veo a Agnes).

9) Ich will das eben wegbringen und dann mit Karl was trinken gehen (I will just discard this and then go for a drink with Karl) (Me limitaré a descartar este y luego ir a tomar una copa con Karl).

10) Die wird auf dem Platz sein, wo wir sie immer hinlegen (It will be in the place where we always store it) (Será en el lugar donde siempre lo guardamos).

La base de datos consta de 535 instancias, de las cuales 127 corresponden al estado de enojado, 81 a aburrido, 46 para disgustado, 69 para miedo, 71 a feliz, 62 a triste y 79 para neutral. Fue grabada con una frecuencia de muestreo de 16kHtz en formato .wav.

3.3. Software

Para la parte de extracción de parámetros que parten de la energía, amplitudes de los picos de la energía y los silencios, fueron calculados con Matlab; para la extracción de los formantes se usó el software de análisis fonético Praat [10], mientras que los demás parámetros fueron extraídos mediante el uso de los paquetes: Detección del pitch toolbox [4] y Auditory Toolbox versión 2 [1].

Praat [10] es un programa que permite hacer análisis espectral (espectogramas), del pitch, de las formantes, de la intensidad, jitter, shimmer.

Matlab es un lenguaje de computación de alto nivel, para desarrollo de algoritmos que permite hacer el análisis y visualización gráfica de los datos.

Para la clasificación se hace uso tanto de Matlab como del software para minería de datos WEKA [3]. WEKA soporta varias tareas de minería de datos, preprocesamiento de datos, clustering, clasificación, regresión, visualización y selección de rasgos. Las técnicas de WEKA se fundamentan en que los datos están disponibles en un archivo de texto plano (arff), en el que se encuentra un número finito de atributos (por lo general numéricos o nominales).

Capítulo 4

Modelo Propuesto

4.1. Parámetros

En el apéndice B, se hace una descripción de la extracción de los parámetros. La extracción de todos estos parámetros se obtuvieron por parte del tesista realizando programas propios en JAVA y se comparo el resultado con lo que entrega el software PRAAT y herramientas de MATLAB. A continuación se listan los parámetros con las aportaciones a la clasificación de manera individual (feature ranking):

1. Moda de los valores de energía (EnergySTMode): 40.18692
2. Mínimo del vector de mínimos de la matriz de valores MFCC's (MFCCMinMin): 39.62617
3. Moda del vector de mínimos de la matriz de valores MFCC's (MFCCMinMode): 39.62617
4. Mínimo del vector de modas de la matriz de valores MFCC's (MFCCModeMin): 38.8785
5. Moda del vector de modas de la matriz de valores MFCC's (MFCCModeMode): 38.8785
6. Máximo del vector de desviaciones estándar de la matriz de valores MFCC's (MFCCMaxSt): 38.69159

7. Promedio del vector de sonoridad (SMean): 37.38318
8. Desviación estándar del vector de sonoridad (SSt): 37.00935
9. Mínimo de los picos positivos del vector de energía (PeakAmplitudesMin): 35.70093
10. Moda de los picos positivos del vector de energía (PeakAmplitudesMode): 35.70093
11. Máximo de los valores de vector Sonoridad (SMax): 35.70093
12. Promedio del vector de los promedios de la matriz de valores MFCC's (MFCCMeanMean): 35.3271
13. Promedio del vector de máximos de la matriz de valores MFCC's (MFCCMaxMean): 34.20561
14. Promedio del vector de mínimos de la matriz de valores MFCC's (MFCCMinMean): 34.01869
15. Desviación estándar del vector de desviaciones estándar de la matriz de valores MFCC's (MFCCStSt): 34.01869
16. Promedio del vector de modas de la matriz de valores MFCC's (MFCCModeMean): 33.83178
17. Mediana de los valores de la energía (EnergySTMedian): 32.71028
18. Moda del vector de desviaciones estándar de la matriz de valores MFCC's (MFCCStMode): 31.96262
19. Mediana del vector de sonoridad (SMedian): 31.96262
20. Mínimo del vector de desviaciones estándar de la matriz de valores MFCC's (MFCCStMin): 31.96262
21. Máximo del vector de promedios de la matriz de valores MFCC's (MFCCMeanMax): 31.21495

22. Mediana del vector de promedios de la matriz de valores MFCC's (MFCC-MeanMedian): 30.84112

23. Promedio del vector de medianas de la matriz de valores MFCC's (MFCCMedianMean): 30.09346

24. Promedio de los valores del vector de la frecuencia fundamental o pitch (PitchVectorMean): 28.41121

25. Mediana del vector de las medianas de la matriz de valores MFCC's (MFCC-MedianMedian: 28.2243

26. Promedio de los valores del vector de energía (EnergySTMean): 28.03738

27. Mediana del vector de desviaciones estandar de la matriz de valores MFCC's (MFCCStMedian): 27.85047

28. Mediana de los valores del vector del pitch (PitchVectorMedian): 27.85047

29. Mediana de las duraciones continuas del pitch (PitchDurationsMedian): 27.85047

30. Promedio de las duraciones continuas del pitch (PitchDurationsMean): 27.47664

31. Desviación estándar de la segunda formante (F2st): 27.47664

32. Máximo del vector de medianas de la matriz de valores MFCC's (MFCCMedianMax): 27.1028

33. Desviación estándar de las duraciones continuas de silencio (SilenceDurationsSt): 26.91589

34. Duración máxima de presencia de pitch continuo (PitchDurationsMax): 26.16822

35. Mediana del vector de mínimos de la matriz de valores MFCC's (MFCCMin-Median): 26.16822

36. Mediana de los picos positivos del vector de energía (PeakAmplitudesMedian): 25.79439

37. Máximo del vector de máximos de la matriz de valores MFCC's (MFCCMaxMax): 25.42056

38. Mínimo del vector de medianas de la matriz de valores MFCC's (MFCCMedianMin): 25.23364

39. Moda del vector de medianas de la matriz de valores MFCC's (MFCCMedianMode): 25.23364

40. Mediana del vector de modas de la matriz de valores MFCC's (MFCCModeMedian): 25.23364

41. Moda del vector de los valores de pitch (PitchVectorMode): 24.6729

42. Mínimo del vector de los valores del pitch (PitchVectorMin): 24.6729

43. Mediana de las duraciones de los silencios (SilenceDurationsMedian): 24.29907

44. Promedio de los picos positivos del vector de energía (PeakAmplitudesMean): 24.29907

45. Máximo del vector de desviaciones estándar de la matriz de valores MFCC's (MFCCStMax): 24.11215

46. Moda de las duraciones de silencio continuo (SilenceDurationsMode): 24.11215

47. Desviación estándar de la cuarta formante (F4st): 23.92523

48. Mínimo de las duraciones continuas del pitch (PitchDurationsMin): 23.73832

49. Moda de las duraciones continuas del pitch (PitchDurationsMode): 23.5514

50. Desviación estándar del vector de modas de la matriz de valores MFCC's (MFCCModeSt): 23.36449

51. Promedio del vector de la tercera formante (F3mean): 23.36449

52. Desviación estándar de las duraciones continuas del pitch (PitchDurationsSt): 23.36449

53. Valor máximo del vector de las duraciones continuas del silencio (SilenceDurationsMax): 23.17757

54. Desviación estándar del vector de mínimos de la matriz de valores MFCC's (MFCCMinSt): 22.99065

55. Mínimo del vector de duraciones continuas del silencio (SilenceDurationsMin): 22.80374

56. Desviación estándar de la tercera formante (F3st): 22.61682

57. Máximo de los valores del vector de energía (EnergySTMax): 22.61682

58. Máximo del vector de los valores del pitch (PitchVectorMax): 22.61682

59. Mínimo del vector de máximos de la matriz de valores MFCC's (MFCCMaxMin): 22.24299

60. Máximo del vector de máximos de la matriz de valores MFCC's (MFCCMaxMode): 22.24299

61. Desviación estándar del vector de valores del pitch (PitchVectorSt): 22.05607

62. Desviación estándar de la primera formante (F1st): 21.49533

63. Mediana del vector de la cuarta formante (F4median): 21.49533

64. Promedio del vector de desviaciones estándar de la matriz de valores MFCC's (MFCCStMean): 21.30841

65. Promedio del vector de la cuarta formante (F4mean): 21.1215

66. Promedio de las duraciones de silencios continuos (SilenceDurationsMean): 20.93458

67. Máximo del vector de modas de la matriz de valores MFCC's (MFCCModeMax): 20.93458

68. Máximo del vector de mínimos de la matriz de valores MFCC's (MFCCMinMax): 20.93458

69. Moda del vector de valores de la tercera formante (F3mode): 20.74766
70. Mínimo del vector de valores de la tercera formante (F3min): 20.74766
71. Desviación estándar del vector de promedios de la matriz de valores MFCC's (MFCCMeanSt): 20.56075
72. Moda del vector de valores de la sonoridad (SMode): 20.37383
73. Desviación estándar del vector de medianas de la matriz de valores MFCC's (MFCCMedianSt): 20.37383
74. Mínimo de los valores del vector de sonoridad (SMin): 20.37383
75. Mediana del vector de máximos de la matriz de valores MFCC's (MFCCMax-Median): 20.18692
76. Mínimo del vector de promedios de la matriz de valores MFCC's (MFCCMean-Min): 20.18692
77. Moda del vector de promedios de la matriz de valores MFCC's (MFCCMean-Mode): 20.18692
78. Promedio del vector de valores de la primera formante (F1median): 20
79. Máximo del vector de valores de la segunda formante (F2max): 19.62617
80. Promedio del vector de valores de la primera formante (F1mean): 19.25234
81. Máximo del vector de valores de la primera formante (F1max): 19.25234
82. Mediana del vector de valores de la segunda formante (F2median): 18.8785
83. Mediana del vector de valores de la tercera formante (F3median): 18.8785
84. Desviación Estándar del vector de energía (EnergySTSt): 18.69159
85. Mínimo de los valores del vector de la primera formante (F1min): 18.69159
86. Moda de los valores del vector de la primera formante (F1mode): 18.69159

87. Máximo de los valores del vector de la tercera formante (F3max): 18.69159
88. Máximo de los picos positivos del vector de energía (PeakAmplitudesMax): 18.31776
89. Máximo del vector de los valores de la cuarta formante (F4max): 18.31776
90. Mínimo del vector de valores de la cuarta formante (F4min): 18.13084
91. Moda del vector de valores de la cuarta formante (F4mode): 18.13084
92. Desviación estándar de los picos positivos del vector de energía (PeakAmplitudesSt): 17.94393
93. Promedio del vector de valores de la segunda formante (F2mean): 17.19626
94. Mínimo del vector de valores de la segunda formante (F2min): 15.3271
95. Moda del vector de valores de la segunda formante (F2mode): 15.3271

Posteriormente al proceso de jerarquizar los parámetros, se realizaron pruebas con diversas estrategias de selección de parámetros usando el software WEKA [3]; en esta parte es importante señalar que la selección de rasgos óptima no es viable en el sentido del costo computacional, una cantidad considerable de pruebas de selección de atributos fueron realizadas para mejorar el desempeño de clasificación.

Usando el clasificador *SimpleLogistic*, se evaluaron los conjuntos de parámetros de prueba con un método wrapper (una validación cruzada con 5 divisiones). La estrategia elegida para la búsqueda del subconjunto de rasgos fue la de búsqueda hacia adelante, dicha búsqueda consiste en hacer un feature ranking en un inicio, se elige el rasgo que mayor aporte de a la clasificación, posteriormente se analizan todas las posibles combinaciones para elegir el segundo rasgo que combinado con el primero aporte más, después se hace lo mismo para el tercer rasgo que junto con los dos primeros de mejor desempeño, este proceso continúa hasta que la combinación del conjunto de rasgos con el siguiente rasgo a buscar empeore la clasificación, una vez sucedido esto, el proceso se detiene, y el conjunto de rasgos ofrecen una solución sub-óptima.

De esta forma, la siguiente lista es el conjunto de parámetros que hasta ahora más ha aportado al índice de clasificación:

1. Promedio de los valores del vector de energía (EnergySTMean)
2. Máximo valor de los picos positivos del vector de energía (PeakAmplitudesMax)
3. Mínimo valor de los picos positivos del vector de energía (PeakAmplitudesMin)
4. Promedio de los valores de los picos positivos del vector de energía (PeakAmplitudesMean)
5. Desviación estándar de los picos positivos del vector de energía (PeakAmplitudesSt)
6. Moda de las duraciones continuas del pitch (PitchDurationsMode)
7. Promedio del vector de sonoridad (SMean)
8. Máximo del vector de promedios de la matriz de valores MFCC's (MFCCMeanMax)
9. Mínimo del vector de mínimos de la matriz de valores MFCC's (MFCCMinMin)
10. Mínimo del vector de promedios de la matriz de valores MFCC's (MFCCMeanMin)
11. Mínimo del vector de desviaciones estándar de la matriz de valores MFCC's (MFCCStMin)
12. Promedio del vector de desviaciones estándar de la matriz de valores MFCC's (MFCCStMean)
13. Mediana del vector de promedios de la matriz de valores MFCC's (MFCCMeanMedian)
14. Desviación estándar de la segunda formante (F2st)

4.2. Modelo

Como se está haciendo uso de las máquinas Alfa-Beta con soporte vectorial y éstas presentan un buen desempeño con el reconocimiento de imágenes binarias [29], esto da lugar a otro enfoque de experimentación, el cual constituye uno de los aportes principales de este trabajo de tesis: se trata de hacer reconocimiento de representaciones bidimensionales que representen la señal de voz.

Figura 4.1: Señal de energía extraída usando Praat. [10]

Las representaciones de la energía (ver Figura 4.1) tienen una dimensión de 178 pixeles de ancho, por 107 de alto. Se seleccionaron experimentalmente esos valores, con el fin de mantener un costo computacional que minimize el tiempo de ejecución sin que impacte esto el desempeño del algoritmo; es decir, que permita hacer un número considerable de pruebas para identificar el trato más satisfactorio de la señal en términos de la clasificación de las emociones.

Figura 4.2: Señal de energía con relleno.

Figura 4.3: Señal de energía

Las representaciones bidimensionales de la intensidad de voz, se realiza un alineamiento en tiempo, la escala en amplitud tiene un valor máximo equivalente a 100 decibeles [10]. Al experimentar directamente con estas imágenes (ver Figura 4.1) no dio resultados satisfactorios, por lo que la primera estrategia que se tomó fue rellenar abajo (Figura 4.2) o arriba (Figura 4.3) de la señal de energía.

Figura 4.4: Señal de energía normalizada en el eje de la amplitud

Posteriormente, se normalizó en el eje de la amplitud (ver Figura 4.4) para homogeneizar la base de datos, esto incrementó el desempeño de la clasificación de emociones. A continuación se explica de forma más detallada el proceso.

En la Figura 4.5 se muestra el esquema de como es el proceso para obtener la representación bidimensional de la energía, se comienza con un preprocesamiento de la señal, se divide la señal de voz en 178 ventanas para posteriormente calcular la energía de cada una de ellas, se obtiene el contorno de la intensidad de la señal de voz con una cota máxima de 100 decibeles.

La representación de la envolvente de energía es una matriz cuyos valores son unos donde se encuentra el valor de la energía y ceros en los demás elementos, de esta representación, los ceros que se encuentran por debajo de cada 1, cambian su valor a 1, quedando así una representación con valores 1 debajo de toda la envolvente

de energía.

Se normalizan todas las columnas de unos que representan la señal de energía con respecto a la columna con más unos, es decir, se normaliza con respecto a la amplitud. Posteriormente se representa esta matriz en un arreglo unidimensional, en el que se respeta el eje del tiempo implícito en la matriz, es decir se concatena columna por columna.

Cada archivo se trata de la misma forma, para poder formar el conjunto fundamental, cabe señalar que en el arreglo unidimensional se respeta el tiempo dentro del proceso de clasificación de las máquinas Alfa-Beta SVM.

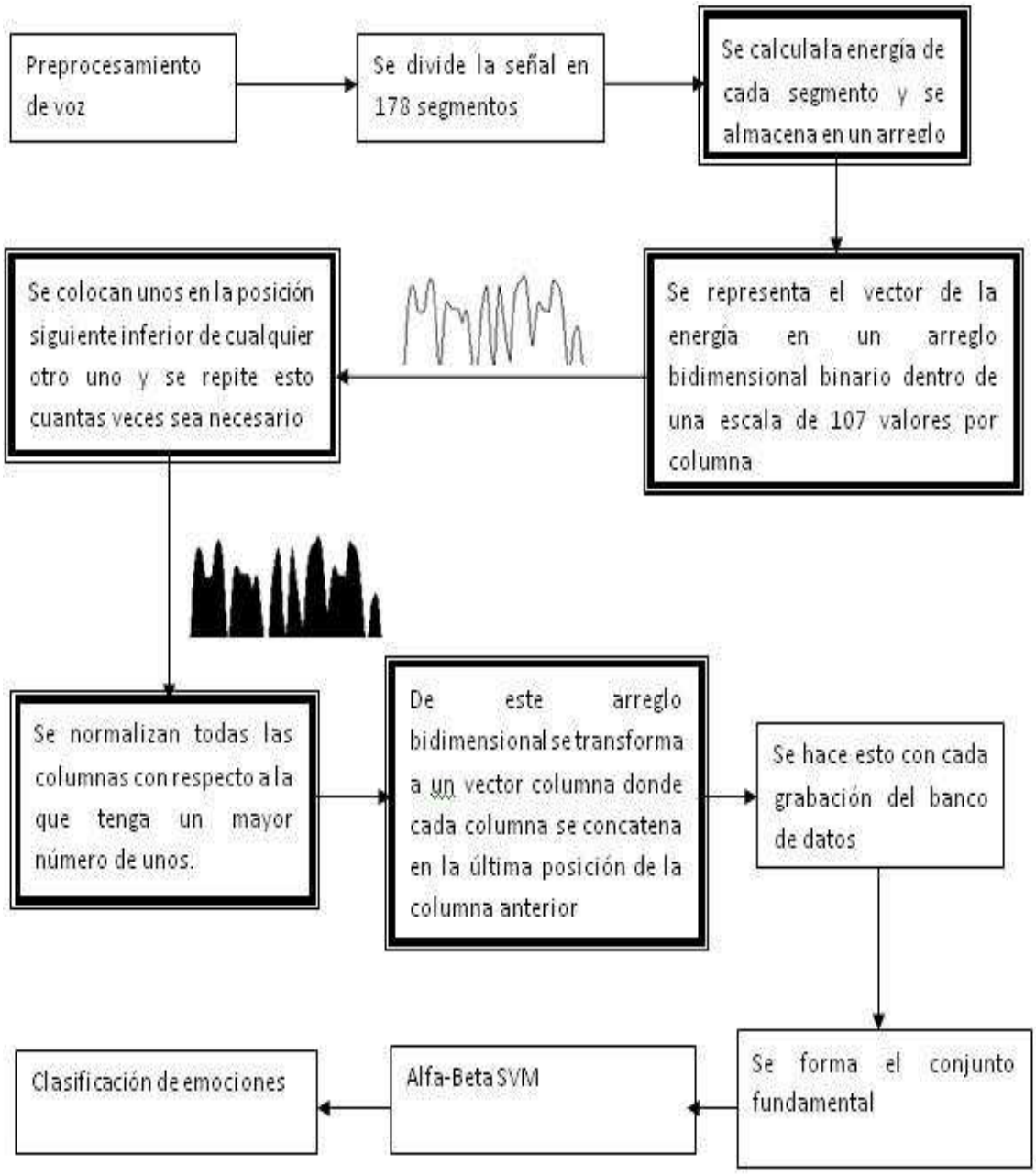


Figura 4.5: Diagrama para representar a la energía como un arreglo bidimensional

Capítulo 5

Resultados

En esta sección se realiza un reporte de los resultados que se han obtenido mediante la clasificación de emociones de la base de datos de Berlín [12].

5.1. Clasificación de emociones

En esta sección se somete a prueba la base de datos con los 14 parámetros. El primer modelo a prueba es el Naive Bayes, se entrena con toda la base de datos y se prueba con toda la base de datos, los resultados son los siguientes:

En la Tabla 5.1 se observa la matriz de confusión obtenida al clasificar toda la base de datos con el modelo Naive Bayes, fueron 327 (61.1215 %) las instancias correctamente reconocidas, y 208 (38.8785 %) las incorrectas.

Tabla 5.1: Matriz de confusión usando el modelo Naive Bayes.

a	b	c	d	e	f	g		Clasificado
118	0	1	0	8	0	0	a	enojado
1	54	6	2	5	2	11	b	aburrido
8	8	21	2	4	1	2	c	disgustado
10	7	7	30	8	4	3	d	miedo
39	1	6	1	20	0	4	e	feliz
0	3	3	0	0	47	9	f	triste
2	23	9	2	2	4	37	g	neutral

Tabla 5.2: Matriz de confusión usando el modelo SimpleLogistic.

a	b	c	d	e	f	g		Clasificado
115	1	0	4	7	0	0	a	enojado
0	70	1	3	0	0	7	b	aburrido
1	3	33	3	2	1	3	c	disgustado
6	1	1	54	1	2	4	d	miedo
25	0	1	5	38	0	2	e	feliz
0	1	0	0	0	60	1	f	triste
1	14	2	2	0	3	57	g	neutral

Tabla 5.3: Matriz de confusión usando Perceptrón Multi-capa.

a	b	c	d	e	f	g		Clasificado
115	0	1	1	9	0	1	a	enojado
0	75	1	0	0	1	4	b	aburrido
0	4	34	3	1	2	2	c	disgustado
3	1	3	52	5	1	4	d	miedo
9	0	3	2	55	0	2	e	feliz
0	0	0	0	0	62	0	f	triste
1	3	3	0	0	2	70	g	neutral

La Tabla 5.2 puede observarse la matriz de confusión de la clasificación de toda la base de datos como prueba y entrenamiento, usando el modelo SimpleLogistic, 427 (79.8131 %) instancias fueron correctamente reconocidas, 108 (20.1869 %) no se reconocieron.

En la Tabla 5.3 se observa la matriz de confusión resultante al usar toda la base de datos como prueba, usando Perceptrón Multi-capa, 463 (86.5421 %) instancias se clasificaron correctamente y 72 (12.3479 %) fueron clasificadas incorrectamente.

Con nuestro modelo se llevaron a cabo pruebas con un 90 % de la base de datos como datos para el entrenamiento, las máquinas Alfa-Beta con soporte vectorial dieron una clasificación de 508 (94.9532 %) instancias correctamente clasificadas y 27 (5.0468 %) instancias incorrectamente clasificadas.

Por otro lado, al usar las imágenes de la energía con relleno (Figura 4.2), como

parámetros, las máquinas Alfa-Beta con soporte vectorial generaron un resultado de 506 (94.5 %) instancias correctamente clasificadas y 29 (5.5 %) instancias incorrectamente clasificadas.

5.2. Clasificación reportada en la literatura

A continuación se muestran los resultados que han sido reportados en la literatura, usando la misma base de datos (de Berlín).

Tabla 5.4: Resultados de clasificación para ambas bases de datos. [13]

Base de datos	Mejor resultado	
	Dependiente del locutor	Independiente del locutor
Polaco	76.30 %	64.18 %
Alemán	74.39 %	72.04 %

Tabla 5.5: Porcentajes de clasificación de emociones para cada caso. [48]

	Hombre	Mujer	Promedio	Mezclado	Mezclado con clasificador del género
Global	81.56	76.76	78.86	75.12	76.95
2 fases	84.17	79.88	81.75	76.22	-
3 fases	-	-	-	-	79.47

En la Tabla 5.4 se observan los resultados alcanzados en la base de datos de Berlín, estos resultados son a partir de la clasificación de 6 emociones (sin tomar en cuenta la emoción de *disgust*), el modelo usado para esta tarea es el de árboles de decisión binarios. [13]

La Tabla 5.5 muestra que el mejor resultado que se obtuvo de la base de datos de Berlín fue del 79.47 %, la validación de estos resultados se hicieron promediando 10 pruebas usando el 50 % de la base de datos para entrenamiento y 50 % para pruebas en orden aleatorio. No se tomó en cuenta la emoción de disgustado, es decir, la clasificación fue para 6 emociones. [48]

En la Tabla 5.6 se muestra una matriz de confusión con los resultados de la clasificación de las 7 emociones de la base de datos de Berlín, la clasificación se llevó a cabo con mixturas Gaussianas, se alcanzó un 50.6 % de precisión con validación cruzada. [42]

Tabla 5.6: Resultados de la clasificación de la base de datos de Berlín. [48]

	Clasificadas como						
	Enojado	Aburrido	Disgustado	Miedo	Feliz	Neutral	Tristeza
Enojado	81	4	10	9	17	6	0
Aburrido	1	37	3	5	1	29	5
Disgustado	9	1	18	9	5	5	6
Miedo	3	12	7	18	14	10	5
Feliz	16	3	8	7	33	4	0
Neutral	1	22	10	8	0	38	0
Tristeza	1	5	2	3	0	5	46

El resultado mas alto encontrado en la literatura se encuentra en el trabajo [51], donde se realizó la clasificación de las 7 emociones con un desempeño del 82.44 %, usando el modelo k-vecinos próximos considerando el costo del error.

Capítulo 6

Conclusiones y Trabajo Futuro

6.1. Conclusiones

En la Tabla 6.1 se puede observar que la clasificación en el estado del arte gira alrededor de aproximadamente del 80 %, así como el modelo asociativo Alfa-Beta SVM es el que mejor se desempeña en la clasificación de emociones de la base de datos de Berlín. Las Memorias Asociativas Alfa-Beta SVM entrenadas con el conjunto fundamental basado en la representación bidimensional de la energía, demuestra experimentalmente que la energía es uno de los parámetros con mayor contenido emotivo de lo que se está hablando.

La selección de rasgos en la que se obtuvo 14 parámetros ha demostrado ser buena, siendo el modelo Alfa-Beta SVM el que mejor desempeño presenta, los modelos asociativos hasta este momento no se habían usado para la clasificación de emociones.

Para el reconocimiento de emociones, el modelo que mejor se desempeña son las máquinas Alfa-Beta con soporte vectorial. Con una eficiencia superior al 90 %.

El parámetro que más información emotiva contiene es la energía, el cual al representarlo en un arreglo bidimensional, se obtiene mayor caracterización de las emociones en la señal de voz que representar la señal de energía con medidas de dispersión como: el valor promedio, máximo, desviación estándar, mediana y moda.

Tabla 6.1: Resultados reportados en la literatura y alcanzados en esta tesis.

Trabajo	Emociones	Desempeño (%)
[13]	6	74.39/72.04
[48]	6	79.47
[42]	7	50.6
[51]	7	82.44
Naive Bayes	7	61.12
SimpleLogistic	7	79.81
Perceptrón Multicapa	7	86.54
Alfa-Beta SVM (14 parámetros)	**7**	**94.95**
Alfa-Beta SVM (Imágenes Energía)	**7**	**94.5**

6.2. Trabajo Futuro

Buscar otros parámetros que caractericen la información afectiva de la voz de las emociones.

Crear un corpus de voz orientada al reconocimiento de emociones, con emociones reales o actuadas.

Desarrollar un modelo para reconocer emociones reales.

Trabajar el análisis de la señal de voz, junto con otros tipos de información, como video o seguimiento del movimiento para clasificar emociones.

Desarrollar una base de datos orientada al reconocimiento de emociones, conteniendo en ella grabaciones de voz y otro tipo de datos como las señales biométricas.

Probar el nuevo modelo con otras bases de emociones como SUSAS y otras más que se puedan adquirir.

6.3. Trabajos publicados y presentados derivados de esta tesis

Publicaciones y presentaciones:

"Sadness Detection in Emotional Acted Speech". Presentado en el WorkShop de MICAI2009, en Guanajuato, Guanajuato.

"Reconocimiento automático de voz emotiva con memorias asociativas Alfa-Beta SVM". Aceptado en la revista POLIBITS, ISSN 1870-9044.

Referencias

[1] *Auditory Toolbox.* URL `http://cobweb.ecn.purdue.edu/~malcolm/interval/1998-010/`.

[2] *Berlin emotional speech database.* URL `http://www.expressive-speech.net/`.

[3] *Data Mining: Practical machine learning tools and techniques.* URL `http://www.cs.waikato.ac.nz/ml/weka/`.

[4] *Detección del pitch toolbox.* URL `http://physionet.cps.unizar.es/~eduardo/docencia/tvoz/Demos/pitchlpc/detpitch.html`.

[5] *Documentation of the Danish emotional speech database des*, 1996. URL `/http://cpk.auc.dk/tb/speech/Emotions/S`.

[6] *University of Pennsylvania Linguistic Data Consortium, Emotional prosody speech and transcripts*, 2002. URL `/http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2002S28S`.

[7] A. Álvarez, I. Cearreta, J. López, A. Arruti, E. Lazkano, Sierra B., y N. Garay. Application of feature subset selection based on evolutionary algorithms for automatic emotion recognition in speech. *Proceedings of NOn LInear Speech Processing*, 2007.

[8] N. Amir, S. Ron, y N. Laor. Analysis of an emotional speech corpus in hebrew based on objective criteria. *Speech Emotion-2000*, págs. 29–33, 2000.

[9] T. Athanaselis, S. Bakamidis, I. Dologlou, R. Cowie, E. Douglas-Cowie, y C. Cox. Asr for emotional speech: clarifying the issues and enhancing performance. *Elsevier Science Ltd. Oxford*, págs. 437–444, 2005.

[10] P. Boersma y D. Weenink. *Praat: doing phonetics by computer Version 5.1.17*, 2009. URL `http://www.praat.org/`.

[11] C. Breazeal y L. Aryananda. Recognition of affective communicative intent in robot-directed speech. *Autonomous Robots*, págs. 83–104, 2002.

[12] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, y B. Weiss. A database of german emotional speech. *Proceedings of Interspeech*, 2005.

[13] J. Cichosz y K. Slot. Emotion recognition in speech signal using emotion extracting binary decision trees. *Proceedings of Affective Computeing and Intelligent Interation*, 2007.

[14] F. Dellaert, Th. Polzin, y A. Waibel. Recognizing emotion in speech. *Proceedings of the ICSLP '96*, 1996.

[15] P. Ekman. An argument for basic emotions. *Cognition and Emotion*, págs. 169–200, 1992.

[16] M. El Ayadi, M. Kamel, y F. Karray. Survey on speech emotion recognition:features, classification schemes, and databases. *Pattern Recognition*, 44:572–587, 2011.

[17] L. Fu, X. Mao, y L. Chen. Speaker independent emotion recognition based on svm/hmms fusion system. *International Conference on Audio, Language and Image Processing, ICALIP2008*, págs. 61–65, 2008.

[18] M. H. Hassoun. Associative neural memories. *Oxford University Press, New York*, 1993.

[19] V. Hozjan y Z. Kačič. Context-independent multilingual emotion recognition from speech signals. *International Journal of Speech Technology*, 6(3):311–320, 2003.

[20] V. Hozjan, Z. Moreno, A. Bonafonte, y A. Nogueiras. Interface databases: design and collection of a multilingual emotional speech database. *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC'02)*, págs. 2019–2023, 2002.

[21] H. Hu, M. Xu, y W. Wu. Dimensions of emotional meaning in speech. *Proceedings of the ISCAITRW on Speech and Emotion*, págs. 25–28, 2000.

[22] H. Hu, M. Xu, y W. Wu. Gmm supervector based svm with spectral features for speech emotion recognition. *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP2007*, 4:IV 413 – IV 416, 2007.

[23] S. Jovičić, Z. Kašić, M. Dordević, y M. Rajković. Serbian emotional speech database: design, processing and evaluation. *Speech and Computer conference*, 2004.

[24] E. Kim, K. Hyun, S. Kim, y Y. Kwak. Speech emotion recognition using eigen-fft in clean and noisy environments. *16th IEEE International Symposium on Robot and Human Interactive Communication. RO-MAN2007*, págs. 689–694, 2007.

[25] R. Kohavi. A study of cross-validation and bootstrap for accuracy estimation and model selection. *International Joint Conference on Artificial Intelligence (IJCAI)*, 1995.

[26] T. Kohonen. Correlation matrix memories. *IEEE Transactions on Computers*, págs. 353–359, 1972.

[27] T. Kohonen. Self-organization and associative memory. *Springer-Verlag, Berlin Heidelberg New York*, 1989.

[28] Y. Li y Y. Zhao. Recognizing emotions in speech using short-term and long-term features. *Proceedings of the ICSLP*, págs. 2255–2258, 1998.

[29] L. López, C. Yáñez, y O. Camacho. *Máquinas asociativas Alfa-Beta con soporte vectorial*. Tesis Doctoral, Instituto Politécnico Nacional, 2008.

[30] L. López-Leyva, C. Yáñez Márquez, y I. López-Yáñez. A new efficient model of support vector machines: Alfa-beta svm. *23rd ISPE International Conference on CAD/CAM*, 2007.

[31] I. Luengo, E. Navas, I. Hernáez, y J. Sánchez. Reconocimiento automático de emociones utilizando parámetros prosódicos. *Natural Language Processing*, 2005.

[32] D. Morrison, R. Wang, y L. DeSilva. Ensemble methods for spoken emotion recognition incall-centres. *Speech Communication*, págs. 98–112, 2007.

[33] T. Nwe, S. Foo, y L. DeSilva. Speech emotion recognition using hidden markov models. *Speech Communication*, págs. 603–623, 2003.

[34] T. Pao, Y. Chen, J. Yeh, y W. Liao. Detecting emotions in mandarin speech. *Computational Linguistics and Chinese Language Processing*, 10(3):347–362, 2005.

[35] T. Polzin y A. Waibel. Detecting emotions in speech. *Proceedings of the CMC*, 1998.

[36] J. Rong, Y. Chen, M. Chowdhury, y L. Gang. Acoustic features extraction for emotion recognition. *Proc. 6th Int. Conf. Computer and Information Science*, págs. 419–424, 2007.

[37] B. Schuller. Towards intuitive speech interaction by the integration of emotional aspects. *IEEE International Conference on Systems, Man and Cybernetics*, 6, 2002.

[38] B. Schuller, S. Reiter, R. Muller, M. Al-Hames, M. Lang, y G. Rigoll. Speaker independent speech emotion recognition by ensemble classification. *IEEE International Conference on Multimedia and Expo., ICME2005*, págs. 864–867, 2005.

[39] M. Slaney y G. McRoberts. Babyears: a recognition system for affective vocalizations. *Speech Communumication*, págs. 367–384, 2003.

[40] Ch. Sobin y A. Murray. Emotion in speech: The acoustic attributes of fear, anger, sadness and joy. *Journal of Psycholinguistic Research*, 28(4), 1999.

[41] R. Tato, R. Santos, R. Kompe, y J. Pardo. Emotional space improves emotion recognition. 7^{th} *International Conference on Spoken Language Processing*, 2002.

[42] K. Truong y D. Leeuwen. An ′open-set′ detection evaluation methodology for automatic emotion recognition in speech. *ParaLing'07, Workshop on Paralinguistic Speech between models and data*, 2007.

[43] D. Ververidids y C. Kotropoulos. A state of the art review on emotional speech databases. *Proceedings of 1st Richmedia Conference*, págs. 109–119, 2003.

[44] D. Ververidis, C. Kotropoulos, y I. Pitas. Automatic emotional speech classification. *Proceedings of International Conference on Acoustics, Speech, and Signal Processing ICASSP*, págs. 593–596, 2004.

[45] T. Vogt y E. Andre. Comparing feature sets for acted and spontaneous speech in view of automatic emotion recognition. *Multimedia and Expo, ICME 2005*, págs. 474–477, 2005.

[46] T. Vogt y E. André. Improving automatic emotion recognition from speech via gender differentiation. *Proceedings of Language Resources and Evaluation Conference*, 2006.

[47] T. Vogt, E. André, y J. Wagner. Automatic recognition of emotions from speech: A review of the literature and recommendations for practical realisation. *Affect and Emotion in Human-Computer Interaction*, 2008.

[48] Z. Xiao, E. Dellandrea, W. Dou, y L. Chen. Hierarchical classification of emotional speech. *IEEE Transactions on Multimedia*, 2007.

[49] C. Yáñez Márquez. *Memorias Asociativas basadas en Relaciones de Orden y Operadores Binarios*. Tesis Doctoral, Centro de Investigación en Computación en el Instituto Politécnico Nacional, México, 2002.

[50] F. Yu, Y. Chang, E.and Xu, y H. Shum. Emotion detection from speech to enrich multimedia content. *Proceedings of the Second IEEE Pacific Rim Conference on Multimedia: Advances in Multimedia Information Processing*, 2195:2255–2258, 2001.

[51] S. Zhang, L. Li, y Z. Zhao. Spoken emotion recognition using kernel discriminant locally linear embedding. *Electronics Letters*, 46:1344–1346, 2010.

[52] J. Zhou, G. Wang, Y. Yang, y P. Chen. Speech emotion recognition based on rough set and svm. *5th IEEE International Conference on Cognitive Informatics, ICCI2006*, 2006.

Apéndice A

Diagrama de flujo de las máquinas Alfa-Beta con soporte vectorial

Figura A.1, se representa el diagrama de flujo de la fase de aprendizaje de las máquinas Alfa-Beta con soporte vectorial.

Figura A.2, representación de la primer parte del diagrama de flujo de la fase de recuperación.

Figura A.3, representación de la segunda parte del diagrama de flujo de la fase de recuperación.

Figura A.4, representación de la tercer parte del diagrama de flujo de la fase de recuperación.

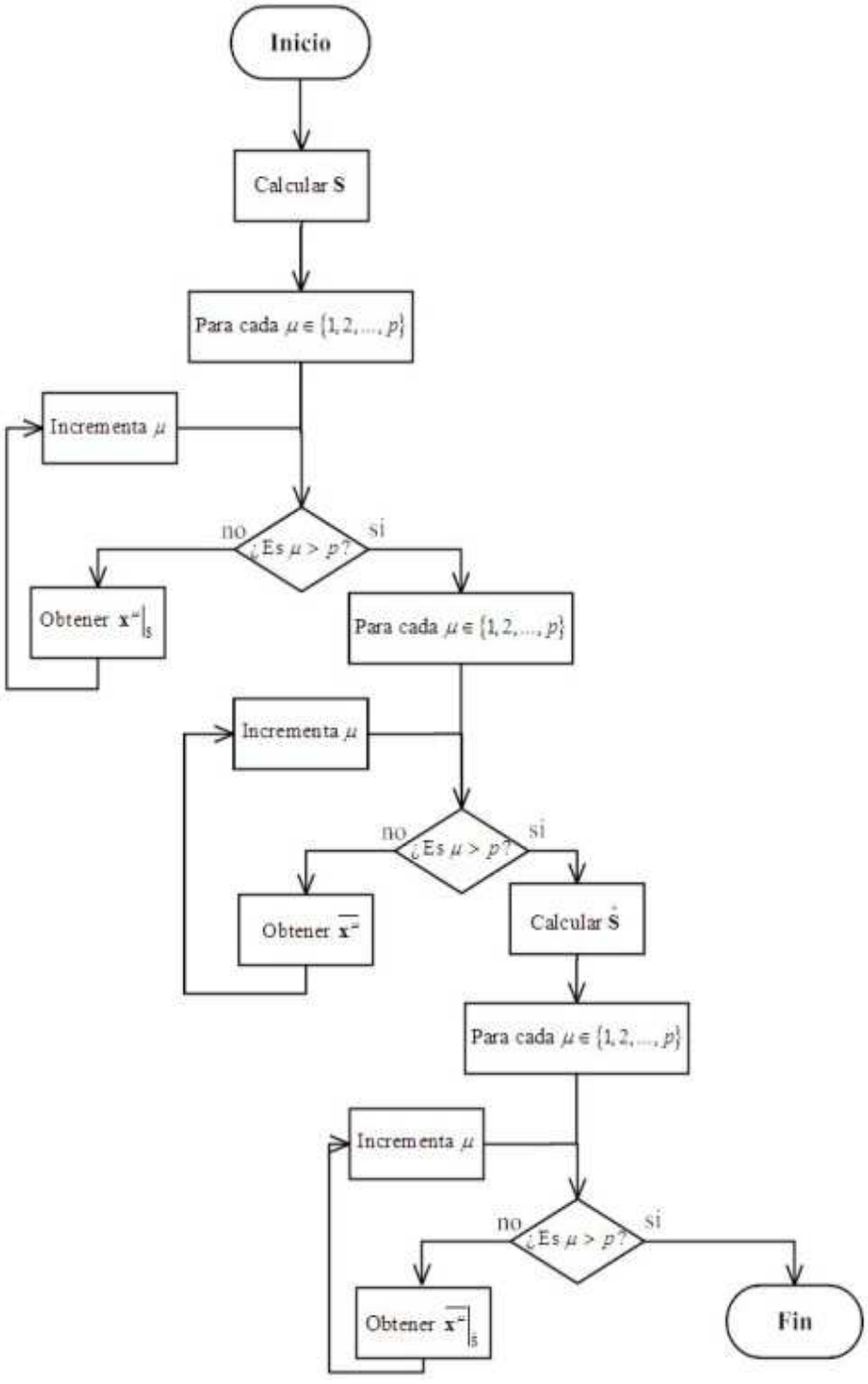


Figura A.1: Fase de aprendizaje de las máquinas Alfa-Beta con soporte vectorial. [29]

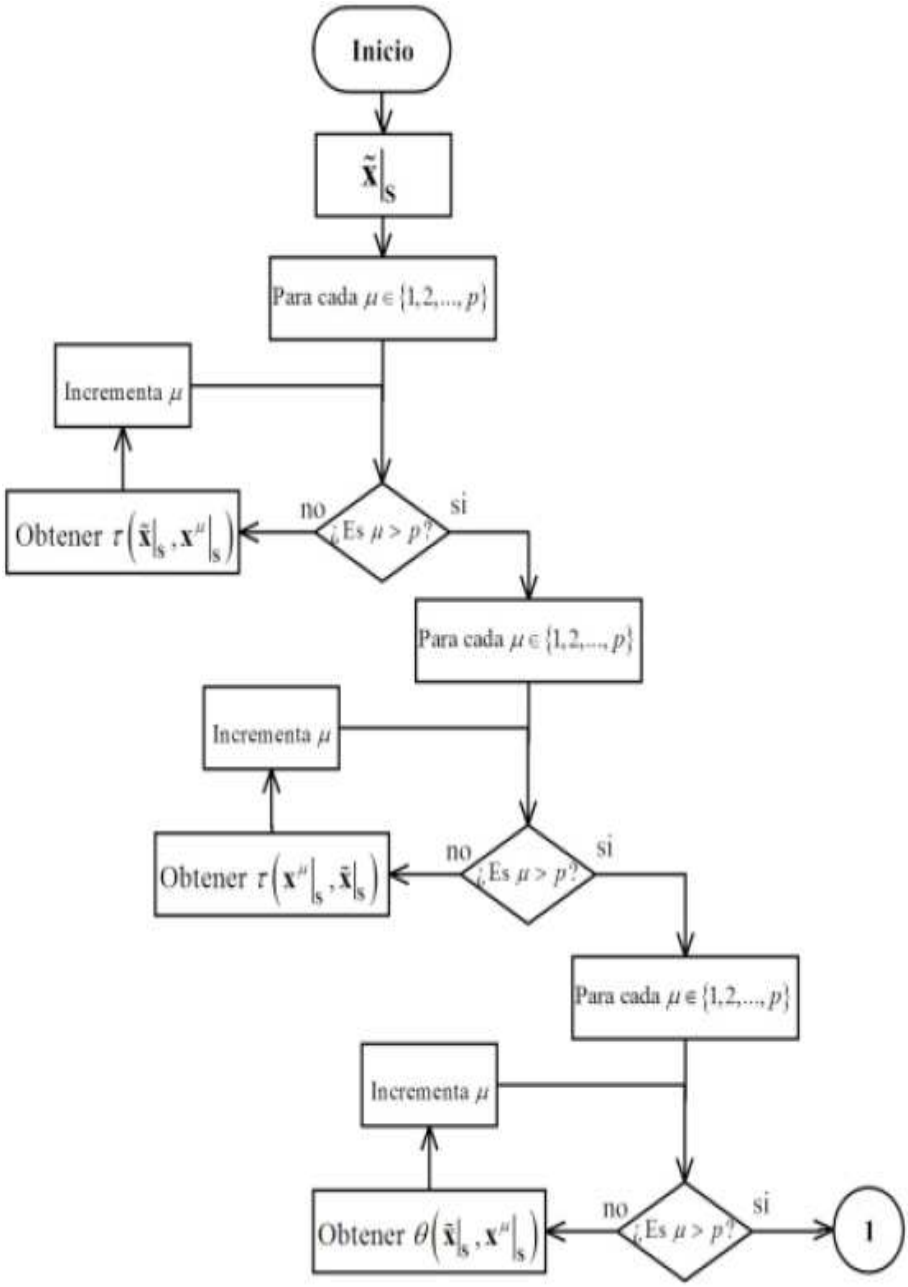


Figura A.2: Fase de recuperación de las máquinas Alfa-Beta con soporte vectorial, parte 1. [29]

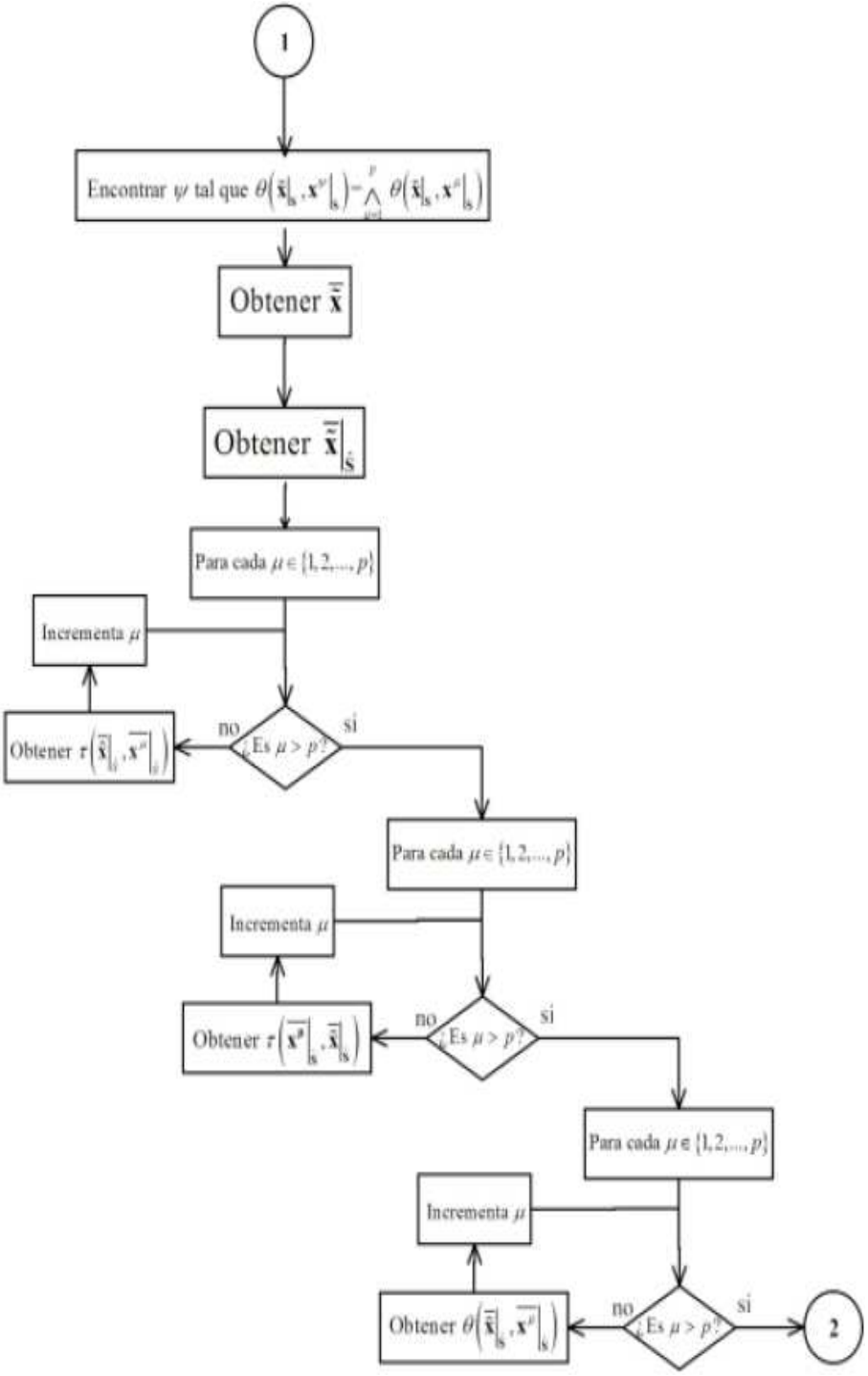


Figura A.3: Fase de recuperación de las máquinas Alfa-Beta con soporte vectorial, parte 2. [29]

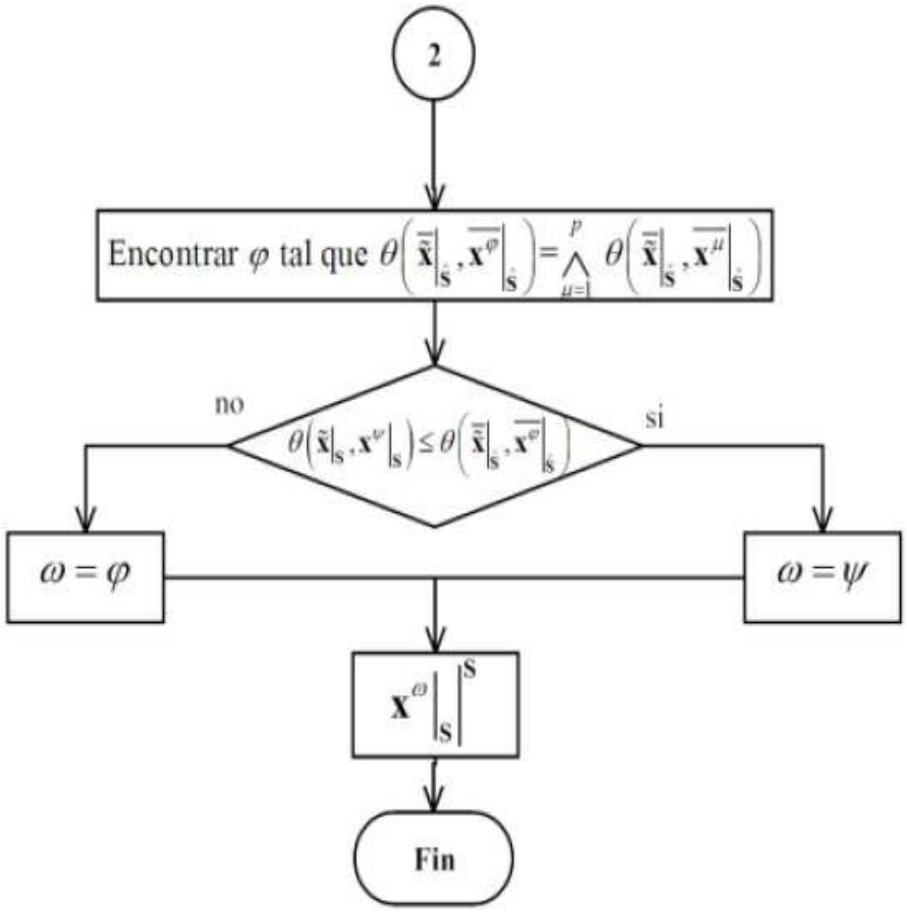


Figura A.4: Fase de recuperación de las máquinas Alfa-Beta con soporte vectorial, parte 3. [29]

Apéndice B

Parámetros

Para la extracción de parámetros, se utilizó una ventana de 301 muestras, que corresponden a 18 msec aproximadamente.

EnergyST - Es el vector que contiene la energía por segmento, de éste vector se extrae el valor máximo, el promedio, mediana, moda y desviación estándar.

PeakAmplitudes - Es un vector que contiene las amplitudes de los picos del vector EnergyST, del vector PeakAmplitudes se obtiene el máximo, mínimo, promedio, mediana, moda y desviación estándar.

SilenceDurations - En este arreglo están almacenados el número de ventanas que dura cada silencio a lo largo de la grabación, del arreglo se extraen los valores máximo, mínimo, promedio, mediana, moda y desviación estándar.

PitchDurations - Este vector contiene el número de ventanas que dura cada segmento de pitch, se calculan el máximo, mínimo, promedio, mediana, moda y desviación estándar.

PitchVector - Es el arreglo que almacena los valores del pitch, se obtienen los valores máximo, mínimo, promedio, mediana, moda y desviación estándar.

S - Es la sonoridad o la evolución de la frecuencia del pitch, se calcula el máximo, mínimo, promedio, mediana, moda y desviación estándar.

Nota: Los parámetros del pitch y la sonoridad fueron obtenidos por medio del toolbox "detección del pitch". [4]

MFCC - Esta es una matriz que contiene los valores de 13 coeficientes a lo largo de la señal, por lo que las medidas estadísticas que dieron mejor resultado experimen-

talmente, fueron las manejadas de la siguiente forma, primero se extraen 6 vectores cada uno representando una medida distinta (máximo, mínimo, promedio, mediana, moda y desviación estándar), y posteriormente a cada vector se le extraen los valores estadísticos máximo, mínimo, promedio, mediana, moda y desviación estándar.

Nota: Los coeficientes MFCC's fueron obtenidos con el toolbox "Auditory Toolbox". [1]

F1, F2, F3 y F4 - Son vectores que representan las gradientes de los primeros 4 formantes, (experimentalmente manejar los valores directos de las formantes no aportó mejora en la tarea de clasificación, por lo que se optó por almacenar únicamente la gradiente de cada vector), a cada vector del gradiente de cada formante, se calculó el máximo, mínimo, promedio, mediana, moda y desviación estándar.

Nota: Los valores de los formantes fueron extraídos mediante el uso del software PRAAT. [10]